Auswertung einer Brustkrebsstudie

mit Hilfe von logistischen Regressionsmodellen

Diplomarbeit von Marco Grzegorczyk

Vorgelegt dem Fachbereich Statistik
an der Universität Dortmund
im Februar 2003.

Angefertigt unter der wissenschaftlichen Betreuung von
Herrn Prof. Dr. Wolfgang Urfer.

Inhaltsverzeichnis:

Kapitel 1. Einleitung

Das Mamma-Karzinom (umgangsprachlich: Brustkrebs) ist der häufigste bösartige Tumor der Frau. Da die Häufigkeit einer solchen Tumorerkrankung zudem in den letzten Jahrzehnten in allen entwickelten Ländern rapide zugenommen hat, ist die Suche nach den Ursachen für diese hohe Erkrankungsrate seit vielen Jahren in vollem Gange (Grundmann 1994). Nach dem saarländischen Krebsregister (Stand: 1989) macht das Mamma-Karzinom in Europa einen Prozentsatz von etwa 23% aller bösartigen Tumorneuerkrankungen bei der Frau aus.
Den Krebsregistern der Vereinigten Staaten kann darüber hinaus entnommen werden, dass die Anzahl Neuerkrankungen pro Jahr (Inzidenz) von der ethnischen Rassenzugehörigkeit abhängt. So zum Beispiel erkrankten in den U.S.A im Jahre 1989 durchschnittlich nur 87,9 von 100.000 Frauen afroamerikanischer Abstammung an einem Mamma-Karzinom, wohingegen im selben Jahr durchschnittlich 108,2 von 100.000 weißen Frauen neu an Brustkrebs erkrankten. Bemerkenswert ist hierbei, dass ein Vergleich altersspezifischer Inzidenzraten zeigt, dass im Gegensatz dazu, speziell die Anzahl Neuerkrankungen vor Erreichen des 40ten Lebensjahres unter den Frauen afroamerikanischer Abstammung größer ist als unter den weißen amerikanischen Frauen.
Nach dem gegenwärtigen Wissenstand der Medizin ist davon auszugehen, dass sowohl genetischen Faktoren als auch Umweltfaktoren von Bedeutung für die Entstehung des Mamma-Karzinoms sind, und deshalb als Ursachen für diese Unterschiede in den altersspezifischen Inzidenzraten weißer und schwarzer Frauen in Frage kommen.

Mit der Zielsetzung zu ergründen, zu welchen Anteilen die unterschiedliche soziökonomische Lebensweise und die genetische Diskrepanz verantwortlich zu machen sind, wurde in den frühen 90er Jahren von der Howard-Universität mit der Durchführung einer mehrstufigen epidemiologischen Brustkrebsstudie begonnen. Die Ergebnisse dieser Studie wurden in Form eines Artikels veröffentlicht (Bonney et al. 1993).

Diese Arbeit beschäftigt sich mit der Suche nach Risikofaktoren für Brustkrebs auf Grundlage zweier Datensätze, die in einer frühen Phase obiger Studie erhoben und von der Howard-Universität zwecks Auswertung zur Verfügung gestellt wurden. Das Datenmaterial umfasst epidemiologische Angaben zu den Angehörigen von Familien afroamerikanischer Abstammung, in denen es jeweils mindestens einen Brustkrebsfall gegeben hat.

Da es sich um Familiendaten handelt, wird einem ersten Auswertungsschritt zunächst mittels einfacher statistischer Methoden nach Anzeichen für eine familiäre Häufung der Brustkrebs-Krankheit gesucht. Daran anschließend geht es im zweiten Auswertungsschritt um die Identifikation nicht-familiärer epidemiologischer Risikofaktoren. Unter Zuhilfenahme logistischer Regressionsmodelle wird nach Faktoren gesucht, die das Brustkrebsrisiko erhöhen. Da keine Daten von Kontrollfamilien, das heißt Familien in denen es keinen Brustkrebsfall gegeben hat, zur Verfügung stehen, müssen in diesem Schritt die brustkrebserkrankten Studienteilnehmer mit den erkrankten Studienteilnehmern im Sinne einer Fall-Kontroll-Studie verglichen werden. Eine Berücksichtigung von Verwandtschaftsbeziehungen zwischen den Angehörigen derselben Familie ist hierbei aufgrund des Studiendesigns nicht möglich.
Speziell die Betrachtung multipler logistischer Regressionsmodelle setzt in Anbetracht vieler fehlender Dateneinträge eine Miteinbeziehung von Personen, über die nicht alle Informationen verfügbar sind, voraus. Im Umgang mit fehlenden Werten werden deshalb in diesem Schritt sowohl die Probability-Imputation-Methode als auch das Indikatorvariablen-Verfahren angewendet. Beide Verfahren ermöglichen es, die gesamte, in den Daten vorhandene, Information in die Regressionsanalysen mit einzubeziehen, machen aber auf der anderen Seite gewissermaßen eine Manipulation fehlender Werte notwendig, was Simulationsstudien zufolge zu Ergebnisverzerrungen führen kann. Dennoch zeigt sich, dass aufgrund des Ausmaßes der Unvollständigkeit des Datenmaterials nicht von dem Einsatz dieser Verfahren abgesehen werden kann.
Erst bei separater Suche nach Risikofaktoren für Frauen vor und nach der Menopause, wird auf eine Berücksichtigung von Personen mit unvollständigen Merkmalswerten verzichtet, was in beiden Fällen einen entsprechenden Informationsverlust zur Folge hat, da mit diesen Personen auch die, über sie verfügbaren, Daten verloren gehen.

Überblick über die folgenden Kapitel:
Nach dieser Einleitung werden im **zweiten Kapitel** wichtige Fachbegriffe erläutert und einige medizinische Angaben zum Mamma-Karzinom gemacht. Darüber hinaus wird das auszuwertende Datenmaterial näher beschrieben.
Daran anschließend gibt **Kapitel 3** einen kurzen Überblick über das Fachgebiet der analytischen Epidemiologie. Insbesondere wird das Design der klassischen Fall-Kontroll-Studie beschrieben und kurz diskutiert.
Kapitel 4 verschafft mathematisches Hintergrundswissen, welches für das Verständnis der im praktischen Teil angewandten Verfahren notwendig ist. Die Theorie des logistischen Regres-

sionsmodells, die in Unterkapitel 4.3 beschrieben wird, bildet den Schwerpunkt dieses Kapitels. Neben Erläuterungen zur Interpretation logistischer Regressionsmodelle wird auch ausführlich auf die Parameterschätzung und auf Parametertests eingegangen. Darüber hinaus wird beschrieben, wie die Anpassung solcher Modelle überprüft werden kann, und es werden drei Regressionsdiagnostiken vorgestellt. Im letzten Unterkapitel werden die bereits angesprochenen Verfahren im Umgang mit fehlenden Werten (Probability-Imputation-Methode und Indikatorvariablen-Verfahren) vorgestellt.

Das **fünfte Kapitel** beschäftigt sich mit der Auswertung des Datenmaterials. Nach einigen ergänzenden Bemerkungen zum vorliegenden Datenmaterial wird in Unterkapitel 5.2 zunächst mittels einfacher statistischer Methoden nach Anzeichen für eine familiäre Häufung der Brustkrebskrankheit gesucht. Daran anschließend geht es im dritten Unterkapitel um die Identifikation nicht-familiärer Risikofaktoren für Brustkrebs mit Hilfe von logistischen Regressionsmodellen. Neben Modellen für die gesamte weibliche Studienpopulation, im Rahmen derer im Umgang mit fehlenden Werten unabhängig voneinander einmal die Probability-Imputation-Methode und einmal das Indikatorvariablen-Verfahren zum Einsatz kommen, wird jeweils ein unabhängiges Modell für die Frauen vor und nach der Menopause generiert.

In **Kapitel 6** werden die Ergebnisse der vorliegenden Arbeit zusammengefasst und diskutiert.

Kapitel 2. Gegenstand der Untersuchung

In diesem Kapitel werden einige medizinische Fachbegriffe erläutert, von denen im Rahmen dieser Arbeit häufiger Gebrauch gemacht wird, und es wird Hintergrundwissen zum Mamma-Karzinom vermittelt. Von besonderem Interesse für die vorliegende Brustkrebsstudie sind die Faktoren, die nach dem heutigen Kenntnisstand der Medizin als Risikofaktoren für Brustkrebs gelten. Im zweiten Unterkapitel folgt eine Beschreibung des für die Auswertung vorliegenden Datenmaterials.

2.1 Medizinischer Hintergrund

Um ein substanzwissenschaftliches Verständnis dieser Arbeit zu ermöglichen, werden im folgenden Abschnitt zunächst einige Fachausdrücke erläutert. Ausführlichere Begriffserklärungen können zum Beispiel Pschyrembel (1998) entnommen werden. Nach den Begriffserklärungen wird im zweiten Abschnitt etwas ausführlicher auf die Entstehung und das Krankheitsbild des Mamma-Karzinoms eingegangen. Die Ausführungen im zweiten Abschnitt orientieren sich vorwiegend an Bleich et al. (1995) und Greskötter (1996).

2.1.1 Fachbegriffe

Tumor: Als Tumor (Geschwulst) werden Gewebsvermehrungen bezeichnet, bei denen das Wachstum nicht mehr mit dem normalen Gewebe koordiniert ist. Es kommt zu einer Verselbstständigung des Wachstumsprozesses und auch des neu entstandenen Gewebes. Dabei auftretende Veränderung der Einzelzellen sind Ausdruck eines vermehrten, gestörten, primär nur auf das Zellwachstum und die Zellvermehrung ausgerichteten Proteinstoffwechsels. Charakteristisch für bösartige Tumore ist ein rasches, invasives und destruierendes Wachstum. Darüber hinaus können sich die Zellen bösartiger Tumoren auch unabhängig vom Primärtumor vermehren. Durch die destruierende Infiltration in Blut- und Lymphgefäße kommt es zur Verschleppung der Tumorzellen, die dann möglicherweise in anderen Bereichen des Körpers Fernabsiedlungen bilden. Man spricht von Metastasen (Tochtergeschwülsten).

Karzinom: Der Sammelbegriff für alle bösartigen epithelialen (=vom Deckgewebe ausgehenden) Tumore lautet: Karzinom (=Krebs).

Fehlgeburt/ Totgeburt: Man spricht in der Medizin von einer Fehlgeburt, wenn eine Schwangerschaft vorzeitig durch die Ausstoßung eines Fetus von unter 500g bei Fehlen aller für eine Lebendgeburt maßgeblichen Lebenszeichen (Herzschlag, Lungenatmung usw.) beendet wird. Erst wenn ein nicht lebensfähiger Fetus bei der Trennung vom Mutterleib ein Gewicht von über 500g aufweist, spricht man von einer Totgeburt.

Menarche/ Menopause: Der Begriff Menarche bezeichnet den Zeitpunkt des ersten Auftretens der Menstruation (monatliche Regelblutung). Dieser Zeitpunkt wird von ethnischen, konstitutionellen und klimatischen Faktoren beeinflusst. In den Industriestaaten beginnt die Menarche der Frau in den meisten Fällen zwischen dem 11. und 14. Lebensjahr. Der Zeitpunkt der letzten spontanen Menstruation, der retrospektiv ein Jahr lang keine Blutung folgt, wird als Menopause bezeichnet. Die mehrjährige Übergangsphase, in der die Blutungen zunehmend unregelmäßiger werden, bezeichnet man als Klimakterium (umgangssprachlich: Wechseljahre). Mit der Menopause beginnt die Phase der Post-Menopause.
In den Industriestaaten erreichen Frauen das Stadium der Post-Menopause durchschnittlich zwischen dem 45. und 50. Lebensjahr.

Tubensterilisation: Spezieller operativer Eingriff, bei dem die Eileiter unterbrochen werden, so dass die Sterilisation der Frau eintritt.

Hysterektomie: Als Hysterektomie wird ein operativer Eingriff bezeichnet, bei welchem die Gebärmutter entfernt wird. Dieser Eingriff, der nicht unbedingt die Entfernung der Eierstöcke (Ovarien) mit einschließt, versetzt die betroffene Frau in den Status der Post-Menopause.
Heutzutage stellen in erster Linie Erkrankungen der Gebärmutter (wie zum Beispiel Gebärmutterkrebs(~10%) oder chronische Unterbauchschmerzen) Indikationen für Hysterektomien dar. Darüber hinaus werden Hysterektomien allerdings in seltenen Fällen auch zur Krebsvorsorge durchgeführt. Als Beispiel wäre die Endometrium-Hyperplasie (Schleimhautvermehrung in der Gebärmutter) zu nennen, die in ihrem weiteren Verlauf zu Gebärmutterkrebs führen kann. In den 60er und frühen 70er Jahren wurden Hysterektomien allerdings auch bei Sterilisationswunsch durchgeführt.

2.1.2 Hintergrundwissen zum Mamma-Karzinom

Die meisten Mamma-Karzinome (85%) gehen von den Epithelzellen der kleinen Milchgänge aus, wohingegen nur 15 % in den Drüsenläppchen entstehen. Die häufigste Lokalisation des Mamma-Karzinoms ist der obere äußere Quadrant der Brust. Der gleichzeitige Befall beider Brüste beobachtet man in weniger als 1% der Fälle.

Typische Früh-Symptome sind:

- Einziehungen der Haut/ Brustwarze
- Ekzemartige Hautveränderungen und Hautverdickungen
- Ausfluss aus der Brustwarze
- tastbare, meist schmerzlose Knoten und Verhärtungen im Bereich der Brust

Lokal führt das destruierende Wachstum des Mamma-Karzinoms zunächst zu Gewebszerstörungen in der betroffenen Brust. Darüber hinaus besteht bereits bei den kleinsten tastbaren Knötchen eine hohes Metastasierungrisiko. Die Metastasierung erfolgt primär in Skelett, Lunge, Leber und Eierstöcke. Wie bei jeder bösartigen Tumorerkrankung fallen mit der Vergrößerung des Tumorgewebes immer mehr toxische Stoffwechselprodukte an, die mehr und mehr zu einer Auszehrung des Patienten führen (Tumorkachexie).

In Deutschland sterben in Folge eines Mamma-Karzinoms jährlich etwa 18.000 Frauen (Statistisches Bundesamt, Wiesbaden 1998). Damit stellt das Mamma-Karzinom nach Herz-Kreislauf-Erkrankungen und Gewalteinwirkungen (Unfälle, Morde, Suizide) die dritthäufigste Todesursache für Frauen in Deutschland dar.

Aufgrund dieser schwerwiegenden Konsequenzen wird bereits beim geringsten Verdacht auf ein Mamma-Karzinom eine eingehende ärztliche Untersuchung durchgeführt. Neben der Erhebung der Krankengeschichte und manuellen Untersuchungen wird eine Röntgenuntersuchung der Brust und eine Ultraschalluntersuchung vorgenommen. Die Mammographie, also die Röntgenuntersuchung der Brust, gestattet eine Beurteilung der Größe des Tumors und eine Abgrenzung von gutartigen Veränderungen. Im Zweifelsfalle kann mit einer feinen Nadel eine Gewebeprobe zur mikroskopischen Untersuchung entnommen werden. Das Ergebnis der mikroskopischen Gewebeuntersuchung gibt Aufschluss über den Tumortyp und den Grad seiner Aggressivität.

Therapie der Wahl bei Diagnose eine Mamma-Karzinoms ist die Operation, eventuell mit vorangehender Chemotherapie zur Verkleinerung des Tumors. Soweit möglich, wird der brusterhaltenden Operation der Vorzug gegeben. Bei ausgedehntem Tumor ist eine Amputation der

Brust nicht zu vermeiden. Die Nachbehandlung besteht aus Strahlentherapie, Chemotherapie und/oder Hormonbehandlung sowie Nachsorgeuntersuchungen.

Im letzten Jahrzehnt wurden zahlreiche Studien durchgeführt, bei denen die Bedeutung vieler Faktoren für die Entstehung von Mamma-Karzinomen Gegenstand der Untersuchung waren. Die Ergebnisse dieser Untersuchungen zeigen, dass das Risiko einer Frau, während ihres Lebens an Brustkrebs zu erkranken, von den unterschiedlichsten Faktoren beeinflusst wird. Die hormonelle Beeinflussung der Brustdrüse während verschiedener Lebensabschnitte (Menstruationsgeschichte, Schwangerschaften, Wechseljahre etc.), Ernährung, Geburtsort, Lebensstil und die medizinische Vorgeschichte spielen genau wie genetische Faktoren eine Rolle (vgl. Kelsey 1993 und Kelsey et al. 1993).
Grundmann (1994) und Kelsey et al. (1993) zufolge ist insbesondere davon auszugehen, dass die folgenden Faktoren das Brustkrebs-Erkrankungsrisiko erhöhen:

Hormonelle Risikofaktoren:

- eine übersteigerte therapeutische Östrogenbehandlung
- ein frühes Eintreten der Menarche
- ein spätes Eintreten der Menopause
- eine späte erste Schwangerschaft
- Kinderlosigkeit

Ernährungsfaktoren:

- Fettreiche und energieüberschüssige Ernährung
- Fettleibigkeit im höheren Alter

Genetische Faktoren:

- Mutter mit Brustkrebs
- Schwester mit Brustkrebs

Geburtsort und Wohnort:

- Mitteleuropa und U.S.A. (im Vergleich zu Asien und Afrika)
- Städtischer Wohnort (im Vergleich zu ländlichem Wohnort)

Lebensstil:

- Familienstand: niemals verheiratet (im Vergleich zum Familienstand: verheiratet)
- Hoher sozioökonomischer Status

Medizinische Vorgeschichte:

- hohe Dosen ionisierender Strahlen (z.B. Röntgenstrahlen) auf den Brustkorb
- eine vorangegangene gutartige Tumorerkrankung an der Brust
- Krebserkrankungen an Gebärmutter oder Eierstock

Zusammenfassend kann festgehalten werden, dass die bisherigen Studienergebnisse zeigen, dass die Entstehung des Mamma-Karzinoms von einer Vielzahl der unterschiedlichsten Faktoren begünstigt wird. Insbesondere kommt auch Faktoren eine Bedeutung zu, für deren Einfluss auf die Brustkrebsentstehung bis heute noch keine biologisch-medizinische Erklärung gefunden werden konnte. In Bezug auf die Auswertung des vorliegenden Datenmaterials bedeutet dies, dass jedes der erhobenen Merkmale bis auf weiteres als potentieller Risikofaktor anzusehen ist.

2.2 Beschreibung des Datenmaterials

Von der Howard-Universität wurden für die statistische Auswertung zwei Datensätze mit epidemiologischen Angaben über die Familienangehörigen von 261 Familien afroamerikanischer Abstammung zur Verfügung gestellt.

Konkrete Datensituation

Während der erste Datensatz zu den insgesamt 3527 Familienangehörigen nur einige wenige Informationen liefert, umfasst der zweite Datensatz ausführlichere epidemiologische Angaben von 1198 dieser Personen. Beide Datensätze sind unvollständig, so dass jeweils eine Diskrepanz zwischen den theoretisch verfügbaren und tatsächlich vorhandenen Informationen vorliegt. Problematisch ist zudem, dass die Angaben in beiden Datensätzen in einigen Fällen im Widerspruch zueinander stehen. Im Hinblick auf die Datenauswertung wurde in widersprüchlichen Fällen stets zu Gunsten der Angaben im zweiten Datensatz entschieden.

Von den 261 Familien afroamerikanischer Abstammung befindet sich jeweils ein Familienmitglied aufgrund einer Brustkrebserkrankung (BCA-Erkrankung) in ärztlicher Behandlung. Da bekanntermaßen Frauen deutlich brustkrebsgefährdeter sind als Männer (vgl. Grundmann

1994) überrascht es nicht, dass es sich bei diesen Fällen vorrangig um Frauen (257 von 261) handelt. Ausgehend von diesen BCA-Fällen in Behandlung, die im Folgenden auch als Indexfälle bezeichnet werden, wurde die Studienpopulation durch Familienangehörige dieser Fälle, über die Angaben verfügbar gemacht werden konnte, vervollständigt. Unabhängig vom Geschlecht kamen hierbei neben den nächsten Familienangehörigen (Eltern, Geschwister und Kinder) auch entferntere (Großeltern, Onkel und Tanten, Enkelkinder, Neffen und Nichten und Halbgeschwister) und angeheiratete Verwandte (Ehepartner, Schwager, angeheiratete Onkel und Tanten und Schwiegerkinder) in Betracht. Nicht vorausgesetzt wurde dabei, dass die Familienangehörigen zum Zeitpunkt des Studienbeginns noch lebten, so dass auch über bereits verstorbene Familienmitglieder Informationen in Erfahrung gebracht wurden. Insgesamt umfasst die Studienpopulation 3527 – nicht notwendigerweise noch lebende - Personen aus 261 Familien, wobei die familiären Beziehungen zwischen den Angehörigen einer jeden Familie theoretisch, das heißt sofern die dazugehörigen Dateneinträge vorhanden sind, bekannt sind.

Informationsgehalt der beiden Datensätze

Wie bereits angemerkt gliedert sich das Datenmaterial in zwei Datensätze.

Offensichtlich zum Zwecke der Untersuchung, ob Brustkrebs familiär gehäuft auftritt, wurde zunächst versucht, von allen Familienmitgliedern in Erfahrung zu bringen, ob sie ebenfalls an Brustkrebs erkrankt sind oder nicht, und dazu einige Zusatzinformationen erhoben. Diese Zusatzinformationen umfassen neben der familiären Beziehung zu dem Indexfall (inklusive Geschlechtsangabe) und dem Geburtsjahr des betreffenden Person einige speziellere Angaben, die als solche aber nicht unbedingt epidemiologisch relevante Risikofaktoren für Brustkrebs darstellen. So zum Beispiel wurden von bereits verstorbenen Personen Todesalter und Todesursache festgehalten und von Brustkrebserkrankten sind Angaben über das Erkrankungsalter verfügbar. Als Risikofaktor von Interesse sind Angaben über Nichtbrustkrebs-Tumorleiden und dem dazugehörige Erkrankungsalter der Personen. Der erste Datensatz setzt sich aus genau diesen Informationen zusammen und kann daher nur dahingehend untersucht werden, ob sich Brustkrebserkrankungen familiär häufen bzw. vererbt werden. Problematisch hinsichtlich einer solchen Untersuchung ist, dass jede Familie durch den Indexfall über mindestens einen sicheren BCA-Fall (Indexfall) verfügt, und dass die Anzahl Studienteilnehmer pro Familie in Abhängigkeit von der tatsächlichen Familiengröße und der Teilnahmebereitschaft der Familienangehörigen stark variiert. Die Extremfälle bilden einerseits 8 Familien, die jeweils nur aus dem Indexfall bestehen, und andererseits Familien der Größen 38, 69 und 110.

Ebenfalls stellt die bereits erwähnte Unvollständigkeit des Datenmaterials ein Problem dar. So zum Beispiel ist von 396 Studienteilnehmern nicht bekannt, ob sie an Brustkrebs leiden oder nicht, und in einigen Familien sind nicht alle familiären Beziehungen ersichtlich.
Dennoch wird in einem ersten Auswertungsschritt (vgl. Unterkapitel 5.2) untersucht, ob sich auf Grundlage dieses Datensatzes ein Anzeichen für eine familiäre Häufung der Brustkrebskrankheit finden lässt.

Der zweite Datensatz umfasst für eine Teilmenge der Studienteilnehmer weitere epidemiologische Informationen, die erst im Anschluss an die erste Datenerhebung eingeholt wurden. Eine erste weniger wesentliche Reduktion der Anzahl Personen ergibt sich aus dem Umstand, dass im Vergleich zum ersten Datensatz aus unbekannten Gründen nur noch 255 der 261 Familien berücksichtigt wurden. Zudem wurden im Gegensatz zum ersten Datensatz in jeder Familie neben dem Indexfall in Behandlung auch nur noch seine nächsten weiblichen Verwandten (Mütter, Großmütter, (Halb-)Schwestern, Tanten und Töchter) als Studienteilnehmer bzw. Familienangehörige ausgewählt. Diese beiden Einschränkungen führen dazu, dass ausführlichere Informationen nur von 1198 der 3527 Studienteilnehmer vorliegen. Unter diesen 1198 Studienteilnehmern befinden sich nur vier männliche Personen, die jeweils der Indexfall ihrer Familie sind.

Die ausführlicheren Informationen umfassen neben biologischen Angaben auch zahlreiche Angaben zu Lebensgewohnheiten und soziokulturellen Gegebenheiten, so dass zahlreiche Expositionen als potentielle Risikofaktoren für Brustkrebs untersucht werden können. Insgesamt wurden soweit verfügbar bis zu 86 Merkmalsinformationen an den 1198 Personen erhoben. Von diesen kommen 68 als potentielle Risikofaktoren für Brustkrebs in Frage. Die anderen 18 Merkmale umfassen zum Beispiel Angaben zu chirurgischen Eingriffen an der erkrankten Brust und betreffen somit ausschließlich die Brustkrebsfälle. Stichwortartige Beschreibungen zu den 68 relevanten Merkmalen, die versucht wurden, an den 1198 Studienteilnehmern zu erheben, können Anhang 4 dieser Arbeit entnommen werden. Aufgrund des Umfangs des Datenmaterials wird bei der Datenauswertung (Kapitel 5.3) zunächst eine Variablen-Vorauswahl getroffen. Anschließend werden ausschließlich die Merkmale näher beschrieben, bei denen die Ergebnisse, der bei der Vorauswahl angewandten statistischen Methoden, auf eine Assoziation mit der Brustkrebserkrankung deuten.

Datenerhebung und Vollständigkeit des Datenmaterials

Konkrete Informationen über die Art der Datenerhebung sind nicht verfügbar. Bekannt ist lediglich, dass mit einigen Personen Interviews (persönlich oder telefonisch) geführt wurden, wohingegen anderen Personen Fragebögen zugeschickt wurden. In Bezug auf die bereits verstorbenen Personen ist in diesem Zusammenhang anzunehmen, dass die Informationen durch zusätzliche Befragungen ihrer Angehörigen in Erfahrung gebracht wurden. Nach welchen Kriterien letztlich entschieden wurde, ob ein Interview geführt oder ein Fragebogen verschickt wird, ist allerdings nicht bekannt.

Eine Überprüfung der Vollständigkeit des Datenmaterials zeigt, dass speziell im zweiten Datensatz viele Merkmale äußerst lückenhaft erhoben wurden. Von welchem Ausmaß die Unvollständigkeit des Datenmaterials ist, verdeutlicht Tabelle 2.2.1. Dieser Tabelle können für 8 Beispielsmerkmale die absoluten und relativen Anteile fehlender Werte entnommen werden. Weitere Vollständigkeitsüberprüfungen zeigen, dass nur bei einigen wenigen Merkmalen unter 200 fehlende Werte vorliegen, wohingegen bei den meisten Merkmalen zwischen 400 und 600 Dateneinträge fehlen. Merkmale, bei denen deutlich über 50% der Werte fehlen, stellen ebenfalls keine Seltenheit dar. Da die Merkmale als Risikofaktoren für Brustkrebs von unterschiedlicher substanzwissenschaftlicher Relevanz sind, wird an dieser Stelle auf genauere Angaben zur Häufigkeitsverteilung der fehlenden Werte verzichtet.

Tabelle 2.2.1: Unvollständigkeit des zweiten Datensatzes (8 Beispielsmerkmale)

Merkmal bzw. Information über…	**Häufigkeit fehlender Werte**	
	absolut	**relativ**
Anzahl Lebendgeburten	43	0,036
Anzahl Fehlgeburten	79	0,066
Fettleibigkeit	481	0,402
Schulbildung	524	0,437
Regelmäßigkeit der Periode	538	0,449
Durchschnittliche Periodendauer	779	0,650
Monate der Kinderstillung	1000	0,835
Taillenumfang	1152	0,962

Besonders problematisch im Hinblick auf die Datenauswertung mit Hilfe logistischer Regressionsmodelle ist, dass die fehlenden Werte im Datensatz nicht zufällig auftreten, sondern einer bestimmten Systematik folgen. Offensichtlich gilt für jedes Merkmal, dass der relative Anteil fehlender Werte bei den Indexfällen deutlich geringer ist als bei ihren Familienangehörigen. Auf diese Problematik wird ausführlich in Kapitel 5.1 dieser Arbeit eingegangen.

Kapitel 3. Epidemiologische Grundlagen

Dieses Kapitel erläutert die typischen Zielsetzungen epidemiologischer Studien. Zunächst wird ein Überblick über das statistische Fachgebiet der (analytischen) Epidemiologie gegeben, bevor dann ein typisches Studiendesign beschrieben wird, welches zum Gewinn epidemiologischer Erkenntnisse herangezogen werden kann.

3.1 Epidemiologie

Die Epidemiologie als Wissenschaftsdisziplin nutzt statistisch-mathematische Methoden zur Gewinnung humanmedizinischer Erkenntnisse. Im Gegensatz zur Humanmedizin, in der vorrangig einzelne Individuen betrachtet werden, zielen epidemiologische Untersuchungen primär darauf, allgemeingültige Aussagen über die Ursachen von Krankheiten herauszuarbeiten. Die Notwendigkeit solcher Untersuchungen ist darin begründet, dass es in der Medizin häufig (noch) nicht möglich ist, die biologischen Mechanismen, die zu bestimmten Krankheiten führen, im Einzelnen substanzwissenschaftlich zu verstehen. Erst durch die Betrachtung der Krankheitsverteilung in ganzen Bevölkerungsgruppen, können Risikofaktoren erkannt und anschließend genauer ergründet werden. Da die Untersuchung ganzer Bevölkerungsgruppen i.a. allerdings nicht praktikabel ist, können Erkenntnisse nur durch Stichprobenerhebungen gewonnen werden. Ein Vorgehen, welches stets mit einer gewissen Unsicherheit verbunden ist. Die Epidemiologie nutzt die Methodik der mathematischen Statistik, die diese Unsicherheit abzuschätzen vermag, und stellt somit das Bindeglied zwischen Medizin und Statistik dar. Sie wird deshalb auch häufig als interdisziplinäre Wissenschaft bezeichnet.

Die analytische Epidemiologie - als Teilgebiet - beschäftigt sich mit der statistischen Erforschung der Bedeutung von verschiedenen Faktoren für die Entstehung von Krankheiten. Der Begriff „Faktoren“ ist in diesem Zusammenhang als Überbegriff für die Gesamtheit aller potentiellen Risikofaktoren zu verstehen, denen ein menschliches Individuum ausgesetzt sein kann. Typische Beispiele sind Lebensgewohnheiten (Ernährung sowie Zigaretten- und Alkoholkonsum etc.) und Umweltbedingungen (Schadstoffbelastungen etc.) allerdings auch unveränderbare Merkmale wie das Geschlecht des Individuums.

Bei epidemiologischen Untersuchungen variieren die interessierenden Faktoren in Abhängigkeit von der zu untersuchenden Krankheit und der Zielsetzung der Studie. Entsprechend können diese nicht allgemeingültig angegeben werden, sondern müssen zu Beginn einer jeden

epidemiologischen Untersuchung sorgfältig definiert werden. Die Faktoren von Interesse werden als Expositionen (Aussetzungen) bezeichnet. Alle Faktoren, deren möglicherweise vorhandenen Einflüsse nicht in unmittelbarem Interesse der Studie stehen, werden unter dem Begriff Störgrößen (Confounder) zusammengefasst.

Für die Zusammenhangsanalyse muss Datenmaterial zur Verfügung gestellt werden. Dieses Material umfasst Angaben zu Expositionen und Krankheitszuständen einer möglichst repräsentativen Stichprobe von Individuen der Bevölkerungsgruppe. Solche Daten können beispielsweise mit Hilfe einer Fall-Kontroll-Studie beschaffen werden. Dieser wichtige Studientyp, der bei epidemiologischen Untersuchungen häufig Anwendung findet, wird im folgenden Unterkapitel vorgestellt.

3.2 Design der Fall-Kontroll-Studie

Zur Gewinnung epidemiologischer Erkenntnisse dienen vorwiegend Beobachtungsstudien, da experimentelle Ansätze, bei denen Probanden geplant bestimmten Expositionen ausgesetzt werden, aus ethischer Sicht nicht vertretbar sind. Bei Beobachtungsstudien wird kein gezielter Eingriff in die Exposition vorgenommen, sondern nur beobachtet, wie Krankheiten und Expositionen in Beziehung zueinander stehen. Im wesentlichen werden drei Typen von Beobachtungsstudien unterschieden:

(a) Querschnittsstudien (“cross-sectional study“)
(b) Kohortenstudien (“follow-up-study“)
(c) Fall-Kontroll-Studien (“case-control-study”)

In diesem Abschnitt wird ausschließlich das Design der Fall-Kontroll-Studie vorgestellt. Ausführliche Beschreibungen zu allen drei Studientypen können zum Beispiel Kreienbrock & Schach (2000) entnommen werden.

Fall-Kontroll-Studie

Eine epidemiologische Fall-Kontroll-Studie weist die Besonderheit auf, dass retrospektiv (rückblickend) und damit indirekt vorgegangen wird. Von der Auswirkung (Krankheit – Ja/Nein) ausgehend, wird nach möglichen Ursachen gesucht. Es stellt sich die Frage, ob und wenn inwieweit sich das Expositionsmuster erkrankter und nichterkrankter Individuen unterscheidet. Eine Frage, deren Beantwortung es voraussetzt, dass eine Gruppe von Erkrankten

mit einer Gruppe von Nichterkrankten hinsichtlich vorausgegangener Expositionen miteinander verglichen wird

Der Ablauf einer Fall-Kontroll-Studie kann entsprechend in zwei Schritte unterteilt werden. Im ersten Schritt werden zwei Gruppen von Studienteilnehmer rekrutiert. Die erste Gruppe, bestehend aus Individuen, die an der zu untersuchenden Krankheit leiden, wird in Anbetracht des Krankheitsstatus als Fallgruppe bezeichnet. Die Kontrollgruppe wird von nichterkrankten Individuen gebildet. Anschließend werden im zweiten Schritt die interessierenden Expositionen aller Probanden in Erfahrung gebracht. Diese können sowohl noch gegenwärtig als auch zeitlich vorausgegangen sein.

Die Aussagekraft einer Fall-Kontroll-Studie ergibt sich unmittelbar aus der Repräsentativität von Fall- und Kontrollgruppe für die jeweilige Teilpopulation erkrankter bzw. krankheitsfreier Personen. Obwohl geeignete Stichprobenverfahren (z.B. geschichtete Zufallsauswahl), mit denen ein hohes Maß an Repräsentativität erreicht werden kann, hinreichend bekannt sind, dienen der Epidemiologie aus Gründen der technischen Realisierbarkeit zumeist deutlich einfachere Auswahlverfahren. Entsprechend sorgfältig ist zu überlegen, inwieweit das Studienergebnis auf die Gesamtpopulation verallgemeinerbar ist.

Auswertung einer Fall-Kontroll-Studie

Bei der Auswertung von Daten, die mit Hilfe einer Fall-Kontroll-Studie erhoben wurden, ist die Art der Datenerhebung zu berücksichtigen. Die Festlegung der Gruppengrößen zu Beginn der Studie determiniert auch die Erkrankungsverteilung der Gesamtheit aller Studienteilnehmer. Das heißt, die Anteile von erkrankten und nichterkrankten Probanden in der Studie unterliegen keinem Zufall. Dies hat zur Konsequenz, dass es nicht möglich ist, Erkrankungswahrscheinlichkeiten (statistisch) zu schätzen. Dies betrifft unabhängig von Anzahl und Messniveau der interessierenden Expositionen die folgenden drei Typen von Wahrscheinlichkeiten:

(1) die unbedingte Erkrankungswahrscheinlichkeit eines Individuums der Gesamtpopulation,
(2) die Wahrscheinlichkeiten des gemeinsamen Auftretens von Krankheit und bestimmten Expositionen
(3) die bedingten Erkrankungswahrscheinlichkeiten gegeben bestimmte Expositionen.

Das Vorliegen jeweils einer Zufallsstichprobe aus der Expositionsverteilung der kranken und krankheitsfreien Teilpopulation ermöglicht ausschließlich die gruppenspezifische Schätzung

von Expositionswahrscheinlichkeiten. Wenngleich diese bedingten Wahrscheinlichkeiten nicht von unmittelbarem epidemiologischem Interesse sind, ermöglicht ihre Schätzbarkeit die Analyse des Zusammenhangs zwischen Krankheit und Expositionen mit Hilfe eines speziellen Assoziationsmaßes. Diese Maß wird als Odds-Ratio (Chancenverhältnis) bezeichnet und in Unterkapitel 4.2 vorgestellt.

Vor- und Nachteile von Fall-Kontroll-Studien

Sowohl bei seltenen Krankheiten als auch bei Krankheiten mit langer Latenzzeit stellt die Fall-Kontroll-Studie die einzige epidemiologische Studienform dar, die praktisch durchführbar ist. In Bezug auf seltene Krankheiten garantiert die explizite Festlegung der Anzahlen von Fällen und Kontrollen zu Beginn der Studie, dass genügend Daten über erkrankte Personen zur Verfügung stehen. Andere Designs, bei denen eine Zufallsauswahl aus der Gesamtpopulation erfolgt (Querschnittsstudie), oder bei denen zu Beginn der Studie Gruppen unterschiedlich exponierter Probanden ausgewählt werden, und dann beobachtet wird, bei welchen Individuen sich die interessierende Krankheit in einer bestimmten Zeitperiode entwickelt (prospektive Kohortenstudie), liefern bei seltenen Krankheiten nur dann genügend Datenmaterial über erkrankte Personen, wenn entsprechend große Stichprobenumfänge gewählt werden. Ein Umstand, der i.a. mit zu hohen Studienkosten verbunden ist, und sich deshalb aus Kostengründen nicht realisieren lässt. Zur Untersuchung von Krankheiten mit einer langen Latenzzeit disqualifiziert sich insbesondere die prospektive Kohortenstudie. Die zeitliche Verzögerung zwischen Exposition und Manifestation solcher Krankheiten, macht entsprechend lange Studiendauern erforderlich, was i.a. als „nicht akzeptabel“ empfunden wird. Ein weiterer großer Vorteil des Fall-Kontroll-Studiendesigns liegt darin, dass in einer einzigen Studie zahlreiche Faktoren als Expositionen berücksichtigt werden können. Stehen Fall- und Kontrollgruppe erst einmal zur Verfügung, können (nachträglich) Informationen über beliebig viele Expositionen eingeholt werden. Eine Festlegung der zu untersuchenden Expositionen vor Beginn der Studie ist im Gegensatz zu Kohortenstudien nicht notwendig, so dass sich Fall-Kontroll-Designs insbesondere eignen, wenn im Rahmen der Studie nach Risikofaktoren für eine bestimmte Krankheit gesucht werden soll.

Diesen Vorteilen steht aus theoretischer Sicht nur der Nachteil gegenüber, dass die Datenauswertung auf Grundlage von Chancenverhältnissen erfolgen muss, was insbesondere impliziert, dass keine Einschätzung der Krankheitshäufigkeit möglich ist. Da die Krankheitshäufigkeit für die Analyse von Zusammenhängen zwischen Expositionen und Krankheit jedoch

nicht von Bedeutung ist und Chancenverhältnisse ein adäquates Mittel zur Beurteilung von Zusammenhängen sind, stellt dieser Aspekt keinen großen Nachteil dar.

Es darf allerdings nicht unberücksichtigt bleiben, dass unter den Beobachtungsstudien speziell das Design von Fall-Kontroll-Studien in Bezug auf die praktische Umsetzung mit Problemen verbunden ist. So zum Beispiel gilt für zeitlich vorausgegangene Expositionen, dass die Datenerhebung nur auf Grundlage von Befragungen der Studienteilnehmer erfolgen kann. Hierbei müssen sich die Studienteilnehmer unter Umständen an weit zurückliegende Ereignisse erinnern, was ihnen erfahrungsgemäß nicht immer möglich ist, so dass keine oder falsche Angaben gemacht werden. Ergebnisverzerrungen, die aus falschen bzw. fehlerhaften Informationen resultieren, werden in der Fachliteratur unter dem Begriff Information-Bias zusammengefasst.
Besonders problematisch ist es, wenn für Fälle und Kontrollen eine unterschiedliche Erinnerungsfähigkeit oder –bereitschaft vorliegt. Diese spezielle Form des Information-Bias wird als Recall-Bias bezeichnet. Zum Beispiel ist insbesondere bei schweren Krankheiten zu erwarten, dass die erkrankten Personen (Fälle) bereits vor der Befragung intensiv über mögliche Ursachen ihrer Erkrankung nachgedacht haben, so dass sie im Vergleich zu den Kontrollpersonen zuverlässigere Angaben zu ihren Expositionen machen können.
Eine weitere Form des Information-Bias, der bei persönlichen Interviews auftritt, ist der sogenannte Interviewer-Bias. Unter diesem Begriff werden in der Epidemiologie alle Ergebnisverzerrungen zusammengefasst, die daraus resultieren, dass Fälle und Kontrollen vom Interviewer nicht gleich behandelt werden. Besteht von seitens des Interviewers zum Beispiel der Verdacht, dass ein bestimmter Faktor einen schädlichen Einfluss hat, kann dies dazu führen, dass er durch bewusstes oder unbewusstes Nachfragen bei den Fällen viel eher in Erinnerung ruft, dass sie im Verlaufe ihres Lebens einer bestimmten Exposition ausgesetzt waren. In Folge dessen ergibt sich für die Kontrollgruppe ein größerer Anteil fälschlicherweise als nicht exponiert eingestufter Personen als in der Fallgruppe und es kommt zu einer Überschätzung der Bedeutung dieses Faktors.

Weitere Fehlerquellen, die die Planung und Durchführung von Beobachtungsstudien betreffen, können der Fachliteratur entnommen werden. Sehr ausführlich beschäftigen sich zum Beispiel Rothman und Greenland (1998) mit dieser Problematik.

Kapitel 4. Methodik

In diesem Kapitel werden die statistischen Methoden beschrieben, die bei der Auswertung des Datenmaterials zur Anwendung kommen.

In Unterkapitel 4.1 wird ein statistisches Testverfahren vorgestellt, mit welchem überprüft werden kann, ob zwei kategoriale Merkmale voneinander abhängig sind. Daran anschließend wird in Unterkapitel 4.2 das Odds-Ratio als Assoziationsmaß vorgestellt. Mit einer ausführlichen Darstellung des logistischen Regressionsmodells im letzten Unterkapitel (4.3) endet der Methodikteil.

4.1 Chi-Quadrat-Test auf Unabhängigkeit

Die gemeinsame empirische Häufigkeitsverteilung zweier kategorialer Merkmale A und B kann in einer Kontingenztafel dargestellt werden. Ausgehend von einer bivariaten Zufallsstichprobe vom Umfang n sind dazu, die absoluten Häufigkeiten sämtlicher Merkmalskombinationen in einem 2-dimensionalen Schema abzutragen.
Davon ausgehend, dass die Merkmale A und B die Ausprägungen $A_1,\ldots,A_k$ bzw. $B_1,\ldots,B_w$ aufweisen können, werden aus der n-Stichprobe zunächst die absoluten Häufigkeit $n_{i,j}$ aller möglichen Ausprägungspaare (A_i,B_j) ermittelt und in einer Kontingenztafel (vgl. Schema 4.1.A) dargestellt ($i\in\{1,\ldots,k\}$, $j\in\{1,\ldots,w\}$).

<u>Schema 4.1.A</u>: Schema einer Kontingenztafel

	$\mathbf{B_1}$	$\mathbf{B_2}$	…	$\mathbf{B_w}$	$\mathbf{\Sigma}$
$\mathbf{A_1}$	$n_{1,1}$	$n_{1,2}$	…	$n_{1,w}$	$n_{1,.}$
$\mathbf{A_2}$	$n_{2,1}$	$n_{2,2}$		$n_{2,w}$	$n_{2,.}$
⋮	⋮	⋮		⋮	⋮
$\mathbf{A_k}$	$n_{k,1}$	$n_{k,2}$	…	$n_{k,w}$	$n_{k,.}$
$\mathbf{\Sigma}$	$n_{.,1}$	$n_{.,2}$	…	$n_{.,w}$	n

Die $n_{.,j}$ werden als Spaltensummen bezeichnet und ergeben sich durch Summation über die Zellen der j-ten Spalte (j=1,…,w). Es gilt die Beziehung:

$$n_{.,j} = \sum_{i=1}^{k} n_{i,j} \text{ (j=1,…,w).}$$

Analog ergeben sich die Zeilensummen $n_{i,.}$ (i=1,…,k) durch Summation über die Zellen der i-te Zeile (i=1,…,k):

$$n_{.i,.} = \sum_{j=1}^{w} n_{i,j} \text{ (i=1,…,k).}$$

Die unter Unabhängigkeit von A und B zu erwartenden Häufigkeiten $N_{i,j}$ der Paare (A_i,B_j) ergeben sich für $(i,j) \in \{1,\ldots,k\} \times \{1,\ldots,w\}$ wie folgt:

$$N_{i,j} = n \cdot \hat{P}(\{A = A_i, B = B_j\}) = n \cdot \hat{P}(\{A = A_i\}) \cdot \hat{P}(\{B = B_j\}) = n \cdot \frac{n_{i,.}}{n} \cdot \frac{n_{.,j}}{n} = \frac{n_{i,.} \cdot n_{.,j}}{n}.$$

Anhand der Abweichungen zwischen den tatsächlich beobachteten Häufigkeiten $n_{i,j}$ und den (unter Unabhängigkeit) zu erwartenden Häufigkeiten $N_{i,j}$ kann beurteilt werden, inwieweit Assoziationen zwischen bestimmten Ausprägungen der beiden Merkmale vorliegen. Aus Gründen der Übersichtlichkeit werden dazu neben den beobachteten Zellhäufigkeiten $n_{i,j}$ üblicherweise auch die unter Unabhängigkeit zu erwartenden Häufigkeiten $N_{i,j}$ (z.B. in Klammern) in die betreffenden Zellen der Kontingenztafel eingetragen.

Formal kann mit Hilfe eines asymptotischen Chi-Quadrat-Tests überprüft werden, ob die Merkmale A und B abhängig voneinander sind. Nach Büning und Trenkler (1994) ist die Statistik:

$$X^2 := \sum_{i=1}^{k} \sum_{j=1}^{w} \frac{(n_{i,j} - N_{i,j})^2}{N_{i,j}}$$

unter der Unabhängigkeit von A und B approximativ chi-quadrat-verteilt mit (k-1)·(l-1) Freiheitsgraden. Da X^2 den Unterschied zwischen den tatsächlich beobachteten und unter Unabhängigkeit zu erwartenden Häufigkeiten misst, sprechen große Werte gegen die Unabhängigkeit von A und B.

Entsprechend kann die Nullhypothese:

H_0: „Die Merkmale A und B sind unabhängig voneinander!“

mit einer Irrtumswahrscheinlichkeit von α zu Gunsten der Alternativhypothesen:

H_1: „Es besteht eine Abhängigkeit (ein Zusammenhang) zwischen den Merkmalen A und B!“

verworfen werden, wenn die realisierte Teststatistik x^2 das $(1-\alpha)$-Quantil der Chi-Quadrat-Verteilung mit (k-1)·(l-1) Freiheitsgraden überschreitet.

Der P-Wert bzw. Überschreitungswert: $p = P(\chi^2_{(k-1)\cdot(w-1)} > x^2)$ stellt eine deskriptive Größe dar, anhand derer der Grad der Abhängigkeit zwischen A und B beurteilt werden kann. In Unterkapitel 4.3.5 wird in Zusammenhang mit dem logistischen Regressionsmodell ausführlicher auf die Bedeutung und Interpretation von P-Werten eingegangen.

Voraussetzungen für die Anwendbarkeit des Chi-Quadrat-Tests auf Unabhängigkeit sind ein hinreichend großer Stichprobenumfang n und die Unabhängigkeit der Untersuchungseinheiten. Erst bei Vorliegen einer hinreichend großen, unabhängigen Zufallsstichprobe aus der gemeinsamen Verteilung von A und B ist die Teststatistik X^2 (unter H_0) approximativ chi-quadrat-verteilt (Büning und Trenkler 1994). Über die notwendige Mindestgröße des Stichprobenumfangs bestehen unterschiedliche Auffassungen. Yarnold (1970) empfiehlt zum Beispiel den Chi-Quadrat-Test nur dann anzuwenden, wenn für alle beobachteten Zellhäufigkeiten $n_{i,j}$ gilt, dass $n_{i,j} \geq 5 \cdot n_0/n$, wobei n_0 die Anzahl Zellen mit Zellhäufigkeiten kleiner oder gleich 5 repräsentiert.

4.2 Das Odds-Ratio Assoziationsmaß

Der Epidemiologie dienen vorrangig zwei Assoziationsmaße, um den Zusammenhang zwischen den Expositionen und einer bestimmten Erkrankung zu beschreiben (Schach und Kreienbrock 2000).
Im folgenden wird der Krankheitszustand eines Individuums durch die Indikatorvariable D beschrieben. Diese kennzeichnet, ob eine Person an der zu untersuchenden Krankheit leidet (D=1) oder nicht (D=0).

Das relative Risiko (RR) ist als Quotient von Erkrankungswahrscheinlichkeiten definiert. Ausgehend von zwei Erkrankungswahrscheinlichkeiten $P(D=1|E_1)$ und $P(D=1|E_2)$, die für unterschiedliche Ausprägungen E_1 und E_2 der zu untersuchenden Expositionen vorliegen, ergibt sich für das relative Risiko ein Wert von:

$$RR(E_1, E_2) = \frac{P(D=1 \mid E_1)}{P(D=1 \mid E_2)}.$$

Das relative Risiko entspricht dem Faktor, um den sich die Erkrankungswahrscheinlichkeit erhöht (RR>1) oder verringert (RR<1), wenn eine Person anstelle des Expositionsmusters E_2 das Muster E_1 aufweist.

Beim Odds-Ratio (OR) werden anstelle der Erkrankungswahrscheinlichkeiten die dazugehörigen Erkrankungschancen zueinander ins Verhältnis gesetzt.
Die Chance (Odds) eines Ereignisses entspricht dem Verhältnis von Ereigniswahrscheinlichkeit (p) zur Gegenwahrscheinlichkeit (1-p): Odds(p) $= \frac{p}{1-p}$.

Für die beiden unterschiedlichen Ausprägungen der Expositionen E_1 und E_2 ergibt sich folglich für das Odds-Ratio:

$$OR(E_1, E_2) = \frac{Odds\{P(D=1 \mid E_1)\}}{Odds\{P(D=1 \mid E_2)\}} = \frac{\frac{P(D=1 \mid E_1)}{1-P(D=1 \mid E_1)}}{\frac{P(D=1 \mid E_2)}{1-P(D=1 \mid E_2)}} = \frac{P(D=1 \mid E_1) \cdot P(D=0 \mid E_2)}{P(D=0 \mid E_1) \cdot P(D=0 \mid E_1)}.$$

Das Odds-Ratio entspricht also dem Faktor, um den sich die Erkrankungschance erhöht (OR>1) oder verringert (OR<1), wenn eine Person anstelle des Expositionsmusters E_2 das Muster E_1 aufweist.

Nach Schach und Kreienbrock (2000) ist das relative Risiko aufgrund seiner einfachen Interpretierbarkeit das bevorzugte Vergleichsmaß zur Beurteilung der Assoziation zwischen Expositionen und Krankheit. Aus den Daten einer Fall-Kontroll-Studie können jedoch ausschließlich Chancenverhältnisse (Odds-Ratios) und keine relativen Risiken geschätzt werden. Entsprechende Bemerkungen wurden bereits in Unterkapitel 3.2 gemacht. Im Zusammenhang mit der logistischen Regression wird in Unterkapitel 4.3.8 noch einmal etwas ausführlicher auf dieses Problem eingegangen.

4.3 Theorie der logistischen Regression

In diesem letzten Unterkapitel wird ausführlich auf die Theorie der logistischen Regression eingegangen. Logistische Regressionsmodelle sind wichtige Instrumentarien in der analytischen Epidemiologie, da diese es ermöglichen, gleichzeitig den Einfluss mehrerer Faktoren auf das Erkrankungsrisiko zu untersuchen. Der wesentliche Vorteil gegenüber anderen Auswerteverfahren besteht darin, dass durch die simultane Betrachtung der Faktoren nicht der Fall auftreten kann, dass das unkontrollierte Wirken, vernachlässigter Faktoren, zu Ergebnisverzerrungen führt. Bei anderen Verfahren, die nur den Einfluss eines einzigen oder einiger weniger Expositionen gleichzeitig untersuchen, besteht stets die Gefahr, dass ein nichtberücksichtigter Faktor sowohl mit den Expositionen als auch der Krankheit vermengt ist, und sich

somit der Beziehung zwischen den Expositionen und der Krankheit in störender Weise überlagert.

Die Abschnitte 4.3.1 bis 4.3.7 sind vorrangig der mathematischen Theorie und Interpretation des logistischen Regressionsmodells gewidmet. Die Darstellungen in diesen Abschnitten orientieren sich vorwiegend an Kleinbaum (1993) und Hosmer und Lemeshow (1989). In Abschnitt 4.3.8 wird erläutert, welche Einschränkungen sich ergeben, wenn die Daten einer Fall-Kontroll-Studie mit Hilfe des logistischen Regressionsmodells auszuwerten sind. Abschließend werden im letzten Abschnitt (4.3.9) zwei spezielle Ad-hoc-Verfahren im Umgang mit fehlenden Dateneinträgen vorgestellt.

4.3.1 Das logistische Regressionsmodell

Mit Hilfe des logistischen Regressionsmodells kann die Wahrscheinlichkeit P für das Auftreten einer Krankheit D in Abhängigkeit von mehreren Einflussvariablen $X_1,\ldots,X_k$ modelliert werden. Die Beziehung zwischen der Erkrankungswahrscheinlichkeit P=P(D=1) und den Einflussvariablen wird dabei durch die folgende Modellgleichung beschrieben:

$$P = P(D=1 \mid X_1,...,X_k) = \frac{\exp(\beta_0 + \sum_{i=1}^{k} \beta_i \cdot X_i)}{1+\exp(\beta_0 + \sum_{i=1}^{k} \beta_i \cdot X_i)}. \qquad (1)$$

Hierbei repräsentieren die β_i unbekannte Modellparameter ($i \in \{0,\ldots,k\}$). Unter Verwendung der Vektornotationen $X:=(1,X_1,\ldots,X_k)^T \in \mathbf{R}^{k+1}$ und $\beta:=(\beta_0,\beta_1,\ldots,\beta_k)^T \in \mathbf{R}^{k+1}$ kann die Modellgleichung (1) in kompakter Vektorschreibweise angegeben werden:

$$P = P(D=1 \mid X) = \frac{\exp(X^T \cdot \beta)}{1+\exp(X^T \cdot \beta)} = \frac{1}{1+\exp(-X^T \cdot \beta)}. \qquad (2)$$

Und durch elementare Umformungen ergibt sich die äquivalente Beziehung:

$$LOGIT(P) := \log_e\left(\frac{P}{1-P}\right) = \log_e\left(\frac{P(D=1 \mid X)}{1-P(D=1 \mid X)}\right) = X^T \cdot \beta. \qquad (3)$$

Das logistische Regressionsmodell unterstellt also für eine Transformation der Wahrscheinlichkeit P(D=1|X) ein gewöhnliches multiples lineares Regressionsmodell. Die Transformation wird als Logit-Transformation von P(D=1|X) bezeichnet und bildet Wahrscheinlichkeiten $P \in (0,1)$ bijektiv auf die reellen Zahlen ab. Anders ausgedrückt, durchläuft Logit(P) für $P \in (0,1)$ alle reellen Zahlen, so dass jeder Erkrankungswahrscheinlichkeit eineindeutig eine reelle Zahl zugeordnet wird. Damit ist ersichtlich, dass im logistischen Regressionsmodell (bei der

Modellschätzung) keine Restriktionen an die Modellparameter gestellt werden müssen. Jeder Wert der Linearkombination $X^T \cdot \beta \in \mathbf{R}$ entspricht der Logit-Transformation einer eindeutigen Wahrscheinlichkeit P. Diese Wahrscheinlichkeit kann unter Zuhilfenahme von Beziehung (2) berechnet werden.

Im Hinblick auf die Auswertung epidemiologischer Studien ist darüber hinaus von Bedeutung, dass die Modellparameter in engem Zusammenhang zum Odds-Ratio stehen. Das logistische Regressionsmodell bietet also nicht nur den Vorteil, mehrere Einflussvariablen simultan betrachten zu können, sondern ermöglicht auch eine einfache Interpretation der Modellparameter. Ausgehend von Gleichung (2) ergibt sich zunächst als Wahrscheinlichkeit für die Nichterkrankung {D=0}:

$$P(D=0 \mid X) = 1 - P(D=1 \mid X) = 1 - \frac{1}{1+\exp(-X^T \cdot \beta)} = \frac{\exp(-X^T \cdot \beta)}{1+\exp(-X^T \cdot \beta)},$$

und damit für die Erkrankungschance Odds(P(D=1|X)) die folgende Produktdarstellung:

$$Odds(P(D=1 \mid X)) := \frac{P(D=1 \mid X)}{P(D=0 \mid X)} = \exp(X^T \cdot \beta) = \exp(\beta_0) \cdot \prod_{i=1}^{k} \exp(X_i \cdot \beta_i).$$

Entsprechend ergibt sich durch Quotientenbildung für zwei unterschiedliche Realisierungen $x=(1,x_1,\ldots,x_k)^T$ und $y=(1,y_1,\ldots,y_k)^T$ von X das folgende Erkrankungschancenverhältnis bzw. Odds-Ratio OR(x,y):

$$OR(x,y) = \frac{Odds\{P(D=1 \mid x)\}}{Odds\{P(D=1 \mid y)\}} = \frac{\exp(\beta_0) \cdot \prod_{i=1}^{k} \exp(\beta_i \cdot x_i)}{\exp(\beta_0) \cdot \prod_{i=1}^{k} \exp(\beta_i \cdot y_i)} = \prod_{i=1}^{k} \exp\{(x_i - y_i) \cdot \beta_i\}. \qquad (4)$$

Aus dieser Darstellung (4) ist bereits ersichtlich, dass sich das Odds-Ratio im logistischen Modell als Produkt von k Einzeleffekten ergibt. Jeder dieser multiplikativen Effekte korrespondiert zu einer Variablen X_i (i=1,…,k) und hängt ausschließlich vom zugehörigen Modellparameter β_i und der Differenz der Variablenwerte (x_i-y_i) ab.

Bei der Interpretation dieser Effekte sind sowohl die Wertebereiche der Einflussvariablen, als auch ggf. vorliegende, funktionale Beziehungen (Abhängigkeiten) zwischen den Einflussvariablen, zu berücksichtigen. Da im nächsten Unterkapitel ausführlicher auf diese Probleme eingegangen wird, soll in diesem Unterkapitel nur noch der einfachste Fall betrachtet werden.
Bei Unabhängigkeit der Einflussvariablen besteht kein Zusammenhang zwischen den Komponenten von X, so dass $X_1,\ldots,X_k$ frei wählbar sind. Das heißt insbesondere, dass die Festle-

gung von X_i keinen Einfluss auf die anderen Variablen hat (i=1,…,k). In dieser Situation kann angenommen werden, dass sich x und y nur in einer einzigen Komponente unterscheiden:
$x_i \neq y_i$ und $x_j = y_j$ $(j \in \{1,\ldots,k\} \setminus \{i\}$ für ein $i \in \{1,\ldots,k\})$.

Durch Einsetzen in (4) ergibt sich dann für das Odds-Ratio:

$$OR(x_i,y_i) := \exp\{(x_i - y_i)\cdot\beta_i\}. \qquad (5)$$

Dieser Ausdruck charakterisiert die Auswirkung der i-ten Variablen auf die Erkrankungschance bei gleichzeitiger Berücksichtigung (Adjustierung) aller anderen im Modell befindlichen Variablen.
Durch $\exp\{(x_i - y_i)\cdot\beta_i\}$ ist also der multiplikative Effekt auf die Erkrankungschance gegeben, wenn - unabhängig von den Werten der anderen Variablen - die i-te Variable anstelle des Wertes y_i den Wert x_i aufweist. In den epidemiologischen Anwendungen kann so der Effekt einer Einflussvariablen (Exposition) auf die Erkrankungschance beurteilt werden, ohne dass die Beziehung durch die anderen Einflussvariablen (Störgrößen) gestört wird.

4.3.2 Variablencodierung und Interpretation der Modellparameter

Im Rahmen epidemiologischer Studien werden Probandenmerkmale erhoben und als potentielle Risikofaktoren für eine bestimmte Krankheit untersucht. Bei diesen Merkmalen kann es sich um kontinuierliche Messgrößen aber auch um kategoriale Daten mit nominalen Ausprägungen handeln. Da i.a. jedoch nicht davon auszugehen ist, dass die Merkmale geeignete Einflussgrößen für ein logistisches Regressionsmodell darstellen, beschäftigt sich dieser Abschnitt mit der Aufbereitung von Rohdaten. Hierbei bestehen in Abhängigkeit vom Messniveau unterschiedliche Möglichkeiten, aus den beobachteten Merkmalen passende Einflussvariablen zu generieren. Auf Wechselwirkungen zwischen Einflussvariablen wird erst in Unterkapitel 4.3.3 eingegangen.

Stetige Merkmale

Ein kardinales Merkmal X (als kontinuierliche Messgröße) kann unmittelbar als Einflussvariable im logistischen Regressionsmodell verwendet werden, wenn eine monotone Beziehung zwischen den Werten des Merkmals und der Erkrankungswahrscheinlichkeit zu erwarten ist. Aus der Beziehung (5) folgt unmittelbar: $OR(x_0+x,x_0) = OR(x) = \exp\{x\cdot\beta_X\}$. Das heißt, ein Anstieg des Merkmalwertes um x Einheiten führt unabhängig vom Ausgangswert x_0 immer

zu demselben multiplikativen Effekt $\exp\{x\cdot\beta(X)\}$auf die Erkrankungschance. Erscheint eine solche Beziehung für ein Merkmal X nicht realistisch, stehen zwei Problemlösungen zur Auswahl.

1) Zusätzliche Transformationen:

Erstens kann neben dem Merkmal X auch eine geeignete Transformation dieses Merkmals Y=T(X) als Einflussvariable ins Modell aufgenommen werden. Dies impliziert eine funktionale Abhängigkeit zwischen den beiden Einflussgrößen X und Y, so dass ihre multiplikativen Effekte nicht mehr unabhängig voneinander betrachtet werden können. Das Odds-Ratio für eine Differenz von x Einheiten im Merkmal X ist dann vom Ausgangswert x_0 abhängig und durch den Ausdruck $OR(x_0+x,x_0)= \exp\{x\cdot\beta_X + [T(x+x_0)-T(x_0)]\cdot\beta_{T(X)}\}$ gegeben. Am häufigsten wird X zusammen mit der Transformation $Y:=T(X)=X^2$ verwendet. Dies führt zu dem multiplikativen Effekt $\exp\{X\cdot\beta_X+X^2\cdot\beta_Y\}$ auf die Erkrankungschance. Der Exponent $X\cdot\beta_X+X^2\cdot\beta_Y$ stellt eine Parabel in X dar, so dass sich in Abhängigkeit von den Modellparametern das Monotonieverhalten ändert. Gilt zum Beispiel $\beta_Y<0$ liegt für den X-Wert$(-0.5)\cdot\beta_X/\beta_Y$ die größte Erkrankungschance vor. In Abhängigkeit vom Abstand zu diesem Wert nimmt die Erkrankungschance in beide Richtungen (kleinere/größere X-Werte) gleichmäßig ab.
Ausführlichere Diskussionen zu Transformationen finden sich z.B. in Kleinbaum et al. (1982).

2) Umwandlung in kategoriales Merkmal:

Zweitens kann das stetige Merkmal Y in ein kategoriales umgewandelt werden. Dazu wird der stetige Wertebereich des Merkmals in $m(\geq 2)$ substanzwissenschaftlich sinnvolle Unterbereiche zerlegt. Diese Unterbereiche bilden die Kategorien und das Y-Merkmal wird durch eine neue Variable X ersetzt, die nur noch angibt, in welcher Kategorie der Y-Wert liegt. Die neue Variable X wird anschließend wie ein kategoriales Merkmal (siehe unten) behandelt. Dies führt dazu, dass den m Unterbereichen unabhängige Effekt zukommen. Abgesehen von dem Informationsverlust, der mit der Kategorisierung einhergeht, ergibt sich der Nachteil, dass kategoriale Variablen mit mehr als zwei Ausprägungen m-2 zusätzliche Modellparameter erforderlich machen, was sich negativ auf die Schätzgenauigkeit auswirkt.

Kategoriale Merkmale

Die Ausprägungen kategorialer Merkmale müssen zuerst in geeignete Zahlenwerte transformiert werden. Ein dichotomes kategoriales Merkmal Y mit den beiden Ausprägungen A und

B kann im Rahmen eines logistischen Regressionsmodells in Form der folgenden Indikatorfunktion X als Einflussvariable berücksichtigt werden:

$$X = X(Y) = \begin{Bmatrix} 0, & Y = A \\ & wenn \\ 1, & Y = B \end{Bmatrix}.$$

Die {0,1}-Kodierung ermöglicht eine einfache Interpretation des zugehörigen Modellparameters β_X. Mit (5) ergibt sich für das Erkrankungschancenverhältnis von Ausprägung B zu A OR(B,A)= exp{1·β_X}. Das heißt, auf die Erkrankungschance wirkt der multiplikative Effekt exp{β_X}, wenn statt A die Merkmalsausprägung B vorliegt. Ein Beispiel für ein dichotomes kategoriales Merkmal ist das Geschlechtsmerkmal mit den beiden Ausprägungen „männlich“ und „weiblich“.

Kategoriale Merkmale mit mehr als zwei nominalen Ausprägungen können im logistischen Regressionsmodell nicht durch eine einzige Einflussvariable repräsentiert werden. Da die m Kategorien $K_1,\ldots,K_m$ nicht quantitativ vergleichbar sind, würde die Transformation in einen Zahlencode (z.B.: 0,1,2,3,…,m) fälschlicherweise eine Ordnung festlegen und somit Abhängigkeiten zwischen den Effekten unterstellen, die realiter nicht gegeben sind. Kodiert man zum Beispiel ein Merkmal Y mit den Ausprägungen A, B und C durch (A,B,C)→(0,1,2) und behandelt die so entstandene Variable anschließend als stetige, ergeben sich aufgrund der Beziehung (5) OR(x,y)=exp{(x-y)·β(Y)} im Modell die Restriktionen:
$OR(C=2,A=0) = OR(B=1,A=0)^2 = OR(C=2,B=1)^2 = \exp\{\beta(Y)\}^2$.
Für die Vermeidung von Restriktionen dieser Art ist es erforderlich, ein Merkmal mit m Kategorien im Modell durch (m-1) künstliche Indikator-Einflussvariablen $X_1,\ldots,X_{m-1}$ wiederzugeben. Bei der sogenannten Referenzkodierung sind die X_i wie folgt definiert:

$$X_i = \begin{Bmatrix} 1, & Y = K_i \\ & wenn \\ 0, & Y \neq K_i \end{Bmatrix} \text{ (i=1,…,m-1).}$$

Die m-te Kategorie wird als Referenzkategorie bezeichnet. Diese Parametrisierung, die eine Verallgemeinerung der {0,1}-Kodierung von dichotomen Merkmalen darstellt, führt zu einer einfachen Interpretierbarkeit der Modellparameter. Die Ausprägung K_m dient als Referenzwert (X_i=0 (i=1,…,m-1)) und für i=1,…,m-1 ergibt sich: $OR(K_i, K_m) = \exp\{\beta(X_i)\}$, so dass sich für Personen mit der Merkmalsausprägung $Y=K_i$ im Vergleich zu Personen mit dem Re-

ferenzwert $Y=K_m$ eine $\exp\{\beta(X_i)\}$-fache Erkrankungschance ergibt. Für zwei beliebige Kategorien K_i und K_j ($i,j \in \{1,\ldots,m-1\}$) ergibt sich dagegen mit (5) ein Odds-Ratio von

$$OR(K_i,K_j)=\exp\{\beta_i-\beta_j\}.$$

Da die Berücksichtigung eines Merkmals mit m Kategorien über (m-1) Einflussvariablen erfolgt, müssen auch (m-1) Parameter geschätzt werden. In Anbetracht der Tatsache, dass die Varianz der Schätzer mit der Anzahl zu schätzender Parameter steigt, sollte insbesondere bei großem m überlegt werden, ob Kategorien sinnvoll zusammengefasst werden können. Durch die Zusammenfassung von zwei Kategorien kann jeweils ein Modellparameter eingespart werden.

4.3.3 Wechselwirkungen

Ganz allgemein liegt eine Wechselwirkung (Interaktion) zwischen zwei Einflussfaktoren genau dann vor, wenn ihr gemeinsames Wirken einen verstärkenden oder abschwächenden Effekt auf eine Zielvariable ausübt. Bezogen auf das logistische Regressionsmodell bedeutet dies, dass sich zwei interagierende Einflussvariablen X_1 und X_2 gegenseitig in ihrer Wirkung auf die Erkrankungschance Odds(P(D=1)) beeinflussen.

Die Beziehung $\text{Odds}(P(D=1|X_1,X_2) = \exp\{\beta_0\} \cdot \exp\{X_1 \cdot \beta(X1)\} \cdot \exp\{X_2 \cdot \beta(X_2)\}$ unterstellt im logistischen Regressionsmodell, dass sich die gemeinsame Wirkung von X_1 und X_2 durch Multiplikation der Einzeleffekte ergibt. Ist nicht davon auszugehen, dass X_1 und X_2 unabhängige multiplikative Effekte auf die Erkrankungschance haben, kann dies im Rahmen von logistischen Regressionsmodellen nur zu einem gewissen Grade berücksichtigt werden. Dazu ist das Produkt der beiden interagierenden Variablen als zusätzliche Einflussvariable $X_3:=X_1 \cdot X_2$ in die Modellgleichung (3) aufzunehmen. Dies führt zu folgender Beziehung zwischen der Erkrankungschance und den beiden Einflussvariablen:

$$\text{Odds}(P(D=1|X_1,X_2))= \exp\{\beta_0\} \cdot \exp\{X_1 \cdot \beta(X_1)\} \cdot \exp\{X_2 \cdot \beta(X_2)\} \cdot \exp\{X_1 \cdot X_2 \cdot \beta(X_3)\}.$$

Der zusätzliche (Wechselwirkungs-)Effekt modifiziert zwar das Produkt der Einzeleffekte, erfasst aber ausschließlich Abweichungen vom multiplikativen Zusammenwirken der Einzeleinflüsse. Ist ein solcher Zusammenhang aus substanzwissenschaftlicher Sicht noch immer nicht problemadäquat, muss die Datenauswertung mit Hilfe umfassenderer Modellklassen vorgenommen werden. Kleinbaum el al. (1982) diskutieren zum Beispiel eine Modellklasse, in der die Effekte der Einflussfaktoren auf die Erkrankungschance additiv zusammenwirken.

Die Berücksichtigung einer Wechselwirkung $X_1 \cdot X_2$ im logistischen Regressionsmodell impliziert, dass die Effekte der Variablen X_1 und X_2 nicht mehr unabhängig voneinander sind. In Abhängigkeit von den Werten der einen Variablen ergeben sich unterschiedliche Effekte der anderen Variablen.

Liegt zum Beispiel für X_2 der Wert z vor, ergibt sich aus (4) bzw. (5) für die Werte x und y von X_1 ein Odds-Ratio $OR(X_1=x,X_1=y|X_2=z)$ von $\exp\{(x-y)\cdot[\beta(X_1)+z\cdot\beta(X_3)]\}$.

Unter Zuhilfenahme der Definition eines „Pseudoparameters“ $\beta(z):= \beta(X_1)+z\cdot\beta(X_3)$ ist ersichtlich, dass dies für jeden festen Wert z eine zu (4) äquivalente Beziehung beschreibt: $OR(X_1=x,X_1=y|X_2=z)= \exp\{(x-y)\cdot\beta(z)\}$. Daher kann für feste z-Werte, die Interpretation analog zu den Ausführungen in Unterkapitel 3.3.2 (in Abhängigkeit vom X_1-Messniveau) erfolgen.

4.3.4 Modellschätzung

Erscheint ein logistisches Modell aus substanzwissenschaftlicher Sicht geeignet, um den Zusammenhang zwischen einer bestimmten Erkrankungswahrscheinlichkeit und k Variablen zu beschreiben, müssen in einem ersten Auswertungsschritt die unbekannten Modellparameter statistisch geschätzt werden. Erst nach der Modellschätzung kann unter Zuhilfenahme weiterer statistischer Methoden beurteilt werden, welche Variablen (im Rahmen eines solchen Modells) tatsächlich Einfluss auf die Erkrankungswahrscheinlichkeit nehmen und, ob die Variablen in ihrer Gesamtheit einen hinreichenden Erklärungswert für die Erkrankungswahrscheinlichkeit haben. Die statistische Auswertung erfolgt auf Datenbasis, so dass in epidemiologischen Anwendungsgebieten entsprechende Probandendaten zu erheben sind. Für jeden Probanden $i\in\{1,\ldots,n\}$ muss der Krankheitszustand $d_i\in\{0,1\}$ und die Werte der k Einflussvariablen ($x_{i,j}$: $j=1,\ldots,k$) bekannt sein. Da die Einflussvariablen unter Zuhilfenahme der in den Unterkapiteln 4.3.2 und 4.3.3 vorgestellten Methoden aus den Rohdaten generiert werden, sind dabei funktionale Beziehungen zwischen den Einflussvariablen nicht ausgeschlossen (z.B. $x_{i,3}=x_{i,1}\cdot x_{i,2}$ (Wechselwirkung) oder $x_{i,4}=(x_{i,3})^2$ (Transformation)). Abhängigkeiten dieser Art müssen zwar bei der Modellinterpretation berücksichtigt werden (vgl. 4.3.2 und 4.3.3), im Hinblick auf die statistische Methodik ergeben sich jedoch keine Unterschiede.

Die Modellparameter $\beta_0,\ldots,\beta_k$ werden mit Hilfe der Maximum-Likelihood-Methode (ML-Methode) geschätzt. Basis für diese Schätzmethode ist die Likelihood-Funktion, die für jeden Parametervektor $\beta=(\beta_0,\ldots\beta_k)^T$ die Wahrscheinlichkeit angibt, dass sich bei den Probanden die beobachteten Krankheitszustände $d=(d_1,\ldots,d_n)$ einstellen. Der Maximum-Likelihood-Schätzer

ist gerade der Vektor $\hat{\beta}$, der die Likelihood-Funktion maximiert. Das bedeutet, der ML-Schätzung ist der Parametervektor, für den die beobachteten Daten am wahrscheinlichsten sind.

Für einen konkreten Modellparametervektor β ergibt sich mit der Notation $x_i=(1,x_{i,1},\ldots,x_{i,k})^T$ zunächst als bedingte Wahrscheinlichkeit, dafür dass der i-te Proband seinen Krankheitsstatus d_i aufweist:

$$P(D=d_i \mid x_i,\beta)=P(D=0 \mid x_i,\beta)^{1-d_i}\cdot P(D=1 \mid x_i,\beta)^{d_i}$$

$$=\left(\frac{\exp(-x_i^T\cdot\beta)}{1+\exp(-x_i^T\cdot\beta)}\right)^{1-d_i}\cdot\left(\frac{1}{1+\exp(-x_i^T\cdot\beta)}\right)^{d_i}$$

$$=\left(\frac{1}{1+\exp(-x_i^T\cdot\beta)}\right)^{1}\cdot\exp(-x_i^T\cdot\beta)^{1-d_i}$$

$$=\left(\frac{\exp(-x_i^T\cdot\beta)}{1+\exp(-x_i^T\cdot\beta)}\right)\cdot\exp(x_i^T\cdot\beta)\cdot\exp(-x_i^T\cdot\beta)^{1-d_i}$$

$$=\left(\frac{\exp(-x_i^T\cdot\beta)}{1+\exp(-x_i^T\cdot\beta)}\right)\cdot\exp(x_i^T\cdot\beta)^{d_i}=(1+\exp(x_i^T\cdot\beta))^{-1}\cdot\exp(x_i^T\cdot\beta)^{d_i}.$$

Davon ausgehend, dass die Probanden unabhängig voneinander erkranken oder nicht, ergibt sich die Likelihood-Funktion als Produkt der n individuellen Erkrankungswahrscheinlichkeiten:

$$L(\beta)=\prod_{i=1}^{n}(1+\exp(x_i^T\cdot\beta))^{-1}\cdot\prod_{i=1}^{n}\exp(x_i^T\cdot\beta)^{d_i}. \qquad (6)$$

Der ML-Schätzung $\hat{\beta}$ ergibt sich durch Maximierung dieser Funktion in Abhängigkeit vom Modellparametervektor β. Aufgrund der funktionalen Struktur von L kann dieses Maximierungsproblems nur durch numerische Verfahren gelöst werden. Üblicherweise kommt das Newton-Raphson-Verfahren zum Einsatz.

Allgemeine Darstellung des Newton-Raphson-Verfahren

Bei dem Newton-Raphson-Verfahren (vgl. Künzi et al. 1962) handelt es sich um ein iteratives numerisches Verfahren zur Lösung eines Maximierungsproblems. Es wird in den folgenden Darstellungen davon ausgegangen, dass eine Zielfunktion L=L(β) in $\beta\in\mathbf{R}^{k+1}$ zu maximieren ist.

Grundliegend für das Newton-Raphson-Verfahren sind Taylor-Entwicklungen zweiter Ordnung, so dass das Verfahren nur dann anwendbar ist, wenn Gradient und Hesse-Matrix der Zielfunktion L in jedem Vektor $\beta \in \mathbf{R}^{k+1}$ berechnet werden können. Ausgehend von einem Startvektor $\beta^{(0)}$, wird beim NR-Verfahren in jedem Iterationsschritt aus dem aktuellen Vektor $\beta^{(i)}$ ein neuer Vektor $\beta^{(i+1)}$ bestimmt. In jeder Iteration wird dafür die Zielfunktion durch ihre Taylorreihe in $\beta^{(i)}$ approximiert. Mit Hilfe der Differentialrechnung wird anschließend die mögliche Extremalstelle der Taylorapproximation bestimmt, und als neuer Parametervektor $\beta^{(i+1)}$ gewählt. Die Iteration endet, wenn ein vorher definiertes Abbruchkriterium erfüllt ist.

Sinnvoll erscheint es, die Iteration abzubrechen, wenn in einem Iterationsschritt nur noch eine unwesentliche Verbesserung des Zielfunktionswertes erzielt wird. Das zugehörige Abbruchkriterium kann mathematisch wie folgt formuliert werden:
Die Iteration bricht ab, wenn für ein vorher festgelegtes $\varepsilon>0$ gilt: $|L(\beta^{(i+1)})-L(\beta^{(i)})|<\varepsilon$.

Im weiteren werden Gradient und Hesse-Matrix von $L=L(\beta)$ in $\beta^{(i)}$ mit γ_i und H_i bezeichnet. In jedem Iterationsschritt i kann $L(\beta)$ in $\beta^{(i)}$ durch ihre Taylorreihe (bis zum quadratischen Term) approximiert werden:

$$L(\beta) \cong L(\beta^{(i)}) + \gamma_i \cdot(\beta-\beta^{(i)}) + 0.5\cdot(\beta-\beta^{(i)})^T\cdot H_i\cdot(\beta-\beta^{(i)})$$

Bedingung erster Ordnung für ein Maximum der Taylorreihe in β ist dann:

$$\gamma_i + H_i\cdot(\beta-\beta^{(i)})=0.$$

Durch Auflösen nach β erhält man eine allgemeine Formel für die NR-Iteration.
Im i-ten Iterationsschritt geht $\beta^{(i+1)}$ wie folgt aus $\beta^{(i)}$ hervor: $\beta^{(i+1)} = \beta^{(i)} - H_i^{-1}\cdot\gamma_i$.
Der Ausdruck $H_i^{-1}\cdot\gamma_i$ wird als Schrittlänge und der normierte Ausdruck $H_i^{-1}\cdot\gamma_i \cdot\| H_i^{-1}\cdot\gamma_i \|^{-1}$ als Schrittrichtung bezeichnet.

Die Effizienz des Newton-Raphson-Verfahrens kann durch geeignete Modifikationen der Schrittlängen verbessert werden. Da die Taylorreihe die Ausgangsfunktion nur lokal, das heißt in der Nähe von $\beta^{(i)}$, hinreichend gut approximiert, kann der Fall auftreten, dass die Schrittlänge über den Maximalpunkt in Schrittrichtung hinausführt. Eine solche Überschreitung führt unter Umständen sogar dazu, dass die Zielfunktion in $\beta^{(i+1)}$ einen kleineren Wert aufweist als in $\beta^{(i)}$. In diesem Fall: $L(\beta^{(i+1)})<L(\beta^{(i)})$ erscheint es sinnvoll, den Iterationsschritt mit der halbierten Schrittlänge $0.5\cdot H_i^{-1}\cdot\gamma_i$ zu wiederholen.

ML-Schätzung der Modellparameter mit Hilfe des Newton-Raphson-Verfahrens

Die Maximum-Likelihood-Schätzer für die Modellparameter eines logistischen Regressionsmodells werden üblicherweise mit Hilfe des Newton-Raphson-Verfahrens bestimmt.

Aus rechentechnischen Gründen wird dabei ausgenutzt, dass die ML-Schätzer auch die Log-Likelihood-Funktion, die sich durch Logarithmieren von L(β) (siehe (6)) ergibt, maximieren. Das führt zu der Zielfunktion:

$$l(\beta) = \log_e\{L(\beta)\} = -\sum_{i=1}^{n} \log_e\{1 + \exp(x_i \cdot \beta)\} + \sum_{i=1}^{n} d_i \cdot x_i^T \cdot \beta \,.$$

Durch partielles Ableiten ergibt sich ausgehend von obiger Log-Likelihood-Funktion zunächst der Gradient:

$$\gamma(\beta) := \frac{\partial l(\beta)}{\partial \beta} = X^T \cdot (d - \hat{p}(\beta)) \,. \qquad (7)$$

Dabei gilt:

$$X=(x_1,...,x_n)^T \in \mathbf{R}^{n,k+1} \text{ mit } x_i=(1,x_{i,1},\ldots,x_{i,k})^T \ (i=1,\ldots,k)$$
$$\text{und } \hat{p}(\beta) = (P(D=1|x_1,\beta),\ldots.,P(D=1|x_n,\beta))^T \in \mathbf{R}^n.$$

Die Matrix X wird als Regressormatrix bezeichnet. Die n Spalten korrespondieren zu den n Probanden und enthalten jeweils nach einem 1-Eintrag die Werte der k Einflussvariablen des zugehörigen Probanden. Die Einträge des n-dimensionalen Vektors $\hat{p}(\beta)$ entsprechen den Erkrankungswahrscheinlichkeiten, die sich bei Vorliegen des Modellparametervektors β für die n Probanden im logistischen Modell ergeben. Für i=1,…,n gilt nach (2):

$$p(\beta)_i = P(D=1|x_i,\beta) = (1+\exp\{-x_i^T \cdot \beta\})^{-1}.$$

Erneutes Ableiten des Gradienten liefert die Hesse-Matrix:

$$H(\beta) = \frac{\partial \gamma(\beta)}{\partial \beta^T} = -X^T \cdot V(\beta) \cdot X \,. \qquad (8)$$

In (7) repräsentiert V(β) eine (n,n)-Diagonalmatrix mit den Diagonalelementen:

$V(\beta)_{i,i} = p(\beta)_i \cdot (1 - p(\beta)_i)$ für i=1,…,n.

Mit (7) und (8) kann das Newton-Raphson-Verfahren für die Schätzung des Modellparametervektors β im logistischen Regressionsmodell formuliert werden. Mit dem Ziel die Effizienz des Verfahrens zu erhöhen (siehe oben), wird in einem Zusatzschritt (Schritt 3) versucht, die Schrittlänge im Iterationsschritt zu optimieren. Bevor mit der Iteration begonnen werden

kann, sind noch ein Startvektor $\beta^{(0)}$ und ein hinreichend kleiner Wert $\varepsilon>0$ für das Abbruchkriterium vorzugeben.

Newton-Raphson-Algorithmus für die Modellparameterschätzung

(1) Initialisierung: Wähle $\beta^{(0)} \in \mathbf{R}^{k+1}$ beliebig

(2) Iterationsschritt: $\beta^{(i+1)} = \beta^{(i)} + (X^T \cdot V(\beta^{(i)}) \cdot X)^{-1} \cdot X^T \cdot (d - \hat{p}(\beta^{(i)}))$

(3) Modifikation der Schrittlänge:

Definiere: $t:=\min\{s \in \mathbf{N}_0: l(\beta^{(i)}) < l(\beta^{(i)} + \{0,5\}^s \cdot \{\beta^{(i+1)} - \beta^{(i)}\})$

Modifiziere: $\beta^{(i+1)} = \beta^{(i)} + (0.5)^t \cdot (\beta^{(i+1)} - \beta^{(i)})$

(4) Abbruchkriterium: Falls $|l(\beta^{(i)}) - l(\beta^{(i+1)})| < \varepsilon$:

Abbruch des Verfahrens mit $\hat{\beta} := \beta^{(i+1)}$

(5) Zurück zu Schritt (2)

Bei der Datenauswertung in Kapitel 5.3 wird das Newton-Raphson-Verfahren ausschließlich wie folgt durchgeführt. $\varepsilon=0{,}0001$ vervollständigt das Abbruchkriterium und als Startvektor fungiert ein Vektor der Form $\beta^{(0)} = (b(d),0,\ldots,0)^T \in \mathbf{R}^{k+1}$. Das heißt, die letzten k Parameter, die jeweils zu einer Einflussvariable korrespondieren, werden im Startvektor auf 0 gesetzt, wohingegen die erste Komponente des Startvektors in Abhängigkeit von den n Krankheitszuständen der Studienteilnehmer gewählt wird. Da die erste Komponente β_0 des Parametervektors zu einem gewissen Grade die „Grunderkrankungschance“ beschreibt, reflektiert ihr Wert im wesentlichen die Krankheitsverteilung der Studienpopulation. Schoenfeld (1982) empfiehlt deshalb die folgende Wahl von $b(d) \in \mathbf{R}$:

$$\beta_1^{(0)} = b(d) = \log_e\left(\sum_{i=1}^{n} \frac{d_i}{n - d_i}\right).$$

Vorteile der Maximum-Likelihood-Parameterschätzung

Die Maximum-Likelihood-Methode ist die am meisten verwendete Methode zur Konstruktion von Punktschätzern, da sie bei regulären Problemen wünschenswerte asymptotische Eigenschaften garantiert (Fahrmeier et al. 1995).

Bereits unter schwachen Regularitätsbedingungen (vgl. zum Beispiel Witting und Nölle 1970) sind Maximum-Likelihood-Schätzungen konsistent und asymptotisch normalverteilt.

Genauer liegt für eine (k+1)-dimensionale ML-Schätzung unter den Regularitätsbedingungen die folgende Verteilungskonvergenz (n→∞) vor:

$$\hat{\beta} \to N_{k+1}(\beta, \frac{1}{n} \cdot E\left[-\frac{\partial^2 l(\beta)}{\partial \beta \cdot \partial \beta^T} \right]^{-1}) = N_{k+1}(\beta, \frac{1}{n} \cdot E[-H(\beta)]^{-1}) .$$

Die Matrix E[-H(β)] wird als Fisher`sche Informationsmatrix zum Schätzen von β bezeichnet. Sie ist die Erwartungswertmatrix der negativen Hessematrix der Log-Likelihood-Funktion, ausgewertet an der Stelle des wahren Parametervektors β. Da der wahre Parametervektor in den Anwendungen unbekannt ist, kann sie nicht berechnet werden, sondern muss geeignet geschätzt werden. Nach Kale (1962) gilt (ebenfalls) unter geeigneten Regularitätsbedingungen, dass der Ausdruck: $-n^{-1} \cdot H(\hat{\beta})$ für n→∞ (nach Wahrscheinlichkeit) gegen E[-H(β)] konvergiert. Das heißt, ein geeigneter Schätzer für die Kovarianzmatrix der ML-Schätzung kann durch die Auswertung der Log-Likelihood-Hesse-Matrix an der Stelle der ML-Schätzung $\hat{\beta}$ konstruiert werden.

Bezogen auf die ML-Schätzung im logistischen Regressionsmodell bedeutet dies, dass $\hat{\beta}$ für hinreichend große n als Realisation einer (k+1)-dimensionalen Normalverteilung mit Erwartungswert $E[\hat{\beta}] = \beta$ und Kovarianzmatrix $Cov(\hat{\beta}) \approx \hat{C}(\hat{\beta}) := -H(\hat{\beta})^{-1} = (X^T \cdot V(\hat{\beta}) \cdot X)^{-1}$ aufgefasst werden kann.

Von besonderer Bedeutung für die Konstruktion asymptotischer Konfidenzintervalle und Tests vom Wald-Typ (vgl. Unterkapitel 4.3.5) ist, dass somit unter geeigneten Regularitätsbedingungen für die einzelnen Komponenten des ML-Schätzers asymptotisch (n→∞) gilt:

$$\hat{\beta}_i \sim N_1(\beta_i, \hat{C}(\hat{\beta})_{i,i}) \text{ (i=1,…,k+1).}$$

4.3.5 Konfindenzintervalle und Wald-Tests

Die Effekte der Einflussvariablen können anhand der Komponenten des geschätzten Parametervektors $\hat{\beta}$ beurteilt werden. Hierbei wird analog zu den Ausführungen in den Unterkapiteln 4.3.2 und 4.3.3 vorgegangen, wobei der unbekannte wahre Modellparametervektor durch die ML-Schätzung zu ersetzen ist. Zu beachten ist allerdings, dass ML-Schätzungen genau wie alle anderen statistische Punktschätzungen Realisationen einer zufälligen Schätzstatistik darstellen und deswegen immer mit einer gewissen Unsicherheit verbunden sind. Bei (asymptotisch) erwartungstreuen Schätzern ist diese Unsicherheit durch ihre Varianz gegeben. Je grö-

ßer die Varianz eines solchen Schätzers ist, desto größer ist seine (durchschnittliche) Schwankung um den tatsächlichen Parameterwert.

Konfidenzintervalle vom Wald-Typ

Zur Erfassung dieser Unsicherheit werden neben Punktschätzungen auch Konfidenzintervalle für die unbekannten Modellparameter berechnet. Ein (1-α)·100%-Konfindenzintervall ist ein Intervall, in dem der unbekannte Parameter mit einer Wahrscheinlichkeit von (1-α)∈(0,1) liegt. Da die ML-Schätzer approximativ normalverteilt sind (vgl. Unterkapitel 4.3.4), können (1-α)·100%-Konfidenzintervalle für die unbekannten Modellparameter wie folgt konstruiert werden:

Für i=1,…,k+1 folgt aus $\hat{\beta}_i \sim N_1(\beta_i, \hat{C}(\hat{\beta})_{i,i})$ unmittelbar: $(\hat{\beta}_i - \beta_i) \cdot \hat{C}(\hat{\beta})_{i,i}^{-1/2} \sim N_1(0,1)$.

Entsprechend gilt für die normierte Zufallsgröße:

$$P(u_{\alpha/2} < (\hat{\beta}_i - \beta_i) \cdot \hat{C}(\hat{\beta})_{i,i}^{-1/2} < u_{1-\alpha/2}) = 1-\alpha,$$

wobei u_δ das δ-Quantil der Standardnormalverteilung repräsentiert.

Durch elementare Umformungen ergibt sich weiter:

$$P(\hat{\beta}_i - u_{1-\alpha/2} \cdot \sqrt{\hat{C}(\hat{\beta})_{i,i}^{-1}} < \beta_i < \hat{\beta}_i + u_{1-\alpha/2} \cdot \sqrt{\hat{C}(\hat{\beta})_{i,i}^{-1}}) = 1-\alpha,$$

so dass das Intervall:

$$[\hat{\beta}_i - u_{1-\alpha/2} \cdot \sqrt{\hat{C}(\hat{\beta})_{i,i}^{-1}} \mid \hat{\beta}_i + u_{1-\alpha/2} \cdot \sqrt{\hat{C}(\hat{\beta})_{i,i}^{-1}})]$$

ein (1-α)·100%-Konfidenzintervall für β_i beschreibt. Üblicherweise werden in der Statistik 95%-Konfindenzintervalle (α=0.05) berechnet.

Variablen, deren Parameterwert bei gleichzeitiger Berücksichtigung der anderen Variablen 0 beträgt, haben überhaupt keinen Einfluss auf die Erkrankungschance. Liegt zum Beispiel für die i-te Variable X_i ein Parameterwert β_i von 0 vor, ergibt sich für zwei beliebige Ausprägungen x_i und y_i aus dem Wertebereich von X_i mit (5) ein Odds-Ratio von $\exp\{(x_i-y_i)\cdot 0\}=1$.
Da einflusslose Variablen nicht von epidemiologischen Interesse sind, stellt sich nach der Modellschätzung die Frage, wie entschieden werden kann, welche Parameter 0 sind und welche nicht. Wenngleich diese Frage aufgrund der bereits oben angesprochenen Schätzungenauigkeit nicht mit Sicherheit beantwortet werden kann, können diesbezügliche Wahrscheinlichkeitsaussagen getroffen werden. Eine erste Möglichkeit bieten Konfidenzintervalle. Ist in ei-

nem (1-α)·100%-Konfidenzintervall der Wert 0 nicht enthalten, kann mit einer Sicherheitswahrscheinlichkeit von 1-α davon ausgegangen werden, dass der zugehörige Parameter nicht 0 ist. Die Irrtumswahrscheinlichkeit beträgt in diesem Fall α.

1-dimensionale Wald-Tests

Wahrscheinlichkeitsaussagen mit größerem Informationsgehalt können auch unter Zuhilfenahme von P-Werten (Überschreitungswahrscheinlichkeiten) statistischer Tests formuliert werden.

Der 1-dimensionale Wald-Test, der sich analog zu den obigen Konfidenzintervallen herleiten lässt, testet die Nullhypothese H_0: „$\beta_i=0$“ gegen die Alternativhypothese H_1: „$\beta_i\neq 0$“. Unter der Nullhypothese ist die Teststatistik $W_i := \hat{\beta}_i \cdot \hat{C}(\hat{\beta})_{i,i}^{-1/2}$ asymptotisch standardnormalverteilt. Entsprechend sind unter H_0 Realisationen w_i außerhalb des Intervalls $(-u_{1-\alpha/2}, u_{1-\alpha/2})$, wobei $u_{1-\alpha/2}$ das (1-α/2)-Quantil der N(0,1) bezeichnet, nur mit einer Wahrscheinlichkeit von α zu erwarten. Wird tatsächlich eine Realisation außerhalb dieses Intervalls beobachtet, kann die Nullhypothese (zu Gunsten der Alternativen) mit einer Irrtumswahrscheinlichkeit von α verworfen werden.

Da eine Verwerfung der Nullhypothese zum Niveau α offensichtlich gleichbedeutend damit ist, dass das (1-α)·100%-Konfidenzintervall für β_i den Wert 0 nicht beinhaltet, liefert der Wald-Test zunächst keinen zusätzlichen Informationsgewinn. Informativer als Konfidenzintervalle sind allerdings die P-Werte solcher Wald-Tests. Ganz allgemein ist der P-Wert eines Tests, bei vorliegender Realisation der Teststatistik, die minimale Irrtumswahrscheinlichkeit, bei der noch eine Entscheidung für die Alternativhypothese möglich ist.

Bei dem oben vorgestellten Waldtest führt eine Realisation w_i der Teststatistik W_i mit einer Irrtumswahrscheinlichkeit von α zu einer Entscheidung für H_1, wenn $w_i \notin (-u_{1-\alpha/2}, u_{1-\alpha/2})$, wobei $u_{1-\alpha/2}$ wieder das (1-α/2)-Quantil der N(0,1) bezeichnet Entsprechend ist der P-Wert die minimale Irrtumswahrscheinlichkeit α, für die das Intervall $(-u_{1-\alpha/2}, u_{1-\alpha/2})$ den Wert w_i nicht enthält:

$$p_i = \min\{\alpha \in (0,1) : w_i \notin (-u_{1-\alpha/2}, u_{1-\alpha/2})\} \Leftrightarrow p_i = \min\{\alpha \in (0,1) :\mid w_i \mid \geq u_{1-\alpha/2}\}.$$

Aufgrund der Beziehung:

$$| w_i | \geq u_{1-\alpha/2} \Leftrightarrow \phi(| w_i |) \geq \phi(u_{1-\alpha/2}) = 1-\alpha/2 \Leftrightarrow 2-2\cdot\phi(| w_i |) \leq \alpha ,$$

wobei ϕ(.) die (streng monoton steigende) Verteilungsfunktion der Standardnormalverteilung bezeichnet, folgt weiter:

$$p_i = \min\{\alpha \in (0,1) :| w_i | \geq u_{1-\alpha/2}\} = \min\{\alpha \in (0,1) : \alpha \geq 2-2\cdot\phi(| w_i |)\} = 2-2\cdot\phi(| w_i |) .$$

Bei der klassischen Wald-Test-Prozedur wird eine Irrtumswahrscheinlichkeit α vorgegeben und der Test liefert als Ergebnis lediglich die Entscheidung, ob diese Irrtumswahrscheinlichkeit bei Verwerfung der Nullhypothese überschritten wird oder nicht. Entweder die Nullhypothese kann zum Testniveau α verworfen werden und der Parameter ist zum Niveau α statistisch signifikant von 0 verschieden, oder nicht. P-Werte hingegen beschreiben gewissermaßen diesen „Grad der statistischen Signifikanz“ und lassen somit zusätzlichen Spielraum für Interpretationen. Je kleiner der P-Wert eines Parameters ist, desto unwahrscheinlicher ist es, dass er 0 ist. Weist ein Parameter einen großen P-Wert auf, legt das die Vermutung nahe, dass die zugehörige Variablen keinen Einfluss auf die Erkrankungschance nimmt. Entsprechend ist zu überlegen, ob nicht ein neues Modell ohne diese Variable geschätzt werden sollte.
Hosmer und Lemeshow (1989) stellen drei automatisierte Verfahren vor, mit denen in logistischen Regressionsmodellen auf Grundlage der P-Werte von Wald-Tests die bedeutsamsten Einflussvariablen selektiert werden können. Von diesen 3 Verfahren kommt nur die automatisierte Rückwärtsauswahl bei der Auswertung des Datenmaterials zum Einsatz. Da dieses Verfahren im Rahmen der Auswertung allerdings nicht nur in der klassischen, sondern auch in einer modifizierten Form benötigt wird, wird erst in Unterkapitel 5.3 dieser Arbeit ausführlicher auf das Rückwärtsauswahlverfahren eingegangen.
Neben dem Wald-Test werden in der Literatur noch andere asymptotische Testverfahren für die Parameter des logistischen Regressionsmodells vorgestellt. Diese werden im vorliegenden Fall bei der Auswertung nicht benötigt, so dass bezüglich des multivariaten Wald-Tests, des Likelihood-Quotienten-Tests und des Score-Tests auf Kleinbaum et al. (1982) verwiesen wird.

4.3.6 Beurteilung der Modellanpassung

Die, im letzten Unterkapitel vorgestellten, Methoden eigenen sich, um nach der Modellschätzung zu beurteilen, welche Variablen einen signifikanten Einfluss auf die Erkrankungschance nehmen. Üblicherweise wird einer Variablen ein statistisch signifikanter Einfluss unterstellt,

wenn sie einen P-Wert kleiner 0,05 aufweist, dass heißt, wenn ihr Parameter mit einer Wahrscheinlichkeit von 0,95 von 0 verschieden ist. Durch wiederholte Entfernung bedeutungsloser Variablen und erneute Modellschätzung kann so ein finales Modell erhalten werden, welches nur noch einflussreiche Variablen enthält. Geht es in der Studie ausschließlich um die Identifikation von Risikofaktoren, sind keine weiteren Untersuchungen notwendig. Darüber hinaus stellt sich allerdings häufig die Frage, ob die signifikanten Variablen in ihrer Gesamtheit einen großen Erklärungswert für die Erkrankungswahrscheinlichkeit haben, so dass das gefundene Modell die vorliegenden Daten gut wiedergibt. Ein großer Erklärungswert bzw. eine gute Modellanpassung liegt vor, wenn das Modell ausschließlich erkrankten Probanden hohe Erkrankungswahrscheinlichkeiten unterstellt. Ein großer Erklärungswert des Modells ist i.a. nur dann zu erwarten, wenn in der Studie alle wichtigen Risikofaktoren für die Krankheit erfasst wurden und, wenn die Entstehung der Krankheit nicht zu einem großen Teil von zufälligen bzw. nichterfassbaren Prozessen abhängt.

Grundidee bei der Beurteilung der Modellanpassung ist ein Vergleich zwischen den beobachteten Krankheitszuständen $d_1,\ldots,d_n$ und den nach Modell zu erwartenden Krankheitszuständen der Studienteilnehmer. Letztere entsprechen den Erkrankungswahrscheinlichkeiten $\hat{p}_1,....,\hat{p}_n$, da gilt:

$$\hat{E}[D \mid x_i] = 1 \cdot \hat{P}(D = 1 \mid x_i) + 0 \cdot \hat{P}(D = 0 \mid x_i) = \hat{P}(D = 1 \mid x_i) =: \hat{p}_i \quad (i=1,\ldots,n).$$

Im Folgenden wird davon ausgegangen, dass das zu beurteilende Modell mindestens einen stetigen Risikofaktor beinhaltet, so dass die n Risikofaktorkonstellationen $x_1,\ldots,x_n$ paarweise verschieden sind. Unter dieser Annahme ist keine Datenzusammenfassung möglich. Zu jeder Risikofaktorkonstellation x_i bzw. geschätzten Erkrankungswahrscheinlichkeit $\hat{p}_i = \hat{p}_i(x_i)$ liegt genau eine Realisierung $d_i \in \{0,1\}$ vor (i=1,…,n). Für viele Modellanpassungskriterien können in diesem Fall keine Aussagen über die (asymptotischen) Verteilungen getroffen werden, so dass sie nicht für formale Anpassungstests geeignet sind. Entsprechend konzentrieren sich die Ausführungen in diesem Unterkapitel primär auf deskriptive Methoden.

Zur Beurteilung des Erklärungswertes von Modellen, bei denen zu jeder Risikofaktorkonstellation x mehrere Realisierungen $d_1(x),\ldots,d_m(x)$ vorliegen, wird auf Hosmer und Lemeshow (1989) verwiesen.

Für den Fall, dass jedes Paar $(d_i, \hat{p}_i)$ (i=1,…,n) eine eigene Einheit darstellt, sind die Pearson-Residuen r_i und die Deviance-Residuen dev_i wie folgt definiert:

$$r_i = r(d_i, \hat{p}_i) = \frac{d_i - \hat{p}_i}{\sqrt{\hat{p}_i \cdot (1-\hat{p}_i)}} \text{ und}$$

$$dev_i = dev(d_i, \hat{p}_i) = \left\{ \begin{array}{ll} -\sqrt{2 \cdot |\log_e(1-\hat{p}_i)|}, & d_i = 0 \\ \textit{wenn} & \\ \sqrt{2 \cdot |\log_e(\hat{p}_i)|}, & d_i = 1 \end{array} \right\}.$$

Die Pearson-Residuen ergeben sich durch Standardisierung der beobachteten Krankheitszustände mit Hilfe der geschätzten Momente. Im Falle einer guten Modellanpassung ist deswegen davon auszugehen, dass jedes Pearson-Residuum einen Erwartungswert von 0 und eine Varianz von 1 aufweist. Bedingt durch die Gewichtung mit dem reziproken der geschätzten Standardabweichung treten die betragsmäßig größten Pearson-Residuen auf, wenn das Modell erkrankten (nichterkrankten) Probanden fälschlicherweise eine kleine (große) Erkrankungswahrscheinlichkeit unterstellt.

Die Quadratsumme X^2 der Pearson-Residuen wird als Pearson-Statistik bezeichnet. Im Falle paarweise verschiedener Risikofaktorkonstellationen $x_1,\ldots,x_n$ kann keine Aussage über die asymptotische Verteilung der Pearson-Statistik getroffen werden, so dass sich kein formaler Anpassungstest konstruieren lässt.

Zumindest eine grobe Beurteilung des Wertes der Pearson-Statistik ermöglichen in diesem Fall die Quantile einer Chi-Quadrat-Verteilung mit n-(k+1) Freiheitsgraden. Hosmer und Lemeshow (1989) vermuten, dass die Pearson-Statistik bei paarweise verschiedenen Risikofaktorkonstellationen $x_1,\ldots,x_n$ bei einer guten Modellanpassung annähernd eine solche Verteilung aufweist, raten in dieser Situation allerdings von einem formalen Anpassungstest ab.

Die Deviance-Residuen stehen in einem engen Zusammenhang zur Log-Likelihood-Funktion. Die Quadratsumme der Deviance-Residuen wird als Deviance D bezeichnet und entspricht dem Produkt aus dem Faktor –2 und der Log-Likelihood-Funktion an der Stelle des ML-Schätzers:

$$D = \sum_{i=1}^{n} dev_i^2 = 2 \cdot \sum_{i=1}^{n} d_i \cdot |\log_e(\hat{p}_i)| + (1-d_i) \cdot |\log_e(1-\hat{p}_i)|$$

$$= -2 \cdot \sum_{i=1}^{n} d_i \cdot \log_e(\hat{p}_i) + (1-d_i) \cdot \log_e(1-\hat{p}_i) = (-2) \cdot l(\hat{\beta}).$$

Die Likelihood-Funktion an der Stelle des ML-Schätzers entspricht der Wahrscheinlichkeit, dass bei Gültigkeit des geschätzten Modells die beobachteten Daten (Krankheitszustände) auftreten. Je größer der Likelihood-Wert $L(\hat{\beta})$ ist, desto plausibler ist das Modell. Die Deviance ergibt sich durch die streng monoton fallende Transformation $(-2)\cdot\log_e\{L(\hat{\beta})\} =: D$, so dass eine umgekehrte Beziehung vorliegt. Folglich deutet ein betragsmäßig großes Deviance-Residuum darauf hin, dass der Krankheitszustand des zugehörigen Probanden im geschätzten Modell unwahrscheinlich und damit unplausibel ist. Treten viele betragsmäßig große Deviance-Residuen auf, bedeutet dies im Umkehrschluss, dass das Modell die Daten nicht gut wiedergibt und somit eine geringe Anpassungsgüte aufweist. Der Grund dafür, dass nicht der anschaulichere Likelihood-Wert sondern die Deviance als Anpassungsmaß verwendet wird, liegt darin, dass unter gewissen Bedingungen asymptotische Verteilungsaussagen über das Deviance-Maß getroffen werden können.

Liegen zu jeder Risikofaktorkonstellation x mehrere Realisierungen $d_1(x),\ldots,d_m(x)$ vor und ist J die Anzahl verschiedener Risikofaktorkonstellation, dann ist die Deviance-Statistik bei guter Modellanpassung chi-quadrat-verteilt mit J-(k+1) Freiheitsgraden.

Hosmer und Lemeshow (1989) vermuten im vorliegenden Fall unabhängiger Paare $(d_i, \hat{p}_i)$ (i=1,…,n) wieder, dass die Quantile der Chi-Quadrat-Verteilung mit n-(k+1) Freiheitsgraden geeignete Richtwerte zur Beurteilung der Modellanpassung sind, raten jedoch auch hier von der Durchführung eines formalen Anpassungstests ab.

Anpassungstest von Hosmer & Lemeshow

Beim Anpassungstest von Hosmer und Lemeshow (1980) werden die Probanden auf Grundlage ihrer geschätzten Erkrankungswahrscheinlichkeiten in 10 Gruppen eingeteilt. Die erste Gruppe wird von den Probanden mit den geringsten Erkrankungswahrscheinlichkeiten gebildet und die zehnte Gruppe besteht aus den Probanden mit den größten Erkrankungswahrscheinlichkeiten. Um zu gewährleisten, dass sich in jeder Gruppe annährend gleich viele Probanden $c_1,\ldots,c_{10}$ befinden, erfolgt die Klassifizierung unter Zuhilfenahme der (empirischen) Dezile q(0.0), q(0.1),…, q(0.9), q(1.0) der geschätzten Erkrankungswahrscheinlichkeiten Eine Klasse besteht jeweils aus den Probanden, deren geschätzte Erkrankungswahrscheinlichkeit zwischen zwei aufeinanderfolgenden Dezilen liegen:

$K_i = \{u \in \{1,...,n\} : q((i-1)\cdot 0.1) \le \hat{p}_u < q(i\cdot 0.1)\}$ (i=1,…,10).

Da die Anzahl nach Modell zu erwartender Fälle von der ersten bis zur zehnten Klasse ansteigt, werden sie auch als Risikogruppen oder als Risikodezile („deciles of risk“) bezeichnet.

Wenn das Modell die Daten gut beschreibt, ist zu erwarten, dass in jeder Risikogruppe K_i die Anzahl kranker Probanden ungefähr, mit der nach dem Modell zu erwartenden, Anzahl übereinstimmt. Ausgehend von dieser intuitiven Überlegung kann die folgende Teststatistik C^2 für einen formalen Anpassungstest verwendet werden:

$$C^2 = \sum_{i=1}^{10} \frac{(o_i - c_i \cdot \overline{\pi}_i)^2}{c_i \cdot \overline{\pi}_i \cdot (1 - \overline{\pi}_i)},$$

wobei $\overline{\pi}_i$ die durchschnittliche Erkrankungswahrscheinlichkeit in der i-ten Risikogruppe

$$\overline{\pi}_i = \frac{1}{c_i} \sum_{u \in K_i} \hat{p}_i$$

und o_i die Anzahl tatsächlicher vorliegender Fälle in der i-ten Risikogruppe

$$o_i = \sum_{u \in K_i} d_i$$

repräsentiert. Die C^2-Statistik ist also eine gewichtete Quadratsumme der Differenzen zwischen tatsächlich beobachteten und zu erwartenden Krankheitsfällen in den 10 Gruppen. Die Gewichte $g_i := (c_i \cdot \overline{\pi}_i \cdot (1 - \overline{\pi}_i))^{-1}$ der Abweichungsquadrate ergeben sich als Reziproke von speziellen Varianzschätzungen in den Gruppen.

Hosmer und Lemeshow (1980) zeigen mit Hilfe von Simulationsstudien, dass bei Modellen, die das Datenmaterial korrekt beschreiben, die Teststatistik C^2 approximativ eine Chi-Quadrat-Verteilung mit 8 Freiheitsgraden aufweist. Dementsprechend kann die Nullhypothese: „Das Modell beschreibt die Daten korrekt!“ mit einer Irrtumswahrscheinlichkeit von α verworfen werden, wenn die realisierte C^2-Statistik das $(1-\alpha)$-Quantil der Chi-Quadrat-Verteilung mit 8 Freiheitsgraden überschreitet.

Im Gegensatz zu den formalen Anpassungstests auf Grundlage der Deviance- bzw. Pearson-Statistik hat der Anpassungstest von Hosmer & Lemeshow nicht zur Voraussetzung, dass für jede Risikofaktorkonstellation mehrere Realisierungen vorliegen.

Klassifikationstafelanalyse

Eine Klassifikationstafel gibt einen tabellarischen Eindruck von der Erklärungsgüte eines logistischen Regressionsmodells (vgl. Hosmer und Lemeshow 1989). Auf Grundlage der geschätzten Erkrankungswahrscheinlichkeiten werden dazu eindeutige Prognosen (krank – ja/nein) für die Probanden erstellt und in Form einer Vier-Feldertafel gegen die tatsächlichen

Krankheitszustände abgetragen. Anschließend können auf Grundlage der realisierten Zellhäufigkeiten epidemiologische Kenngrößen berechnet werden.

Zur Erstellung einer eindeutigen Prognose des Krankheitszustandes bedient man sich einer einfachen Entscheidungsregel. Diese ordnet jeder geschätzten Erkrankungswahrscheinlichkeit eindeutig eine Prognose zu:

$$E_c:(0,1)\text{->}\{0,1\} \text{ mit } E_c(x) = \begin{cases} 1, & x > c \\ & wenn \\ 0, & x \le c \end{cases}.$$

Der Entscheidungsregel liegt ein Diskriminanzwert c zugrunde. In Abhängigkeit davon, ob die geschätzte Erkrankungswahrscheinlichkeit $\hat{p}_i$ eines Probanden diesen Wert übersteigt oder nicht, wird der Zustand „erkrankt“ bzw. „nichterkrankt“ prognostiziert. Wird ein Diskriminanzwert c von 0,5 gewählt, ergibt sich die intuitivste Entscheidungsregel. Diese prognostiziert für jeden Probanden den Krankheitszustand, der nach dem geschätzten Modell für ihn am wahrscheinlichsten ist. Soll hingegen die Verteilung der tatsächlichen Krankheitszustände $d_1,\ldots,d_n$ berücksichtigt werden, empfiehlt es sich, als Diskriminanzwert die relative Häufigkeit von Erkrankungen zu verwenden.
Unabhängig von der Wahl des Diskriminanzwertes ist zu bedenken, dass der Übergang von den Erkrankungswahrscheinlichkeiten $\hat{p}_1,...,\hat{p}_n$ zu den Prognosen $E_c(\hat{p}_1),...,E_c(\hat{p}_n)$ mit einem nicht unerheblichen Informationsverlust verbunden ist. Dafür aber sind nur 4 Kombinationen der Paare $(d_i,\hat{d}_i)$ (i=1,…,n) möglich, so dass das Klassifikationsergebnis nach Auszählung der absoluten Häufigkeiten übersichtlich in Form einer Vier-Feldertafel dargestellt werden kann. Eine solche Klassifikationstafel hat das äußere Erscheinungsbild:

Tabelle 4.A: Schema einer Klassifikationstafel

Prognose E	**Tatsächlicher Krankheitszustand D**		$\sum$
	D=0	**D=1**	
E=0	Anzahl richtiger „nichterkrankt“ Prognosen (**a**)	Anzahl falscher „nichterkrankt“ Prognosen (**b**)	**a+b**
E=1	Anzahl falscher „erkrankt“ Prognosen (**c**)	Anzahl richtiger „erkrankt“ Prognosen (**d**)	**c+d**
$\sum$	**a+c**	**b+d**	**n**

Mit Hilfe der Klassifikationstafel kann ein Überblick über die gemeinsame Verteilung von Krankheitszuständen und Prognosen gewonnen werden. Neben den Anteilen richtiger und falscher Prognosen sind insbesondere die folgenden beiden Maßzahlen von Bedeutung für die Beurteilung der Erklärungsgüte:

Die Sensitivität $\hat{P}(E=1 \mid D=1)$ entspricht dem Anteil der Erkrankten, für die eine richtige Prognose erstellt wurde. Die Sensitivität ist also als Schätzung für die Wahrscheinlichkeit, mit der ein Erkrankter auf Grundlage des Modells als solcher erkannt wird, zu verstehen:

$$\hat{P}(E=1 \mid D=1) = \frac{d}{b+d}.$$

Die Spezifität $\hat{P}(E=0 \mid D=0)$ hingegen schätzt die Wahrscheinlichkeit, mit der für einen Nichterkrankten eine richtige Prognose erstellt wird:

$$\hat{P}(E=0 \mid D=0) = \frac{a}{a+c}.$$

Ab welchen Werten von Sensitivität und Spezifität von einer guten Modellanpassung gesprochen werden kann, hängt von der jeweiligen Fragestellung ab. Grundsätzlich gilt, dass nur dann von einer hohen Prognosegüte des Modells auszugehen ist, wenn beide Maßzahlen eine „zufriedenstellende“ Größe aufweisen.

Abschließend soll noch einmal angemerkt werden, dass bei der Klassifikationstafelanalyse ein erheblicher Informationsverlust daraus resultiert, dass nicht mit den geschätzten, kontinuierlichen Wahrscheinlichkeiten selber, sondern mit (binären) Prognosen der Krankheitszustände gearbeitet wird. Insbesondere für Individuen, denen im Modell eine Erkrankungswahrscheinlichkeit nahe dem Diskriminanzwert zukommt, ergeben sich daher sehr unsichere Prognosen.

4.3.7 Lokale Anpassungsgüte und Regressionsdiagnostik

Bereits im letzten Unterkapitel wurden die Deviance- und Pearson-Residuen vorgestellt. Diese beiden Residuentypen bilden die Grundlage für die Beurteilung der lokalen Anpassungsgüte eines Modells. In beiden Fällen kann zum Beispiel anhand von Indexplots, d.h. Streudiagrammen der Residuen r_j bzw. dev_j gegen den zugehörigen Beobachtungsindex j (j=1,…,n), graphisch veranschaulicht werden, inwieweit die tatsächlichen Krankheitszustände und die geschätzten Erkrankungswahrscheinlichkeiten bei den einzelnen Probanden voneinander abweichen.

Neben diesen Aspekten der lokalen Anpassungsgüte ist häufig von Interesse, ob es einzelne Individuen in der Studienpopulation gibt, die eine ungewöhnliche Konstellation von Einflussvariablen x_j und Krankheitszustand d_j aufweisen und daher einen großen Einfluss auf die ML-Schätzung $\hat{\beta}$ und/oder Anpassungsgüte des Modells nehmen.

Zur Quantifizierung des diesbezüglichen Einflusses der Studienteilnehmer erscheint es sinnvoll, zu untersuchen, welchen Veränderungen die drei Modellgrößen X^2, D und $\hat{\beta}$ unterliegen, wenn die betreffende Person im Modell nicht berücksichtigt wird. Pregibon (1981) hat unter Zuhilfenahme approximativer Methoden für alle drei Größen gezeigt, dass die Veränderungen, die sich bei Vernachlässigung des j-ten Studienteilnehmers ergeben, zum einen von den zugehörigen Pearson- oder Deviance-Residuen andererseits aber auch von den Diagonalelementen einer speziellen Matrix $H(\hat{\beta})$ abhängen.

Diese Matrix $H(\hat{\beta})$ kann wie folgt berechnet werden

$$H(\hat{\beta}) := V(\hat{\beta})^{1/2} \cdot X \cdot (X^T \cdot V(\hat{\beta}) \cdot X)^{-1} \cdot X^T \cdot V(\hat{\beta})^{1/2}$$

und es ist leicht ersichtlich, dass für ihre Diagonalelemente h_j gilt:

$$h_j := H(\hat{\beta})_{j,j} = \hat{p}_j \cdot (1-\hat{p}_j) \cdot x_j^T \cdot (X^T \cdot V(\hat{\beta}) \cdot X)^{-1} \cdot x_j \text{ (j=1,…,n).}$$

Pregibon (1981) zeigt, dass der idempotenten und symmetrischen Matrix $H(\hat{\beta})$ approximativ eine ähnliche Bedeutung zukommt wie der Projektionsmatrix im gewöhnlichen linearen Modell. Genauer gilt, dass der Vektor der geschätzten Erkrankungswahrscheinlichkeiten approximativ durch die Projektion des Beobachtungsvektors $d=(d_1,\ldots,d_n)^T$ auf den Spaltenraum von $H(\hat{\beta})$ gegeben ist:

$$\hat{p} \approx H(\hat{\beta}) \cdot d \,.$$

Wie im linearen Modell wird das j-te Diagonalelemente h_j der (approximativen) Projektionsmatrix $H(\hat{\beta})$ als Leverage-Werte (leverage) der j-ten Beobachtung x_j bezeichnet (j=1,…,n) und es zeigt sich (Pregibon 1981), dass diese fast in Analogie zu den Leverage-Werten (leverages) im linearen Modell interpretiert werden können.

Auch im logistischen Regressionsmodell wirken sich Beobachtungen (bzw. Studienteilnehmer) mit großen Leverage-Werten in besonderem Maße auf die Parameterschätzung und damit auch auf die Modellanpassung aus. Der wesentliche Unterschied besteht in der inhaltlichen Interpretation großer Leverage-Werte. In der linearen Regression verhalten sich die Größen der Leverage-Werte h_j erfahrungsgemäß proportional zu der Distanz der zugehörigen

Beobachtungsvektoren x_j vom Mittelpunktsvektor aller Beobachtungen $x_1,\ldots,x_n$. Im logistischen Regressionsmodell ist diese Proportionalität nur zu einem gewissen Grade gegeben. Genauer verhalten sich nach Hosmer und Lemeshow (1989) lediglich die Faktoren $b_j := x_j^T \cdot (X^T \cdot V(\hat{\beta}) \cdot X)^{-1} \cdot x_j$ proportional zu den Distanzen der Vektoren x_j vom Mittelpunkt aller Beobachtungsvektoren $x_1,\ldots,x_n$.
Aufgrund der Beziehung: $h_j = \hat{p}_j \cdot (1 - \hat{p}_j) \cdot b_j$ überträgt sich diese Eigenschaft jedoch nicht unmittelbar auf die Leverage-Werte.

Die Veränderungen der drei Modellgrößen D, X^2 und $\hat{\beta}$, die sich bei Vernachlässigung des j-ten Probanden ergeben, werden von Pregibon (1981) linear approximiert. Er kommt zu dem Ergebnis, dass sich die Pearson-Statistik und Deviance bei Vernachlässigung des j-ten Probanden (approximativ) um die folgenden Beträge verringern:

$$\Delta X^2_{(-j)} \approx (r_j)^2 \cdot \frac{1}{1-h_j}$$

$$\Delta D_{(-j)} \approx (dev_j)^2 \cdot \frac{1}{1-h_j}.$$

Für die Quantifizierung der Auswirkung auf die vektorielle Größe $\hat{\beta}$ ist zunächst ein geeignetes skalares Maß festzulegen. Pregibon (1981) empfiehlt eine Diagnostik, die ausnutzt, dass durch die Menge

$$C = \{\beta \in R^{k+1} \mid (\hat{\beta} - \beta)^T \cdot (X^T \cdot V(\hat{\beta}) \cdot X) \cdot (\hat{\beta} - \beta) \leq c\}$$

ein asymptotisches Konfidenzellipsoid für den wahren Parametervektor β gegeben ist. Pregibons Überlegungen zufolge muss, da durch C ein Konfidenzbereich beschrieben wird, die quadratische Form

$$(\hat{\beta} - \beta)^T \cdot (X^T \cdot V(\hat{\beta}) \cdot X) \cdot (\hat{\beta} - \beta)$$

(asymptotisch) geeignet sein, um die Distanz zwischen der Schätzung $\hat{\beta}$ und einem beliebigen Vektor β unter Berücksichtigung der (geschätzten) Kovarianzstrukur von $\hat{\beta}$ zu messen. Entsprechend ist es nahe liegend, auch die Distanz zwischen $\hat{\beta}$ und dem ML-Schätzer $\hat{\beta}_{(-j)}$, der sich bei Nichtberücksichtigung der j-ten Beobachtung ergibt, mit Hilfe dieser quadratischen Form zu messen:

$$\Delta\hat{\beta}_{(-j)} := (\hat{\beta} - \hat{\beta}_{(-j)})^T \cdot (X^T \cdot V(\hat{\beta}) \cdot X) \cdot (\hat{\beta} - \hat{\beta}_{(-j)}).$$

Pregibon (1981) zeigt, dass approximativ für j=1,…,n gilt:

$$\Delta\hat{\beta}_{(-j)} \approx (r_j)^2 \cdot \frac{h_j}{(1-h_j)^2}.$$

Bei allen drei vorgestellten Diagnosemaßen sind hohe Werte ein Hinweis auf Beobachtungen bzw. Personen mit großem Einfluss auf die entsprechende Modellgröße. Hosmer & Lemeshow (1989) empfehlen zur Identifikation einflussreicher Beobachtungen (x_j,d_j) die Regressionsdiagnostiken $\Delta X^2_{(-j)}$, $\Delta D_{(-j)}$ und $\Delta\hat{\beta}_{(-j)}$ in Streudiagrammen gegen die geschätzten Wahrscheinlichkeiten $\hat{p}_j$ (j=1,…,n) abzutragen. Da die Diagnostiken ganz wesentlich von den Erkrankungswahrscheinlichkeiten beeinflusst werden, haben diese Darstellungen gegenüber herkömmlichen Indexplots den Vorteil, dass unmittelbar deutlich wird, inwieweit hohe Werte auf extreme Erkrankungswahrscheinlichkeiten (zum Beispiel $\hat{p}_j > 0.9$ oder $\hat{p}_j < 0.1$) zurückzuführen sind.

4.3.8 Auswertung von Fall-Kontroll-Studien

In logistischen Regressionsmodellen werden Erkrankungswahrscheinlichkeiten in Abhängigkeit von mehreren Expositionen X modelliert. Eine Schätzung dieser Wahrscheinlichkeiten P(D=1|X=x) ist auf Grundlage von Daten einer Fall-Kontroll-Studie jedoch nicht möglich. Aufgrund des retrospektiven Designs können in Fall-Kontroll-Studien für Fall- und Kontrollgruppe ausschließlich Expositionswahrscheinlichkeiten geschätzt werden. Diese Wahrscheinlichkeiten P(X=x|D=1) bzw. P(X=x|D=0) sind für alle möglichen Ausprägungen x des Expositionsvektors X schätzbar.

Breslow und Day (1980) zeigen, dass das logistische Regressionsmodell auch zur Auswertung von Fall-Kontroll-Studien geeignet ist. Die einzige Einschränkung bei Fall-Kontroll-Studien ist, dass dem konstanten Absolutglied (Intercept) β_0 eine inhaltlich andere Bedeutung zukommt. Die Modellparameter $\beta_1,\ldots,\beta_k$ sind hingegen weiterhin einer einfachen Interpretation zugänglich (vgl. Unterkapitel 4.3.1 und 4.3.2).

Nach Breslow und Day (1980) können Fall-Kontroll-Studien wie folgt interpretiert werden. Aus der Grundgesamtheit aller Individuen werden bei der Fall-Kontroll-Studie sowohl Erkrankte (Fälle) als auch Nichterkrankte (Kontrollen) unabhängig von ihrer Ausprägung des Expositionsvektors X mit unbekannten Wahrscheinlichkeiten p_1 bzw. p_0 zufällig ausgewählt. Kennzeichnet die Indikatorvariable Z, ob ein Individuum ausgewählt wird (Z=1) oder nicht

(Z=0), können die Auswahlwahrscheinlichkeiten p_1 (Fälle) und p_0 (Kontrollen) wie folgt dargestellt werden:

p_1 := P(Z=1|D=1, X=x) =P(Z=1|D=1) und p_0 :=P(Z=1|D=0, X=x) =P(Z=1|D=0).

Durch zweifache Anwendung der Bayesschen Formel (vgl. Bosch 1995) folgt dann für die Erkrankungswahrscheinlichkeit, einer für die Fall-Kontroll-Studie ausgewählten Person (Z=1) bei gegebenem Expositionsvektor X:

$$P(D=1 \mid Z=1, X) = \frac{P(Z=1 \mid D=1, X) \cdot P(D=1 \mid X)}{P(Z=1 \mid D=1, X) \cdot P(D=1 \mid X) + P(Z=1 \mid D=0, X) \cdot P(D=0 \mid X)}$$

$$= \frac{p_1 \cdot P(D=1 \mid X)}{p_1 \cdot P(D=1 \mid X) + p_0 \cdot P(D=0 \mid X)} \cdot \frac{(p_0 \cdot P(D=0 \mid X))^{-1}}{(p_0 \cdot P(D=0 \mid X))^{-1}}$$

$$= \frac{(p_1 / p_0) \cdot Odds\{P(D=1 \mid X)\}}{(p_1 / p_0) \cdot Odds\{P(D=1 \mid X)\} + 1}.$$

$$= \frac{\exp\{\log_e(p_1 / p_0)\} \cdot \exp\{\beta_0 + \sum_{i=1}^{k} X_i \cdot \beta_i\}}{\exp\{\log_e(p_1 / p_0)\} \cdot \exp\{\beta_0 + \sum_{i=1}^{k} X_i \cdot \beta_i\} + 1}.$$

Durch die Definition eines neuen Parameters: $\alpha := \log_e\left(\frac{p_1}{p_0}\right) + \beta_0$ ergibt sich wieder ein gewöhnliches logistisches Regressionsmodell (vgl. 4.3.1 (1)) für die Erkrankungswahrscheinlichkeit:

$$P(D=1 \mid Z=1, X) = \frac{\exp\{\alpha + \sum_{i=1}^{k} X_i \cdot \beta_i\}}{\exp\{\alpha + \sum_{i=1}^{k} X_i \cdot \beta_i\} + 1}.$$

Als Fazit kann also festgehalten werden, dass zumindest die Erkrankungswahrscheinlichkeiten der Studienteilnehmer (Z=1) in Abhängigkeit von dem Expositionsvektor X mit Hilfe des logistischen Regressionsmodells modelliert bzw. geschätzt werden können. Diese Wahrscheinlichkeiten hängen von α und damit von den beiden unbekannten Auswahlwahrschein-

lichkeiten ab. Folglich entsprechen sie nicht den Erkrankungswahrscheinlichkeiten von Individuen der Gesamtpopulation mit Expositionsvektor X.

Analog zu den Ausführungen in Unterkapitel 4.3.1 folgt allerdings für das Odds-Ratio zweier unterschiedlich exponierter Studienteilnehmer OR(x,y|Z=1) mit $x=(x_1,\ldots,x_k)^T$ und $y=(y_1,\ldots,y_k)^T$:

$$OR(x, y \mid Z = 1) = \frac{Odds\{P(D = 1 \mid Z = 1, X = x)\}}{Odds\{P(D = 1 \mid Z = 1, X = y)\}} = \ldots = \exp\{(x_1 - y_1) \cdot \beta_1 + \ldots + (x_k - y_k) \cdot \beta_k\},$$

so dass dieses nicht von α abhängt, und genau dem Erkrankungschancenverhältnis OR(x,y) zweier beliebiger Individuen der Gesamtpopulation mit den Expositionen x bzw. y entspricht:

$$OR(x, y \mid Z = 1) = OR(x, y).$$

Für das relative Risiko (vgl. Unterkapitel 3.2) liegt eine solche Eigenschaft nicht vor. Es ist leicht ersichtlich, dass aufgrund der jeweiligen Abhängigkeit von α bzw. β_0 im allgemeinen gilt:

$$RR(x, y \mid Z = 1) = \frac{P(D = 1 \mid Z = 1, X = x)}{P(D = 1 \mid Z = 1, X = y)} \neq \frac{P(D = 1 \mid X = x)}{P(D = 1 \mid X = y)} = RR(x, y).$$

Folglich können unter Zuhilfenahme logistischer Regressionsmodelle, die auf Grundlage von Daten einer Fall-Kontroll-Studie geschätzt wurden, ausschließlich Odds-Ratios und keine relativen Risiken für die Gesamtpopulation geschätzt werden.

Zusammenfassend gilt für die Auswertung von Fall-Kontroll-Studien mit Hilfe eines logistischen Regressionsmodells, dass lediglich das konstante Absolutglied α bzw. β_0 nicht interpretierbar ist. Dies führt dazu, dass Erkrankungswahrscheinlichkeiten und relative Risiken nur für die Studienpopulation und nicht für die eigentlich im Interesse stehende Gesamtpopulation geschätzt werden können. Nicht nur für die Studienpopulation, sondern auch für die interessierende Gesamtpopulation, gültig sind ausschließlich Odds-Ratio-Schätzungen.

4.3.9 Umgang mit fehlenden Variableneinträgen

Wurde das auszuwertende Datenmaterial (x_i,d_i) (i=1,…,n) nur unvollständig erhoben, stellt sich die Frage wie im Rahmen des logistischen Regressionsmodells mit fehlenden Variableneinträgen verfahren werden soll. Die folgenden Darstellungen konzentrieren sich auf den Fall, dass ausschließlich die Einflussvariablen $x_i=(x_{i,1},\ldots,x_{i,k})^T$ (i=1,…,n) lückenhaft sind.

Bereits die in Unterkapitel 4.3.4 vorgestellte Maximum Likelihood-Schätzung der unbekannten Parameter ist nur möglich, wenn die Regressormatrix X keine fehlenden Einträge aufweist. Das bedeutet, dass ohne eine Auffüllung der fehlenden Werte alle Studienteilnehmer, von denen nur ein einziger Wert der k Einflussvariablen fehlt, vollständig von der Analyse ausgeschlossen werden müssen. Insbesondere gehen damit auch die verfügbaren Merkmalswerte dieser Personen verloren, so dass unter Umständen ein erheblicher Informationsverlust resultiert.
In der Literatur (vgl. z.B. Little und Rubin 1987) werden Verfahren im Umgang mit fehlenden Werten vorgestellt, die eine Miteinbeziehung von Studienteilnehmern mit unvollständigen Merkmalswerten ermöglichen. Neben sogenannten Ad-hoc-Verfahren, die sich durch eine einfache Praktikabilität auszeichnen, werden für den Umgang mit fehlenden Werten in logistischen Regressionsmodelle mit nur zwei Einflussvariablen, von denen nur eine unvollständig erhoben wurde, auch fundiertere Ansätze vorgestellt. So zum Beispiel stellt Ibrahim (1990) einen Maximum-Likelihood-Ansatz vor und Pepe et al. (1991) beschreiben einen Pseudo-Maximum-Likelihood-Ansatz. Obwohl letztere Ansätze im Vergleich zu den Ad-hoc-Verfahren bei Simulationsstudien i.a. bessere Ergebnisse erzielen, sind sie für Modelle mit deutlich mehr als 2 Einflussvariablen nach Schumacher und Vach (1993) jedoch nicht mehr implementierbar. Schemper und Smith (1990) empfehlen deshalb in Modellen mit vielen Einflussvariablen auf einfachere Ad-hoc Verfahren zurückzugreifen.

Im Hinblick auf die Fülle und Unvollständigkeit des auszuwertenden Datenmaterials (vgl. Unterkapitel 2.2) werden deshalb in diesem Unterkapitel nur zwei praktikable Ad-hoc-Verfahren im Umgang mit fehlenden Werten vorgestellt, die im Rahmen der Datenauswertung angewendet werden können.

Für alle Ad-hoc-Verfahren gilt, dass insbesondere dann mit Ergebnisverzerrungen gerechnet werden muss, wenn die fehlenden Daten nicht zufällig auftreten, sondern einer bestimmten Systematik folgen. Eine Systematik bzw. ein Mechanismus für das Fehlen von Werten einer Einflussvariablen liegt nach Little und Rubin (1987) vor, wenn die Wahrscheinlichkeit für einen fehlenden Wert in Abhängigkeit von den Werten der anderen Einflussvariablen und/oder Zielvariablen variiert. Die Situation, in der die Wahrscheinlichkeit für fehlende Werte in Abhängigkeit von dem Wert der Zielvariable abhängt, wird in der Literatur als Missing-Randomly-At-Outcome-Mechanismus bezeichnet.

1) Indikatorvariablen-Verfahren - „Eigene Kategorie für fehlende Werte“

Ein in der Epidemiologie häufig angewandtes Verfahren im Umgang mit fehlenden Werten ist das sogenannte Indikatorvariablen-Verfahren (vgl. Yuen Fung und Wrombel 1989). Bei diesem Verfahren werden fehlende Werte einer Einflussvariablen als eigene Kategorie aufgefasst.

Bei kategorialen Variablen mit den Ausprägungen {1,…,K} wird den fehlenden Werten einfach die Ausprägung (bzw. Kategorie) K+1 zugewiesen. Anschließend kann die herkömmliche Referenzgruppen-Kodierung (vgl. Unterkapitel 4.3.2) verwendet werden, wobei die zusätzliche Kategorie „Fehlender Wert“ die Einführung einer zusätzlichen Dummy-Variablen erforderlich macht. Beachtet werden sollte lediglich, dass die zusätzliche Kategorie K+1 fehlende Werte repräsentiert, und deshalb nicht von inhaltlicher Bedeutung ist. Von Interesse sind die Odds-Ratios zwischen den regulären K Kategorien.

Bei einer lückenhaften kontinuierlichen Variablen X steht die zusätzliche Kategorie „fehlender Wert“ in keinem Verhältnis zu ihren regulären Ausprägungen. Die Variable verfügt in dieser Betrachtungsweise über die beiden Kategorien „Fehlender Wert“ und „Vorhandener Wert“, wobei in letzterer Kategorie noch zwischen den konkreten Ausprägungen zu unterscheiden ist. Dementsprechend kann genau wie bei der Referenzgruppenkodierung auch hier nur mit Hilfe einer zusätzlichen Indikator-Variablen I=I(X) eine sinnvolle Parametrisierung erreicht werden.

Zunächst wird eine Indikator-Variable I definiert, die genau dann den Wert 1 annimmt, wenn die Ausgangsvariable X einen fehlenden Wert aufweist, also nicht beobachtet wurde:

$$I = \begin{Bmatrix} 1, & & \textit{nicht} \quad \textit{beobachtet} \\ \textit{wenn} \quad X & & \\ 0, & & \textit{regulär} \quad \textit{beobachtet} \end{Bmatrix}.$$

Anschließend werden alle fehlenden Werte der Ausgangsvariablen X mit Nullen aufgefüllt.

Es entsteht eine neue Variable Z=Z(X):

$$Z = \begin{Bmatrix} X, & & I = 0 \\ & \textit{wenn} & \\ 0, & & I = 1 \end{Bmatrix}.$$

Werden nun Z und I als kontinuierliche Variablen ins Regressionsmodell aufgenommen, ergibt sich die Modellgleichung:

$$\text{Logit}(p(X)) = \alpha + Z\cdot\beta_Z + I\cdot\beta_I = \alpha + (1\text{-}I)\cdot X\cdot\beta_Z + I\cdot\beta_I.$$

Tabelle 4.B kann entnommen werden, zu welchen Logit-Werten unterschiedliche X-Werte bei dieser Parametrisierung führen. Die ersten beiden Zeilen zeigen, dass für regulär beobachtete Ausprägungen von X die übliche Parametrisierung kontinuierlicher Variablen vorliegt. Nichtvorliegende Beobachtungen von X (Zeile 3) bilden eine davon unabhängige Ausprägungskategorie. Ihr Einfluss auf den Logit wird entsprechend durch einen eigenständigen Parameter β_I charakterisiert.

Tabelle 4.B: Logit-Werte in Abhängigkeit von den X-Werten

x	**I(x)**	**Z(x)**	**Logit(p(x))**
0	0	0	α
$x \in R\backslash\{0\}$	0	x	$\alpha + x\cdot\beta_Z$
Nicht beobachtet	1	0	$\alpha + \beta_I$

Diskussion:

Vach und Blettner (1991) haben für logistische Regressionsmodelle mit wenigen (2-3) Einflussvariablen gezeigt, dass das Indikatorvariablen-Verfahren insbesondere dann zu Ergebnisverzerrungen führt, wenn die Wahrscheinlichkeit für einen fehlenden Wert nicht nur von der Zielvariablen, sondern gleichzeitig auch von anderen Einflussvariablen, abhängt. Die Ergebnisse von Simulationsstudien (vgl. Yuen Fung und Wrombel 1989) zeigen, dass das Indikatorvariablen-Verfahren zumindest beim zufälligen Fehlen von Merkmalswerten im Vergleich zu anderen ad-hoc Verfahren sehr gute Ergebnisse erzielt. In Bezug auf logistische Modelle mit vielen lückenhaften Einflussvariablen, deren Werte zudem nicht zufällig fehlen, waren in der durchgesehenen Literatur keine Angaben zur Effizienz des Indikatorvariablen-Verfahrens oder anderer Ad-hoc-Verfahren zu finden.

Theoretische Überlegungen in Anhang 5 dieser Arbeit zeigen, dass bei Vorliegen eines Missing-Randomly-At-Outcome-Mechanismus für fehlende Werte Ergebnisverzerrungen nicht ausgeschlossen werden können.

2) Unbedingte Probability-Imputation Verfahren

Bei dem unbedingten Probability-Imputation Verfahren werden fehlende Einträge einer Einflussvariablen durch den Mittelwert der vorhandenen Werte dieser Variablen ersetzt. Bei stetigen Variablen ist dies der Mittelwert der Originalmerkmalswerte. Bei kategorialen Merkmalen hingegen wird zunächst die übliche Referenzgruppenkodierung verwendet, wobei ein fehlender Merkmalswert dazu führt, dass alle zugehörigen Dummy-Indikatorvariablen einen fehlenden Wert aufweisen. Anschließend werden unabhängig für alle Dummy-Variablen die fehlenden Werte durch die Mittelwerte der vorhandenen 0/1-Dummy-Einträge ersetzt. Diese Mittelwerte entsprechen somit den relativen Häufigkeiten der beobachteten Merkmalsausprägungen.

Diskussion:

Auch das unbedingte Probability-Imputation-Verfahren wird in epidemiologischen Anwendungen im Umgang mit fehlenden Werten häufig angewendet (vgl. Yuen Fung und Wrombel 1989). In Modellen mit vielen unkorrelierten Einflussvariablen stellt das unbedingte Probability-Imputation-Verfahren für Schemper und Smith (1990) aus Gründen der Praktikabilität die Methode der Wahl dar.

In der durchgesehenen Literatur wurden keine Angaben zur Effizienz des (unbedingten) Probability-Imputation-Verfahrens im Zusammenhang mit logistischen Regressionsmodellen, bei denen viele lückenhafte Einflussvariablen vorliegen, deren Werte nicht zufällig fehlen, gemacht.

Kapitel 5. Auswertung des Datenmaterials

Das zur Verfügung gestellte Datenmaterial wird in zwei Schritten ausgewertet.
Im ersten Auswertungsschritt (vgl. 5.2) wird untersucht, ob dem Datenmaterial zu entnehmen ist, dass Brustkrebs familiär gehäuft auftritt. Da keine Daten von Kontrollfamilien zur Verfügung stehen, können die eigentlich notwendigen Vergleiche von Erkrankungshäufigkeiten innerhalb bestimmter Verwandtschaftsgruppen (z.B. Schwestern und Kinder) von Brustkrebsfällen und Nichtbrustkrebsfällen jedoch nicht vollzogen werden. Stattdessen kommen deskriptive Ansätze zum Zuge, die für eine solche Untersuchung zumindest bedingt geeignet sind. Bei Vernachlässigung nicht-familiärer Einflussgrößen können in diesem Auswertungsschritt auch die Individuen berücksichtigt werden, von denen keine ausführlicheren epidemiologische Angaben verfügbar sind.
Daran anschließend wird im zweiten Auswertungsschritt (vgl. 5.3) unter Zuhilfenahme logistischer Regressionsmodelle nach nicht-familiären Risikofaktoren für Brustkrebs gesucht. Da ausführlichere epidemiologische Angaben nur von 1198 der insgesamt 3527 Studienteilnehmer vorliegen, muss die Studienpopulation in diesem Auswertungsschritt entsprechend eingeschränkt werden. Als problematisch erweist sich in diesem Auswertungsschritt die Unvollständigkeit des Datenmaterials.

Bevor mit der Datenauswertung begonnen werden kann, ist noch zu überprüfen, inwieweit das vorliegende Datenmaterial in den beiden Auswertungsschritten sinnvoll genutzt werden kann. Im ersten Unterkapitel (5.1) werden deshalb zunächst die geschlechtsspezifischen Brustkrebsverteilungen untersucht und die Vollständigkeit des Datenmaterials überprüft. Dieses Unterkapitel stellt somit in gewisser Weise eine Ergänzung zur Beschreibung des Datenmaterials in Unterkapitel 2.2 dar.

5.1 Datenmaterial

Im Hinblick auf den ersten Auswertungsschritt, in welchem nach Anzeichen für eine familiäre Häufung der Brustkrebskrankheit gesucht wird, stellt sich zunächst die Frage nach der Brustkrebshäufigkeitsverteilung in der Studienpopulation. Tabelle 5.A können neben der Brustkrebshäufigkeitsverteilung aller 3527 Studienteilnehmer auch die geschlechtsspezifischen Verteilungen entnommen werden.

Tabelle 5.A Brustkrebshäufigkeitsverteilungen in der Studienpopulation:

Brustkrebs-Status:	**Studienpopulation**		
	männlich	**weiblich**	**gesamt**
unbekannt	237	137	374
0 - kein BCA	1451	1342	2793
1 - BCA	4	356	360
Σ	**1692**	**1835**	**3527**

Da nur 4 der 1692 männlichen Studienteilnehmer an einer Brustkrebserkrankung leiden, können die Männer in keinem der beiden Auswertungsschritte sinnvoll berücksichtigt werden. Wenngleich im ersten Auswertungsschritt noch denkbar wäre, die Untersuchungen unabhängig für beide Geschlechter durchzuführen, liegen zu wenige männliche Brustkrebsfälle vor, als dass aussagekräftige Ergebnisse zu erwarten wären.

Im zweiten Auswertungsschritt ist eine Berücksichtigung vollkommen ausgeschlossen, da es sich bei den vier männlichen Brustkrebsfällen um die einzigen Männer handelt, von denen zusätzliche Informationen über nichtfamiliäre Expositionen vorliegen. Die erste Problemlösung, das Geschlecht zu vernachlässigen und somit für die vier Männer dieselben nichtfamiliären Risikofaktoren bzw. biologischen Mechanismen zu unterstellen wie für Frauen, ist aus medizinischer Sicht nicht vertretbar (Grundmann 1994). Auf der anderen Seite würde die Aufnahme des Geschlechts als binäre Einflussvariable ebenfalls nicht zu sinnvollen Ergebnissen führen. Wenn die einzigen vier Männer, deren Berücksichtigung möglich ist, an Brustkrebs erkrankt sind, würde der Geschlechtsstatus „maskulin“ eindeutig Brustkrebs implizieren.

Fazit:

Bereits ein erster Blick auf die Daten (vgl. Tabelle 5.A) zeigt, dass lediglich die Daten der weiblichen Studienteilnehmer sinnvoll ausgewertet werden können. Aus den Daten der 1692 männlichen Individuen können keine aussagekräftigen Informationen gewonnen werden. Für den ersten Auswertungsschritt (Unterkapitel 5.2) bedeutet dies, dass die Analysen auf die Daten der 1835 weiblichen Studienteilnehmer beschränkt werden müssen. Von den Analysen im zweiten Auswertungsschritt (vgl. Unterkapitel 5.3) müssen von den 1198 Individuen, über die

ausführlichere Informationen vorliegen, die 4 männlichen Brustkrebsfälle ausgeschlossen werden.

Problematisch im Hinblick auf den zweiten Auswertungsschritt, in welchem es um die Identifikation nicht-familiären Risikofaktoren geht, ist zudem, dass aufgrund der Beschaffenheit des Datenmaterials keine sinnvolle Möglichkeit besteht, die familiären bzw. genetischen Abhängigkeiten zwischen den Angehörigen derselben Familie in den Regressionsmodellen zu berücksichtigen. Insbesondere das Fehlen von Kontrollfamilien ohne sicheren Brustkrebsfall führt dazu, dass der einzige sinnvolle Untersuchungsansatz darin besteht, die brustkrebserkrankten mit den nichtbrustkrebserkrankten Frauen hinsichtlich ihrer Expositionen zu vergleichen. In Anbetracht der notwendigen Vernachlässigung der familiären Abhängigkeiten erinnert dies an die Vorgehensweise bei der Auswertung einer epidemiologischen Fall-Kontroll-Studie (vgl. 3.2).

Wenngleich das Design der vorliegenden Studie (vgl. 2.2) ganz sicher nicht dem einer klassischen Fall-Kontroll-Studie entspricht, erscheint es deshalb notwendig, die epidemiologischen Daten als Ergebnis einer Fall-Kontroll-Studie aufzufassen. Bei bewusster Vernachlässigung der familiären Abhängigkeiten bilden in dieser Betrachtungsweise die 249 weiblichen Indexfälle zusammen mit ihren 73 brustkrebserkrankten Familienangehörigen eine „Fallgruppe“ der Größe n_F=326, die hinsichtlich verschiedener Expositionen mit den n_K=868 nichtbrustkrebserkrankten Familienangehörigen („Kontrollgruppe“) zu vergleichen ist. Bei Auswertung dieser „Fall-Kontroll-Studie“ stellen die familiären Abhängigkeiten unkontrollierte Störgrößen dar, deren Einflüsse sich den Beziehungen zwischen Expositionen und BCA-Krankheit möglicherweise in störender Weise überlagern.

Obwohl damit die Vorgehensweise für den zweiten Auswertungsschritt festgelegt ist, soll bereits in diesem Kapitel darauf eingegangen werden, dass die Unvollständigkeit des Datenmaterials die Auswertung erheblich verkompliziert. Das größte Problem in diesem Zusammenhang ist, dass die fehlenden Dateneinträge nicht zufällig auftreten, sondern einer bestimmten Systematik folgen. Genauer gilt, dass fehlende Werte vorwiegend bei den Familienangehörigen der Indexfälle zu finden sind, wohingegen die Angaben zu den Indexfällen weitgehend vollständig sind. In der oben motivierten Betrachtungsweise als Fall-Kontroll-Studie führt dies zu dem folgenden, ungünstigen, Sachverhalt. In der „Fallgruppe“, die sich zum größten Teil (249 von 326) aus den Indexfällen zusammensetzt, liegen hinsichtlich aller relevanten

Merkmale, nur wenige fehlende Werte vor. In der Kontrollgruppe, die ausschließlich von Familienangehörigen der Indexfälle gebildet wird, liegen hingegen deutlich größerer Anteile fehlender Merkmalswerte vor.

Tabelle 5.B können für 6 Beispielsmerkmale die relativen Anteile fehlender Werte in Fall- und Kontrollgruppe entnommen werden.

Tabelle 4.1: Relative Anteile fehlender Werte in Fall- und Kontrollgruppe für 6 epidemiologische Beispielsmerkmale

Merkmal bzw. Angabe über…	**Fallgruppe (n_F=326)**	**Kontrollgruppe (n_K=868)**
Anzahl Lebendgeburten	0,015	0,044
Anzahl Fehlgeburten	0,039	0,076
Fettleibigkeit	0,148	0,498
Schulbildung	0,152	0,546
Regelmäßigkeit der Periode	0,170	0,555
Durchschnittliche Periodendauer	0,330	0,772

Die zu erkennende Assoziation zwischen der Gruppenzugehörigkeit und der Häufigkeit von fehlenden Werten kann auch formal nachgewiesen werden. Für jede Frau ist bekannt, ob sie zur Fall- oder Kontrollgruppe gehört, und in Bezug auf jedes Merkmal kann bei ihr entweder ein regulärer Datenwert vorliegen oder kein Wert vorliegen. Entsprechend kann mit Hilfe eines Chi-Quadrat-Tests auf Unabhängigkeit (vgl. 4.1) für jedes Merkmal überprüft werden, ob eine Abhängigkeit zwischen der Gruppenzugehörigkeit und dem Nichtvorliegen von Merkmalswerten besteht.

Tatsächlich kann die Nullhypothese der Unabhängigkeit für jedes Merkmal (vgl. Anhang 4) zum Niveau 5% verworfen werden.

Für das Merkmal „Anzahl Lebendgeburten“ ergibt sich zum Beispiel die auf der folgenden Seite dargestellte Kontingenztafel (Tafel 4.2).

Tafel 4.2: Gruppenzugehörigkeit – Merkmalswert (vorhanden/ nicht vorhanden)

Merkmal „Fehlender Wert“	Gruppenzugehörigkeit:		
	Fallgruppe	**Kontrollgruppe**	**Σ**
Merkmalswert fehlt	4 (11)	38 (31)	**42**
Wert vorhanden	322 (315)	830 (837)	**1152**
Σ	**326**	**868**	**1198**

Ein Vergleich der beobachteten Zellhäufigkeiten mit den unter Unabhängigkeit zu erwartenden (in der Tafel in Klammern angegeben) zeigt, dass in der Fallgruppe (Kontrollgruppe) weniger (mehr) fehlende Werte vorliegen, als unter Unabhängigkeit von Gruppenzugehörigkeit und fehlenden Werten zu erwarten ist. Ein deskriptiv durchgeführter Chi-Quadrat-Test auf Unabhängigkeit (vgl. 4.1) liefert einen P-Wert von 0,0085, so dass die Nullhypothese der Unabhängigkeit zum Niveau 5% verworfen werden könnte.

Da die fehlenden Werte somit nachweislich nicht zufällig auftreten, sondern einer bestimmten Systematik folgen, die in der Fachliteratur auch als Missing-Randomly-At-Outcome-Mechanismus bezeichnet wird, kommen im zweiten Auswertungsschritt im Umgang mit fehlenden Werten zwei spezielle Verfahren zum Einsatz. Beide Ad-hoc-Verfahren wurden im Methodik-Kapitel dieser Arbeit (vgl. 4.3.9) vorgestellt.

5.2 Auswertungsschritt 1: Familiäre Häufung

Der erste Auswertungsschritt beschäftigt sich mit der Frage, ob familiäre bzw. genetische Vorbelastungen zur Entstehung von Brustkrebs beitragen. Da das vorliegende Datenmaterial keine Daten von Kontrollfamilien umfasst, ist nicht unbedingt zu erwarten, dass diese Frage zufriedenstellend beantwortet werden kann. Vielmehr zielt dieser erste Schritt lediglich darauf, Anzeichen zu finden, die die Vermutung, einer familiären Häufung von Brustkrebs, nahe legen. Hierzu werden ausschließlich einfache deskriptive statistische Methoden eingesetzt.

5.2.1 Ungeeignete Ansätze

Dieses erste Unterkapitel diskutiert zwei intuitiv naheliegende Untersuchungsansätze, die jedoch im vorliegenden Fall aufgrund des Studiendesigns nicht zu sinnvollen Ergebnissen führen.

Klassifizierung der Familien

Eine erste Idee wäre es, die 253 Familien in Abhängigkeit von der Anzahl aufgetretener Brustkrebsfälle in zwei Klassen aufzuteilen. Eine erste Klasse, bestehend aus Familien mit nur wenigen Brustkrebsfällen, und eine zweite Klasse mit einer erhöhten Anzahl von Brustkrebsfällen. Diesem Ansatz liegt die Idee zu Grunde, dass der Brustkrebs, des in Behandlung befindlichen Familienmitglieds (Indexfalls), in einigen Familien nicht durch familiäre Dispositionen hervorgerufen wurde, in anderen Familien hingegen die Konsequenz einer solchen genetischen Vorbelastung darstellt. Mit Hilfe einer solchen Klassifizierung der Familien könnte bei späterer Betrachtung von Regressionsmodellen eine binäre Kovariable definiert werden, die jeder Frau in Abhängigkeit von ihrer Familienzugehörigkeit eine genetische Belastung unterstellt oder nicht. Im Hinblick auf die Untersuchung nichtfamiliärer Risikofaktoren im zweiten Auswertungsschritt könnte so für eine gewisse Adjustierung gesorgt werden.

Wenngleich das Studiendesign in diesem Zusammenhang kein unmittelbares Problem darstellt, zeigt sich, dass aufgrund der unterschiedlichen Familiengrößen keine sinnvolle Klassifizierung vorgenommen werden kann. Bei Verwendung von absoluten Häufigkeiten von Brustkrebsfällen pro Familie kann die Familiengröße nicht berücksichtigt werden, so dass nicht klar wird, inwieweit große absolute Häufigkeiten von BCA-Fällen mit der Familiengröße assoziiert sind. Denn auch wenn Brustkrebs nicht familiär gehäuft auftritt, ist zu erwarten, dass große Familien durchschnittlich mehr (absolute) Brustkrebsfälle aufweisen als kleine Familien. Bei Verwendung von relativen Häufigkeiten ergibt sich zunächst das Problem, dass die Indexfälle in jeder Familie einen sicheren Fall darstellen und somit gerade bei den kleinen Familien zu einer hohen relativen Häufigkeit von Brustkrebsfällen führen. Tabelle 5.C zeigt, dass relative Häufigkeiten auch bei Vernachlässigung der 253 Indexfälle keine adäquate Kenngröße darstellen.

Der zweiten und dritten Spalte dieser Tabelle können die absoluten bzw. prozentualen Häufigkeiten von Familien entnommen werden, in denen die in der ersten Spalte angegebene Anzahl zusätzlicher BCA-Fälle (außer Indexfall) vorliegt. In der vierten Spalte ist für die betref-

fenden Familien jeweils die durchschnittliche Anzahl Angehöriger der Indexfälle angegeben. Offensichtlich gilt, dass auch bei Vernachlässigung der Indexfälle weiterhin mit der durchschnittlichen Familiengröße die absolute Häufigkeit von zusätzlichen BCA-Fällen steigt. Eine große relative Häufigkeit von zusätzlichen BCA-Fällen ist hingegen nur bei entsprechend kleinen Familien zu erwarten. Die Familie mit 6 zusätzlichen BCA-Fällen kommt zum Beispiel lediglich auf eine relative Häufigkeit von 6/36=1/6. Eine genauere Untersuchung zeigt, dass diese relative Häufigkeit von 16 der 55 Familien mit nur einem zusätzlichen BCA-Fall (bedingt durch die geringe Familiengröße) überschritten wird.

Tabelle 5.C: Häufigkeitstabelle hinsichtlich zusätzlicher BCA-Fälle in den Familien:

Zusätzliche BCA-Fälle in der Familie	Anzahl Familien	Prozentualer Anteil Familien	Durchschnittliche Anzahl Angehöriger
0	178	70,92 %	5,056
1	55	21,91 %	7,109
2	15	5,98 %	13,800
3	2	0,00 %	16,000
4	0	0,00 %	---
5	0	0,00 %	---
6	1	0,40 %	36,000
Σ	**251**	**100,00 %**	**---**

Generationsübergreifende Häufigkeitsanalysen

Da nach gegenwärtigem Wissenstand in der Humanmedizin davon ausgegangen wird, dass Töchter und Schwestern von Frauen mit Brustkrebs ebenfalls ein erhöhtes Erkrankungsrisiko haben (vgl. Grundmann 1994), erscheint es sinnvoll, zu untersuchen, ob sich diese Expositionen (Schwester bzw. Mutter an BCA erkrankt) auch mit Hilfe der vorliegenden Daten als Risikofaktoren für Brustkrebs bestätigen lassen.

Bei diesen beiden Expositionen kann im Gegensatz zu den nichtfamiliären Expositionen (vgl. 5.3) jedoch keineswegs unberücksichtigt bleiben, dass es sich um Familiendaten handelt. Bei Vernachlässigung dieses Aspekts bzw. bei Betrachtung der Daten als (generationsübergreifende) Fall-Kontroll-Studie, in der die brustkrebskranken Frauen die Fallgruppe und die nichtbrustkrebskranken Familienangehörigen die Kontrollgruppe bilden, ergibt sich bei bei-

den Expositionen eine Scheinassoziation zwischen Exposition und Krankheit, die nur Konsequenz des tatsächlichen Studiendesigns ist. Von besonderer Bedeutung in diesem Zusammenhang ist, dass die Exposition eines Familienmitglieds automatisch auch den Krankheitsstatus eines anderen Familienmitglieds darstellt und umgekehrt.

Betrachtet man zum Beispiel n Schwestern, von denen nur eine an Brustkrebs leidet, ergeben sich bei Vernachlässigung der familiären Beziehung (n-1) nichtkranke Frauen mit der Exposition („erkrankte Schwester“) und eine kranke Frau ohne Exposition („erkrankte Schwester“). Die Konstellation „exponiert und erkrankt“ kann nur auftreten, wenn mindestens 2 der Schwestern an Brustkrebs erkrankt sind. Die Konstellation „nichtexponiert und nichterkrankt“ hingegen setzt voraus, dass alle n Schwestern gesund sind, und tritt dann automatisch bei allen n Schwestern auf.
Letzterer Punkt stellt im vorliegenden Fall das Hauptproblem bei Untersuchung der Exposition „erkrankte Schwester“ dar. Von den 1835 Frauen stammen 895 aus der Generation der Indexfälle. Diese Generation setzt sich aus den Indexfällen und ihren Schwestern und Cousinen zusammen. Bei Vernachlässigung der Familien, in denen der Indexfall ein Einzelkind ist oder sich aufgrund fehlender Angaben zu den Eltern keine Schwestern des Indexfalls ermitteln lassen, besteht diese Generation noch aus 242 Indexfällen sowie 516 Schwestern und 128 Cousinen dieser sicheren BCA-Fälle. Die 516 Schwestern der Indexfälle sind als solche natürlich exponiert. Dies betrifft neben den 29 an Brustkrebs erkrankten insbesondere auch die 487 nichtbrustkrebserkrankten Schwestern. Dies führt dazu, dass insgesamt bei 758 Frauen (Indexfälle und ihre Schwestern) die Konstellation „nichtexponiert und nichterkrankt“ bedingt durch das Studiendesign nicht auftreten kann. Für diese 758 Frauen ergibt sich die folgende Vierfeldertafel.

Tafel 5.D: Vierfeldertafel: Brustkrebs-Status gegen Exposition „erkrankte Schwester“ für die Indexfälle und ihre Schwestern

BCA-Status: / **Expositions-Status:**	**0** **(kein Brustkrebs)**	**1** **(Brustkrebs)**	**Σ**
0 - nicht exponiert	0	214	**214**
1- erkrankte Schwester	487	57	**544**
Σ	**487**	**271**	**758**

Bei den 214 brustkrebserkrankten und nichtexponierten Frauen kann es sich bedingt durch das Studiendesign nur um Indexfälle handeln, da alle Schwestern der BCA-Indexfälle als solche natürlich exponiert sind. Die 544 exponierten Frauen setzen sich aus den 516 Schwestern der Indexfälle und 28 Indexfällen, von denen auch eine Schwester an BCA erkrankt ist, zusammen. Die 28 exponierten Indexfälle gehören - als sichere BCA-Fälle - genauer zu den 57 erkrankten exponierten Frauen..
Bei Durchführung einer analogen Untersuchung für alle weiblichen Studienteilnehmer, das heißt bei Betrachtung aller brustkrebserkrankten Frauen als Fallgruppe und aller nichtbrustkrebserkrankten Frauen als Kontrollgruppe, ist somit für über ein Drittel der 1895 Frauen die Konstellation „nichtexponiert und nichterkrankt“ ausgeschlossen. Berücksichtigt man nun, dass gerade das gemeinsame gehäufte Auftreten der Konstellationen „nichtexponiert und nichterkrankt“ und „exponiert und erkrankt“ für eine Erhöhung des Erkrankungsrisikos bei vorliegender Exposition sprechen, ist offensichtlich, dass eine solche Analyse studiendesignbedingt zu keinem sinnvollen Ergebnis führen kann. Bedingt durch die Überlagerung mit der Indexfall-Generation kommt es unabhängig vom wahren Zusammenhang zwischen Exposition und Krankheit bei gleichzeitiger Untersuchung der 1835 Frauen in jedem Fall zu einer deutlichen Unterschätzung des Erkrankungschancenverhältnisses zwischen exponierten und nicht-exponierten Frauen.

Ein ähnliches Problem ergibt sich hinsichtlich der Untersuchung der Exposition „erkrankte Mutter“. Von besonderer Bedeutung ist hier, dass Brustkrebs genau wie alle anderen Krebsarten eine eher seltene Krankheit darstellt, so dass selbst bei ungünstigster Risikofaktorkonstellation zwar ein entsprechend erhöhtes aber immer noch ein verhältnismäßig geringes Erkrankungsrisiko besteht. Folglich liegen in der Generation der Indexfälle deutlich mehr Brustkrebsfälle vor als in allen anderen Generationen.

Im Folgenden wird die Generation der Indexfälle als dritte Generation bezeichnet. Diese setzt sich aus den Indexfällen sowie den (Halb-)Schwestern und Cousinen dieser Fälle zusammen. Die Eltern und Tanten der Indexfälle bilden die zweite und die Großmütter und Großtanten der Indexfälle die erste Generation. Die jüngeren Generationen setzen sich aus den Töchtern und Nichten (Generation 4) bzw. Enkeltöchtern (Generation 5) der Indexfälle zusammen.

Die Indexfälle zusammen mit ihren (Halb-)Schwestern und Cousinen bilden die dritte Generation, so dass es zu einer Ballung von Brustkrebsfällen in der dritten Generation kommt, wohingegen es in den anderen Generationen (insbesondere der vorangegangenen zweiten und der nachfolgenden vierten Generation) deutlich weniger Brustkrebsfälle gibt. Dies führt zu dem ungünstigen Umstand, dass die vorwiegend nichtbrustkrebserkrankten Frauen der vierten Generation, zum größten Teil Töchter der Indexfälle aus der dritten Generation sind. Was unter den Frauen der vierten Generation zu einer entsprechenden Häufung der Konstellation „exponiert und nichterkrankt“ führt. In der zweiten Generation, zu der auch die Mütter der Indexfälle gehören, liegt ebenfalls eine deutlich geringere Anzahl von Brustkrebsfällen vor als in der dritten Indexfall-Generation. Entsprechend haben die Indexfälle größtenteils nichtbrustkrebserkrankte Mütter, was zu einer Häufung der Konstellation „nichtexponiert und erkrankt“ in der dritten Generation führt.

Das durchs Studiendesign bedingte gehäufte Auftreten dieser beiden Konstellationen spricht gegen eine Mutter-Tochter-Vererbung der Brustkrebskrankheit. Unter Berücksichtigung, dass unabhängig vom Expositions-Status in jedem Fall von einem geringen Risiko, an Brustkrebs zu erkranken, auszugehen ist, kommt es bei Überlagerung des wahren Zusammenhangs durch diese gehäuften Konstellationen zu einer deutlichen Unterschätzung des Chancenverhältnisses. Tatsächlich präsentiert sich fälschlicherweise bzw. durchs Studiendesign bedingt die Exposition „brustkrebskranke Mutter“ als deutlicher Schutzfaktor vor Brustkrebs.

Fazit

Die Untersuchungen zeigen, dass die Einflüsse der Expositionen „erkrankte Schwester“ und „erkrankte Mutter“ auf das Brustkrebsrisiko auf Grundlage des vorliegenden Datenmaterials nicht sinnvoll untersucht werden können.

5.2.2 Brustkrebsverteilung in den Generationen

Nachdem im letzten Unterkapitel verdeutlicht wurde, dass das Datenmaterial aufgrund der familiären Abhängigkeiten und der sicheren BCA-Indexfälle in der 3ten Indexfall-Generation eine ungünstige Struktur aufweist, soll nun ein Überblick über die Brustkrebsverteilung in den fünf Generationen gegeben werden. Tabelle 5.E zeigt die generationsspezifischen Brustkrebshäufigkeitsverteilungen.

Tabelle 5.E: Brustkrebsverteilung in den fünf Generationen (G-1 bis G-5)

BCA-Status (0,1)-kodiert	Generation				
	Vorfahren		Indexfälle	Nachkommen	
	G-1	**G-2**	**G-3**	**G-4**	**G-5**
Unbekannt	57	55	22	3	0
0	76	433	580	250	3
1	3	54	293	6	0

Die 293 Brustkrebsfälle der dritten Generation setzen sich aus 253 Brustkrebsfällen in Behandlung sowie 29 Schwestern und 11 Cousinen dieser Fälle zusammen. Die 54 Brustkrebsfälle der zweiten Generation umfassen 22 Mütter und 32 Tanten der Indexfälle. Bei den 3 BCA-Fällen der 1ten Generation handelt es sich um Großmütter der Indexfälle. Unter den 6 BCA-Fällen der 4ten Generation befinden sich drei Töchter der Indexfälle. Bei den anderen 3 BCA-Fällen dieser Generation ist aufgrund fehlender Angaben zu den Familienbeziehungen nicht klar, ob es sich um Töchter oder Nichten der Indexfälle handelt.
Da nur in der 2ten (54) und 3ten (293) Generation „hinreichend“ viele Brustkrebsfälle vorliegen, erscheint es sinnvoll, sich im weiteren auf diese beiden Generationen zu konzentrieren. Aufgrund der Beschaffenheit des Datenmaterials können lediglich spezielle Untersuchungen durchgeführt werden. Stellvertretend für einige Ansätze, mit denen versucht wurde, Anzeichen für eine familiäre Häufung der Brustkrebskrankheit zu finden, soll an dieser Stelle lediglich auf zwei näher eingegangen werden.

Ansatz 1

Ansatz 1 konzentriert sich ausschließlich auf die 253 Mütter der Indexfälle. Es soll untersucht werden, ob die Töchter (3te Generation) der brustkrebserkrankten Mütter (2te Generation) zu einem größeren Anteil an Brustkrebs erkrankt sind als die Töchter der gesunden Mütter. Da die 253 Mütter unterschiedlich viele Töchter haben und, da sich unter diesen mit dem Indexfall stets ein sicherer BCA-Fall befindet, treten grundsätzlich dieselben Probleme auf, die schon in Unterkapitel 5.2.1 im Zusammenhang mit der Klassifizierungsidee erörtert wurden. Im vorliegenden Fall ist es allerdings nicht notwendig die Familien einzeln zu betrachten, sondern es besteht die Möglichkeit jeweils die Töchter der brustkrebserkrankten und die Töchter der gesunden Indexfall-Mütter zusammenzufassen, so dass die Gesamtheiten größer und damit die relativen Anteile aussagekräftiger werden.

Konkret wird wie folgt vorgegangen:
Die Töchter der brustkrebserkrankten Indexfall-Mütter werden als erste Grundgesamtheit und die Töchter der gesunden Indexfall-Mütter als zweite Grundgesamtheit aufgefasst. Anschließend werden die Anteile brustkrebserkrankter Frauen für beide Grundgesamtheiten berechnet und miteinander verglichen. Zu beachten ist allerdings, dass die Mütter unterschiedlich viele Töchter haben und sich unter diesen mit dem Indexfall jeweils ein sicherer Brustkrebsfall befindet. Setzt sich eine Grundgesamtheit festen Umfangs aus den Töchtern vieler Indexfall-Mütter zusammen, befinden sich somit zwangsläufig auch entsprechend viele Indexfälle in der Grundgesamtheit, was automatisch einen größeren Anteil Erkrankter zur Konsequenz hat. Entsprechend empfiehlt es sich nicht, die sicheren BCA-Indexfälle in den Grundgesamtheiten zu berücksichtigen. Darüber hinaus können 9 Familien nicht berücksichtigt werden, bei denen keine Angaben über Schwestern oder Mütter der Indexfälle verfügbar sind.

Die Tabellen 5.F zeigt die Ergebnisse, die sich bei Vernachlässigung der Indexfälle durch (computergestützte) Auszählungen ergeben. Die zweite Spalte der Tabelle gibt an, wie viele Indexfall-Mütter den entsprechenden Brustkrebs-Status (Spalte 1) aufweisen. Der dritten Spalte kann entnommen werden, wie viele Töchter N_i diese Mütter zusammen haben, wobei die Indexfälle als sichere BCA-Fälle nicht mitgezählt werden. Die vierte Spalte zeigt, wie viele der Töchter aus der dritten Spalte an Brustkrebs erkrankt sind. Welchen relativen Anteilen von BCA-Erkrankungen unter den Schwestern der Indexfälle das entspricht, kann der letzten Spalte entnommen werden.

Tabelle 5.F: Relativer Anteil BCA-Fälle unter den Schwestern der Indexfälle

BCA-Status der Indexfall-Mutter	Absolute Häufigkeit	Anzahl Nicht-Indexfall-Töchter N_i	davon BCA-Fälle p_i	Relativer Anteil p_i/N_i
Unbekannt	4	16	0	0,000
0 – kein BCA	218	468	25	0,053
1 – BCA erkrankt	22	51	2	0,039

Von Interesse sind die letzten beiden Zeilen von Tabelle 5.F. Diesen kann entnommen werden, dass eine BCA-Erkrankung der Mutter offensichtlich nicht zu einer Erhöhung des Erkrankungsrisikos ihrer Töchter führt. Die 22 erkrankten Indexfall-Mütter haben neben den 22

sicher an Brustkrebs erkrankten Indexfällen zusammen noch 51 weitere Töchter. Von diesen sind lediglich 2 und damit ein relativer Anteil von 0,039 an Brustkrebs erkrankt. Von den 468 Schwestern der Indexfälle mit gesunden Mütter sind 25 und damit ein relativer Anteil von 0,053 erkrankt. Ein Vergleich dieser relativen Häufigkeiten zeigt, dass kein Anzeichen für eine familiäre Häufung - im Sinne einer Mutter-Tochter-Vererbung – gegeben ist.

Ansatz 2

Ansatz 2 stellt eine Weiterführung des ersten Ansatzes dar. Da neben den 22 Müttern auch noch 32 Tanten der Indexfälle an Brustkrebs erkrankt sind, soll bei diesem Ansatz auch noch berücksichtigt werden, ob blutsverwandte Tanten der Indexfälle an Brustkrebs leiden. Da nur in wenigen Familien mehr als eine blutsverwandte Tante an Brustkrebs erkrankt ist, wird bezüglich der Tanten lediglich unterschieden, ob es mindestens einen Fall gegeben hat oder nicht. Entsprechend sind lediglich die relativen Anteile von vier Grundgesamtheiten miteinander zu vergleichen. Tabelle 5.G zeigt die Ergebnisse der Auszählung, die sich bei Vernachlässigung der Indexfälle ergeben:

Die ersten beiden Spalten legen in Abhängigkeit von den Erkrankungszuständen der Mütter und Tanten der Indexfälle die vier Grundgesamtheiten fest. Analog zu Tabelle 5.F kann den nachfolgenden beiden Spalten entnommen werden, wie viele Indexfälle über eine entsprechende Mutter und Tante verfügen und wie viele Schwestern N_i diese Indexfälle jeweils zusammen haben. Die fünfte bzw. sechste Spalte zeigt, wie viele der Schwestern (4te Spalte) an Brustkrebs erkrankt sind bzw., welchen relativen Anteilen von BCA-Erkrankungen unter den Schwestern der Indexfälle das entspricht.

Tabelle 5.G: Erkrankungshäufigkeiten in Abhängigkeit von den Brustkrebszuständen der Mütter und Tanten

BCA-Status der Mutter	Tanten mit BCA	Anzahl Familien	Anzahl zusätzlicher Töchter	Zusätzliche BCA-Fälle	Relativer Anteil
0	**0**	201	431	24	0,056
0	**1**	17	37	1	0,027
1	**0**	15	35	1	0,029
1	**1**	7	16	1	0,063

Offensichtlich liefert auch der zweite Ansatz keinen Anhaltspunkt für eine familiäre Häufung von Brustkrebs. Obwohl die 16 Frauen, die sowohl eine BCA-erkrankte Mutter als auch Tante

haben, den größten relativen Anteil aufweisen, erscheint es fragwürdig, ob sich die 4 Anteile p_i (i=1,…,4) statistisch signifikant unterscheiden.

Die Nullhypothese: H_0: $p_1=p_2=p_3=p_4$ kann nach Hartung (1995) mit Hilfe eines Chi-Quadrat-Tests auf Unabhängigkeit (vgl. 4.1) überprüft werden. Die Nullhypothese ist nämlich gleichbedeutend damit, dass das Merkmal BCA-Status (Ja/ Nein) der Töchter von der Grundgesamtheit (1,2,3,4) unabhängig ist. Die dazugehörige Kontingenztafel ergibt sich durch Auszählen der Häufigkeiten von BCA- und Nicht-BCA-Fällen in den vier Grundgesamtheiten G(i,j), wobei i den Brustkrebsstatus der Mutter wiedergibt und j kennzeichnet, ob mindestens eine blutverwandte Tante an BCA erkrankt ist (j=1) oder nicht (j=0) (i,j=1,2).

Nach Yarnold (1970) ist der Chi-Quadrat-Test trotz dreier (n_0=3) Zellhäufigkeiten von 1(≤5) anwendbar, da für alle 8 Zellhäufigkeiten gilt, dass sie größer 5·no/n= 5·3/519=0,03 sind (vgl. 4.1). Nach anderen Autoren ist die Anwendbarkeit des Chi-Quadrat-Tests in dieser Situation nicht zulässig (vgl. zum Beispiel Cochran 1954), so dass das Testergebnis in Anbetracht von drei sehr geringen Zellhäufigkeiten in jedem Fall mit äußerster Vorsicht zu betrachten ist. Dennoch wird ein deskriptiver Chi-Quadrat-Test auf Unabhängigkeit durchgeführt.

Tabelle 5.H: Kontingenztafel BCA-Status – Grundgesamtheit

BCA	G(0,0)	G(0,1)	G(1,0)	G(1,1)	Σ
0 – Nein	407	36	34	15	492
1 – Ja	24	1	1	1	27
Σ	**431**	**37**	**35**	**16**	**519**

Der P-Wert von 0,80 des Chi-Quadrat-Tests, gibt keinen Anhaltspunkt dafür, dass die Nullhypothese falsch und damit die Anteile signifikant verschieden sind. Folglich konnte auch mit dem zweiten Ansatz kein Anzeichen für eine familiäre Häufung der Brustkrebskrankheit gefunden werden

5.2.3 Diskussion der bisherigen Ergebnisse

Die Frage, ob Brustkrebs familiär gehäuft auftritt, kann auf Grundlage der vorliegenden Daten nicht beantwortet werden.

Das Hauptproblem ist, dass das Datenmaterial lediglich die Brustkrebszustände der Familienangehörigen von 257 Brustkrebsfällen umfasst und somit keine Daten über Kontrollfamilien beinhaltet. Da das Nichtvorhandensein von Kontrollfamiliendaten dazu führt, dass die eigentlich notwendigen statistischen Vergleiche, von Erkrankungshäufigkeiten innerhalb bestimm-

ter Verwandtschaftsgruppen (Geschwister, Kinder usw.) von Brustkrebs- und Nichtbrustkrebsfällen (vgl. Cohen et al. 1993), nicht durchführbar sind, konnte lediglich mittels speziellerer Vergleiche nach Anzeichen für eine familiäre Häufung gesucht werden. Allerdings liefern auch die Vergleiche, die im vorliegenden Fall sinnvoll erscheinen, keine auffälligen Ergebnisse und damit nichts Sachdienliches, zur Beantwortung der Frage, ob Brustkrebs familiär gehäuft auftritt.

Zur endgültigen Beantwortung der Frage, ob Brustkrebs familiär gehäuft auftritt, ist in jedem Fall eine erneute Datenerhebung erforderlich. Wie eine solche Datenerhebung aussehen könnte, wird im folgenden Unterkapitel kurz beschrieben.

5.2.4 Ausblick auf weiterführende Untersuchungen

Eine bestimmte Krankheit tritt definitionsgemäß genau dann familiär gehäuft auf, wenn Individuen, die in einer bestimmten Verwandtschaftsbeziehung zu einer erkrankten Person stehen, ein erhöhtes Erkrankungsrisiko aufweisen. Zur Untersuchung, ob eine Krankheit familiär gehäuft auftritt, wird deshalb die Verwandtschaft eines Individuums mit einer erkrankten Person als Exposition aufgefasst und wie üblich mit Hilfe einer Fall-Kontroll-Studie oder Kohortenstudie untersucht, ob diese Exposition einen Risikofaktor darstellt.

In einer Kohortenstudie wird die Erkrankungshäufigkeit von Individuen, die in einer bestimmten Verwandtschaftsbeziehung zu einer erkrankten Person stehen, mit der Erkrankungshäufigkeit von Individuen verglichen, die nicht in einer solchen Verwandtschaftsbeziehung zu einer erkrankten Person stehen. Als Maß für die Assoziation zwischen Exposition und Krankheit kann das relative Risiko verwendet werden.
In einer Fall-Kontroll-Studie hingegen wird eine Gruppe Erkrankter mit einer Gruppe Nichterkrankter hinsichtlich der Exposition „bestimmte Verwandtschaftsbeziehung zu einem Erkrankten“ verglichen und als Maß für die Assoziation zwischen Exposition und Krankheit kann ausschließlich das Odds-Ratio verwendet werden.

In Abhängigkeit davon, welche Verwandtschaftsbeziehung zu einem Erkrankten als Exposition untersucht werden soll, muss das vorliegende Datenmaterial entweder zu einer Fall-Kontroll-Studie oder zu einer Kohortenstudie ergänzt werden. Da sich aus dem vorliegenden Datenmaterial bereits einige Auffälligkeiten erkennen lassen, sollten ergänzende Datenerhebungen gezielt vorgenommen werden.

So zum Beispiel sind von den 516 Schwestern, der 253 BCA-Indexfälle, 29 ebenfalls an Brustkrebs erkrankt, was die Vermutung nahe legt, dass die Exposition „Schwester eines BCA-Falls zu sein“ einen Risikofaktor für Brustkrebs darstellt. Für diese Risikogruppe von 516 exponierten Frauen ergibt sich eine Erkrankungshäufigkeit von 29/516=0,056. Im Sinne einer Kohortenstudie könnte ergänzend die BCA-Erkrankungshäufigkeit einer Kontrollgruppe von Schwestern gesunder afroamerikanischer Frauen ermittelt werden. Eine hinreichende Vergleichbarkeit von Risiko- und Kontrollgruppe erscheint bereits gegeben, wenn die Altersverteilung in beiden Gruppen ähnlich ist. Dies kann unter Umständen schon dadurch gewährleistet werden, indem zu jedem BCA-Indexfall eine gleichaltrige gesunde Frau ausgewählt wird, deren Schwestern dann als Kontrollen dienen. Soll die Kontrollgruppe größer werden als die Risikogruppe, sollten zu jedem Indexfall mehrere gleichaltrige Frauen ausgewählt werden und anschließend die Gesamtheit aller Schwestern als Kontrollgruppe dienen.

Darüber hinaus erscheint ebenfalls auffällig, dass 22 der 253 Indexfälle eine brustkrebserkrankte Mutter haben. Zur Untersuchung des Einflusses der Exposition „Mutter ist an BCA erkrankt“ auf das Brustkrebsrisiko können die Daten zu einer Fall-Kontroll-Studie ergänzt werden. Da die Fallgruppe mit den 253 Indexfällen bereits zur Verfügung steht, müssen nur noch Daten einer Kontrollgruppe beschafft werden. Auch hier scheint eine hinreichende Vergleichbarkeit von Fall- und Kontrollgruppe gegeben, wenn die Altersverteilung ähnlich ist. Zu jedem Indexfall sollte deshalb eine gleichaltrige, gesunde Frau als Kontrolle rekrutiert und nach dem Brustkrebszustand ihrer Mutter (als Exposition) befragt werden. Anschließend dient das Odds-Ratio als Maß für den Zusammenhang zwischen Exposition „erkrankte Mutter“ und der Brustkrebs-Krankheit.

Ausführlichere Informationen zur Planung und Auswertung von Studien, bei denen die familiäre Häufung einer Krankheit nachgewiesen werden soll, können Cohen et al. (1993) entnommen werden.

5.3 Auswertungsschritt 2: Nichtfamiliäre Risikofaktoren

In diesem Auswertungsschritt geht es um die Identifikation nichtfamiliärer Faktoren, die das Risiko, an Brustkrebs zu erkranken, erhöhen. Die Untersuchungen basieren auf den ausführlicheren Informationen, die von den 251 weiblichen Brustkrebsfällen in Behandlung (Indexfällen) und 947 ihrer weiblichen Familienangehörigen zur Verfügung stehen. Diese Informationen umfassen neben biologischen Angaben auch zahlreiche Angaben zu Lebensgewohnheiten und soziokulturellen Gegebenheiten, so dass zahlreiche Expositionen als potentielle Risikofaktoren für Brustkrebs untersucht werden können. Die Datenauswertung erfolgt unter Zuhilfenahme von logistischen Regressionsmodellen (vgl. 4.3), da diese es erlauben, eine Beziehung zwischen mehreren Risikofaktoren und dem Brustkrebs-Erkrankungsrisiko herzustellen.

5.3.1 Beschreibung und Aufbereitung des Datenmaterials

Das vorliegende Datenmaterial (vgl. 2.2, 5.1 und Anhang 4) ist umfangreich und lückenhaft, so dass es für eine sinnvolle Datenauswertung unumgänglich ist, sich zunächst einen Überblick über die Daten zu verschaffen. Bereits in Unterkapitel 5.1 wurde motiviert, dass eine Datenauswertung nur dann möglich ist, wenn die Daten als Ergebnis einer Fall-Kontrollstudie aufgefasst werden. Tabelle 5.3.1 zeigt, dass sich in dieser Betrachtungsweise die Fallgruppe aus 326 brustkrebserkrankten Frauen zusammensetzt, wohingegen 868 nichterkrankte Frauen die Kontrollgruppe bilden. Tabelle 5.3.2 kann entnommen werden, welchen Anteil die Indexfälle an der Fallgruppe haben. Die 253 Brustkrebsfälle in Behandlung (Indexfälle) machen einen Anteil von 77,61% der Fallgruppe aus.

Tabelle 5.3.1: Brustkrebsverteilung in den Daten

BCA-Status	Anzahl Frauen
0 – nicht BCA erkrankt	868
1 – BCA erkrankt	326

Tabelle 5.3.2: Zusammensetzung der Fallgruppe

Fallgruppe	Absolute Häufigkeit	Prozentualer Anteil
BCA-Indexfälle	253	77,61 %
Familienangehörige	73	22,39 %

Obwohl es sich um Familiendaten handelt, das heißt genetische Abhängigkeiten zwischen den Angehörigen einer Familie zu erwarten sind, besteht keine Möglichkeit diesen Aspekt bei der Auswertung zu berücksichtigen (vgl. 5.1). Entsprechend müssen die 1194 Studienteilnehmerinnen als unabhängige Individuen angesehen werden. Eine Betrachtungsweise, die zu einem

gewissen Grade dadurch gerechtfertigt werden kann, dass im ersten Auswertungsschritt (vgl. 5.2) kein Anzeichen für eine familiäre Häufung der Brustkrebskrankheit gefunden wurde.

Definition der Variablen „Lebensalter“

Da davon auszugehen ist, dass das Lebensalter das Brustkrebsrisiko entscheidend beeinflusst, soll in einem ersten Schritt untersucht werden, inwieweit die BCA-Verteilungen in unterschiedlichen Altersgruppen variieren. Im Hinblick auf eine solche Untersuchung stellt sich die Frage, in wie weit in einer generationsübergreifenden Studie ein sinnvolles Maß für das Lebensalter festgelegt werden kann. Da auch Daten von Frauen, die zum Studienbeginn bereits verstorben waren, zur Verfügung gestellt wurden, muss in diesem Zusammenhang zunächst zwischen 887 noch lebenden und 307 bereits verstorbenen Studienteilnehmerinnen unterschieden werden. Tabelle 5.3.3 zeigt, wie sich die 307 Todesfälle auf die drei Untergruppen verteilen.

Tabelle 5.3.3: Lebensstatus in den drei Untergruppen

Gruppe:	**Kontrollgruppe**	**Fallgruppe**	
Status	**Familienangehörige**	**Familienangehörige**	**Indexfälle**
lebendig	606	41	240
bereits verstorben	262	32	13
Σ	**868**	**73**	**253**

Theoretisch, das heißt bei Vernachlässigung der Fehlenden-Werte-Problematik, liegen unabhängig vom BCA-Status die folgenden Altersinformationen (in Jahren) von den Studienteilnehmerinnen vor:

Von den 307 bereits verstorbenen Frauen ist bekannt, wie alt sie geworden sind, so dass bei ihnen das Lebensalter nur als Todesalter definiert werden kann. Von den 887 noch lebenden Frauen hingegen wurde das Alter zum Zeitpunkt des Studienbeginns als Lebensalter festgehalten. Ihre Todesalter sind nicht verfügbar. Verwendet man nun als Maß für das Lebensalter in Abhängigkeit vom Lebensstatus entweder das Todesalter oder das Alter zum Zeitpunkt der Datenerhebung, ergeben sich die in Tabelle 5.3.4 dargestellten altersspezifischen BCA-Häufigkeitsverteilungen.

Den ersten drei Spalten können für jede Altersgruppe i neben der Anzahl Frauen (N_i), die absolute (H_i) und relative (h_i) Häufigkeit von BCA-Fällen entnommen werden. Darüber hinaus liefert die letzte Spalte der Tabelle für jede Altersklasse den relativen Anteile p_i der Indexfälle

an allen BCA-Fällen (H_i). Letztere Anteile sind von Bedeutung für die Einschätzung, ob sich das Lebensalter der Indexfälle systematisch von dem Lebensalter der anderen BCA-Fälle unterscheidet.

Tabelle 5.3.4: Altersspezifische BCA-Häufigkeitsverteilung

Altersklasse i	N_i	H_i	$h_i=H_i/N_i$	p_i
00 – 09 Jahre	1	0	0	---
10 – 19 Jahre	4	0	0	---
20 – 29 Jahre	81	5	0.062	1.000
30 – 39 Jahre	178	31	0.174	0.871
40 – 49 Jahre	205	50	0.244	0.880
50 – 59 Jahre	209	83	0.397	0.723
60 – 69 Jahre	184	67	0.364	0.806
70 – 79 Jahre	181	55	0.304	0.836
80 – 89 Jahre	83	27	0.325	0.630
90 – 99 Jahre	11	0	0	---
100 – 109 Jahre	3	0	0	---
110 – 119 Jahre	0	0	0	---
Keine Altersangabe	54	8	8/54	0.000

Bemerkenswert ist es, dass die Anteile p_i in den Altersklassen nur unwesentlich variieren, so dass die Indexfälle in jeder Altersklasse ungefähr denselben Anteil der BCA-Fälle ausmachen. Da ein Blick auf die Daten zeigt, dass die Lebensalter der Indexfälle nur unwesentlich von ihren Erkrankungsaltern abweichen, bestand zunächst der Verdacht, dass nur solche BCA-Fälle als Indexfälle ausgewählt wurden, die nur wenige Jahre zuvor an Brustkrebs erkrankt waren, zum Zeitpunkt des Studienbeginns also Neuerkrankungen der letzten Jahre darstellten. Dies hätte zur Konsequenz gehabt, dass das Lebensalter der Indexfälle im Gegensatz zu dem Lebensalter der anderen BCA-Fälle als Erkrankungsalter hätte interpretiert werden müssen, und damit zu erheblichen Komplikationen bei der Interpretation geführt. Die Anteile p_i zeigen jedoch, dass sich die Lebensaltersverteilungen von Indexfällen und anderen BCA-Fällen nicht systematisch unterscheiden.
Auffallend ist lediglich, dass es sich bei den BCA-Fällen der Altersklasse „20-29 Jahre“ ausschließlich aus Indexfälle handelt, wohingegen in der höchsten Altersklasse mit BCA-Fällen ($H_i>0$) „80-89 Jahre“ der geringste Anteil Indexfälle vorzufinden ist.

Davon ausgehend, dass sich die p_i nicht systematisch unterscheiden, kann anhand der relativen Anteile h_i beurteilt werden, inwieweit das Erkrankungsrisiko mit dem Lebensalter variiert. In den beiden schwach besetzten Altersklassen „<20 Jahre“ ist keine Frau an Brustkrebs

erkrankt. Anschließend steigen die relativen Häufigkeiten an, bis der Häufigkeitsgipfel von 0,397 im Intervall „50-59 Jahre“ erreicht wird. Danach nehmen die relativen Anteile im Bereich „60-89 Jahre“ wieder etwas ab, ohne dabei aber den Wert 0,3 zu unterschreiten. In den drei schwach besetzten, höchsten Altersklassen „>89 Jahre“ findet sich überhaupt kein BCA-Fall mehr.

Als Fazit dieser ersten Untersuchung lässt sich also festhalten, dass das Lebensalter einen entscheidenden Einfluss auf das Brustkrebsrisiko nimmt. Bis etwa zum 55ten Lebensjahr scheint das Erkrankungsrisiko mit dem Lebensalter anzusteigen, anschließend fällt es wieder etwas ab. Da das Lebensalter somit als bedeutsame Einflussvariable anzusehen ist, so dass die Bedeutung anderer potentieller Risikofaktoren nur bei gleichzeitiger Adjustierung hinsichtlich des Lebensalters geschätzt werden kann, wurde entschieden, die 54 Frauen ohne verfügbare Altersangabe, nicht länger zu berücksichtigen. Ebenfalls werden die 5 Frauen unter 20 Jahren von der Analyse ausgeschlossen, da aus substanzwissenschaftlicher Sicht (Grundmann 1994) für Frauen dieses Alters Brustkrebserkrankungen höchst unwahrscheinlich sind. Für die Auswertung verbleiben noch 1135 Studienteilnehmerinnen, die sich wie in Tabelle 5.3.6 angegeben auf die Fall- und Kontrollgruppe verteilen.

Tabelle 5.3.6: Brustkrebsverteilung nach Reduktion der Studienpopulation

BCA-Status	Anzahl Frauen
0 (Kontrollgruppe)	817
1 (Fallgruppe)	318

Modellierung der Erkrankungswahrscheinlichkeit in Abhängigkeit vom Lebensalter

Unter Zuhilfenahme eines logistischen Modells, welches ausschließlich Transformationen des Lebensalters als Einflussvariablen beinhaltet, kann dieser Zusammenhang quantifiziert werden. Da die altersspezifischen Häufigkeitsverteilungen zeigen, dass keine monotone Beziehung zwischen dem Lebensalter und dem Erkrankungsrisiko zu erwarten ist, wird für die Logit-Transformation der Erkrankungswahrscheinlichkeit ein quadratisches Modell im Lebensalter unterstellt.

Unter Verwendung der Bezeichnungen D für den Krankheitsstatus und X für das Lebensalter wird den Daten also ein Modell der folgenden Form angepasst:

$$Logit\{P(D=1 \mid X)\} = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X^2 .$$

Für die 1135 vorliegenden Daten (d_i,x_i) ergeben sich die in Tabelle 5.3.7 dargestellten Ergebnisse.

Die Wald-Test P-Werte beziehen sich auf die Testprobleme H_0:„$\beta_i=0$“ gegen H_1:¬H_0 (i=1,2,3). Die Überschreitungswahrscheinlichkeiten aller drei Parameter liegen unter 0,0001, so dass in diesem Modell alle drei Parameter signifikant von Null verschieden sind.

Tabelle 5.3.7: Ergebnisse der ML-Schätzung

Parameter β_i	**ML-Schätzwert $\hat{\beta}_i$**	**Wald-Test P-Wert**
β_0	-6,4570	<0,0001
$\beta_1=\beta(X)$	0,1882	<0,0001
$\beta_2=\beta(X^2)$	-0,00148	<0,0001

Die zusätzliche Aufnahme weiterer Transformationen von X zeigt, dass der quadratische Effekt X^2 ausreicht, um die nichtmonotone Beziehung zwischen Lebensalter und Erkrankungswahrscheinlichkeit zu beschreiben. Die Parameter von Effekten der Form X^3 oder $\log_e(X)$ sind nicht einmal zum Niveau 0,1 (signifikant) von Null verschieden.

Die geschätzte Regressionsgleichung für die Erkrankungswahrscheinlichkeit gegeben das Lebensalter lautet:

$$P(D=1\mid X)=\frac{1}{1+\exp\{-(-6{,}457+0{,}188\cdot X-0{,}00148\cdot X^2)\}}.$$

Da der Wert des Parameter β_0 aufgrund des Studiendesigns nicht interpretierbar (vgl. 4.3.8) ist, muss beachtet werden, dass aus der Modellgleichung ausschließlich Erkrankungschancenverhältnisse für unterschiedlich alte Personen berechnet werden können. Aussagen über konkrete Erkrankungswahrscheinlichkeiten können nicht gemacht werden. Gemäß den Ausführungen in Unterkapitel 4.3.2 ergibt sich für Personen des Alters x und y ein Erkrankungschancenverhältnis OR(x,y) von $\exp\{0{,}188\cdot(x-y)-0{,}00148\cdot(x^2-y^2)\}$ und die maximale Erkrankungschance liegt nach Modellgleichung für das Lebensalter 63,6 Jahre vor.

Tabelle 5.3.8 können die Odds-Ratio-Verhältnisse zwischen den Lebensaltern 20,30,…,80 Jahre entnommen werden. Die Odds-Ratio Werte zeigen, dass der Einfluss des Lebensalters nicht nur statistisch signifikant, sondern auch inhaltlich relevant ist.

Tabelle 5.3.8: Odds-Ratios OR(x,y) für unterschiedliche Lebensalter

x / y	20	30	40	50	60	70	80
20	**1**	3,135	7,301	12,658	16,314	15,385	11,156
30	0,319	**1**	2,331	4,032	5,208	5,000	3,571
40	0,137	0,429	**1**	1,733	2,234	2.141	1,528
50	0,079	0,248	0,577	**1**	1,289	1,236	0,882
60	0,061	0,192	0,448	0,776	**1**	0,959	0,684
70	0,065	0,200	0,467	0,809	1,043	**1**	0,713
80	0,090	0,280	0,654	1,134	1,462	1,402	**1**

Das größte Odds-Ratio ergibt sich für die Lebensalter 63.6 und 20 Jahre: OR(63.6;20) =16,63. Die Chance an Brustkrebs zu erkranken für eine Frau des Alters 63,6 Jahre ist also schätzungsweise fast 17 mal so groß wie die Chance einer 20jährigen Frau.
Bei der Interpretation dieser Odds-Ratio Werte ist allerdings zu beachten, dass die anderen Personenmerkmale im Modell nicht berücksichtigt wurden und daher genau wie möglicherweise vorliegende familiäre Abhängigkeiten als potentielle Störgrößen anzusehen sind.

Als Fazit kann festgehalten werden, dass das Lebensalter offensichtlich eine sehr große Bedeutung für das Brustkrebsrisiko hat und deshalb in allen weiteren Modellen berücksichtigt werden sollte. In den folgenden Modellen werden daher grundsätzlich das Lebensalter und das quadrierte Lebensalter als Kovariablen aufgenommen. Da die Auswertung auf den, auf 1135 Frauen reduzierten Datensatz beschränkt wurde, stehen für alle Studienteilnehmerinnen Altersangaben zur Verfügung. Damit gehört das Lebensalter zu den wenigen Variablen, deren Aufnahme in ein umfangreicheres Modell keine „Manipulation“ (im Sinne einer Aufbereitung) von fehlenden Werten bzw. die Nichtberücksichtigung von Frauen mit fehlenden Werten erforderlich macht. Dies ist von Bedeutung, weil die Miteinbeziehung von Frauen mit fehlenden Werten zu Ergebnisverzerrungen führen kann (vgl. 4.3.8), wohingegen andererseits aber auch jeder Informationsverlust vermieden werden sollte.
Aufgrund der vollständig vorhandenen Altersangaben können die Lebensalter-Kovariablen in jedes Modell - zwecks Adjustierung bzw. Störgrößenkontrolle - integriert werden, ohne dass eine der beiden Gefahren auftritt. Damit nimmt das Lebensalter eine Sonderstellung unter den Merkmalen ein. Die folgenden Untersuchungen zeigen, dass fast alle anderen Merkmale nur unvollständig erhoben wurden, so dass eine Adjustierung hinsichtlich dieser Merkmale nicht unbedingt sinnvoll ist. Dem Vorteil der Adjustierung bei simultaner Betrachtung von Merk-

malen steht bei unvollständigen Daten immer der Nachteil eines Informationsverlustes oder einer Ergebnisverzerrung durch die Miteinbeziehung (im Sinne von Manipulation) fehlender Werte gegenüber.

5.3.2 Variablen-Vorauswahl

Da das Datenmaterial insgesamt 68 Merkmale umfasst und zudem unvollständig ist, muss vor dem Einsatz multipler logistischer Regressionsmodelle eine Variablenvorauswahl getroffen werden. Erst nach Ausschluss aller mutmaßlich unwichtigen Merkmale kann entschieden werden, inwieweit - bedingt durch die Unvollständigkeit - eine simultane Betrachtung der verbliebenen Merkmale sinnvoll ist.
Eine solche Vorauswahl, bei der jedes Merkmal unabhängig von allen anderen und/oder in Verbindung mit dem Lebensalter betrachtet wird, ist notwendig, da die sofortige simultane Betrachtung vieler Einflussvariablen aufgrund der fehlenden Werte zu den bereits oben angesprochenen Problemen führt. Entweder die simultane Betrachtung kann ausschließlich auf Grundlage der Teilmenge der Studienpopulation mit vollständigen Variableneinträgen durchgeführt werden (Informationsverlust), oder es ist zuvor eine nicht fundierte Manipulation (Aufbereitung) der fehlenden Werte (vgl. 4.3.8) vorzunehmen, was zu Ergebnisverzerrungen führen könnte.
Speziell in Bezug auf einflusslose, unvollständige Variablen ist intuitiv klar, dass ihre Aufnahme nur einen Informationsverlust oder bei Miteinbeziehung der fehlenden Werte die Gefahr einer Ergebnisverzerrung mit sich bringen kann, niemals aber zu einer substanzwissenschaftlichen Modellverbesserung führt. Durch die Vorauswahl wird die Gefahr gebannt, dass solche Variablen ins Modell integriert werden.

Strategie bei der Variablen-Vorauswahl

Die Variablenvorauswahl erfolgt unabhängig für jedes Merkmal unter Zuhilfenahme von Kontingenztafeln und einfachen logistischen Regressionsmodellen.
Die Kontingenztafel vermittelt einen groben Eindruck der gemeinsamen Verteilung des Merkmals und der Zielvariablen Brustkrebszustand. (Stetige Merkmale müssen zuvor geeignet in kategoriale umgewandelt werden.) Zudem kann auf Grundlage einer solchen Tafel ein Chi-Quadrat-Test auf Unabhängigkeit durchgeführt werden. Anhand des P-Wertes dieses Tests kann beurteilt werden, inwieweit eine Abhängigkeit zwischen dem Merkmal und den Brustkrebszuständen vorliegt. Wenngleich im Rahmen solcher (einfachen) Kontingenztafelanalysen nicht einmal bezüglich des Lebensalters für eine Adjustierung gesorgt werden kann,

kommt dieser Untersuchungsform des Zusammenhangs eine besondere Bedeutung zu. Bei Anwendung logistischer Regressionsmodelle muss a priori festgelegt werden, in welcher Beziehung Merkmal und Erkrankungsrisiko stehen. Dies führt zu der ungünstigen Eigenschaft, dass der Grad der Abhängigkeit – quantifiziert durch Odds-Ratio-Schätzungen – im Gegensatz zum P-Wert eines Chi-Quadrat-Tests von der Parametrisierung des Merkmals abhängt. Liefert der Chi-Quadrat-Test einen kleinen P-Wert, kann zudem durch einen Vergleich der realisierten Kontingenztafel mit der unter Unabhängigkeit zu erwartenden Tafel entschieden werden, welche Abhängigkeitsstruktur vorliegt und eine entsprechende Parametrisierung für das logistische Modell gewählt werden. Eine ähnliche Vorgehensweise hat zum Beispiel oben dazu geführt, dass für den Logit der Erkrankungswahrscheinlichkeit kein lineares Modell im Lebensalter unterstellt wurde. Nach Kategorisierung des Lebensalters wurde unter Zuhilfenahme einer Häufigkeitstabelle erkannt, dass keine monotone Beziehung zwischen dem Lebensalter und dem Erkrankungsrisiko vorliegt und deshalb ein quadratisches Modell für den Logit gewählt.

Nach Festlegung einer geeigneten Parametrisierung kann durch den Einsatz logistischer Regressionsmodelle das interessierende Merkmal simultan mit dem Lebensalter untersucht werden, so dass im Gegensatz zur Kontingenztafelanalyse noch eine gewisse Adjustierung erreicht wird. Darüber hinaus können Odds-Ratio-Aussagen getroffen werden, so dass die Auswirkungen der unterschiedlichen Merkmalsausprägungen quantifiziert werden. Zudem können unterschiedliche Verfahren im Umgang mit den fehlenden Werte hinsichtlich ihrer Folgen im Sinne von Ergebnisverzerrungen untersucht werden. Dazu sind die Ergebnisse (Parameterschätzungen), die ausschließlich auf den vorhanden Daten basieren, mit den Ergebnissen zu vergleichen, die sich bei Miteinbeziehung von Studienteilnehmerinnen mit nicht vorhandenen Merkmalswerten ergeben.

Variablen-Vorauswahl

Da bei der Variablen-Vorauswahl sämtliche Variablen unabhängig von den Anderen nach dem, im letzten Abschnitt vorgestellten, Schema untersucht werden, wird die Vorgehensweise in diesem Abschnitt lediglich anhand der Beispielsvariablen „Fettleibigkeit“ demonstriert. Insbesondere wird untersucht, inwieweit die unterschiedlichen Verfahren der Manipulation fehlender Werte zu Ergebnisverzerrungen führen. Auf andere Merkmale wird nur eingegangen, sofern auf Grundlage von Vorauswahl-Ergebnissen entschieden wurde, aus den ursprünglichen Merkmalen neue Variablen zu generieren.

Erst im nächsten Abschnitt 5.3.3 werden die Ergebnisse der Variablen-Vorauswahl zusammenfassend dargestellt.

Untersuchung des Merkmals „Fettleibigkeit“

Als erstes wird das Merkmal „Fettleibigkeit“ untersucht, dessen Verteilung im Datensatz Tabelle 5.3.9 entnommen werden kann. Der Fettleibigkeitsstatus von 431 Frauen ist nicht verfügbar, so dass im Hinblick auf multiple Regressionsmodelle zu entscheiden ist, wie mit fehlenden Werten umgegangen werden soll.

Tabelle 5.3.9: Verteilung des Merkmals „Fettleibigkeit

Fettleibigkeits-Status	Häufigkeit (absolut)
Unbekannt	*431*
Nicht fettleibig	467
Fettleibig	237

Die Tabellen 5.3.10a und 5.3.10b zeigen Kontingenztafeln der gemeinsamen Verteilung von Fettleibigkeits- und Brustkrebsstatus. In Tabelle 5.3.10a wurden fehlende Werte als eigene Merkmalsausprägung aufgefasst, wohingegen die 431 Frauen mit fehlenden Werten in Tabelle 5.3.10b überhaupt nicht berücksichtigt wurden.

Das verwendete Codierungsschema lautet:
0 – „nicht fettleibig“, 1 – „fettleibig“ und FW – „fehlender Wert“.

Tabelle 5.3.10a: Kontingenztafel „Fettleibigkeit“ mit Kategorie „fehlender Wert“

BCA-Status	Fettleibigkeits-Status		
	FW	**0**	**1**
0	390 (310)	278 (336)	149 (171)
1	41 (121)	189 (131)	88 (66)

Tabelle 5.3.10b: Kontingenztafel „Fettleibigkeit“ (ohne fehlende Werte)

Obesity – Fettleibigkeit BCA-Status:	0	1
0	278 (283)	149 (144)
1	189 (184)	88 (93)

Ein Vergleich der beobachteten Häufigkeiten mit den unter Unabhängigkeit zu erwartenden Häufigkeiten (in Klammern) in Tabelle 5.3.10a bestätigt noch einmal die Missing-Randomly-At-Outcome-Annahme für fehlende Werte (vgl. 5.1).
In der Fallgruppe (Kontrollgruppe) liegt ein unbekannter Fettleibigkeitsstatus 41 (390) mal vor, was in Anbetracht der Tatsache, dass unter Unabhängigkeit eine Häufigkeit von 121 (310) zu erwarten gewesen wäre, deutlich zu selten (häufig) der Fall ist
Tabelle 5.3.10a zeigt zudem, dass sich für die beiden regulären Ausprägungen 0 und 1 des Merkmals bedingt durch die ungleiche Verteilung der fehlenden Werte in beiden Gruppen entsprechend verzerrte erwartete Häufigkeiten ergeben. Der Expositionsstatus 1 (bzw. 0) wurde in Kontroll- und Fallgruppe jeweils deutlich häufiger (bzw. seltener) beobachtet als unter Unabhängigkeit zu erwarten gewesen wäre. Konsequenterweise entsteht bei Miteinbeziehung der fehlenden Werte der Eindruck, dass die beiden regulären Ausprägungen das Brustkrebsrisiko erhöhen, wohingegen unbekannte Merkmalswerte das Brustkrebsrisiko senken. Diese substanzwissenschaftlich nicht interpretierbare Abhängigkeit stellt allerdings nur ein Artefakt des Studiendesigns bzw. der daraus resultierten Fehlenden-Werte-Systematik dar. Genauer ausgedrückt, kann dieser Zusammenhang den vorliegenden Daten zwar tatsächlich entnommen werden, ohne dabei aber von allgemeingültiger Natur bzw. substanzwissenschaftlichem Wert zu sein.

In Tabelle 5.3.10b werden die Frauen mit fehlenden Variableneinträgen nicht berücksichtigt, so dass die Tafel ausschließlich auf regulären Beobachtungen beruht. Eine Verzerrung, bedingt durch die ungleiche Verteilung der fehlenden Werte, ist damit ausgeschlossen. Dieser Kontingenztafel kann entnommen werden, dass sich die Fettleibigkeit in den vorliegenden Daten nicht als Risikofaktor präsentiert. Der Chi-Quadrat-Test auf Unabhängigkeit (vgl. 4.1) zwischen Brustkrebszustand und Fettleibigkeitsstatus liefert einen P-Wert in Höhe von 0,39, so dass davon auszugehen ist, dass der Fettleibigkeitsstatus keinen Einfluss auf das Brustkrebsrisiko nimmt.

Dieses Ergebnis bestätigt sich, wenn im Rahmen eines logistischen Regressionsmodells zusätzlich das Lebensalter und das quadrierte Lebensalter als Kovariablen betrachtet werden. Das Modell:

$$Logit\{P(D=1 \mid X,Y)\} = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X^2 + \beta_3 \cdot Y,$$

wobei X das Lebensalter und Y den 0/1-codierten Fettleibigkeitsstatus repräsentiert, liefert bei Vernachlässigung der Frauen mit unbekannten Fettleibigkeitszuständen für die 704 verblei-

benden Frauen die in Tabelle 5.3.11 dargestellten Parameterschätzungen und Wald-Test-P-Werte. Da die Lebensalter-Kovariablen der Adjustierung dienen, ist von den drei aufgeführten Parametern primär $\beta_3=\beta(Y)$ von Interesse.

Der dritten Spalte „Odds-Ratio“ kann entnommen werden, dass Fettleibigkeit (Y=1) das Brustkrebsrisiko schätzungsweise ver-0,9-facht. Angesichts eines P-Wertes in Höhe von 0,58 ist dieser Einfluss jedoch nicht statistisch signifikant, so dass davon auszugehen ist, dass der Fettleibigkeitsstatus auch bei Adjustierung hinsichtlich des Lebensalters aus statistischer Sicht nicht von Bedeutung für das Brustkrebsrisiko ist.

Tabelle 5.3.11: Parameterschätzung basierend auf 704 Studienteilnehmerinnen

Parameter $\beta(.)$	**ML-Schätzwert**	**Wald-Test-P-Wert**	**Odds Ratio** $\exp\{\hat{\beta}(.)\}$
$\beta(X)$: Lebensalter	0,145	<0,0001	1,167
$\beta(X^2)$: quadriertes Lebensalter	-0.001	0,0005	0,999
$\beta(Y)$: Fettleibigkeit	-0.094	0,5844	0,910

Obwohl damit im Hinblick auf die Variablen-Vorauswahl feststeht, dass der Fettleibigkeitsstatus in multiplen Modellen keine Berücksichtigung finden sollte, wird abschließend noch untersucht, zu welchen Verzerrungen unterschiedliche Verfahren im Umgang mit fehlenden Werte führen. Zunächst wird die Unzulänglichkeit eines intuitiven Verfahrens im Umgang mit fehlenden Werten demonstriert. Anschließend werden die beiden in Unterkapitel 4.3.9 vorgestellten Ad-hoc-Verfahren angewendet.

Davon ausgehend, dass die Frauen mit fehlenden Einträgen für die Variable „Fettleibigkeit“ nur fettleibig sein können oder nicht, ist zu vermuten, dass ihr Brustkrebsrisiko zwar in Abhängigkeit von der genauen Verteilung der unbekannten Zustände variiert, definitiv aber zwischen dem der Fettleibigen und Nichtfettleibigen liegt. Entsprechend erscheint es naheliegend, die Variable Y „Fettleibigkeit“ wie in Tabelle 5.3.12 dargestellt zu codieren.

Tabelle 5.3.12: Intuitive Manipulation der fehlenden Werte für „Fettleibigkeit“

Merkmalsausprägung	**Y-Codierung**
Nicht fettleibig	-1
unbekannt	0
Fettleibig	1

Diese Kodierung erweckt den Eindruck, dass der zugehörige Parameter $\beta(Y)$ ein Odds-Ratio-Vergleich zwischen den fettleibigen und nichtfettleibigen Frauen ermöglicht, ohne dass dieses Odds-Ratio dabei von Frauen mit unbekanntem Fettleibigkeitszustand beeinflusst wird.

Obwohl in der Tat die Beziehung: $OR(Y=1,Y=-1)= \exp\{2\cdot\beta(Y)\}$ gilt, muss allerdings berücksichtigt werden, dass für $\beta(Y)$ ebenfalls gilt: $OR(Y=1,Y=0) = OR(Y=0,Y=-1) = \exp\{\beta(Y)\}$. Der Parameter charakterisiert also zusätzlich auch die Chancenverhältnisse zwischen den beiden Regulären und der nicht-beobachteten Merkmalsausprägung.
Für die 1135 Frauen führt die Verwendung dieser (-1/0/1)- Codierung bei gleichzeitiger Berücksichtigung der beiden Lebensalter-Kovariablen zu einem ML-Schätzwert von -0,30 für $\beta(Y)$. Der dazugehörige Wald-Test-P-Wert beträgt 0,0013, so dass dieser Parameter sogar statistisch signifikant von Null verschieden ist. Es resultieren die folgenden geschätzten Odds-Ratio-Beziehungen:

$$OR(0,-1) = OR(+1,0) = 0{,}744 \text{ und } OR(+1,-1) = 0{,}554.$$

Naiv interpretiert besagt dieses Ergebnis u.a., dass die Erkrankungschance fettleibiger Frauen nur etwa halb so groß ist wie die Chance nichtfettleibiger Frauen. Ein Ergebnis, welches in Anbetracht der Ergebnisse, die sich bei Vernachlässigung der Frauen mit fehlenden Werten ergaben (siehe oben), nur ein Resultat der „Manipulation“ fehlender Werte darstellen kann.
Dieses Ergebnis zeigt, dass eine ungeeignete Miteinbeziehung fehlender Werte zu erheblichen Verzerrungen und damit verbundenen Fehlinterpretationen führen kann. In Anhang 1 dieser Arbeit wird etwas ausführlicher motiviert, warum die (-1/0/1)-Kodierung ungeeignet ist.

Im Folgenden werden die beiden in Unterkapitel 4.3.8 vorgestellten Ad-hoc-Verfahren im Umgang mit fehlenden Werten angewendet.

<u>Verfahren 1</u>: Indikatorvariablen-Verfahren: „Eigene Kategorie für fehlende Werte“:
Beim ersten Verfahren werden fehlende Variablenwerte als eigene Ausprägung der dazugehörigen Variablen aufgefasst. In dieser Betrachtungsweise handelt es sich bei der Variablen „Fettleibigkeit“ Y um eine kategoriale Variable mit den drei Ausprägungen „fettleibig“, „nichtfettleibig“ und „fehlender Wert“, die im Regressionsmodell durch zwei Dummy-Variablen $I_1(Y)$ und $I_2(Y)$ parametrisiert werden kann (Referenzgruppencodierung).
Tabelle 5.3.13 zeigt das dazugehörige Codierungsschema. Ohne Beschränkung der Allgemeinheit wurde der Fettleibigkeitsstatus „nicht fettleibig“ als Referenzkategorie ausgewählt.

Tabelle 5.3.13: Referenzgruppencodierung für Fettleibigkeit bei Verfahren 1

Y: Fettleibigkeitsstatus	**$I_1=I_1(Y)$**	**$I_2=I_2(Y)$**
Nicht fettleibig	0	0
Fettleibig	1	0
unbekannt	0	1

Bei der Parameterschätzung des Modells:

$$Logit\{P(D=1 \mid X, I_1(Y), I_2(Y))\} = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X^2 + \beta_3 \cdot I_1(Y) + \beta_4 \cdot I_2(Y)$$

können alle 1135 Studienteilnehmerinnen berücksichtigt werden. Es resultieren die in Tabelle 5.3.14 dargestellten Ergebnisse.

Tabelle 5.3.14: Ergebnisse der Parameterschätzung – Indikatorvariablen-Verfahren

Modellparameter:	**P-Wert**	**OR =exp{β(.)}**
β(X): Lebensalter	<0,0001	1,210
$\beta(X^2)$: quadriertes Lebensalter	<0,0001	0,998
$\beta(I_1)$: Indikator: "fettleibig"	0,511	0,893
$\beta(I_2)$: Indikator: "unbekannt"	<0.0001	0,134

Die letzte Spalte der Tabelle zeigt, dass das geschätzte Odds-Ratio zwischen den fettleibigen und nichtfettleibigen Frauen 0,893 beträgt. Dieser Schätzwert weicht damit nur geringfügig von der Odds-Ratio-Schätzung 0,910 im Modell, welches nur die 704 Frauen mit bekannter Merkmalsausprägung berücksichtigt, ab. Entsprechend führt das Indikatorvariablen-Verfahren im vorliegenden Fall zu keiner wesentlichen Verzerrung des relevanten Parameters. Die Bedeutung des Merkmals „Fettleibigkeitsstatus" wird fast unverzerrt wiedergegeben. Problematisch ist allerdings das Vorhandensein eines hochsignifikanten Parameters $\beta(I_2)$, welcher das Erkrankungschancenverhältnis zwischen den Studienteilnehmerinnen mit unbekanntem Fettleibigkeitsstatus und den nichtfettleibigen Frauen charakterisiert. Wenngleich im Datensatz bedingt durch das ungünstige Studiendesign tatsächlich gilt, dass die Erkrankungschance einer Frau mit unbekanntem Fettleibigkeitsstatus sehr gering ist, besitzt dieser Sachverhalt keinen allgemeingültigen bzw. substanzwissenschaftlichen Wert. Gewissermaßen also basiert das geschätzte Modell in seiner Gesamtheit sowohl auf biologischen Zusammenhängen als auch auf Zusammenhängen, die lediglich als Artefakt der Datenerhebung anzusehen sind. Folglich ergibt sich auch der Erklärungswert des Modells durch eine Überlagerung von tat-

sächlich vorhanden biologischen Mechanismen und inhaltlich nutzlosen Aspekten. Der Anteil des Erklärungswertes, der auf biologischen Aspekten beruht und damit inhaltlich relevant ist, ist hierbei unbekannt. Die in Unterkapitel 4.3.6 vorgestellten Kriterien, zur Beurteilung der Anpassungsgüte eines logistischen Modells, vermitteln nur einen Eindruck von der Gesamterklärungsgüte, so dass der Anteil, den die biologischen Mechanismen an dieser Erklärungsgüte haben, nicht quantifizierbar ist. Da aus medizinischer Sicht jedoch ausschließlich dieser Anteil von Relevanz ist, hat dies zur Konsequenz, dass die substanzwissenschaftliche Bedeutung des Modells nicht beurteilt werden kann.

Speziell bei logistischen Modellen mit mehreren Einflussvariablen besteht darüber hinaus die Gefahr, dass die substanzwissenschaftlich relevante Information, die in den Daten steckt, teilweise oder ganz durch die inhaltlich nutzlose Information, die auf der Datenstruktur gründet, verdeckt wird (vgl. Anhang 5).

Verfahren 2: „Unbedingte Probability-Imputation“

Bei dem Verfahren der unbedingten Probability-Imputation werden die fehlenden Werte einer Variablen Y durch den Mittelwert der vorhandenen Y-Einträge ersetzt. Die neue Variable Y_2 wird unabhängig davon, ob es sich bei der ursprünglichen Variablen Y um eine stetige, dichotome kategoriale oder Dummy-Variable gehandelt hat, als stetige Variable ins Modell aufgenommen. Ausführlichere Informationen zu diesem Verfahren können Unterkapitel 4.3.9 entnommen werden.

Zunächst wird das dichotome Merkmal „Fettleibigkeitsstatus“ Y in eine Indikatorvariable I(Y) transformiert. Die fehlenden Variablenwerte I(Y), die aus unbeobachteten Merkmalsausprägungen resultieren, werden in einem zweiten Schritt durch den Mittelwert der vorhandenen I(Y)-Werte aufgefüllt. Tabelle 5.3.15 veranschaulicht die Vorgehensweise. Der Mittelwert von 0,336 entspricht der relativen Häufigkeit bzw. empirischen Wahrscheinlichkeit für den Fettleibigkeitsstatus „fettleibig“ im vorliegenden Datensatz.

Tabelle 5.3.15: Codierung für Fettleibigkeit bei Probability-Imputation

Y: Fettleibigkeitsstatus	Absolute Häufigkeit	I(Y)	$I(Y)_2$
Nicht fettleibig	237	0	0
Fettleibig	467	1	1
unbekannt	431	Nicht definiert	237/(237+467) =0,336

Bedingt durch diese Mittelwert-Auffüllung kann bei der Modellschätzung die gesamte Studienpopulation berücksichtigt werden. Für das Modell:

$$Logit\{P(D=1 \mid X, Y_2)\} = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X^2 + \beta_3 \cdot I(Y)_2$$

resultieren die folgenden Ergebnisse (Tabelle 5.3.16).

Tabelle 5.3.16: Ergebnisse der Parameterschätzung - Probability-Imputation-Verfahren

Modellparameter:	P-Wert	OR =exp{β(.)}
β(X): Lebensalter	<0,0001	1,208
β(X^2): quadriertes Lebensalter	<0,0001	0,999
β_3=β($I(Y)_2$): Fettleibigkeitsstatus	0,3723	0,848

Obwohl dem Probability-Imputation-Verfahren eine ähnliche Idee zugrunde liegt wie dem intuitiven (–1/0/1)-Codierungsansatz (siehe oben), kann Tabelle 5.3.16 entnommen werden, dass die Verzerrung deutlich geringer ausfällt.

Der Schätzwert 0,848 für β_3 weicht nur geringfügig vom ursprünglichen Schätzwert 0,910 (bei Vernachlässigung der 431 Frauen mit unbekanntem Merkmalswert) ab.

Im Gegensatz zum (–1/0/1)-Codierungsansatz (siehe oben) werden fehlende Werte nämlich nicht durch die „Mitte“ der beiden regulären Ausprägungen, sondern durch eine Punktschätzung für ihren (unbedingten) Erwartungswert ersetzt. Diese Punktschätzung entspricht der relativen Häufigkeit p der Exposition und reflektiert somit die (empirische) Verteilung des Merkmals im Datensatz.

In Anhang 2 dieser Arbeit werden theoretische Überlegungen dazu angestellt, warum das Verfahren der Probability-Imputation im vorliegenden Fall zu besseren Schätzwerten führt als das naive (-1/0/1)-Kodierungsverfahren.

Aus den ML-Schätzwerten resultieren die folgenden Odds-Ratio-Beziehungen:

OR(“fettleibig”, ”nicht fettleibig”) =OR(1) =exp(β_3) =0,848,

OR(“fettleibig”, ”fehlender Wert“) =OR(0,664) =exp(0,664·β_3) =0,896 und

OR(“fehlender Wert”, “nicht fettleibig”) =OR(0,336) =exp(0,336·β_3) =0,946.

Neben dem Gesichtspunkt, dass das substanzwissenschaftlich relevante Odds-Ratio OR(1) nur unwesentlich verzerrt ist, ist von Bedeutung, dass die anderen beiden Odds-Ratios nahe bei Eins liegen. Für die Frauen mit fehlenden Merkmalsausprägungen ergeben sich also nur geringfügig andere Erkrankungswahrscheinlichkeiten als für die exponierten bzw. nichtexpo-

nierten Frauen. Das bedeutet, dass die Anwendung der Probability-Imputation-Methode bei Vorliegen des Fehlende-Werte-Mechanismus „Missing-Randomly-At-Outcome“ nicht so sehr zu einer durchs Studiendesign bedingten Vergrößerung des Modellerklärungswertes führt wie das Indikatorvariablen-Verfahren. Trotzdem kann bei Überprüfung der Modellanpassung nicht unberücksichtigt bleiben, dass sich die Erklärungsgüte auch hier zumindest zum Teil aus datenspezifischen Aspekten ergibt, und deshalb nicht ausschließlich aus substanzwissenschaftlichen Aspekten resultiert.

Weitere Resultate der Variablen-Vorauswahl:

Im letzten Abschnitt wurde die Vorgehensweise bei der Variablen-Vorauswahl anhand der Beispielsvariablen „Fettleibigkeitsstatus“ demonstriert. Für die anderen potentiellen Einflussvariablen werden jeweils analoge Untersuchungen durchgeführt, auf die im Einzelnen allerdings nicht mehr eingegangen wird. Für eine simultane Untersuchung im Rahmen logistischer Regressionsmodelle werden genau die Variablen ausgewählt, bei denen nicht zu viele fehlende Variableneinträge (>50%) vorliegen und, deren Untersuchungsergebnisse auf eine gewisse Bedeutung für das Brustkrebsrisiko schließen lassen (siehe unten). Mit dem Ziel, einen unnötigen Informationsverlust bei simultaner Betrachtung unvollständiger Variablen oder eine Ergebnisverzerrung durch die Miteinbeziehung fehlender Werte zu vermeiden, werden alle anderen Variablen von den weiterführenden Untersuchungen ausgeschlossen.

Noch einmal erinnert werden soll daran, dass in Anbetracht der Unvollständigkeit des Datenmaterials bei der Vorauswahl jedes Merkmal unabhängig für sich betrachtet werden muss, so dass nicht ausgeschlossen werden kann, dass die anderen Variablen als nicht kontrollierte Störgrößen zu Ergebnisverzerrungen führen. Ausschließlich das Lebensalter kann im Rahmen der einfachen logistischen Modelle stets simultan als Kovariable berücksichtigt werden und wird daher als Störgröße kontrolliert. Angesichts dieser unzureichenden Störgrößenkontrolle ist zu beachten, dass keine Gewährleistung dafür gegeben werden kann, dass im Rahmen solcher Vorauswahlverfahren mit Sicherheit die bedeutsamen von den unbedeutsamen Merkmalen getrennt werden. Grundsätzlich steht jedoch fest, dass mit dem endgültigen Ausschluss einer bedeutsamen Variablen unwiderruflich Information verloren geht, wohingegen einflusslose Variablen auch im Rahmen komplexerer Modelle noch als solche erkannt werden können. Deshalb ist es i.a. sinnvoll, die Ausschlussbedingungen für Variablen so festzulegen, dass möglichst wenige bedeutsame Variablen ausgeschlossen werden. Im Hinblick auf die

durchzuführenden Untersuchungen, wird im vorliegenden Fall die folgende Ausschlussbedingung definiert:

Eine Variable wird genau dann ausgeschlossen, wenn weder eine „Auffälligkeit“ in der Kontingenztafel beobachtet werden kann, noch im Regressionsmodell mit dem Lebensalter ein Wald-Test-P-Wert kleiner 0,2 resultiert.

Der Begriff „Auffälligkeit“ kann in diesem Zusammenhang nicht genauer präzisiert werden. Gemeint ist, dass wann immer auf Grundlage der realisierten Kontingenztafel eine gewisse Abhängigkeit zwischen dem Merkmal und der Zielvariablen zu vermuten ist, untersucht wird, ob sich aus dem Merkmal ein Neues generieren lässt, welches von Bedeutung für das Brustkrebsrisiko ist. Denkbar wären zum Beispiel Umwandlungen von stetigen Merkmalen in Kategoriale oder Zusammenfassungen von Kategorien kategorialer Merkmale, sofern sie substanzwissenschaftlich gerechtfertigt werden können. Anschließend kann untersucht werden, ob die neu generierte bzw. modifizierte Variable bei simultaner Betrachtung mit den Lebensalter-Kovariablen zum Niveau 0,2 signifikant ist.

Die Verwendung eines Signifikanzniveaus von 0,2 anstelle des üblichen Niveaus 0,05 wird von Hosmer & Lemeshow (1989) für Vorauswahlen empfohlen, bei denen gewährleistet werden soll, dass keine bedeutsamen Variablen ausgeschlossen werden. Durch die Heraufsetzung des Niveaus soll verhindert werden, dass bereits geringfügige Ergebnisverzerrungen - bedingt durch den Einfluss der nichtberücksichtigten Variablen – zu einem Ausschluss bedeutsamer Variablen führen.

Im Folgenden wird auf 4 Merkmale eingegangen, bei denen den dazugehörigen Kontingenztafeln entnommen werden konnte, dass lediglich spezielle Ausprägungen das Brustkrebsrisiko erhöhen. Aus jedem dieser 4 Merkmale kann auf Grundlage der Kontingenztafel durch die Zusammenlegung von Kategorien oder eine geeignete Gruppierung eine neue Variable generiert werden, die inhaltlich sinnvoll ist und einen deutlich geringeren Wald-Test-P-Wert aufweist als das ursprüngliche Merkmal. Abschließend wird noch auf das Merkmal „Menopause-Status“ eingegangen, da es im weiteren Verlauf dieser Arbeit genutzt wird, um die Studienpopulation in zwei Subpopulationen zu unterteilen.

1. Familienstand

Für die kategoriale Variable „Familienstand“ ergibt sich bei Vernachlässigung der 119 fehlenden Werte die in Tabelle 5.3.17 dargestellte Kontingenztafel. Der dazugehörige Chi-

Quadrat-Test weist einen P-Wert in Höhe von 0,0001 auf, so dass zunächst der Verdacht besteht, dass der Familienstand Einfluss auf das Brustkrebsrisiko nimmt. Der Kontingenztafel kann genauer entnommen werden, dass von den geschiedenen und verwitweten Frauen deutlich mehr an Brustkrebs erkrankt sind, als unter Unabhängigkeit zu erwarten ist.

Tabelle 5.3.17 : Kontingenztafel Familienstand – BCA-Status

BCA-Status	**Familienstand**			
	1 - ledig	**2- verheiratet**	**3 - geschieden**	**4 – verwitwet**
0 – kein BCA	135 (120)	395 (373)	76 (103)	103 (119)
1 - BCA	37 (52)	140 (162)	71 (44)	59 (49)
$\sum$	**172**	**535**	**146**	**162**

Da allerdings davon auszugehen ist, dass geschiedene und verwitwete Frauen im Durchschnitt älter sind als ledige oder verheiratete Frauen kommt dem Lebensalter eine besondere Bedeutung als Confounding-Variable zu. Es liegt die Vermutung nahe, dass in den 4 Familienstandsgruppen unterschiedliche Lebensaltersverteilungen vorliegen und somit das nichtkontrollierte Lebensalter Mitverursacher der unterschiedliche BCA-Häufigkeit ist. Tatsächlich gilt, dass bei Verwendung der Referenzgruppencodierung für den Familienstand sich lediglich bei Vernachlässigung des Lebensalters signifikante Unterschiede zwischen den 4 Gruppen beobachten lassen. Bei simultaner Betrachtung von Familienstand und Lebensalter im Rahmen eines logistischen Modells, in welchem fehlende Werte als eigene Kategorie aufgefasst werden, ergibt sich das extremste Odds-Ratio zwischen den Gruppen „geschieden“ und „ledig“. Dieses beträgt 1,64 und der dazugehörige P-Wert ist 0,078, so dass davon auszugehen ist, dass der Familienstand zumindest einen gewissen Einfluss auf das Brustkrebsrisiko nimmt. Der nächstkleinste P-Wert von 0,158 ergibt sich für das Odds-Ratios zwischen verheirateten und ledigen Frauen. Dieses beträgt 0,720. Die beiden Familienstände „geschieden“ und „ledig“ weisen bei Adjustierung hinsichtlich des Lebensalters in etwa dasselbe Brustkrebsrisiko auf. Der P-Wert des dazugehörigen Odds-Ratios von 1,204 beträgt 0,483.

Wenngleich das Merkmal „Familienstand“ damit bereits für die Betrachtung in komplexeren Modellen in Frage kommt, erscheint es aus inhaltlicher Sicht vertretbar, jeweils zwei Kategorien zusammenzufassen. Durch die Zusammenlegung der Ausprägungen „verwitwet“ und „geschieden“ einerseits und „ledig“ und „verheiratet“ andererseits kann eine neue Variable definiert werden, welche nur noch zwischen ehemals verheirateten Frauen und noch gar nicht oder noch immer verheirateten Frauen unterscheidet. Inhaltliche Grundidee für diese Neude-

finition ist, dass der Verlust des Ehemanns unabhängig davon, ob er durch Scheidung oder Versterben herbeigeführt wurde, seelische Probleme und/oder hormonelle Umstellungen bedingen kann, die den eigentlichen Risikofaktor für Brustkrebs darstellen.
Für das Regressionsmodell:

$$Logit\{P(D=1 \mid X, I)\} = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X^2 + \beta_3 \cdot I(Y) + \beta_4 \cdot I_2(Y),$$

wobei X das Lebensalter und Y den {1,2,3,4} codierten Familienstand repräsentiert und die Indikatorvariablen I(Y) bzw. I_2(Y) die Familienstände 3 und 4 bzw. fehlende Werte kennzeichnen, ergeben sich die in Tabelle 5.3.18 dargestellten Ergebnisse.

Tabelle 5.3.18: Ergebnisse der ML-Schätzung

Modellparameter	ML-Schätzwert	Odds-Ratio	P-Wert
β(Lebensalter)	0,1811	1,198	<0,0001
β(Lebensalter2)	-0,00142	0,999	<0,0001
β(Verlust des Ehemanns)	0,6061	1,833	<0,0001
β(fehlender Wert)	-1,3200	0,267	<0,0001

Der Tabelle ist zu entnehmen, dass bei Adjustierung hinsichtlich des Lebensalters der Verlust des Ehemanns die Brustkrebserkrankungschance bei Frauen ver-1,833-facht und damit erheblich beeinflusst. Aufgrund des äußerst geringen P-Werts (<0,0001) wurde im Hinblick auf komplexere Modelle entschieden, die ursprüngliche Variable „Familienstand“ durch die binäre Variable „Verlust des Ehemanns“ zu ersetzen. Ebenfalls zeigt Tabelle 5.3.18, dass fehlende Familienstandsangaben – bedingt durch das Studiendesign – wieder mit einem geringen Brustkrebsrisiko assoziiert sind.

2. Schuljahre

Die ordinale Variable „Schuljahre“ besitzt 17 unterschiedliche Ausprägungen, deren Bedeutung im Folgenden erläutert wird. Die Ausprägungen 1 bis 12 entsprechen den Schuljahren an einer Primary- bzw. Secondary-School. Wurde (nach 12 Schuljahren) noch eine Highschool bzw. Universität besucht, werden anschließend noch die bis zu 4 dort absolvierten Jahre hinzuaddiert. Die Ausprägung 17 repräsentiert noch weiter führende akademischen Ausbildungen („postgraduate education“) .
Aufgrund der hohen Anzahl unterschiedlicher Ausprägungen und des ordinalen Messniveaus dieses Merkmal wird die Variable „Schuljahre“ in einem ersten Schritt als Kontinuierliche in ein Regressionsmodell integriert. Die 461 fehlenden Merkmalsausprägungen werden dazu

durch eine Indikatorvariable I(Y) gekennzeichnet und im Ausgangsmerkmal Y auf den Wert 0 gesetzt. Anschließend kann das Merkmal „Schuljahre“ - parametrisiert durch die Variablen Y und I(Y) – simultan mit den beiden Lebensaltervariablen X und X^2 betrachtet werden. Für den Parameter der Variablen Y resultiert ein Schätzwert von 0,0206 und der dazugehörige P-Wert beträgt 0,4638. Entsprechend ist nicht davon auszugehen, dass das Brustkrebsrisiko in monotoner Weise mit den absolvierten Schuljahren variiert. Die anschließende Betrachtung einer Kontingenztafel der gemeinsamen Verteilung der Variablen „Schuljahre“ und „BCA-Status“ (Tabelle 5.3.20) zeigt eine seltsame Form der Abhängigkeit, die sich auch in einem Chi-Quadrat-Test P-Wert von 0,0128 äußert.

Die letzten Spalte „Grafik“ der Tabelle dient als einfaches Hilfsmittel, um einen Überblick zu geben, ob die Differenzen zwischen den beobachteten und zu erwartenden Zellhäufigkeiten einer gewissen Systematik folgen. Jede Differenz (bzgl. der zweiten Spalte) wird in Abhängigkeit von ihrem Vorzeichen durch entsprechend viele „*“ oder „+“ Symbole repräsentiert. Darüber hinaus werden zwecks Erhöhung der Übersichtlichkeit die Symbole in Abhängigkeit vom Vorzeichen der Differenz links- oder rechtsbündig in die Zelle gelegt.

Eine Systematik in Bezug auf diese Differenzen ist nicht zu erkennen.

Tabelle 5.3.20: Kontingenztafel: Schuljahre – BCA-Status

(ohne 468 fehlende Werte)

Schuljahre	0 – kein BCA	1 - BCA	Grafik
1	0 (1)	1 (0)	+
2	1 (1)	1 (1)	
3	2 (1)	0 (1)	*
4	0 (1)	1 (0)	+
5	3 (4)	4 (3)	+
6	5 (6)	5 (4)	+
7	5 (7)	7 (5)	++
8	10 (12)	10 (8)	++
9	8 (11)	10 (7)	+++
10	**15 (22)**	**22 (15)**	+++++++
11	27 (25)	16 (18)	**
12	**148 (130)**	**74 (92)**	**********
13	21 (23)	19 (17)	++
14	37 (34)	21 (24)	+++
15	11 (10)	6 (7)	*
16	**33 (45)**	**44 (32)**	++++++++
17	**65 (59)**	**35 (41)**	******

Am auffälligsten ist, dass bei der Ausprägung „12“, das heißt bei Beendigung der Schullaufbahn mit einem Secondary-School-Abschluss, die zu erwartende Zellhäufigkeit die Beobachtete um 18 übersteigt, was darauf schließen lässt, dass sie mit einem geringen Brustkrebsrisiko assoziiert ist. Darüber hinaus bestehen lediglich noch bei den Ausprägungen „10“, „16“ und „17“ nennenswerte Diskrepanzen (>5) zwischen den beobachteten und zu erwartenden Zellhäufigkeiten. Eine inhaltlich interpretierbare Systematik ist nicht zu erkennen.
Eine willkürliche Zusammenlegung von Kategorien führt zu neuen Variablen, die inhaltlich nicht interpretierbar sind. Inhaltlich sinnvoll erscheint es, aus dem Merkmal „Schuljahre“ eine neue Indikatorvariable I(Y) zu generieren, die lediglich angibt, ob eine weiterführende Schule (Highschool oder Universität) besucht wurde (Y>12) oder nicht (Y≤12).
Bei simultaner Betrachtung des Lebensalters und der Einführung einer zusätzlichen Kategorie für fehlende Werte ergibt sich als Ergebnis der ML-Schätzung, dass der Besuch einer weiterführenden Schule I(Y)=1 das Erkrankungsrisiko ver-1,36-facht. Der dazugehörige P-Wert beträgt 0,072, so dass tatsächlich davon ausgegangen werden kann, dass der Besuch einer weiterführenden Schule einen Risikofaktor für Brustkrebs darstellt. Im Hinblick auf komplexere Modelle wird die ursprüngliche Variable „Schuljahre“ durch die binäre Variable „Weiterführende Schule“ ersetzt.

3. Alter zum Zeitpunkt der 1ten Schwangerschaft „Alter 1te SS“

Die Variable „Alter 1te SS“ (Y) gibt an, in welchem Lebensalter die jeweilige Frau zum ersten mal eine vollständige Schwangerschaft (Lebend- oder Totgeburt) hinter sich gebracht hat. Fehlgeburten und Schwangerschaftsabbrüche werden in diesem Zusammenhang nicht als (vollständige) Schwangerschaften betrachtet.
Eine sinnvolle Verwendung dieser Variablen setzt deshalb voraus, dass neben einer Indikatorvariablen für fehlende Werte $I_1(Y)$ auch noch eine zweite Indikatorvariable $I_2(Y)$ definiert wird, die Frauen kennzeichnet, die zum Zeitpunkt des Studienbeginns noch nie (vollständige) schwanger gewesen waren. Bei Verwendung dieser Kodierung (vgl. Tabelle 5.3.21) ergeben sich im Modell mit den beiden Lebensalter-Variablen die in Tabelle 5.3.22 dargestellten Ergebnisse der ML-Schätzung.

Tabelle 5.3.21: Kodierung der Variablen „Alter 1te SS“

Y „Alter 1te SS“	Y_2	$I_1(Y_2)$	$I_2(Y_2)$
Unbekannt	0	1	0
Noch keine SS	0	0	1
Altersangabe(1te SS)	Alter 1te SS	0	0

Tabelle 5.3.22: Ergebnisse der ML-Schätzung

Modellparameter	ML-Schätzwert	P-Wert (Wald-Test)	Odds Ratio
β(Lebensalter)	0,1942	<0,0001	1,214
$\beta(\text{Lebensalter}^2)$	-0,00152	<0,0001	0,998
$\beta(Y_2)$	0,0368	0,0071	1,039
$\beta(I_{\text{fehlender Wert}})$	-0,2678	0,5462	0,765
$\beta(I_{\text{noch keine SS}})$	0,9596	0,0078	2,611

Den Ergebnissen der ML-Schätzung kann entnommen werden, dass das Brutkrebsrisiko signifikant mit dem Alter der ersten Schwangerschaft steigt. Beginnend bei der geringsten Y_2-Ausprägung (12) steigt das Risiko mit jedem Lebensjahr um den Faktor 1,039 an. Der dazugehörige Wald-Test-P-Wert beträgt 0,0071.

4. Regelmäßigkeit der Periode

Das kategoriale Merkmal „Regelmäßigkeit der Periode“ Y gibt an, inwieweit die Perioden der jeweiligen Studienteilnehmerin in regelmäßigen Intervallen stattgefunden haben. Tabelle 5.3.23 zeigt die Kontingenztafel, die sich bei Vernachlässigung der 482 fehlenden Werte für dieses Merkmal ergeben. Der dazugehörige Chi-Quadrat-P-Wert beträgt 0,0157.

Tabelle 5.3.23: Kontingenztafel Regelmäßigkeit der Periode – BCA-Status

Regelmäßigkeit der Periode	Brustkrebs-Status	
	0 – kein BCA	**1 - BCA**
1 – immer regelmäßig	133 (118)	69 (84)
2 – fast immer regelmäßig	183 (199)	158 (142)
3 – häufiger unregelmäßig	45 (47)	36 (34)
3 – stets unregelmäßig	21 (17)	8 (12)

Ein Vergleich der unter Unabhängigkeit erwartenden (in Klammern) mit den beobachteten Zellhäufigkeiten zeigt, dass die bedeutsamsten Unterschiede zwischen den ersten beiden Kategorien „immer regelmäßig“ und „fast immer regelmäßig“ lokalisiert sind. Die Abweichungen in den letzten beiden schwächer besetzten Kategorien sind weniger deutlich.
Entsprechend ist nicht anzunehmen, dass das Brustkrebsrisiko monoton mit Abnahme der Regelmäßigkeit der Periode variiert. Am geeignetsten erscheint es aus dem Merkmal Y eine neue binäre Variable Y_2 zu generieren, die lediglich noch angibt, ob die Periode immer regelmäßig stattgefunden hat (Y=1) oder nicht (Y=2,3,4). Bei Anwendung des Indikatorvariablen-Verfahrens und Verwendung der Referenzgruppencodierung für die neue Variable Y_2 ergibt sich im logistischen Regressionsmodell mit den beiden Lebensalter-Variablen, dass für Frauen mit einer unregelmäßigen Periode (Y=2,3,4) ein 1,526-mal so hohes Erkrankungsrisiko wie für Frauen mit einer „immer regelmäßigen“ Periode (Y=1). Der Wald-Test-P-Wert des dazugehörigen Modellparameters beträgt 0,0108.

***5. Menopause-Status**:*

Das Merkmal „Menopause-Status“ gibt an, ob die Menopause der betreffenden Frau noch nicht, gerade oder bereits eingetreten ist, und unterscheidet zudem zwischen regulär, das heißt durch biologische Abläufe, eingetreten und operativ herbeigeführten Menopausen. Der Kontingenztafel (Tabelle 5.3.24) kann entnommen werden, dass offensichtlich ein bedeutsamer Zusammenhang (Chi-Quadrat-Test-P-Wert: 0,0001) zwischen dem Menopause- und dem Brustkrebs-Status besteht. Für 145 Frauen liegen keine Informationen über dieses Merkmal vor.

Tabelle 5.3.24: Kontingenztafel: Menopause-Status gegen BCA-Status

Menopause-Status	**BCA-Status**	
Menopause…	**0 – kein BCA**	**1 – BCA erkrankt**
1 – …noch nicht erreicht	262 (222)	61 (101)
2 – …stellt sich gerade ein	32 (36)	21 (17)
3 – …regulär erreicht	249 (245)	107 (111)
4 – …operativ herbeigeführt	138 (177)	120 (81)

Die Kontingenztafel zeigt, dass in der ersten Kategorie die erwartete Häufigkeit von Brustkrebs deutlich unterschritten und in der vierten Kategorie deutlich überschritten wird. Aller-

dings liegt die Vermutung nahe, dass das Lebensalter als Confounding-Variable wirkt. Da die Menopause der Frau aus biologischen Gründen i.a. im Alter zwischen 45 und 50 Jahren eintritt (vgl. Pschyrembel 1998), ist davon auszugehen, dass sich die Lebensalterverteilung und damit auch das Durchschnittsalter in allen 4 Menopause-Status-Kategorien deutlich unterscheidet. Tabelle 5.3.25 zeigt die Lebensalter-Mittelwerte für alle 4 Kategorien und bestätigt obige Vermutung.

Tabelle 5.3.25: Durchschnittsalter nach Menopause-Status

Menopause-Status Menopause…	**Anzahl Frauen**	**Durchschnittsalter (in Jahren)**
1 – …noch nicht erreicht	323	36,0
2 – …stellt sich gerade ein	53	55,1
3 – …regulär erreicht	356	67,9
4 – …operativ herbeigeführt	258	59,0

Bei Verwendung der Referenzgruppencodierung, wobei fehlende Werte wieder als eigene Kategorie aufgefasst werden, ergeben sich im logistischen Regressionsmodell mit den beiden Lebensalter-Variablen tatsächlich weniger deutliche Odds-Ratio-Beziehungen. Diese und die dazugehörigen P-Werte (in Klammern angegeben) können Tabelle 5.3.26 entnommen werden.

Tabelle 5.3.26:Odds-Ratio-Beziehungen zwischen den Menopause-Status-Ausprägungen

OR{x,y} (P-Wert)	**x=1**	**x=2**	**x=3**	**x=4**
y=1	**1,000 (-----)**	1,023 (0,95)	0,601 (0,07)	1,364 (0,23)
y=2	0,978 (0,95)	**1,000 (-----)**	0,588 (0,09)	1,333 (0,36)
y=3	1,664 (0,07)	1,702 (0,09)	**1,000 (------)**	2,269 (0,00)
y=4	0,733 (0,23)	0,750 (0,36)	0,441 (0,00)	**1,000 (-----)**

Bei Adjustierung hinsichtlich des Lebensalters ergibt sich das extremste Odds-Ratio also zwischen der dritten und vierten Kategorie. Die Erkrankungschance für Frauen, deren Menopause operativ herbeigeführt wurde, ist schätzungsweise 2,269-mal so groß wie die Erkrankungschance von Frauen, die die Menopause auf biologischem Wege erreicht haben.

In Anbetracht dieser Untersuchungsergebnisse ist davon auszugehen, dass die Variable „Menopause-Status“ einen großen Einfluss auf das Brustkrebsrisiko nimmt. Neben der daraus resultierenden Notwendigkeit, sie in komplexeren Modellen zu berücksichtigen, kommt ihr im weiteren Verlauf dieser Arbeit allerdings eine noch größere Bedeutung zu.
In Unterkapitel 5.3.8 wird das Merkmal Menopause genutzt, um die Studienteilnehmerinnen in zwei Subpopulationen aufzuteilen. Die erste Subpopulation besteht aus allen Frauen, die die Menopause noch nicht erreicht haben oder gerade in den Wechseljahren sind (Y=1,2) und die zweite Subpopulation wird von den Frauen gebildet, bei denen die Menopause bereits eingetreten ist (Y=3,4).
Eine Unterteilung auf Grundlage dieses Merkmals „Menopause-Status“ ist sinnvoll, da medizinische Untersuchungen zeigen, dass Brustkrebserkrankungen verstärkt in den Lebensjahren nach der Menopause auftreten. Insbesondere gilt ein spätes Menopause-Alter nach Grundmann (1994) als eigenständiger Risikofaktor für Brustkrebs.

Für die 614 Frauen der „Post-Menopause“ kann das Merkmal „Menopause-Status“ noch als binäres untersucht werden. Es ergibt sich die in Tabelle 5.3.27 dargestellte Kontingenztafel.

Tabelle 5.3.27: Kontingenztafel „Post-Menopause“

Menopause-Status	**BCA-Status**	
Menopause…	**0 – kein BCA**	**1 – BCA erkrankt**
3 – …regulär erreicht	249 (224)	107 (132)
4 – …operativ herbeigeführt	138 (163)	120 (95)

Ein Vergleich der beobachteten mit den erwarteten Zellhäufigkeiten in obiger Kontingenztafel zeigt, dass eine operativ herbeigeführte Menopause deutlich mit Brustkrebs assoziiert ist. Der P-Wert des dazugehörigen Chi-Quadrat-Tests ist kleiner als 0,0001. Bei Adjustierung hinsichtlich des einfachen und quadrierten Lebensalters ergibt sich ein Schätzwert von 2,326 für das Odds-Ratio zwischen Frauen, deren Menopause operativ herbeigeführt wurde, und Frauen, deren Menopause durch biologische Vorgänge eingeleitet wurde. Dieser Wert ist vergleichbar mit der Odds-Ratio Schätzung von 2,269, die sich oben zwischen diesen beiden Gruppen ergab, als die gesamte Studienpopulation berücksichtigt wurde.

5.3.4 Ergebnisse der Variablen-Vorauswahl

In diesem Unterkapitel werden die Ergebnisse der Variablen-Vorauswahl zusammengefasst. Zunächst folgt zu jeder selektierten Variablen eine Kurzbeschreibung. Im Anschluss werden die Odds-Ratio Schätzungen und Wald-Test-P-Werte, die sich für diese Variablen ergaben, tabellarisch aufgelistet. Die Angaben beziehen sich jeweils auf logistische Regressionsmodelle, in denen fehlende Werte als eigenständige Ausprägungskategorie aufgefasst und die beiden Lebensalter-Kovariablen berücksichtigt werden.

Kurzbeschreibung der selektierten Variablen

1) Einkommen - EK

Die kategoriale Variable „Einkommen“ weist die Ausprägungen 1 bis 6 auf. Die Ausprägungen ergeben sich durch Gruppierung des durchschnittlichen jährlichen Familieneinkommens der jeweiligen Frau. Tabelle 5.3.28 kann das zugrunde gelegte Gruppierungsschema entnommen werden. Aufgrund des ordinalen Messniveaus wurde entschieden, die Variable im Rahmen logistischer Modelle als Stetige zu behandeln.

Tabelle 5.3.28: Gruppierungsschema „familiäres Jahreeinkommen“

Familiäres Jahreseinkommen	„Einkommen“
<5.000 $	1
5.000 – 15.0000 $	2
15.000 – 25.000 $	3
25.000 – 35.000 $	4
35.000 – 50.000 $	5
>50.000 $	6

2) Weiterführende Schule - WS

Die binäre Variable „Weiterführende Schule“ gibt an, ob die Frau über einen Secondary-School-Abschluss (12 Schuljahre) hinausgekommen ist (WS=1) oder nicht (WS=0). Diese Variable wurde im letzten Unterkapitel aus dem Merkmal „Schuljahre“ generiert.

3) Schwangerschaft - SS

Die binäre Variable „Schwangerschaft“ gibt an, ob die Frau schon einmal schwanger war (SS=1) oder nicht (SS=0). Ob die Schwangerschaft zu der Geburt eines Kindes geführt hat

oder nicht, ist dabei nicht von Bedeutung. Diese Variable ist im Modell nicht von Signifikanz, wird aber dennoch ausgewählt, da sie ggf. zur Adjustierung eingesetzt werden kann bzw. muss.

4) Anzahl Schwangerschaften - ASS

Die Variable „Anzahl Schwangerschaften“ gibt an, wie oft die jeweilige Frau schwanger war. Diese Variable wird im Rahmen logistischer Modelle als Stetige aufgefasst.

5) Anzahl Lebendgeburten – ALG

Die Variable „Anzahl Lebendgeburten“ gibt für jede Frau an, wie viele ihrer Schwangerschaften zu Lebendgeburten führten.

6) Tubensterilisation – TS

Die Variable „Tubensterilisation“ gibt an, ob bei der betreffenden Studienteilnehmerin eine Tubensterilisation durchgeführt wurde (TS=1) oder nicht (TS=0).

7) Fehlgeburt - FG

Die binäre Variable „Fehlgeburt“ gibt an, ob die Studienteilnehmerin schon einmal eine Fehlgeburt erlitten hat (FG=1) oder nicht (FG=0). Ob es eine oder mehrere Fehlgeburten gegeben hat, ist nicht von Bedeutung.

8) Schwangerschaftsabbruch - SA

Die binäre Variable „Schwangerschaftsabbruch“ gibt für jede Frau an, ob sie schon einmal eine Schwangerschaft durch Abtreibung beenden ließ (SA=1) oder nicht (SA=0). Die Anzahl durchgeführter Abtreibungen ist nicht von Bedeutung.

9) Hysterektomie – HT

Die binäre Variable „Hysterektomie“ gibt für jede Frau an, ob ihr die Gebärmutter operativ entfernt wurde (HT=1) oder nicht (HT=0).

10) Ovar-Entfernung - OE

Die binäre Variable „Ovar-Entfernung“ gibt an, ob der jeweiligen Studienteilnehmerin ein Eierstock operativ entfernt wurde (OE=1) oder nicht (OE=0). Auch wenn beide Eierstöcke entfernt wurden, wird die Variable OE auf den Wert 1 gesetzt.

11) Tumor - TU

Die binäre Variable „Tumor“ gibt an, ob die jeweilige Frau an einer Nichtbrustkrebs-Tumorerkrankung leidet (TU=1) oder nicht (TU=1). Von Bedeutung sind ausschließlich Primärtumoren, die keine Folge (Metastasen) von Brustkrebstumoren sind.

12) Menopause-Status – MS

Die kategoriale Variable „Menopause-Status“ gibt an, ob die betreffende Frau die Menopause noch nicht (MS=1), gerade erst (MS=2) oder schon erreicht hat. (MS=3,4). Wurde die Menopause bereits erreicht, wird zusätzlich unterschieden, ob die Menopause durch biologische Prozesse eingetreten ist (MS=3) oder auf operativem Wege herbeigeführt wurde (MS=4). In Tabelle 5.3.29 beziehen sich Odds-Ratio und P-Wert auf den Vergleich zwischen den Kategorien 3 und 4.

13) Unregelmäßige Periode – URP

Die binäre Variable „Regelmäßige Periode“ wurde im letzten Unterkapitel generiert und gibt für jede Frau an, ob ihre Periode immer in regelmäßigen Abständen eingetreten ist (URP=0) oder nicht (URP=1).

14) Antibabypille – AP

Die binäre Variable Antibabypille gibt für jede Frau an, ob sie im Verlaufe ihres Lebens die Antibabypille eingenommen hat (AP=1) oder nicht (AP=0).

15)Verlust des Ehemanns – VE

Die binäre Variable „Verlust des Ehemanns“ wurde im letzten Unterkapitel aus dem Merkmal „Familienstand“ generiert und gibt für jede Frau an, ob sie ihren letzten Ehemann durch Versterben oder Scheidung verloren hat (VE=1) oder nicht, das heißt gegenwärtig verheiratet ist bzw. nie verheiratet war (VE=0).

16) Alter 1te Geburt – AG

Die stetige Variable „Alter 1te Geburt“ gibt an, in welchem Lebensjahr die Studienteilnehmerin zum ersten mal eine vollständige Schwangerschaft, das heißt eine Schwangerschaft die zur Geburt eines lebendigen Kindes oder einer Totgeburt geführt hat, durchlebt hat. Die Bedeutung dieser Variablen kann nur beurteilt werden, wenn eine zusätzliche Adjustierungsvariable

I_{AG} ins Modell integriert wird, welche angibt, ob die Frau überhaupt schon mal eine vollständige Schwangerschaft durchlebt hat (I_{AG}=0) oder nicht (I_{AG}=1) (vgl. 5.3.2). Für Frauen, die noch keine vollständige Schwangerschaft hinter sich haben (I_{AG}=1), wird die Variable „Alter 1te Geburt“ auf den Wert 0 gesetzt (AG=0).

17) Menarche – ME

Die binäre Variable „Menarche“ gibt für jede Frau an, ob bei ihr die Menarche nach dem 12ten Lebensjahr eingetreten ist (ME=1) oder nicht (ME=0).

18) Menopause – MP

Die binäre Variable „Menopause“ wurde aus Angaben über das Lebensalter zum Zeitpunkt des Eintretens der Menopause generiert. Die Variable „Menopause“ gibt für jede Frau an, ob sie die Menopause nach dem 50ten Lebensjahr erreicht hat (MP=1) oder nicht (MP=0). Da einige Frauen die Menopause noch gar nicht erreicht haben, muss ins Modell eine zusätzliche Indikatorvariable I_{MP} integriert werden, welche angibt, ob die Probandin die Menopause schon erreicht hat (I_{MP}=0)oder nicht (I_{MP}=1).
Anschließend wird die Variable „Menopause“ für diese Frauen auf Null gesetzt (MP=0).

Tabelle 5.3.29: Tabellarische Zusammenfassung der Ergebnisse – Teil 1

Variable	Variablentyp und Wertebereich	Anzahl fehlender Werte	OR	P-Wert
Einkommen	Stetig - {1,2,…,6}	597	1,123	0,0473
Weiterführende Schule	Binär - 0/1	468	1,360	0,0715
Schwangerschaft	Binär - 0/1	10	0,944	0,7622
Anzahl Schwangerschaften	Stetig – {0,1,…,17}	66	0,912	0,0007
Anzahl Lebendgeburten	Stetig – {0,1,…,17}	38	0,843	<0,0001
Tubensterilisation	Binär – 0/1	520	1,415	0,1854
Fehlgeburt	Binär - 0/1	72	2,170	<0,0001
Schwangerschaftsabbruch	Binär - 0/1	72	2,585	<0,0001
Hysterectomy	Binär - 0/1	108	1,881	<0,0001
Ovar-Enfernung	Binär - 0/1	191	1,525	0,0480
Tumor	Binär – 0/1	55	0,565	0,0469
Menopause-Status	Kategorial {1,…,4}	145	2,269	<0,0001

Tabelle 5.3.29: Tabellarische Zusammenfassung der Ergebnisse – Teil 2

Variable	Variablentyp und Wertebereich	Anzahl fehlender Werte	OR	P-Wert
Antibabypille	Binär – 0/1	416	1,417	0,0890
Verlust des Ehemannes	Binär – 0/1	119	1,833	<0,0001
Alter 1te Geburt	Stetig – {0,12,…,43}	99	1,039	0,0071
Menarche	Binär - 0/1	389	1,636	0,0025
Menopause	Binär – 0/1	379	0,531	0,0067
Unregelmäßige Periode	Binär – 0/1	482	1,543	0,0195

Bevor die 18 Variablen simultan betrachtet werden können, ist noch zu klären, ob Abhängigkeiten zwischen diesen vorliegen. Liegt zwischen zwei Variablen eine Abhängigkeit im Sinne einer Korrelation vor, ist es nicht sinnvoll, beide gleichzeitig als Einflussvariablen in logistischen Regressionsmodellen zu betrachten. Hosmer und Lemeshow (1989) zeigen anhand von Beispielen, dass bei Vorliegen korrelierter Variablen die dazugehörigen Varianzen der Schätzstatistiken sehr groß werden. Da weiter große Varianzen bzw. Varianzschätzwerte zwangsläufig auch zu großen Wald-Test-P-Werten führen, entsteht bei Betrachtung dieser P-Werte der Eindruck, dass die korrelierten Variablen keinen signifikanten Einfluss auf die Zielvariable nehmen. Folglich stellen P-Werte nur dann ein adäquates Mittel zur Beurteilung des Signifikanzgrades von Variablen dar, wenn hochgradige Korrelationen ausgeschlossen werden können. Entsprechend sind vor der Betrachtung multipler Modelle Korrelationsanalysen durchzuführen. Wird zwischen zwei Variablen eine Korrelationen entdeckt, kann nur eine der beiden ins Modell aufgenommen werden. Je nachdem, welche Ursache für die Abhängigkeit vorliegt, ist entweder aus mathematisch-statistischer oder substanzwissenschaftlicher Sicht zu entscheiden, welche Variable besser geeignet ist.

Zum einen wäre denkbar, dass von zwei hochgradig korrelierten Variablen nur eine tatsächlich mit der Zielvariablen assoziiert ist, wohingegen zwischen der Anderen und der Zielvariablen nur eine Scheinassoziation vorliegt. In diesem Fall nimmt nur die eine Variable Einfluss auf das Erkrankungsrisiko, wohingegen die Andere nur indirekt, das heißt über die Korrelation mit der einflussreichen Variablen, mit dem Erkrankungsrisiko assoziiert ist. Da dieser somit keine kausale Bedeutung für das Erkrankungsrisiko zukommt, ist es nicht sinnvoll sie als Exposition zu untersuchen. Da solche Abhängigkeitsstrukturen aus mathematischer Sicht nicht zu erkennen sind, kann bei Vorliegen einer Korrelation nur auf Grundlage substanzwis-

senschaftlicher Überlegungen entschieden werden, ob die beschriebene Situation vorliegt, und welche der beiden Variablen folglich im Modell Berücksichtigung finden sollte.
Zum anderen wäre aber auch denkbar, dass die Korrelation zwischen zwei Variablen nur dadurch zustande kommt, dass von beiden dasselbe Merkmal auf unterschiedliche Weise beschrieben wird. Wenngleich wieder nur inhaltlich zu erkennen ist, dass zwei Variablen nur alternative Quantifizierungen eines interessierenden Merkmals sind, ist es in diesem Fall legitim, aus mathematischer Sicht zu entscheiden, welche der beiden Variablen im Modell berücksichtigt werden soll. Bei vergleichbarer substanzwissenschaftlicher Interpretierbarkeit und Aussagekraft sollte in diesem Fall zu Gunsten der Variablen entschieden werden, der im Modell eine größere Bedeutung zukommt. Sinnvoll erscheint es, zwei Alternativmodelle zu betrachten, die sich nur dahingehend unterscheiden, dass die korrelierten Variablen gegeneinander ausgetauscht wurden. Auf Grundlage der P-Werte kann entschieden werden, welche der beiden Variablen im Modell einen höheren Signifikanzgrad aufweist und damit das interessierende Merkmal besser repräsentiert.

Suche nach Abhängigkeiten zwischen den Einflussvariablen

Unter Zuhilfenahme von Kontingenztafeln (kategoriale Variablen) bzw. durch die Berechnung von Korrelationskoeffizienten (stetige Variablen) konnten im vorliegenden Fall 2 deutliche Abhängigkeitsstrukturen entdeckt werden. Beide Abhängigkeiten können inhaltlich gedeutet werden.

Zunächst zeigen die folgenden Kontingenztafeln (Tabelle 5.3.30 und 5.3.31), dass die Variable „Menopause-Status“ sowohl mit der Variablen „Ovar-Entfernung“ als auch mit der Variablen „Hysterektomie“ korreliert ist.

Tabelle 5.3.30: Kontingenztafel Ovar-Entfernung gegen Menopause-Status

Menopause-Status	**Ovar-Entfernung**	
Menopause…	**Ovar nicht entfernt** **OE=0**	**Ovar entfernt** **OE=1**
1 – …noch nicht erreicht	310 (276)	5 (40)
2 – …stellt sich gerade ein	45 (42)	3 (6)
3 – …regulär erreicht	307 (275)	7 (39)
4 – …operativ herbeigeführt	110 (179)	94 (25)

Tabelle 5.3.30: Kontingenztafel Hysterektomie gegen Menopause-Status

Menopause-Status **Menopause…**	**Hysterektomie**	
	Gebärmutter nicht entfernt **HT=0**	**Gebärmutter entfernt** **HT=1**
1 – …noch nicht erreicht	323 (235)	0 (88)
2 – …stellt sich gerade ein	47 (36)	3 (14)
3 – …regulär erreicht	305 (233)	15 (87)
4 – …operativ herbeigeführt	14 (185)	240 (69)

In beiden Fällen ist der P-Wert des Chi-Quadrat-Tests auf Unabhängigkeit (bzw. Unkorreliertheit) kleiner als 0,0001. Die genaue Abhängigkeitsstruktur ist jeweils durch einen Vergleich der beobachteten mit den unter Unabhängigkeit zu erwartenden Zellhäufigkeiten zu erkennen.

Besonders auffällig ist in beiden Fällen, dass die beobachtete Häufigkeit der Kombination „Menopause operativ erreicht“ und „Ovar entfernt“ bzw. „Gebärmutter entfernt“ deutlich über dem Erwartungswert liegt. Offensichtlich sind Hysterektomie-Eingriffe und/oder Operationen bei denen Ovarien entfernt werden, die Hauptverursacher von „operativ herbeigeführten“ Menopausen.

Unklar bleibt in beiden Fällen, ob der operative Eingriff oder das unnatürliche Einsetzen der Menopause von Bedeutung für das Brustkrebsrisiko ist. Im Hinblick auf die folgenden Regressionsmodelle wurde der Variablen „Menopause-Status“ der Vorzug gegeben.

Ebenfalls besteht ein ungünstiger Zusammenhang zwischen der binären Variablen „Menopause-Alter“ und dem „Menopause-Status“. Es gilt, dass Frauen, deren Menopause operativ herbeigeführt wurde, die Menopause deutlich früher erreicht haben als die Frauen, bei denen die Menopause auf biologischem Wege eingetreten ist. Da somit eine frühe Menopause (<51 Jahre) mit einer operativ herbeigeführten Menopause assoziiert ist, was in Anbetracht der bisherigen Untersuchungsergebnisse offensichtlich einen deutlichen Risikofaktor für Brustkrebs darstellt, präsentiert sich die Ausprägung „frühe Menopause“ im vorliegenden Fall ebenfalls als Risikofaktor für Brustkrebs. Dies steht im Widerspruch dazu, dass in der Medizin bereits als empirisch gesichert gilt, dass eine späte Menopause einen Risikofaktor für Brustkrebs darstellt (vgl. Grundmann 1994). Folglich kann die Variable „Menopause-Alter“ nicht ins Modell aufgenommen werden. Die folgende Kontingenztafel (Tabelle 5.3.31) verdeutlicht den ungünstigen Zusammenhang zwischen den beiden Variablen.

Tabelle 5.3.31: Kontingenztafel Menopause-Status – Menopause-Alter

Menopause…	**Menopause-Alter**	
	<50 Lebensjahr	**>50 Lebensjahr**
„…biologisch eingetreten“	94 (124)	62 (32)
„…operativ herbeigeführt“	204 (174)	16 (46)

Darüber hinaus ergibt sich für die Variablen „Lebendgeburten“ und „Anzahl Schwangerschaften“ ein Korrelationskoeffizient von 0,93. Intuitiv ist klar, dass diese Korrelation dadurch zustande kommt, dass die meisten Schwangerschaften zur Geburt eines lebendigen Kindes führen. Schwangerschaftsabbrüche und Komplikationen bei der Geburt (Fehl- und Totgeburten) stellen seltene Ereignisse dar, die die Beziehung „Anzahl Schwangerschaften“ =„Lebend-Geburten“ nur unwesentlich stören.
In der humanmedizinischen Literatur ist umstritten, ob „häufiges Stillen“ das Brustkrebserkrankungsrisiko senkt oder nicht (vgl. Bleich et al. 1995). Davon ausgehend, dass beide Variablen mit dem Ziel erhoben wurden, die Häufigkeit des Stillens zu quantifizieren, stellt die Variable „Anzahl Lebendgeburten“ eine geeignetere Größe dar.

Ebenfalls zeigen weiterführende Untersuchungen, dass die Variable „Anti-Baby-Pille“ aufgrund des Studiendesigns keine sinnvolle Exposition darstellt. Die Frauen, die die Anti-Baby-Pille eingenommen haben, sind im Durchschnitt nur 41,7 Jahre alt, wohingegen die Frauen, die die Pille nicht genommen haben, im Durchschnitt 62,5 Jahre alt sind. Dieser Altersunterschied ist inhaltlich darauf zurückzuführen, dass die Antibabypille erst in den 70er Jahren auf den Markt gekommen ist. Berücksichtigt man nun, dass die Studienteilnehmerinnen aus unterschiedlichen Generationen stammen und, dass davon auszugehen ist, dass die Pille zur Empfängnisverhütung ausschließlich vor Erreichen der Menopause eingenommen wird, ist die Einnahme der Pille zwangsläufig mit den jüngeren Generationen assoziiert. Da sich weiter - bedingt durch das Studiendesign - die Fallgruppe fast ausschließlich aus Frauen der dritten Generation zusammensetzt, so dass für die Fälle die Pille bereits verfügbar war, wohingegen für die deutlich seltener an Brustkrebs erkrankten Mütter dieser Fälle (Generation 2) diese Art der Empfängnisverhütung noch nicht möglich war, ergibt sich eine studiendesignbedingte Assoziation zwischen der Einnahme der Pille und der Brustkrebskrankheit.

Als Fazit, der in diesem Unterkapitel durchgeführten Untersuchungen, kann festgehalten werden, dass die 5 Variablen „Hysterektomie“, „Ovar-Entferung“, „Anzahl Schwangerschaften“, „Menopause-Alter“ und „Anti-Baby-Pille“ als solche nicht berücksichtigt werden können.

5.3.4 Multiple logistische Regressionsmodelle

Die bisherigen Ergebnisse wurden mit Hilfe einfacher Regressionsmodelle erzielt, in denen jeweils ein interessierendes Merkmal bzw. eine daraus generierte Variable (als Exposition) simultan mit den beiden Lebensalter-Variablen betrachtet wurde. Auf Grundlage dieser Ergebnisse konnten 18 Variablen selektiert werden (vgl. Tabelle 5.3.29), die mutmaßlich von Bedeutung für das Brustkrebsrisiko sind. Weiterführende Analysen zeigten, dass von diesen 18 Variablen höchstens 13 sinnvoll simultan als Expositionen betrachtet werden können.
Ein wesentliches Problem ist nun, dass nicht ausgeschlossen werden kann, dass die bisherigen Ergebnisse - bedingt durch die unkontrollierten Einflüsse der anderen Variablen – fehlerhaft bzw. verzerrt sind. Durch die Aufnahme der beiden Lebensalter-Variablen wurde bei jeder untersuchten Exposition ausschließlich hinsichtlich des Lebensalters adjustiert. Davon ausgehend, dass darüber hinaus auch einige der anderen Variablen das Erkrankungsrisiko beeinflussen (Störvariablen), stört die Nichtberücksichtigung dieser die Beziehung zwischen Exposition und Brustkrebsrisiko. Speziell, wenn die Störvariablen mit der Exposition vermengt sind, führt dies dazu, dass die unterschiedlichen Expositionslevel bereits mit unterschiedlichen Grunderkrankungsrisiken assoziiert sind. Bedingt durch die Nichtberücksichtigung der Störvariablen kommt es unweigerlich zu einer Effektüberlagerung und damit zu einer fehlerhaften Einschätzung des Einflusses der unterschiedlichen Expositionslevel. Zur Vermeidung von Verzerrungen dieser Art („Confounding-Bias“) müssen die 13 Variablen im Rahmen logistischer Regressionsmodelle simultan betrachtet werden. Bedingt durch die damit verbundene Adjustierung kann erst in einem solchen Modell anhand der Wald-Test-P-Werte entschieden werden, welche der selektierten Variablen tatsächlich Einfluss auf das Erkrankungsrisiko nehmen.

Im vorliegenden Fall muss aufgrund der Unvollständigkeit des Datenmaterials zunächst entschieden werden, wie mit fehlenden Variableneinträgen verfahren werden soll. Aufgrund der großen Anzahl fehlender Werte setzt die gleichzeitige Betrachtung mehrerer Variablen eine Miteinbeziehung von Frauen mit unvollständigen Merkmalswerten voraus. Im Folgenden werden deshalb (unabhängig voneinander) die beiden Ad-hoc-Verfahren (vgl. 4.3.9) im Umgang mit fehlenden Werten eingesetzt.

In Unterkapitel 5.3.5 werden fehlende Werte durch Anwendung des Indikatorvariablen-Verfahrens als eigenständige Kategorie der zugehörigen Variablen aufgefasst und in Unterkapitel 5.3.6 kommt das (unbedingte) Probability-Imputation-Verfahren zum Einsatz. In Unterkapitel 5.3.7 werden die Ergebnisse, die mit den beiden Verfahren erzielt wurden, verglichen.

5.3.5 Eigene Kategorie für fehlende Werte

In einem ersten Regressionsmodell werden fehlende Werte der 13 Variablen als eigenständige Ausprägungskategorie aufgefasst. Da fehlende Werte bzgl. der Variablen „Schwangerschaftsabbruch“ und „Fehlgeburt“ ausschließlich gemeinsam auftreten, sind nicht 13, sondern nur 12 zusätzliche Indikatorvariablen, zur Kennzeichnung fehlender Werte, im Modell zu berücksichtigen. In den Originalvariablen werden fehlende Einträge jeweils durch den Wert 0 aufgefüllt. Einzige Ausnahme bildet das kategoriale Merkmal „Menopause-Status“, welches unter Zuhilfenahme der Referenzgruppencodierung parametrisiert werden muss. Fehlende Menopause-Zustände werden durch den Status 3: „Menopause biologisch eingetreten“ aufgefüllt, da dieser Status als Referenzkategorie dient.

Darüber hinaus muss zur Adjustierung eine Indikatorvariable „Vollständige Schwangerschaft“ ins Modell integriert werden, welche genau dann 1 ist, wenn die betreffende Studienteilnehmerin noch keine vollständige Schwangerschaft hinter sich gebracht hat. Die eigentliche Variable „Alter 1te Schwangerschaft“ wird in diesem Fall auf 0 gesetzt. Zusammen mit den beiden Lebensalter-Variablen beinhaltet das Modell insgesamt 30 Variablen, von denen allerdings nur 18 von substanzwissenschaftlicher Bedeutung sind, das heißt keine Indikatorvariablen für fehlende Werte darstellen.

Tabelle 5.3.32 zeigt das Ergebnis (Odds-Ratio-Schätzungen und P-Werte) der ML-Modellschätzung.

Tabelle 5.3.32: Ergebnisse bei simultaner Betrachtung der 14 Variablen und Einführung von 12 Indikatorvariablen für fehlende Werte

Die 14 regulären Merkmale (parametrisiert durch 18 Variablen)…		
Merkmal bzw. Variable	**Odds-Ratio**	**P-Wert**
Lebensalter…	------	------
…einfache (stetig)	1,225	**<0,0001**
…quadriert (stetig)	0,999	**<0,0001**
Einkommen (stetig)	1,048	0,5043
Weiterführende Schule (binär)	1,045	0,8334
Schwangerschaft (binär)	0,544	0,1609
Anzahl Lebendgeburten (stetig)	0,880	**0,0031**
Unregelmäßige Periode (binär)	1,519	**0,0372**
Verlust des Ehemannes (binär)	0,851	0,3864
Späte Menarche (binär)	1,341	0,1119
1te vollständige Schwangerschaft…	------	------
…Alter 1te Schwangerschaft (stetig)	1,019	0,3191
…Vollständige Schwangerschaft (binär)	0,738	0,6121
Tumor (binär)	0,863	0,6627
Tubensterilisation (binär)	1,385	0,2471
Menopause… (kategorial)	------	------
…noch nicht erreicht (binär)	0,902	0,7599
…Menopause gerade erreicht (binär)	0,725	0,3817
…Menopause operativ erreicht (binär)	1,223	0,3609
Schwangerschaftsabbruch (binär)	1,380	0,2091
Fehlgeburt (binär)	1,640	**0,0445**

…und die 12 Indikatorvariablen für fehlende Werte		
Indikatorvariable I(.)	**Odds-Ratio**	**P-Wert**
I(Einkommen)	0,886	0,7281
I(Weiterführende Schule)	0,579	0,1988
I(Schwangerschaft)	0,443	0,5759
I(Anzahl Lebendgeburten)	0,338	0,2356

Tabelle 5.3.23: Fortführung

Indikatorvariable I(.)	Odds-Ratio	P-Wert
I(Unregelmäßige Periode)	1,224	0,6430
I(Verlust des Ehemannes)	0,748	0,4868
I(Späte Menarche)	0,839	0,4737
I(Alter 1te Schwangerschaft)	2,386	0,1339
I(Tumor)	1,081	0,8563
I(Tubal Ligation)	0,199	<0,0001
I(Menopause-Status)	0,582	0,2021
I(Fehlgeburt und SS-Abbruch)	2,918	0,0230

Die Ergebnisse der inhaltlich relevanten Variablen werden durch die Anwesenheit substanzwissenschaftlich irrelevanter Indikatorvariablen, die ausschließlich die zugrunde liegende Fehlende-Werte-Systematik des Datensatzes zur Erklärung der Zielvariablen heranziehen, verzerrt. Folglich ist die Berücksichtigung von Indikatorvariablen, die zu nicht bedeutsamen bzw. nicht-signifikanten, inhaltlich interpretierbaren Variablen korrespondieren, nicht sinnvoll. Im Hinblick auf eine Entfernung dieser Indikatorvariablen ist allerdings zu beachten, dass jedes Merkmal aufgrund der fehlenden Merkmalswerte nur durch die gleichzeitige Anwesenheit von Substanz- und Indikatorvariable sinnvoll parametrisiert ist. Da fehlende Werte in den Substanzvariablen mit dem Ausprägungswert 0 aufgefüllt wurden, würde die alleinige Entfernung einer Indikatorvariablen zu einer inadäquaten Parametrisierung des dazugehörigen Merkmals führen. In jeder Substanzvariablen nämlich führt die Auffüllung fehlender Werte dazu, dass diese in Bezug gesetzt werden zu den regulär beobachteten Ausprägungen. Erst durch die simultane Betrachtung der dazugehörigen Indikatorvariablen, welche die fehlenden Werte kennzeichnet, kommt den fehlenden Werten ein eigener Effekt zu. Dieser zusätzliche „Fehlende Werte“-Effekt löst die Beziehung zwischen den regulären Ausprägungen und den, in den Substanzvariablen auf 0 gesetzten, fehlenden Werten wieder auf. Konsequenterweise können zusammengehörige Substanz- und Indikatorvariablen nur gemeinsam aus dem Modell entfernt werden, wenn eine substanzwissenschaftliche Interpretierbarkeit erhalten bleiben soll. Davon ausgehend, dass die Indikatorvariablen, die zu unbedeutenden Substanzvariablen korrespondieren, nur unnötige Verzerrungen hervorrufen, stellt die Entfernung solcher Paare (einflusslose Substanzvariable und dazugehörige Indikatorvariable) kein Problem dar.

Durch die Anwendung einer speziellen Prozedur (siehe unten), kann schrittweise jeweils ein Paar, bestehend aus einer unbedeutenden Substanzvariablen und einer verzerrenden Indikatorvariablen, entfernt werden, bis das Modell ausschließlich bedeutsame Substanzvariablen und die dazugehörigen Indikatorvariablen beinhaltet. Da es sich bei dieser Prozedur um eine modifizierte Version der klassischen Rückwärtsauswahl handelt, wird zunächst kurz auf diese eingegangen.

Klassisches automatisiertes Rückwärtsauswahlverfahren

Die klassische schrittweise Rückwärtsauswahl zielt primär darauf, unnötige Modellparameter einzusparen, da mit der Anzahl von Modellparametern die Schätzvarianz steigt. Schrittweise wird jeweils die unwichtigste Variable aus dem Modell entfernt und an die verbleibenden Variablen ein neues Modell angepasst. Das Verfahren endet, wenn sich nur noch bedeutsame Variablen im Modell befinden. Als Gradmesser für die Wichtigkeit bzw. Bedeutsamkeit dienen i.a. die Wald-Test-P-Werte. Das heißt, in jedem Schritt wird die Variable entfernt, die im aktuellen Modell den größten Wald-Test-P-Wert aufweist. Einzelheiten zu der klassischen Rückwärtsauswahl können Hosmer & Lemeshow (1989) entnommen werden.
Da bei der klassischen Rückwärtsauswahl alle Einflussvariablen „gleichberechtigt“ behandelt werden, im vorliegenden Fall allerdings zwischen Substanz- und Indikatorvariablen zu unterscheiden ist, ist das Verfahren in obiger Form nicht anwendbar. Sinnvoll erscheint allerdings die folgende modifizierte Variante der Rückwärtsauswahl:

Spezielle Rückwärts-Auswahl-Prozedur

Schrittweise wird jeweils das substanzwissenschaftlich unwichtigste Paar, bestehend aus der Substanz- und der dazugehörigen Indikatorvariablen, aus dem Modell entfernt und den verbleibenden Variablen ein neues Modell angepasst. Als Gradmesser für die Bedeutung eines jedes Paares dient in jedem Schritt allerdings ausschließlich der Wald-Test-P-Wert der Substanzvariablen, da lediglich diese Variable einen substanzwissenschaftlich relevanten Aspekt beschreibt, wohingegen die dazugehörige Indikatorvariable – unabhängig von ihrem P-Wert – nicht von inhaltlichem Interesse ist. Das Verfahren endet, wenn sich im aktuellen Modell ausschließlich Paare befinden, bei denen der P-Wert der Substanzvariablen einen vorher festgelegten Wert unterschreitet.

Durch die Anwendung dieser speziellen Prozedur wird gewährleistet, dass nacheinander die substanzwissenschaftlich unbedeutsamsten Merkmale aus dem Modell entfernt werden. Jedes

Merkmal ist im Modell durch eine Substanz- und eine Indikatorvariable parametrisiert. Der P-Wert der Substanzvariablen vermittelt einen Eindruck von der Bedeutung des Merkmals für das Erkrankungsrisiko und der P-Wert der Indikatorvariablen zeigt an, inwieweit – studiendesignbedingt – fehlende Werte dieses Merkmals mit Brustkrebs assoziiert sind.

Da der studiendesignbedingte Erklärungswert fehlender Werte inhaltlich nicht von Relevanz ist, ganz im Gegenteil sogar zu erwarten ist, dass dieser Ergebnisverzerrungen hervorruft, kann auf Grundlage des P-Wertes der Substanzvariablen die Bedeutung des Merkmals für das Brustkrebsrisiko beurteilt werden. Diesen Aspekt nutzt das modifizierte Rückwärtsauswahl-Verfahren. Zu beachten ist lediglich, dass in jedem Schritt der Ausschluss der unbedeutsamsten Substanzvariablen mit dem Ausschluss einer Indikatorvariablen einhergeht, welche möglicherweise einen studiendesignbedingten Erklärungswert für die Zielvariable hat. Im Gegensatz zur klassischen Rückwärtsauswahl ist es daher nicht ungewöhnlich, wenn die Odds-Ratio-Schätzwerte der verbleibenden Variablen von Schritt zu Schritt Veränderungen unterliegen. Solche Veränderungen sind dann nämlich nicht auf den Ausschluss der nicht-signifikanten Substanzvariablen, sondern auf den Ausschluss der signifikanten Indikatorvariablen, zurückzuführen. Die Entfernung einer zwar signifikanten, aber inhaltlich wertlosen, Indikatorvariablen führt gewissermaßen zu einer „Entzerrung“ der Odds-Ratio-Werte, und die resultierenden Veränderungen stellen folglich eher ein Argument für als gegen den Ausschluss des Variablenpaares dar. Dieser Sachverhalt gilt allerdings ausschließlich für das modifizierte Verfahren. Ergibt sich bei Anwendung der klassischen Rückwärtsauswahl nach dem Ausschluss einer nicht-signifikanten Variablen eine wesentliche Veränderung der Odds-Ratio-Schätzwerte, ist dies immer als Indiz dafür zu werten, dass die Variable doch einen gewissen Einfluss auf die Zielvariable hat (vgl. Hosmer und Lemeshow 1989).

Bei Anwendung des modifizierten Rückwärtsauswahl-Verfahrens auf das obige Modell (vgl. Tabelle 5.3.32) ist im ersten Schritt die binäre Variable „Weiterführende Schule“ inklusive der dazugehörigen Indikatorvariable für fehlende Werte aus dem Modell zu entfernen, da diese unter den Substanzvariablen den größten Wald-Test-P-Wert (0,8334) aufweist. Anschließend kann ein neues Modell an die verbleibenden 28 Variablen angepasst werden und erneut ein Variablenpaar entfernt werden. Dieser Schritt wird solange wiederholt, bis alle Substanzvariablen einen hinreichend kleinen Wald-Test-P-Wert aufweisen. In der Literatur (vgl. Hosmer & Lemeshow 1989)wird für die klassische Rückwärtsauswahl das in der Statistik übliche Niveau von 0,05 als Grenzwert empfohlen. Dieses ist im vorliegenden Fall jedoch nicht prob-

lemadäquat, da in jedem Fall Indikatorvariablen im Modell verbleiben, so dass angesichts ihrer studiendesignbedingten Assoziation mit der Zielvariablen Ergebnisverzerrungen hinsichtlich der Substanzvariablen zu erwarten sind. Testläufe, denen verschiedene Niveaus zugrunde gelegt wurden, zeigen, dass bei der Wahl von 0,1 eine angemessene Anzahl von Substanzvariablen im Modell verbleibt.
Tabelle 5.3.33 zeigt für jeden Schritt, welches substanzwissenschaftlich wertlose Variablenpaar bei Verwendung dieses Grenzwertes aus dem Modell entfernt werden muss. Sämtliche P-Wert-Angaben beziehen sich auf das aktuelle Modell, aus welchem bereits die in den vorangegangenen Schritten genannten Variablenpaare entfernt wurden. Bei Durchführung des Verfahrens sind drei Besonderheiten zu beachten:

1) Zunächst ist das Merkmal „Menopause-Status“ durch 3 Substanzvariablen parametrisiert, welche deshalb nur gemeinsam ausgeschlossen werden können. Als Gradmesser für die Bedeutung des Merkmals wird in jedem Schritt der kleinste der drei P-Werte verwendet (vgl. Schritt 4).

2) Darüber hinaus wird die zur Variablen „Alter 1te Schwangerschaft“ gehörige Indikatorvariable, die kennzeichnet ob die zugehörige Studienteilnehmerin überhaupt schon mal vollständig schwanger war, ebenfalls wie eine Indikatorvariable für fehlende Werte behandelt. Da sie nur zur Adjustierung ins Modell aufgenommen wurde, kann sie folglich - unabhängig von ihrem P-Wert - nur zusammen mit der Variablen „Alter 1te Schwangerschaft“ aus dem Modell entfernt werden (vgl. Schritt 5).

3) Da für die beiden Variablen „Fehlgeburt“ und „Schwangerschaftsabbruch“ nur eine Indikatorvariable für fehlende Werte definiert werden konnte (siehe oben), kann diese erst mit der zweiten dieser Variablen aus dem Modell entfernt werden (vgl. Schritt 8).

Tabelle 5.3.33: Zwischenschritte der speziellen Rückwärts-Auswahl-Prozedur

Schritt	Entferne – zusammen mit der Indikatorvariablen	P-Wert der Variablen	P-Wert der Indikatorvariablen
1	Weiterführende Schule	0,8334	0,1988
2	Tumor	0,6323	0,8944
3	Verlust des Ehemannes	0,4666	0,4738
4	Menopause-Status	0,3758 (Minimum)	0,1216
5	Alter 1te Schwangerschaft + dazugehörige Indikatorvariable	0,3255 0,6029	0,1519 ---
6	Schwangerschaft	0,9897	0,9479
7	Tubensterilisation	0,2333	<0,0001
8	Fehlgeburt	0,2059	---
9	Einkommen	0,1316	0,5141

Nach 9 Schritten (vgl. Tabelle 5.3.33) verbleiben neben den beiden Lebensalter-Variablen noch 4 Variablenpaare im Modell, bei denen die Substanzvariablen jeweils einen P-Wert kleiner 0,1 aufweist. Die genauen Ergebnisse können Tabelle 5.3.34 entnommen werden. Neben dem Lebensalter sind 4 Merkmale aus statistischer Sicht von Bedeutung für das Brustkrebsrisiko.

Tabelle: 5.3.34: Finales Modell nach der speziellen Rückwärtsauswahl

Variable	Odds-Ratio	P_Wert
Lebensalter einfach	**1,229**	**<0,0001**
Lebensalter quadriert	**0,999**	**<0,0001**
Anzahl Lebendgeburten	**0,893**	**0,0006**
Schwangerschaftsabbruch	**1,520**	**0,0766**
Unregelmäßige Periode	**1,473**	**0,0401**
Späte Menarche	**1,380**	**0,0658**
I(Lebendgeburten)	0,241	0,0492
I(SS-Abbrüche)	2,032	0,0902
I(Unregelmäßige Periode)	1,473	<0,0001
I(Späte Menarche)	0,867	0,5306

Aus den Odds-Ratio-Schätzwerten lassen sich die folgenden Aussagen ableiten:
Mit jeder Lebendgeburt verringert sich die Erkrankungschance um den Faktor 0,893, so dass das Brustkrebsrisiko mit der Anzahl Lebendgeburten sinkt. Die späte Einsetzung der Menarche, Schwangerschaftsabbrüche und eine nichtregelmäßige Periode stellen hingegen offensichtlich Risikofaktoren dar. Jede dieser drei Expositionen hat unabhängig von den Anderen einen multiplikativen Effekt auf die Brustkrebs-Erkrankungschance. Um welchen Faktor sich die Erkrankungschance jeweils schätzungsweise erhöht, kann den Odds-Ratio-Schätzwerten in Tabelle 5.3.34 entnommen werden.

Eine Beurteilung, welchen substanzwissenschaftlichen Erklärungswert das Modell für die Entstehung der Brustkrebskrankheit hat, ist nicht möglich, da sich 4 Indikatorvariablen für fehlende Werte im Modell befinden, deren P-Werten entnommen werden kann, dass zumindest drei von ihnen einen studiendesignbedingten Erklärungswert für die Zielvariable haben. Die Güte der Modellanpassung ergibt sich folglich zu unbekannten Anteilen aus den substanzwissenschaftlich-relevanten (Substanzvariablen) und den inhaltlich-wertlosen (Indikatorvariablen) Erklärungswerten der Variablen und stellt somit kein adäquates Maß für den substanzwissenschaftlichen Wert des Modells dar.
Zudem kann nicht abgeschätzt werden, in welchem Ausmaß die Indikatorvariablen, deren Erklärungswerte nur ein Artefakt des Studiendesigns darstellen, zu Ergebnisverzerrungen hinsichtlich der Substanzvariablen führen. Daher ist ebenfalls nicht klar, ob die bedeutsamen Merkmale selektiert wurden und inwieweit die geschätzten Odds-Ratio-Werte die tatsächlichen Erkrankungschancenverhältnisse realistisch (bzw. unverzerrt) beschreiben.

Um trotzdem zumindest einen gewissen Eindruck davon zu gewinnen, ob das finale Modelle die bedeutsamen Variablen beinhaltet, werden weitere Modelle generiert. Die Generierung erfolgt durch Anwendung der speziellen Rückwärtsauswahl auf verschiedene Teilmengen der Variablen. Grundidee für diese Vorgehensweise ist die Folgende:
Wenn tatsächlich nur die 4 Merkmale das Erkrankungsrisiko beeinflussen, ist nach Ausschluss bedeutungsloser Merkmale aus dem Ausgangsmodell zu erwarten, dass die Anwendung der Prozedur trotzdem wieder zu obigem Finalmodell führt.

Wenngleich die durchgeführten Untersuchung zeigen, dass tatsächlich in den meisten Fällen dasselbe finale Modelle resultiert, und ansonsten zumindest immer eine weitgehende Überein-

stimmung beobachtet werden kann, deutet die folgende Untersuchung hingegen darauf hin, dass das finale Modelle möglicherweise doch nicht alle bedeutsamen Variablen beinhaltet.

Da davon auszugehen ist, dass sich für die Merkmale mit vielen fehlenden Werten am ehesten Indikatorvariablen ergeben, die in statistisch signifikanter Weise studiendesignbedingt mit Brustkrebs assoziiert sind, werden in einer weiterführenden Untersuchung alle Merkmale mit mehr als 200 fehlenden Werten a priori aus dem Ausgangsmodell ausgeschlossen. Neben den Merkmalen „Weiterführende Schule“, „Einkommen“ und „Tubensterilisation“ sind damit auch die beiden Merkmale „Unregelmäßige Periode“ und „Späte Menarche“ auszuschließen. Die spezielle Rückwärtsauswahl-Prozedur führt in diesem Fall zu dem in Tabelle 5.3.35 dargestellten Modell.

Tabelle 5.3.35: Ergebnisse der Prozedur nach Modifikation des Ausgangsmodells

Variable	Odds-Ratio	P-Wert
Lebensalter	1,237	<0,0001
Lebensalter quadriert	0,998	<0,0001
Lebendgeburten	0,870	0,0004
Schwangerschaftsabbruch	1,860	0,0096
Fehlgeburt	1,507	0,0489
Verlust des Ehemannes	1,589	0,0040
Alter 1te Schwangerschaft	1,033	0,0415
Indikator 1te Schwangerschaft	1,622	0,2788
Menopause-Status OR(4,3)	1,852	0,0012
I(Lebendgeburten)	0,345	0,1438
I(SS-Abbruch, Fehlgeburt)	1,349	0,4739
I(Verlust des Ehemannes)	0,554	0,1094
I(Alter 1te Schwangerschaft)	1,316	0,5821
I(Menopause-Status)	0,264	0,3666

Tabelle 5.3.35 kann zunächst entnommen werden, dass die beiden Merkmale „Lebendgeburten“ und „Schwangerschaftsabbruch“ wiederum im finalen Modell verbleiben und damit offensichtlich wirklich von Bedeutung für die Brustkrebs-Krankheit sind. Problematisch ist allerdings, dass das Modell darüber hinaus 4 neue Merkmale umfasst. Es stellt sich die Frage, wie zu interpretieren ist, dass diese 4 bei Berücksichtigung der Merkmale „Unregelmäßige

Periode“ und „Späte Menarche“ als bedeutungslos eingestuft wurden, nun aber von statistischer Signifikanz sind. Die folgenden Überlegungen zeigen, dass aus mathematischer Sicht keine Antwort auf diese Frage gegeben werden kann.

Die beiden Merkmale „Unregelmäßige Periode“ und „Späte Menarche“ sind in ihrer Parametrisierung durch jeweils eine Substanz- und eine Indikatorvariable im ursprünglichen Modell sowohl substanzwissenschaftlich als auch studiendesignbedingt mit der Brustkrebskrankheit assoziiert. Das heißt, sie sorgen für eine Adjustierung (Substanzvariablen) und zugleich zu einer Ergebnisverzerrung (Indikatorvariablen). Entsprechend kommen sowohl Adjustierung als auch Ergebnisverzerrung als Ursache für den Ausschluss dieser 4 Merkmale im ursprünglichen Modell in Betracht. Allerdings besteht keine Möglichkeit zu entscheiden, ob die Ursache des Ausschlusses die substanzvariablen-induzierte Adjustierung oder die indikatorvariablen-induzierte Verzerrung ist. Da der Ausschluss dieser Merkmale somit nicht eindeutig auf die Adjustierung zurückgeführt werden kann, muss bis auf weiteres davon ausgegangen werden, dass auch diese 4 Merkmale einen Einfluss auf das Brustkrebsrisiko haben.

Zusammenfassung der Ergebnisse:

Als Fazit dieses Unterkapitels lässt sich festhalten, dass die Methode, fehlende Werte durch Indikatorvariablen zu kennzeichnen, aufgrund der zugrundeliegenden fehlenden Werte Systematik im vorliegenden Fall nur bedingt geeignet ist. Die Untersuchungsergebnisse zeigen, dass neben dem Lebensalter ziemlich sicher davon ausgegangen werden kann, dass die Merkmale „Lebendgeburten“ und „Schwangerschaftsabbruch“ Einfluss auf das Brustkrebserkrankungsrisiko nehmen. Darüber hinaus ist anzunehmen, dass auch die Merkmale „Späte Menarche“ und „Unregelmäßige Periode“, welche beide enorm lückenhaft erhoben wurden, Risikofaktoren darstellen. Aufgrund der fehlenden Werte Problematik ist nicht entscheidbar, ob es sich bei den Merkmalen „Fehlgeburt“, „Verlust des Ehemannes“, „Späte erste Schwangerschaft“ und „Menopause-Status“ ebenfalls um Risikofaktoren für Brustkrebs handelt.

Angemerkt werden soll noch, dass nicht ausgeschlossen werden kann, dass zudem weitere Merkmale bedeutsam für die Brustkrebs-Krankheit sind. Wenngleich die Ergebnisse der durchgeführten Untersuchungen keinen Anhaltspunkt dafür geben, machten Umfang und Unvollständigkeit des Datenmaterials es erforderlich, eine Auswertungsstrategie zu verfolgen, die zwei nicht zu vernachlässigende Nachteile aufweist. Zunächst musste eine Variablenvorauswahl getroffen werden, so dass möglicherweise noch vor Betrachtung multipler Modelle bedeutsame Variablen aufgrund mangelhafter Adjustierung fälschlicherweise ausgeschlossen

wurden. Anschließend wurden in den multiplen Modellen fehlende Werte durch Indikatorvariablen gekennzeichnet. Eine Ad-hoc-Vorgehensweise, die möglicherweise zu deutlich verzerrten Schätzungen geführt haben könnte (vgl. 4.3.9).

5.3.6 Probability-Imputation für fehlende Variablenwerte

In diesem Unterkapitel wird die Datenauswertung alternativ unter Zuhilfenahme des (unbedingten) Probability-Imputation-Verfahrens vorgenommen. Das heißt, fehlende Werte der 13 selektierten Merkmale werden durch den Mittelwert der vorhandenen Merkmalswerte ersetzt. Das kategoriale Merkmal „Menopause-Status“ wird dabei durch 3 Dummy-Variablen parametrisiert und eine zusätzliche Indikatorvariable muss kennzeichnen, ob die betreffende Studienteilnehmerin schon eine vollständige Schwangerschaft hinter sich hat. Durch Letzteres wird wieder eine Adjustierung bzgl. der Variablen „Alter 1te Schwangerschaft“ erreicht. Zusammen mit den zwei Lebensalter-Variablen beinhaltet das Modell somit 18 Einflussvariablen. Die Ergebnisse der ML-Schätzung können Tabelle 5.3.36 entnommen werden.

Tabelle 5.3.36: Ergebnisse nach Probability-Imputation – Teil 1

Die 14 regulären Merkmale (parametrisiert durch 18 Variablen)…		
Merkmal bzw. Variable	**Odds-Ratio**	**P-Wert**
Lebensalter…	------	------
…einfach (stetig)	1,195	**<0,0001**
…quadriert (stetig)	0,999	**<0,0001**
Einkommen (stetig)	1,104	0,1607
Weiterführende Schule (binär)	1,063	0,7684
Schwangerschaft (binär)	0,737	0,4450
Anzahl Lebendgeburten (stetig)	0,870	**0,0004**
Unregelmäßige Periode (binär)	1,670	**0,0131**
Verlust des Ehemannes (binär)	1,791	**0,0004**
Späte Menarche (binär)	1,519	**0,0084**
1te vollständige Schwangerschaft…	------	------
…Alter 1te Schwangerschaft (stetig)	1,021	0,2132
…Vollständige Schwangerschaft (binär)	1,021	0,9693
Tumor (binär)	0,606	0,1004
Schnittentbindung (binär)	1,451	0,2033

Tabelle 5.3.36: Ergebnisse nach Probability-Imputation – Teil 2

Merkmal bzw. Variable	Odds-Ratio	P-Wert
Menopause… (kategorial)	------	------
…noch nicht erreicht (binär)	0,914	0,7631
… gerade erreicht (binär)	1,263	0,4923
… operativ erreicht (binär)	1,814	**0,0020**
Schwangerschaftsabbruch (binär)	1,916	**0,0088**
Fehlgeburt (binär)	1,982	**0,0028**

Unter Zuhilfenahme des klassischen Rückwärtsauswahlverfahrens (vgl. 5.3.5) können schrittweise die bedeutungslosen Variablen aus dem Modell entfernt werden. In jedem Schritt wird ein Merkmal entfernt und ein neues Modell an die verbleibenden Variablen angepasst. Ausgeschlossen wird jeweils das Merkmal, das den größten Wald-Test-P-Wert aufweist. Als Abbruchkriterium wird definiert, dass das Verfahren endet, wenn alle Variablen im Modell einen P-Wert kleiner 0,1 (vgl. 5.3.5) aufweisen.

Zu beachten ist, dass die Adjustierungsvariable für das Merkmal „Alter erste Schwangerschaft“ außer Konkurrenz läuft, das heißt nur zusammen mit der eigentlichen Substanzvariablen aus dem Modell entfernt werden kann, und dass auch die drei „Menopause-Status“-Variablen nur gemeinsam ausgeschlossen werden können. Als Vergleichswert für das Ausschlusskriterium dient hier wieder der kleinste der drei Wald-Test-P-Werte. Tabelle 5.3.37 zeigt für jeden Schritt, welches Merkmal ausgeschlossen wird. Der dritten Spalte kann entnommen werden, mit welchen P-Werten die jeweiligen Variablen ausgeschlossen werden. In den einzelnen Schritten der Rückwärtsauswahl sind keine Auffälligkeiten, das heißt nennenswerte Veränderungen der Odds-Ratio-Schätzwerte der verbleibenden Variablen, zu beobachten.

Tabelle 5.3.37: Zwischenschritte der klassischen Rückwärtsauswahl-Prozedur

Schritt	ausgeschlossene Variable	P-Wert der Variablen (im aktuellen Modell)
1	Weiterführende Schule	0,7684
2	Schwangerschaft	0,4413
3	Alter 1te Schwangerschaft + dazugehörige Indikatorvariable	0,1778 0,5411
4	Tubensterilisation	0,1612
5	Tumor	0,1012

Bereits nach dem fünften Schritt liegen alle P-Werte unter 0,1, so dass das Verfahren endet.

Das finale Modell kann Tabelle 5.3.38 entnommen werden. Die Ergebnisse zeigen, dass bei Anwendung des Probability-Imputation-Verfahrens neben dem Lebensalter noch 8 weitere Merkmale im Modell verbleiben. Diese weisen (sogar) alle P-Werte kleiner 0,05 auf und haben somit - zumindest nach Mittelwert-Auffüllung der fehlenden Werte - einen statistisch-signifikanten Erklärungswert für die Brustkrebskrankheit.
Unklar bleibt allerdings, inwieweit sich die Mittelwert-Auffüllung fehlender Werte bei den einzelnen Variablen auf ihren Erklärungswert auswirkt. Nicht ausgeschlossen werden kann, dass die geschätzten Beziehungen zwischen den Variablen und der Erkrankungschance zu einem gewissen Grade auch durch die Auffüllung der fehlenden Werte verursacht ist, und somit nicht ausschließlich auf empirischen Beobachtungen gründet.

Tabelle 5.3.38: Finales Modell nach Rückwärtsselektion (bei Probability-Imputation)

Variable	**Odds-Ratio**	**P-Wert**
Lebensalter einfache (stetig)	1,207	**<0,0001**
Lebensalter quadriert (stetig)	0,999	**<0,0001**
Einkommen (stetig)	1,153	**0,0302**
Anzahl Lebendgeburten (stetig)	0,881	**<0,0001**
Unregelmäßige Periode (binär)	1,604	**0,0214**
Verlust des Ehemannes (binär)	1,836	**0,0002**
Späte Menarche (binär)	1,633	**0,0052**
Menopause… (kategorial)	------	------
…noch nicht erreicht (binär)	1,032	0,9137
…Menopause gerade erreicht (binär)	1,372	0,3446
…Menopause operativ erreicht (binär)	1,771	**0,0029**
Schwangerschaftsabbruch (binär)	1,979	**0,0045**
Fehlgeburt (binär)	1,813	**0,0046**

Davon ausgehend, dass das Probability-Imputation-Verfahren im vorliegenden Fall zu vernachlässigbar geringen Verzerrungen führt, lassen sich die Ergebnisse wie folgt zusammenfassen.

Genau wie bei dem Modell, in welchem fehlende Werte durch Indikatorvariablen gekennzeichnet wurden, sinkt das Brustkrebsrisiko mit der Anzahl Lebendgeburten und die drei Merkmale „Unregelmäßige Periode“, „Späte Menarche“ und „Schwangerschaftsabbruch“ stellen Risikofaktoren für Brustkrebs dar. Welche unabhängigen multiplikativen Effekte auf die Erkrankungschance diesen Merkmalen (bei Anwendung des Probability-Imputation-Verfahrens) zukommen, kann den Odds-Ratio-Schätzungen in Tabelle 5.3.37 entnommen werden.
Darüber hinaus verbleiben aber auch die vier Merkmale „Einkommensklasse“, „Verlust des Ehemannes“, „Fehlgeburt“ und „Menopause-Status“ im Modell. Die dazugehörigen Odds-Ratios zeigen, dass das Risiko mit dem jährlichen Einkommen steigt und, dass eine operativ herbeigeführte Menopause zu einer Erhöhung des Risikos führt. Ebenfalls stellen Fehlgeburten und der Verlust des Ehemannes im Modell Risikofaktoren für Brustkrebs dar.

5.3.7 Vergleich der Modelle

Abschließend sollen die Ergebnisse der beiden letzten Unterkapitel (5.3.5 und 5.3.6) miteinander verglichen werden. Zu beachten ist, dass sowohl das Indikatorvariablen-Verfahren als auch das Probability-Imputation-Verfahren mit Ergebnisverzerrungen verbunden sein kann. Beim Probability-Imputation-Verfahren werden zwar keine Indikatorvariablen für fehlende Werte ins Modell integriert, so dass keine offensichtlichen Verzerrungen durch eigenständige Fehlende-Werte-Effekte auftreten können, dafür aber kann die Mittelwert-Auffüllung fehlender Werte eine verzerrende Wirkung auf Parameterschätzungen haben. Entsprechend schwierig gestalten sich Interpretation und Vergleich der Ergebnisse.

Tabelle 5.3.38 stellt die Ergebnisse der Vorauswahl den Odds-Ratio-Schätzwerten, der drei ausführlicher diskutierten Modelle, gegenüber. Bei der Interpretation der Ergebnisse muss berücksichtigt werden, dass die Odds-Ratio-Schätzwerte von Modell zu Modell variieren und darüber hinaus in allen drei Fällen verzerrt sein können. Entsprechend ist es nicht möglich, die Bedeutungen der Merkmale für das Brustkrebsrisiko zuverlässig zu quantifizieren. Tabelle 5.3.38 kann allerdings entnommen werden, dass zumindest keine Widersprüche vorliegen.

Tabelle 5.3.38: Vergleich der OR-Schätzungen in den drei Modellen

Variable	Vor-Auswahl	Unterkapitel 5.3.3.2		5.3.3.3
		Modell 1	Modell 2	Modell 3
Lebensalter einfach	-----	**1,229**	**1,237**	**1,207**
Lebensalter quadriert	-----	**0,999**	**0,998**	**0,999**
Anzahl Lebendgeburten	**0,843**	**0,893**	**0,870**	**0,881**
Unregelmäßige Periode	1,543	1,473	-----	1,604
Späte Menarche	1,636	1,380	-----	1,633
Schwangerschaftsabbruch	**2,585**	**1,520**	**1,860**	**1,979**
Fehlgeburten	2,170	-----	1,507	1,813
Menopause-Status OR(4,3)	2,269	-----	1,852	1,771
Verlust des Ehemannes	1,833	-----	1,589	1,836
Einkommen	1,123	-----	-----	1,153
Alter 1te Schwangerschaft	1,039	-----	1,033	-----
+ Indikator (Alter 1te SS)	-----	-----	1,622	-----

Der Fall, dass ein Merkmal in einem Modell die Erkrankungschance erhöht und in einem anderen die Erkrankungschance verringert, tritt nicht auf. Jedes Merkmal präsentiert sich bei der Variablen-Vorauswahl und in den Modellen, in denen es von statistischer Signifikanz für die Erkrankungschance ist, stets einheitlich entweder als Risikofaktor oder als Schutzfaktor.

Neben dem Lebensalter kommt lediglich den beiden Merkmalen „Anzahl Lebendgeburten“ und „Schwangerschaftsabbruch“ in allen drei Modellen ein signifikanter Erklärungswert für die Erkrankungschance zu. Den dazugehörigen Odds-Ratio-Schätzwerten kann entnommen werden, dass sich das Brustkrebsrisiko mit jeder Lebendgeburt in etwa um den Faktor 0,88 verringert und, dass ein Schwangerschaftsabbruch in Abhängigkeit vom Modell zu einer Ver-1,5- bis Ver-2-fachung der Erkrankungschance führt.

Darüber hinaus ist davon auszugehen, dass auch die Merkmale „Späte Menarche“ und „Unregelmäßige Periode“ von Bedeutung sind. Während sie aus dem zweiten Modell, in Anbetracht der Tatsache, dass sie beide nur sehr lückenhaft erhoben wurden, a priori ausgeschlossen wurden, stellen sie im ersten und dritten Modell Risikofaktoren für Brustkrebs dar. Die Schätzwerte deuten darauf hin, dass sowohl eine unregelmäßige Periode als auch ein später Eintritt der Menarche (nach dem 12ten Lebensjahr) jeweils ungefähr mit einer Ver-1,5-Fachung der Erkrankungschance einhergehen.

Weniger offensichtlich ist, ob die vier Merkmale „Fehlgeburt“, „Menopause-Status“, „Verlust des Ehemannes“ und „Alter 1te Schwangerschaft“ bedeutsame Risikofaktoren für Brustkrebs darstellen oder nicht. Bei Anwendung des Indikatorvariablen-Verfahrens ergaben sich für diese vier Variablen erst nach Ausschluss aller Merkmalen mit mehr als 200 fehlenden Einträgen (Modell 2) signifikante P-Werte. Da in Modell 2 somit auch zwei signifikante Merkmale („Unregelmäßige Periode“ und „Späte Menarche“) inklusive ihrer Indikatorvariablen ausgeschlossen wurden, stellte sich im letzten Unterkapitel die Frage, wie dieses Ergebnis zu bewerten ist. Die Überlegungen zeigten, dass als Ursache für die Nicht-Signifikanz im ersten Modell sowohl eine Indikatorvariablen-induzierte Verzerrung als auch eine Substanzvariablen-induzierte Adjustierung denkbar ist. Erst die Ergebnisse, die sich bei Anwendung der Probability-Imputation-Methode ergeben, ermöglichen diesbezüglich differenziertere Aussagen. Obwohl auch die Anwendung der Probability-Imputation-Methode mit Ergebnisverzerrungen einhergehen kann, kommen den fehlenden Werten hier zumindest keine eigenständigen Effekte in Form von Indikatorvariablen zu. Bei jeder Variablen führt das Auffüllen fehlender Werte mit dem Mittelwert der vorhandenen Variablenwerte dazu, dass diese zu den regulären Beobachtungen in Beziehung gesetzt werden. Entsprechend ist davon auszugehen, dass fehlende Werte eines Merkmals primär das zu diesem Merkmal gehörige Odds-Ratio verzerren und nur sekundär, das heißt bedingt durch die Primärverzerrung, zu Ergebnisverzerrungen bezüglich der Parameter anderer Variablen führen.

Tabelle 5.3.38 kann entnommen werden, dass die Odds-Ratio-Schätzwerte bzgl. der beiden lückenhaften Merkmale „Unregelmäßige Periode“ und „Späte Menarche“ in den Modellen 1 und 3 nur unwesentlich voneinander abweichen (1,543 und 1,604 bzw. 1,636 und 1,633). Dieser Vergleich zeigt, dass das Probability-Imputation-Verfahren im vorliegenden Fall hinsichtlich der beiden Variablen „Unregelmäßige Periode“ und „Späte Menarche“ zu keiner größeren Verzerrung führt als das Indikatorvariablen-Verfahren. Das heißt, dass zwischen den regulären Beobachtungen dieser Merkmale in beiden Modellen (bzw. bei beiden Verfahren) dieselbe Beziehung vorliegt. Das bedeutet, dass es in den Modellen 1 und 3 gleichermaßen zu einer Adjustierung hinsichtlich der regulären Beobachtungen der Merkmale „Unregelmäßige Periode“ und „Späte Menarche“ kommt. Folglich gilt für jedes Merkmal, welchem in Modell 3 eine signifikante Bedeutung zukommt, dass es nicht an der Adjustierung bezüglich der Merkmale „Unregelmäßige Periode“ und „Späte Menarche“ liegen kann, wenn ihm in Modell 1 keine Signifikanz zukommt.

Da für die drei Merkmale „Menopause-Status“, „Fehlgeburt“ und „Verlust des Ehemannes“ genau diese Situation vorliegt, scheidet als Ursache für ihre Nichtsignifikanz in Modell 1 eine

Adjustierung bzgl. der Substanzvariablen der Merkmale „Unregelmäßige Periode“ und „Späte Menarche“ aus. Als Ursache für ihre Nichtsignifikanz ist also von einer Indikatorvariablen-induzierten Verzerrung auszugehen.

Unter Berücksichtigung, dass diesen drei Merkmalen zudem auch bei Anwendung des Indikatorvariablen-Verfahrens eine signifikante Bedeutung zukommt, wenn die besonders lückenhaften Merkmale ausgeschlossen werden (Modell 2), liegt der Schluss nahe, dass sie in der Tat Risikofaktoren für Brustkrebs darstellen.

Die Schätzwerte in Tabelle 5.3.38 zeigen, dass Frauen, bei denen auf operativem Wege die Menopause herbeigeführt wurde gegenüber Frauen, bei denen die Menopause auf biologischen Wege eingetreten ist, eine circa 1,8-mal so große Brustkrebs-Erkrankungschance aufweisen. Das Erleiden einer Fehlgeburt und der Verlust des Ehemannes durch Scheidung oder Versterben führen in Abhängigkeit vom Modell jeweils ungefähr zu einer Ver-1,5- bis Ver-1,85-Fachung der Erkrankungschance.

Dem Merkmal „Alter 1te Schwangerschaft“ kommt hingegen nur in Modell 2, das heißt insbesondere nach Ausschluss der Merkmale „Unregelmäßige Periode“ und „Späte Menarche“, eine signifikante Bedeutung zu. Es ist deswegen davon auszugehen, dass es nicht von unmittelbarer Bedeutung für das Brustkrebsrisiko ist. Der Umstand, dass ausschließlich in Modell 2 das Risiko mit dem Alter der ersten Schwangerschaft steigt, kann vermutlich auf die nichtvorhandene Adjustierung bezüglich der Merkmale „Späte Menarche“ und „Unregelmäßige Periode“ zurückgeführt werden und stellt somit – mutmaßlich - nur eine Scheinassoziation dar.

Das Merkmal „Einkommensklasse“, welches aufgrund der Unvollständigkeit aus Modell 2 ausgeschlossen wurde, ist nur in Modell 3, das heißt bei Anwendung des Probability-Imputation-Verfahrens, von signifikanter Bedeutung für die Erkrankungschance. Wird dieses Merkmal, parametrisiert durch eine Substanzvariable und eine Indikatorvariable für fehlende Werte, nachträglich in Modell 2 aufgenommen, zeigt sich, dass der Substanzvariablen auch hier keine signifikante Bedeutung zukommt (P-Wert: 0,3207). Die Einkommensklasse ist also auch nach Ausschluss aller anderen, unvollständig erhobenen Merkmale bei Anwendung des Indikatorvariablen-Verfahrens nicht von signifikanter Bedeutung. Da diesem Merkmal allerdings in Modell 3 (Probability-Imputation-Verfahren) eine signifikante Bedeutung zukommt, kann nicht ausgeschlossen werden, dass es dennoch von Bedeutung für das Brustkrebsrisiko ist. Nach den Ergebnissen der ML-Schätzung in Modell 3 steigt die Erkrankungschance mit der (ordinalen) Einkommensklasse.

Fazit

Die durchgeführten Untersuchungen zeigen, dass die Unvollständigkeit des Datenmaterials ein erhebliches Problem bei der Datenauswertung darstellt. Nur bei Miteinbeziehung fehlender Werte unter Zuhilfenahme zweier spezieller Ad-hoc-Verfahren konnten Ergebnisse erzielt werden. Da die Anwendung dieser Verfahren zu Ergebnisverzerrungen führen kann, wird im nächsten Unterkapitel versucht, die Fehlende-Werte-Problematik zu umgehen.

Auf Grundlage des Merkmals „Menopause-Status“ wird die Studienpopulation dazu zunächst in zwei Untergruppen unterteilt. Für beide Subpopulationen wird anschließend jeweils ein logistisches Regressionsmodell generiert, welches ausschließlich solche Einflussvariablen umfasst, die in der jeweiligen Subpopulation nur wenige fehlende Werte aufweisen. Folglich kommt es in diesen Modellen bei Ausschluss der Studienteilnehmerinnen, von denen nicht alle benötigten Merkmalswerte vorhanden sind, nur zu geringen Informationsverlusten. Der einzige wesentliche Nachteil bei dieser Vorgehensweise besteht darin, dass die Merkmale mit vielen fehlenden Werte a priori ausgeschlossen werden müssen und deshalb als nicht-kontrollierte Störgrößen anzusehen sind. Entsprechend stellt in diesem Fall nicht der, den Daten zugrunde liegende „Fehlende-Werte-Mechanismus“, sondern eine möglicherweise unzureichende Adjustierung, die Hauptgefahr für Ergebnisverzerrungen dar.

5.3.8 Modelle für Frauen vor und nach der Menopause

Auf Grundlage des Merkmals „Menopause-Status“ kann die Studienpopulation in zwei Subpopulationen unterteilt werden. Alle Frauen, bei denen die Menopause bereits eingetreten ist, bilden die erste Subpopulation, welche im Folgenden mit „Post-Menopause“ bezeichnet wird. Diese Population umfasst die 614 Studienteilnehmerinnen, bei denen die Menopause auf biologischem oder operativem Wege eingetreten ist. Die zweiten Subpopulation „Prae-Menopause“ wird von den Frauen gebildet, die die Menopause noch nicht erreicht haben oder sich in den Wechseljahren befinden. Tabelle 5.3.39 gibt einen Überblick über die Brustkrebs-Verteilung unterteilt nach dem Menopause-Status und in den beiden Subpopulationen. Der Tabelle kann entnommen werden, dass die relative Häufigkeit von BCA-Fällen in der Subpopulationen Prae-Menopause mit 0,218 deutlich unter der Häufigkeit 0,370 in der Subpopulation Post-Menopause liegt. Dies zeigt, dass eine Unterteilung der Studienpopulation auf Grundlage dieses Merkmals durchaus sinnvoll ist. Es stellt sich insbesondere die Frage, ob sich die Risikofaktoren für Brustkrebs in beiden Subpopulationen unterscheiden.

Tabelle 5.3.39: BCA-Verteilung nach Menopause-Status und in den Populationen

Menopause-Status: / **Kenngröße:**	**Noch nicht**	**Gegen-wärtig**	**Biologisch erreicht**	**Operativ erreicht**
Anzahl Frauen	323	53	356	258
Anzahl BCA-Erkrankungen	61	21	107	120
Relativer BCA-Anteil	0,189	0,396	0,301	0,465
Subpopulation:	**Prae-Menopause**		**Post-Menopause**	
Frauen	**376**		**614**	
BCA-Erkrankungen	**82**		**227**	
Relativer Anteil	**0,218**		**0,370**	

Der erste Teil der Tabelle zeigt zudem, dass sich die relativen Häufigkeiten von BCA-Fällen innerhalb beider Subpopulationen in Abhängigkeit vom genauen Menopause-Status unterscheiden. Dieser stellt in beiden Subpopulationen nunmehr ein binäres Merkmal dar. Dementsprechend sollte in beiden Subpopulationen neben dem Lebensalter auch der Menopause-Status als Kovariable dienen. Die 145 Studienteilnehmerinnen , von denen die Information über ihren Menopause-Status nicht verfügbar ist, müssen von den weiteren Analysen ausgeschlossen werden.

5.3.8.1 Prae-Menopause-Modell

Als erstes wird nun analog zur Variablen-Vorauswahl in Unterkapitel 5.3.2 untersucht, welche Merkmale in den beiden Subpopulationen als Expositionen für komplexere Modelle in Frage kommen. Dazu werden logistische Regressionsmodelle betrachtet, in denen neben dem zu untersuchenden Merkmal auch die beiden stetigen Lebensalter-Variablen sowie der binäre Menopause-Status als Kovariablen berücksichtigt werden. Ausgeschlossen werden die Merkmale, bei denen zu viele fehlende Werte vorliegen (>20%) oder, deren Wald-Test P-Wert im Rahmen des Modells nicht zum Niveau 15% signifikant ist. Auch hier wird im Rahmen der Vorauswahl überprüft, ob sich aus nicht-signifikanten Merkmalen in geeigneter Weise neue Variablen generieren lassen, die in obigem Modell von Signifikanz sind.

Tabelle 5.3.40 zeigt, dass bei diesen Ausschlusskriterien nur 4 Merkmale für die Subpopulation Prae-Menopause im Rahmen multipler Modelle untersucht werden können. Diese werden im Folgenden kurz beschrieben.

1) Späte Menarche – Men

Die binäre Variable „Späte Menarche“ gibt an, ob bei der betreffenden Frau die Menarche nach dem 13ten Lebensjahr eingetreten ist (Men=1) oder nicht (Men=0).

2) Probleme – Pro

Die binäre Variable „Probleme“ gibt für jede Studienteilnehmerin an, ob sie schon einmal eine Fehlgeburt oder einen Schwangerschaftsabbruch erlitten hat (Pro=1) oder nicht (Pro=0).

3) Alter erste Schwangerschaft – A1te

Die stetige Variable „Alter erste Schwangerschaft“ gibt an, in welchem Lebensalter die Frau zum ersten Mal eine vollständige Schwangerschaft, das heißt eine Schwangerschaft, die zu einer Lebend- oder Totgeburt führte, durchlebt hat.

4) Keine Geburt – KGeb

Die binäre Variable „Keine Geburt“ kennzeichnet die Frauen, die noch kein lebendiges Kind zur Welt gebracht haben und noch keine Totgeburt erlitten haben (KGeb=1).

Tabelle 5.3.40: Ergebnisse der Variablen-Vorauswahl für die 314 Frauen der Prae-Menopause

Merkmal	Fehlende Werte	Odds-Ratio	Wald-Test-P-Wert
Späte Menarche	71	1,838	0,0360
Probleme	11	2,103	0,0109
Alter 1te SS	15	1,054	0,0070
Keine Geburt	15	0,331	0,0103

Bemerkenswert ist, dass die Ergebnisse der Vorauswahl-Untersuchung darauf deuten, dass Frauen, die noch keine vollständige Schwangerschaft hinter sich haben, deutlich weniger brustkrebsgefährdet sind. Dies bestätigt auch das Ergebnis bei der Variablen „Alter erste Schwangerschaft“. Der OR-Schätzwert zeigt, dass sich auch hier das Merkmal „Keine Geburt“ indirekt als Risikofaktor präsentiert. Für zwei gleichaltrige Frauen mit demselben Menopause-Status, die sich nur dahingehend unterscheiden, dass die Eine noch keine vollständige Schwangerschaft hinter sich gebracht hat (A1te=0), wohingegen die Andere (zum Beispiel) im Alter von 20 Jahren zum ersten Mal schwanger wurde (A1te=20), gilt in diesem Modell schätzungsweise die folgende Odds-Ratio-Beziehung:

$$\text{OR}(\text{„Noch keine Geburt“}, \text{„Erste Geburt im Alter von 20 Jahren“}) = 1{,}054^{-20} = 0{,}349.$$

Speziell bei gleichzeitiger Betrachtung der beiden Variablen „Alter erste Schwangerschaft“ und „Keine Geburt“, weist letztere einem P-Wert von 0,765 auf, was bestätigt, dass die Variable „Keine Geburt“ nicht zur Adjustierung bezüglich des Merkmals „Alter erste Schwangerschaft“ benötigt wird.

Generierung eines multiplen Modells für die Frauen der Prae-Menopause

Im Modell für die Frauen der „Prae-Menopause“ wird die Variable „Späte Menarche“ zunächst nicht berücksichtigt, da diese mit Abstand die meisten fehlenden Werte (71 Stück) aufweist und ihre Berücksichtigung somit zu einem großen Informationsverlust (Ausschluss von Studienteilnehmerinnen) führt. Werden neben den Kovariablen (Lebensalter, quadriertes Lebensalter und binärer Menopause-Status) lediglich die vier anderen selektierten Variablen berücksichtigt, müssen insgesamt nur 21 Frauen aufgrund unvollständiger Variableneinträge ausgeschlossen werden. Der damit verbundene Informationsverlust erscheint vernachlässigbar gering.

Die klassische Rückwärtsauswahl-Prozedur führt hier nacheinander zum Ausschluss der Variablen „Keine Geburt“ (P-Wert: 0,845) und „Menopause-Status“ (P-Wert: 0,341), bis im dritten Schritt alle Modellvariablen (sogar) zum Niveau 5% von Bedeutung für die Erkrankungschance sind. Die resultierenden Schätzwerte und P-Werte können Tabelle 5.3.41 entnommen werden.

Tabelle 5.3.41: Ergebnisse der Rückwärtsauswahl (Prae-Menopause-Modell)

Variable	Odds-Ratio	P-Wert
Alter: Lebensalter	1,331	0,0015
Alter^2: Lebensalter quadriert	0,998	0,0090
A1TE: Alter erste Schwangerschaft	1,053	0,0011
PRO: Probleme	1,827	0,0441

Ausgehend von diesem Modell stellt sich als nächstes die Frage, wie mit der Variablen „Späte Menarche“ verfahren werden soll. Während diese Variable bei Nichtberücksichtigung als nichtkontrollierte Störgröße anzusehen ist, setzt ihre Berücksichtigung wieder die Miteinbeziehung von Studienteilnehmerinnen mit fehlenden Werten voraus.

Tabelle 5.3.42 zeigt die Ergebnisse der Parameterschätzung von drei Modellen, in denen auf unterschiedliche Weise das Merkmal „Späte Menarche“ integriert wurde. Die erste Spalte korrespondiert zunächst zu dem Modell, in welchem Frauen mit unvollständigen Variablen-

werten von der Analyse ausgeschlossen wurden. Die folgenden zwei Spalten zeigen, welche Schätzergebnisse sich bei Anwendung des Indikatorvariablen- bzw. Probability-Imputation-Verfahrens für fehlende Werte der Variablen „Späte Menarche“ ergeben.

Tabelle 5.3.42: „Späte Menarche“ als zusätzliche Einflussvariable

Modell:	**Ausschluss, falls MEN-Eintrag fehlt**		**Indikatorvariablen-Verfahren**		**Probability-Imputation-Verfahren**	
Variable	**OR**	**P-Wert**	**OR**	**P-Wert**	**OR**	**P-Wert**
Alter	1,295	0,0075	1,301	0,0023	1,312	0,0015
Alter2	0,998	0,0357	0,998	0,0143	0,998	0,0087
A1TE	1,046	0,0081	1,052	0,0015	1,051	0,0016
PRO	1,663	0,1107	1,727	0,0716	1,823	0,0462
MEN	**1,669**	**0,0771**	**1,691**	**0,0807**	**1,562**	**0,1391**
I(MEN=.)	-----	-----	0,774	0,4840	-----	-----

Die Ergebnisse der ML-Schätzung zeigen, dass das Merkmal „Späte Menarche“ in keinem der drei Modelle einen Wald-Test-P-Wert kleiner 0,05 aufweist. Darüber hinaus weist dieses Merkmal im zweiten und dritten Modell sogar den größten Wald-Test-P-Wert auf.
Da die Nichtberücksichtigung der 71 Frauen mit fehlendem Wert für die Variable „Späte Menarche“ (Modell 1) mit einem zu großen Informationsverlust verbunden ist, muss bei Berücksichtigung des Merkmals ein Ad-hoc-Verfahren im Umgang mit fehlenden Werten angewendet werden. Von Bedeutung ist dabei, dass sowohl im zweiten als auch im dritten Modell die Anwendung des Rückwärtsauswahlverfahrens (zum Niveau 0,05) zum Ausschluss der Variablen „Späte Menarche“ und damit wieder zum Ausgangsmodell führt (Tabelle 5.3.41).
Entsprechend ist davon auszugehen, dass das Ausgangsmodell, welches weder einem Informationsverlust noch einer Verzerrung durch die Miteinbeziehung von fehlenden Werten unterliegt, für die Daten am besten geeignet ist. Lediglich auf eine Adjustierung hinsichtlich des Merkmals „Späte Menarche“ muss verzichtet werden, was aber in Anbetracht der P-Werte dieses Merkmals (>0,05) durchaus gerechtfertigt werden kann.

Da keine der drei Produktvariablen (MEN·Lebensalter, PRO·Lebensalter bzw. MEN·PRO) bei ihrer Aufnahme ins Modell (Tabelle 5.3.31) einen Wald-Test-P-Wert kleiner 0,2 aufweist, kann davon ausgegangen werden, dass keine (Zweifach-)Wechselwirkungen zwischen den

drei Faktoren „Lebensalter“, „Alter erste Schwangerschaft“ und „Schwangerschaftsprobleme“ vorliegen.

Im Folgenden werden die Schätzergebnisse des Prae-Menopause-Modells (Tabelle 5.3.31) inhaltlich gedeutet. Offensichtlich variiert die Brustkrebserkrankungschance für Frauen der Prae-Menopause nicht nur mit dem Lebensalter, sondern wird darüber hinaus von zwei Faktoren beeinflusst. Zunächst steigt die Erkrankungschance mit dem Lebensjahr zum Zeitpunkt der ersten Geburt an, so dass (insbesondere) eine (späte erste) Schwangerschaft als Risikofaktor für eine frühe Brustkrebserkrankung, das heißt eine Erkrankung noch vor Eintreten der Menopause, anzusehen ist. Darüber hinaus sind Frauen bereits vor Eintreten der Menopause brustkrebsgefährdet, wenn sie eine Fehlgeburt erlitten haben oder, wenn bei ihnen eine Schwangerschaft abgebrochen wurde. Trift (mindestens) eine der beiden Aussagen auf eine Frau zu, erhöht sich ihre Erkrankungschance schätzungsweise um den Faktor 1,8. Die 6 möglichen Wechselwirkungen

Bereits der Ausschluss von 21 Studienteilnehmerinnen führt dazu, dass keine Manipulationen an fehlenden Werten mehr vorgenommen werden müssen, so dass das Modell ausschließlich auf substanzwissenschaftlich-relevanten Beziehungen beruht. Entsprechend kann die Modellanpassung mit Hilfe, der in Unterkapitel 4.3.6 beschriebenen, Methoden überprüft werden.

Überprüfung der Modellanpassung

Zunächst wird der Anpassungstest von Hosmer und Lemeshow (vgl. 4.3.6) durchgeführt. Dazu werden die 355 Frauen der Prae-Menopause auf Grundlage der Dezile ihrer geschätzten Erkrankungswahrscheinlichkeiten in 10 Risikogruppen eingeteilt. Tabelle 5.3.43 zeigt für jede der 10 Risikogruppen, welche Anzahl von BCA-Fällen nach dem geschätzten Modell zu erwarten ist und, welche Anzahl tatsächlich vorliegt. Die größte Absolutabweichung von 2,28 liegt in der achten Risikogruppe vor.

In den 9 anderen Risikogruppen sind die Abweichungen deutlich geringer. Eine bestimmte Struktur ist folglich nicht zu erkennen, so dass von einer relativ gleichmäßigen Anpassung gesprochen werden kann.

Tabelle 5.3.43: Tafel zum Anpassungstest von Hosmer & Lemeshow

Risiko-gruppe	Gruppen-größe	Erwartete An-zahl BCA-Fälle	Beobachtete An-zahl BCA-Fälle
1	37	0.86	1
2	38	2.11	1
3	36	3.47	5
4	36	5.10	4
5	36	6.49	7
6	36	8.42	8
7	36	11.01	11
8	36	12.72	15
9	36	15.01	14
10	28	14.81	14

Für die, im Falle einer guten Modellanpassung, mit 8 Freiheitsgraden chi-quadrat-verteilte Teststatistik ergibt sich ein Wert von 2.590, was einem P-Wert von 0.957 entspricht. Folglich deutet das Ergebnis des Anpassungstests von Hosmer und Lemeshow nicht darauf hin, dass das Modell die Daten nicht gut beschreibt.

Ebenfalls soll die Modellanpassung unter Zuhilfenahme von Klassifikationstafeln beurteilt werden. Für jede der 355 Studienteilnehmerinnen wird daher auf Grundlage ihrer Erkrankungswahrscheinlichkeit prognostiziert, ob sie an Brustkrebs leidet oder nicht. Zunächst wird ein Diskriminanzwert von 0,5 gewählt. Das heißt, ausschließlich für die Frauen, deren Erkrankungswahrscheinlichkeit im geschätzten Modell über 0,5 liegt, wird eine BCA-Erkrankung prognostiziert. Es resultiert, die im Folgenden dargestellte Klassifikationstafel (Tabelle 5.3.44).

Tabelle 5.3.44: Klassifikationstafel - Diskriminanzwert p=0,5

BCA-Status: Vorhersage:	BCA=0	BCA=1	Gesamt
BCA=0	267	70	337
BCA=1	8	10	18
Gesamt:	275	80	355

Den Zellhäufigkeiten kann entnommen werden, dass nur für 8 der 275 nicht BCA-erkrankten Frauen (Spalte 1) eine falsche Prognose erstellt wird. Auf der anderen Seite führt diese Entscheidungsregel dazu, dass 70 der 80 BCA-Fälle (Spalte 2) als nicht BCA-erkrankt und damit falsch eingestuft werden. Dies entspricht einer Spezifität von 97,1% und einer Sensitivität von 12,5%.
Der Anteil richtiger Vorhersagen von 78% ist somit fast ausschließlich darauf zurückzuführen, dass für die nicht BCA-erkrankten Frauen richtige Prognosen erstellt werden. Das bedeutet, dass das Klassifikationsverfahren ungeeignet ist, BCA-Erkrankungen richtig zu prognostizieren.

Wird als Diskriminanzwert der relative Anteil von BCA-Fällen verwendet, ergibt sich die folgende Klassifikationstafel (Tabelle 5.3.44).

Tabelle 5.3.44: Klassifikationstafel - Diskriminanzwert p=80/355=0,225

BCA-Status: **Vorhersage:**	**BCA=0**	**BCA=1**	**Gesamt**
BCA=0	178	23	201
BCA=1	97	57	154
Gesamt:	275	80	355

Wenngleich der Anteil korrekter Vorhersagen hier nur 66,2% beträgt, ergeben sich für die Sensitivität und Spezifität Werte von 71,3% bzw. 64,7%. Da somit auf Grundlage dieser Entscheidungsregel sowohl für die erkrankten als auch für die nichterkrankten Frauen mehr richtige als falsche Prognosen erstellt werden, ist sie definitiv besser geeignet als die Erste. Insbesondere, da Sensitivität und Spezifität relativ nahe beieinander liegen, kann die Eignung des Verfahrens anhand des Anteils korrekter Vorhersagen von 66,2% beurteilt werden.
Als Ergebnis der Klassifikationsanalyse lässt sich also festhalten, dass unter Zuhilfenahme des geschätzten Modells die BCA-Zustände von 66,2% der Studienteilnehmerinnen richtig prognostiziert werden konnten. Berücksichtigt man weiter, dass die Tumorentstehung ein komplexes Geschehen darstellt (vgl. Grundmann 1994), an dem auch viele nicht erfassbare Faktoren beteiligt sind, und, dass das geschätzte Modell ausschließlich auf 3 Informationen (Lebensalter, Alter 1te Schwangerschaft und Schwangerschaftsprobleme) beruht, kann von einer verhältnismäßig guten „Vorhersagekraft des Modells für die Studienpopulation“ (siehe unten) und damit von einer guten Modellanpassung gesprochen werden.

Unbedingt angemerkt werden muss, dass sich die Angaben zu Erkrankungswahrscheinlichkeiten ausschließlich auf die Studienpopulation beziehen (vgl. 4.3.9). Da die BCA-Häufigkeit im Datensatz deutlich über der wahren Häufigkeit von Brustkrebs in der Bevölkerung liegt, können das Auftreten eines Schwangerschaftsproblems (Fehlgeburt oder Schwangerschaftsabbruch) und/oder eine späte erste Schwangerschaft zwar als Risikofaktoren für Brustkrebs angesehen werden, ermöglichen allerdings bezogen auf die Gesamtbevölkerung definitiv keine sicheren Krankheitsprognosen. Das Ergebnis der Klassifikationsanalyse kann ausschließlich dahingehend interpretiert werden, dass das Modell den Daten der Studienpopulation gut angepasst ist.

Regressionsdiagnostik

Bei den anschließend durchgeführten Regressionsdiagnosen (vgl. Anhang A 3.1) wurden lediglich zwei erwähnenswerte Besonderheiten entdeckt.

Zum einen weisen 7 Frauen überdurchschnittlich hohe und damit möglicherweise aus biologischer Sicht unplausible Lebensalter auf. Die Berücksichtigung dieser Frauen stellt allerdings kein großes Problem dar, da sie keinen unverhältnismäßig großen Einfluss auf die Parameterschätzungen nehmen.

Darüber hinaus zeigt sich, dass einzelne Beobachtungen schlecht vom Modell beschrieben werden, was zu einer Verschlechterung der globalen Modellanpassung führt. Hierbei handelt es sich genauer um Studienteilnehmerinnen, bei denen trotz dafür untypischer Merkmalswerte Brustkrebserkrankungen aufgetreten sind. Da diese Beobachtungen jedoch aus biologischer Sicht plausibel sind, kann ein Ausschluss dieser Frauen inhaltlich nicht gerechtfertigt werden.

5.3.8.2 Post-Menopause-Modell

Bei der Generierung eines logistischen Regressionsmodells für die 614 Frauen der Post-Menopause wird in vollkommener Analogie zum Prae-Menopause-Modell (vgl. 5.3.8.1) vorgegangen. In einem ersten Schritt wird eine Variablen-Vorauswahl getroffen, bei der sowohl zu lückenhaft erhobene, als auch solche Merkmale von der Betrachtung in multiplen Regressionsmodellen ausgeschlossen werden, die im Rahmen eines Modells, in welchem zur Adjustierung simultan die beiden Lebensalter-Variablen und der binären Menopause-Status betrachtet werden, nicht von signifikanter Bedeutung für die Erkrankungschance sind.

Im zweiten Schritt werden die verbleibenden Variablen simultan betrachtet. Durch die Anwendung der klassischen Rückwärtsauswahl-Prozedur wird anschließend die Anzahl der Einflussvariablen weiter reduziert.

Variablen-Vorauswahl

Für die Variablen-Vorauswahl werden analoge Ausschlusskriterien wie beim Prae-Menopause-Modell festgelegt. Das heißt, für die Betrachtung in multiplen Regressionsmodellen werden ausschließlich die Merkmale ausgewählt, für die gilt, dass höchstens 20% (122) der Werte fehlen und, die bei gleichzeitiger Betrachtung der drei Kovariablen (Lebensalter, quadriertes Lebensalter und Menopause-Status) einen Wald-Test-P-Wert kleiner 0,15 aufweisen.

Die 7 Merkmale bzw. Variablen, die diese beiden Bedingungen erfüllen, werden im Folgenden kurz beschrieben. Die konkreten Ergebnisse, die jeweils im Rahmen des Modells mit 3 Kovariablen (siehe oben) beobachtet wurden, können Tabelle 5.3.45 entnommen werden.

1) Verlust des Ehemannes – VE

Die binäre Variable „Verlust des Ehemannes“ gibt für jede Frau an, ob sie schon verheiratet war und ihren Ehemann verloren hat (VE=1) oder nicht (VE=0). Als Grund für den Verlust des Ehemannes kommen Scheidung und Versterben in Frage.

2) Geburt – GE

Die binäre Variable „Geburt“ kennzeichnet, ob die betreffende Studienteilnehmerin schon einmal ein lebendiges oder totes Kind geboren haben (LG=1) oder nicht (LG=0). Die Indikatorvariable I:=1-GE kennzeichnet die Frauen, die noch kein Kind (lebendig oder tot) geboren haben (I=1), und wird zwecks Adjustierung bei Untersuchung der Variablen „Alter 1te Schwangerschaft“ (siehe unten) benötigt.

3) Anzahl Lebendgeburten – ALG

Die stetige Variable „Anzahl Lebendgeburten“ gibt für jede Frau an, wie viele ihrer Schwangerschaften zur Geburt eines lebendigen Kindes führten. Der Wertebereich liegt zwischen 0 und 15.

4) Schwangerschaftsabbruch - SSA

Die binäre Variablen „Schwangerschaftsabbruch“ kennzeichnet, ob bei der betreffenden Frau schon einmal eine Schwangerschaft operativ abgebrochen wurde (SSA=1) oder nicht (SSA=0).

5) Fehlgeburt – FG

Die binäre Variable „Fehlgeburt“ gibt für jede Frau an, ob sie schon einmal eine Fehlgeburt erlitten hat (FG=1) oder nicht (FG=1).

6) Alter erste Schwangerschaft – A1te

Die Variable Alter erste Schwangerschaft gibt für jede Studienteilnehmerin an, in welchem Lebensalter sie zum ersten Mal eine vollständige Schwangerschaft, das heißt eine Schwangerschaft die mit der Geburt eines lebendigen oder toten Kindes endete, hinter sich gebracht hat. Zur Adjustierung werden durch eine zusätzliche Indikatorvariable I die Frauen gekennzeichnet, die noch nicht „vollständig“ schwanger waren (I=1). Für diese wird die Variable „Alter 1te Schwangerschaft“ auf 0 gesetzt.
Bei Betrachtung eines multiplen Modells im nächsten Abschnitt ersetzt die Variable „Geburt“ (GE) diese Indikatorvariable (I). Es gilt die funktionale Beziehung: GE =1-I.

7) Tumor - TU

Die binäre Variable „Tumor“ kennzeichnet für die betreffende Frau, ob sie schon einmal eine nicht Brustkrebs Tumorerkrankung erlitten hat (TU=1) oder nicht (TU=0).

Tabelle 5.3.45: Ergebnisse der Variablen-Vorauswahl (Post-Menopause)

Variable	Fehlende Werte	Odds-Ratio	Wald-Test-P-Wert
Verlust des Ehemannes (binär)	42	1,986	0,0002
Geburt (binär)	10	0,533	0,0035
Anzahl Lebendgeburten (stetig)	9	0,844	<0,0001
Schwangerschaftsabbruch (binär)	27	3,105	0,0012
Fehlgeburt (binär)	28	2,053	0,0041
Alter 1te Schwangerschaft (stetig)	51	1,047	0,0154
+ Indikator (noch keine SS)		4,611	0,0009
Tumor (binär)	18	0,609	0,1128

Klassische Rückwärtsauswahl

Um die Problematik, der Miteinbeziehung von fehlenden Werten zu umgehen, wird akzeptiert, dass bei simultaner Betrachtung der 7 Einflussvariablen zunächst 108 der 614 Studienteilnehmerinnen aufgrund unvollständiger Merkmalswerte von der Analyse ausgeschlossen werden müssen. Allerdings werden die Frauen, die im Verlaufe der Auswahlprozedur, das heißt nach Ausschluss bestimmter Variablen, lückenlose Variableneinträge aufweisen, nachträglich ins Modell integriert. Folglich ist zu erwarten, dass die Anzahl berücksichtigter Studienteilnehmerinnen mit jedem Schritt der Auswahlprozedur ansteigt.

Wird für das finale Modell gefordert, dass jede Einflussvariable einen Wald-Test-P-Wert kleiner 0,05 aufweist, endet die Rückwärtsauswahl nach 3 Schritten.

Die Ergebnisse der ML-Schätzung (Tabelle 5.3.45) zeigen, dass im ersten Schritt die Variable „Tumor“ mit einem P-Wert von 0,381 auszuschließen ist.
Nach Neuanpassung eines Modells an die verbleibenden 6 Variablen wird im zweiten Schritt die Variable „Fehlgeburt“ (P-Wert: 0,102) aus dem Modell entfernt.
Die Ergebnisse der ML-Schätzung im dritten Schritt können Tabelle 5.3.46 entnommen werden. Das Rückwärtsauswahlverfahren endet, da alle Wald-Test-P-Werte kleiner als 0,05 sind.

Tabelle 5.3.46: Finales Modell für die Frauen der Post-Menopause

Variable	Odds-Ratio	Wald-Test-P-Wert
Lebensalter	1,236	0,0034
Lebensalter quadriert	0,999	0,0077
Menopause-Status	2,018	0,0008
Geburt	0,329	0,0486
Anzahl Lebendgeburten	0,867	0,0032
Schwangerschaftsabbruch	2,820	0,0071
Alter 1te Schwangerschaft	1,042	0,0483
Verlust des Ehemannes	1,975	0,0007

Bei dieser Kombination von Einflussvariablen weisen lediglich 96 Frauen unvollständige Variableneinträge auf, so dass das Modell an 518 der 614 Post-Menopause Studienteilnehmerinnen angepasst ist. Signifikante (Zweifach-)Wechselwirkungen liegen nicht vor. Sämtliche 20

(inhaltlich sinnvollen) Produktvariablen (alle außer „Geburt“·“Anzahl Lebendgeburten“) weisen nach ihrer Aufnahme ins Modell einen P-Wert größer 0,15 auf.

Interpretation des finalen Modells

Den Odds-Ratio-Schätzungen (vgl. Tabelle 5.3.46) kann entnommen werden, dass die Brustkrebserkrankungschance von Frauen, die die Menopause bereits erreicht haben, neben dem Lebensalter noch von 6 weiteren Faktoren beeinflusst wird.

Zunächst ist für diese Frauen von Bedeutung, wie sie die Menopause erreicht haben. Frauen, bei denen die Menopause operativ herbeigeführt wurde, haben schätzungsweise eine ungefähr zweimal so große Erkrankungschance wie Frauen, die die Menopause auf biologischem Wege erreicht haben. Folglich ist davon auszugehen, dass operative Eingriffe (Hysterektomie), die den Eintritt der Menopause zur Konsequenz haben, einen ersten Risikofaktor für Brustkrebs darstellen.
Darüber hinaus kommt 4 Merkmalen, die die Fortpflanzung betreffen, eine Bedeutung für das Brustkrebsrisiko zu. Zu beachten ist hierbei, dass drei dieser Variablen („Geburt“, „Anzahl Lebendgeburten“ und „Alter erste Schwangerschaft“) voneinander abhängig sind, so dass ihre Auswirkung auf die Erkrankungschance nur gemeinsam untersucht werden kann.

Zunächst gilt, dass sich das Brustkrebsrisiko für jede Frau, die eine (vollständige) Schwangerschaft hinter sich hat, um den Faktor 0,329 verringert und zusätzlich für jede Lebendgeburt um den Faktor 0,867 sinkt. Allerdings gilt ebenfalls, dass die Brustkrebschance zudem mit dem Lebensalter zum Zeitpunkt der ersten „vollständigen“ Schwangerschaft steigt. Pro Lebensjahr erhöht sie sich schätzungsweise um den Faktor 1,042. Das heißt, die Auswirkung einer (vollständigen) Schwangerschaft auf das Brustkrebsrisiko kann nur in Zusammenhang mit den beiden Variablen „Alter 1te Schwangerschaft“ (ASS) und „Anzahl Lebendgeburten“ (ALG) sinnvoll angegeben werden. Genauer gilt schätzungsweise die folgende Beziehung:
Eine Frau die im Alter ASS zum ersten Mal eine vollständige Schwangerschaft hinter sich gebracht hat (GEB=1) und im weiteren Verlauf ihres Lebens ALG lebendige Kinder geboren hat, hat gegenüber einer gleichaltrigen Frau, die noch nie schwanger war (GEB=0, ASS=0 und ALG=0)– unabhängig von den anderen Risikofaktoren – eine um den Faktor

$$0{,}329^{GEB} \cdot 1{,}042^{ASS} \cdot 0{,}876^{ALG} \quad (GEB=1)$$

erhöhte bzw. verringerte Brustkrebserkrankungschance.

Um weiter festzustellen, in welcher Beziehung die Variablen ASS und ALG stehen müssen, damit die „Fortpflanzung“ (GEB=1) das Brustkrebsrisiko verringert, ist die Gleichung

$$0{,}329 \cdot 1{,}042^{ASS} \cdot 0{,}876^{ALG} < 1$$

nach ASS aufzulösen. Elementare Umformungen führen zu dem Ergebnis, dass die „Fortpflanzung“ das Brustkrebsrisiko nur dann verringert, wenn gilt:

$$ASS < 27{,}0 + ALG \cdot 3{,}2.$$

Daraus ergibt sich zum Beispiel, dass die Geburt eines einzigen Kindes (ALG=1) nur dann das Brustkrebsrisiko der Mutter verringert, wenn diese vor dem 30,2ten Lebensjahr ASS schwanger wird. Andernfalls stellt die Geburt dieses Einzelkindes – bedingt durch die (relativ) späte erste Schwangerschaft – einen Risikofaktor für Brustkrebs dar.

Tabelle 5.3.47 zeigt, welche Auswirkung verschiedene Ausprägungen der beiden Variablen „Anzahl Lebendgeburten“ (ALG) und „Alter 1te Schwangerschaft“ (ASS) auf die Erkrankungschance haben. Als Referenzperson (vgl. erste Zeile) dient eine Frau, die noch keine vollständige Schwangerschaft hinter sich gebracht hat ($\Rightarrow$ GEB=0, ASS=0 und ALG=0), ansonsten aber eine identische Risikofaktorkonstellation aufweist. Da im vorliegenden Datensatz keine Frau, die zum ersten Mal mit 40 Jahren (vollständig) schwanger wurde, insgesamt mehr als 3 Kinder geboren hat, ergibt sich das Ergebnis der Zelle (40 Jahre, 4) nur durch Expolierung des tatsächlich vorhanden Datenmaterials, und muss deshalb mit Vorsicht betrachtet werden.

Tabelle 5.3.47: Auswirkung der Variablen ALG und ASS auf die Erkrankungschance

Alter erste Schwangerschaft	**Anzahl Lebendgeburten**				
	0	**1**	**2**	**3**	**4**
Geburt=0	1	-----	-----	-----	-----
18 Jahre	**0,690**	**0,598**	**0,519**	**0,450**	**0,390**
25 Jahre	**0,920**	**0,798**	**0,692**	**0,600**	**0,520**
30 Jahre	1,130	**0,980**	**0,850**	**0,737**	**0,639**
35 Jahre	1,389	1,204	1,044	**0,905**	**0,785**
40 Jahre	1,706	1,479	1,282	1,112	**(0,964)**

In Zusammenhang mit der Fortpflanzungsmerkmalen ist ebenfalls von Relevanz, dass ein Schwangerschaftsabbruch, unabhängig von den 3 bisher diskutierten Fortpflanzungsmerkmalen, schätzungsweise zu einer Ver-2,8-fachung der Erkrankungschance führt. Besonders

brustkrebsgefährdet sind folglich Frauen, die noch keine vollständige Schwangerschaft hinter sich haben, wohl aber eine oder mehrere Schwangerschaften durch Abbrüche beenden ließen. Der letzte Risikofaktor des finalen Modells ist durch den Verlust des Ehemannes (Scheidung oder Versterben) gegeben. Für Frauen, die ihren Ehemann verloren haben, kommt es im Modell ungefähr zu einer Verdopplung der Erkrankungschance.

Zusammenfassend lässt sich somit festhalten, dass für Frauen der Post-Menopause das Brustkrebsrisiko insbesondere dann steigt, wenn diese

- nicht eine einzige vollständige Schwangerschaft hinter sich haben **oder**
  spät zum ersten Mal (vollständig) schwanger wurden
- durch einen operativen Eingriff in den Status der Menopause eingetreten sind
- ihren Ehemann durch Scheidung oder Versterben verloren haben.

Für Frauen der Post-Menopause, die eine vollständige Schwangerschaft hinter sich haben, gilt unabhängig davon, dass sich das Brustkrebsrisiko mit der Anzahl Lebendgeburten verringert. Darüber hinaus variiert das Brustkrebsrisiko in nicht monotoner Weise mit dem Lebensalter der Frauen.

Überprüfung der Modellanpassung

Zur Überprüfung der globalen Modellanpassung wird genau wie beim Prae-Menopause-Modell zunächst der Anpassungstest von Hosmer und Lemeshow (vgl. 4.3.6) durchgeführt. Tabelle 5.3.48 zeigt, wie viele BCA-Fälle in den einzelnen Risikogruppen nach Modell zu erwarten sind und, wie viele tatsächlich vorliegen.

Für die dritte (4,25), sechste (5,04) und achte (5,13) Gruppe ergeben sich die größten Absolutabweichungen, wohingegen die Abweichung zwischen den beobachteten und erwarteten Häufigkeiten in den extremen Risikogruppen (1 und 2 bzw. 9 und 10) deutlich geringer ausfällt. Dies legt die Vermutung nahe, dass das Modell das Erkrankungsrisiko speziell für die Studienteilnehmerinnen gut beschreibt, die entweder nur wenigen oder aber gleichzeitig vielen Risikofaktoren ausgesetzt sind. In den mittleren Risikogruppen, das heißt für Frauen, die einer mittleren Anzahl von Risikofaktoren ausgesetzt sind, liegt hingegen offensichtlich eine schlechtere Anpassung vor.

Tabelle 5.3.48: Tafel zum Anpassungstest von Hosmer & Lemeshow

Risiko-gruppe	Gruppen-größe	Erwartete An-zahl BCA-Fälle	Beobachtete An-zahl BCA-Fälle
1	52	6.36	8
2	53	11.41	9
3	52	14.75	19
4	52	17.66	16
5	52	20.35	17
6	53	23.04	18
7	54	25.36	26
8	53	27.87	33
9	52	31.39	32
10	45	32.81	33

Die Teststatistik für den globalen Anpassungstest weist einen Wert von 7,99 auf, was einem P-Wert von 0,43 entspricht. Somit kann trotz der ungleichmäßigen Modellanpassung (vgl. Tabelle 5.3.48) insgesamt davon ausgegangen werden, dass das Modell die Daten hinreichend gut beschreibt.

Bei der Klassifikationstafelanalyse zeigt sich, dass auch hier die Wahl der relativen BCA-Häufigkeit von 0,407 als Diskriminanzwert zu einem besseren Ergebnis führt als der Standard-Diskriminanzwert von 0,5.
Tabelle 5.3.49 zeigt die Klassifikationstafel, die resultiert, wenn genau für die Studienteilnehmerinnen eine Brustkrebskrankheit unterstellt wird, für die sich im Modell eine Brustkrebserkrankungswahrscheinlichkeit größer 0,407 ergibt.

Tabelle 5.3.49: Klassifikationstafel - Diskriminanzwert p=211/518 =0.407

BCA-Status: Vorhersage:	BCA=0	BCA=1	Gesamt
BCA=0	178	69	247
BCA=1	129	142	271
Gesamt:	307	211	518

Den Zellhäufigkeiten kann entnommen werden, dass unter Zuhilfenahme des Post-Menopause-Modells für 61,8% der 518 Studienteilnehmerinnen der richtige BCA-Zustand prognostiziert werden kann. Sensitivität und Spezifität betragen 67,3% bzw. 58% und sind somit näherungsweise von derselben Größenordnung. Daher kann die Eignung des Verfahrens genau wie beim Prae-Menopause-Modell durch den Anteil korrekter Vorhersagen von 61,8% wiedergegeben werden.
Bereits im Zusammenhang mit dem Prae-Menopause-Modell wurde ausführlicher erläutert, dass trotz eines eigentlich eher geringen Anteils korrekter Prognosen - in Anbetracht der Komplexität der Tumorentstehung - von einer verhältnismäßig guten Vorhersagekraft und damit Modellanpassung ausgegangen werden kann. Ebenfalls soll noch einmal daran erinnert werden, dass sich die Vorhersagekraft ausschließlich auf das vorliegende Datenmaterial bezieht und daher nicht auf andere Daten bzw. die Gesamtpopulation übertragen lässt.

Zusammenfassend kann festgehalten werden, dass die Ergebnisse des Hosmer und Lemeshow Anpassungstests und der Klassifikationsanalyse nicht im Widerspruch zu einer guten globalen Modellanpassung stehen. Bis auf Weiteres kann deshalb auch hier davon ausgegangen werden, dass das Modell die vorliegenden Daten hinreichend gut beschreibt.

Regressionsdiagnose

Bei der Regressionsdiagnose für das Post-Menopause-Modell (vgl. Anhang A.3.2) wurden keine besonderen Auffälligkeiten entdeckt.
Lediglich einige wenige brustkrebserkrankte Studienteilnehmerinnen weisen für Brustkrebs untypische Merkmalswerte auf, so dass ihnen im Modell geringe Erkrankungswahrscheinlichkeiten zukommen. Ein Vergleich der Merkmalswert-Kombinationen dieser Frauen zeigt, dass diese keinen Zusammenhang aufweisen.
Da diese Frauen keine großen Leverage-Werte zukommen, das heißt ihre Merkmalswerte für die Studienpopulation keine Besonderheit darstellen, verschlechtern sie als solche zwar die globale Modellanpassung, nehmen aber nur einen geringfügigen Einfluss auf die Parameterschätzung.

5.3.8.3 Vergleich der Ergebnisse

Als Fazit des Unterkapitels 5.3.8, in welchem für die Frauen vor (5.3.8.1) und nach (5.3.8.2) der Menopause separate Modelle generiert wurden kann festgehalten werden, dass in beiden Subpopulationen neben dem Lebensalter auch Fortpflanzungsmerkmale von Bedeutung sind.

In beiden Subpopulationen stellen eine späte erste Schwangerschaft und Schwangerschaftsprobleme (Fehlgeburten und/oder Schwangerschaftsabbrüche) Faktoren dar, die das Brustkrebsrisiko erhöhen. Für die Frauen der Post-Menopause gilt zudem, dass das Erkrankungsrisiko mit der Anzahl Lebendgeburten sinkt.
Ein wesentlicher Unterschied besteht jedoch darin, dass für Frauen der Prae-Menopause jede vollständige Schwangerschaft bereits als solche einen Risikofaktor darstellt, wohingegen für die Frauen der Post-Menopause ausschließlich späte erste Schwangerschaften das Brustkrebsrisiko erhöhen.
Nicht verwunderlich ist, dass der Lebensumstand „Verlust des Ehemannes“ (durch Scheidung oder Versterben) nur für Frauen der Post-Menopause von signifikanter Bedeutung als Risikofaktor ist. Bedingt durch den Altersunterschied in den beiden Populationen haben nur sehr wenigen Studienteilnehmerinnen der Prae-Menopause ihren Ehemann verloren, so dass in dieser Population alternativ die binäre Variable „Verheiratet“ (Ja/ Nein) betrachtet wurde. (Letztere wurde bereits im Rahmen der Vorauswahl-Untersuchungen als bedeutungslos eingestuft.) Analoges gilt auch für das binäre Merkmal „Menopause-Status“, da dieses in beiden Subpopulationen nicht dieselbe inhaltliche Bedeutung hat.

Ein Vergleich dieser Ergebnisse mit den Ergebnissen, die bei Betrachtung der gesamten Studienpopulation erzielt wurden, wird dadurch erschwert, dass einige Merkmale aufgrund ihrer Unvollständigkeit in den separaten Modellen unberücksichtigt blieben. Hierzu gehören die Merkmale „Einkommensklasse“ und „Regelmäßigkeit der Periode“. Das Merkmal „Späte Menarche“ konnte ausschließlich im Prae-Menopause-Modell betrachtet werden. Wenngleich es letztlich in Anbetracht von 71 fehlenden Werten im finalen Ergebnis nicht berücksichtigt wurde, zeigen die Zwischenergebnisse (Tabelle 5.3.42), dass dieses Merkmal auch für die Frauen der Prae-Menopause von Bedeutung sein könnte.
Abschließend soll noch angemerkt werden, dass für beide Subpopulationen ausschließlich Merkmale von Bedeutung sind, die auch in Modellen für alle Studienteilnehmerinnen zu finden sind (vgl. Tabelle 5.3.38). Inhaltlich nicht zu deuten ist, warum sich die Geburt eines Kindes für Frauen der Prae-Menopause grundsätzlich als Risikofaktor präsentiert, wohingegen für die Frauen der Post-Menopause (und dadurch bedingt auch in der gesamten Studienpopulation) nur späte erste Schwangerschaften das Brustkrebsrisiko erhöhen.

Kapitel 6. Diskussion und Zusammenfassung der Ergebnisse

In der vorliegenden Arbeit wurde mit Hilfe statistischer Methoden auf Grundlage von Daten, die von der Howard-Universität im Rahmen einer umfassenderen Brustkrebsstudie erhoben wurden, nach epidemiologischen Risikofaktoren für Brustkrebs gesucht.

In einem ersten Schritt wurde das Datenmaterial dahingehend untersucht, ob sich Anzeichen für eine familiäre Häufung der Brustkrebskrankheit finden lassen. Die Untersuchungen gründen auf den Verwandtschaftsbeziehungen und Brustkrebszuständen von 1835 afroamerikanischen Frauen aus 261 Familien. Aufgrund der unterschiedlichen Familiengrößen und dem studiendesign-bedingten Umstand, dass sich in jeder Familie mindestens eine brustkrebserkrankte Frau befindet, gelang es nicht, besonders brustkrebsgefährdete Familien zu identifizieren. Ebenfalls war es aufgrund der Verwandtschaftsbeziehungen zwischen den Studienteilnehmern nicht möglich, zu untersuchen, ob Frauen, deren Mütter oder Schwestern an Brustkrebs erkrankt sind, ein erhöhtes Risiko aufweisen. Lediglich untersucht werden konnte, ob sich die durchschnittlichen Anteile von Brustkrebserkrankungen unter den Schwestern der sicheren Fälle in Abhängigkeit von den Erkrankungszuständen ihrer Mütter und blutsverwandten Tanten unterscheiden. Bei dieser Untersuchung zeigten sich allerdings keine signifikanten Unterschiede in den relativen Häufigkeiten, so dass kein Nachweis für eine familiäre Häufung der Brustkrebskrankheit erbracht werden konnte.

Im zweiten Auswertungsschritt wurde auf Grundlage der ausführlicheren epidemiologischen Informationen, die von 1135 afroamerikanischen Frauen zur Verfügung stehen, nach nichtfamiliären Risikofaktoren für Brustkrebs gesucht. Da keine Daten von Kontrollfamilien zur Verfügung stehen und da im ersten Auswertungsschritt kein Anzeichen für eine familiäre Häufung gefunden werden konnte, wurde entschieden, die familiären Abhängigkeiten zwischen den Angehörigen derselben Familien zu vernachlässigen und die Daten im Sinne einer Fall-Kontroll-Studie auszuwerten. In dieser Betrachtungsweise konnten die Studienteilnehmerinnen in Abhängigkeit von ihren Brustkrebszuständen in eine Fallgruppe der Größe 318 und eine Kontrollgruppe der Größe 817 aufgeteilt werden.
Die anschließende Suche nach Risikofaktoren wurde durch die Unvollständigkeit des Datenmaterials erschwert. Im Rahmen einer Variablen-Vorauswahl mittels einfacher logistischer Regressionsmodelle wurde jedes Merkmal nur zusammen mit zwei Lebensalter-Kovariablen betrachtet, so dass der jeweilige Ausschluss von Frauen mit fehlenden Merkmalswerten zu keinem Informationsverlust hinsichtlich anderen Variablen führte. Bei der anschließenden

Betrachtung multipler logistischer Regressionsmodelle war eine Miteinbeziehung fehlender Werte unumgänglich, da in diesem Fall, der Ausschluss von Frauen mit unvollständigen Merkmalswerten gleichzeitig auch den Verlust, der über diese Frauen zur Verfügung stehenden Informationen, bedeutet hätte.
Im Umgang mit den fehlenden Werten, die im Datensatz nicht zufällig auftreten, sondern einem Missing-Randomly-At-Outcome-Mechanismus folgen, kamen unabhängig voneinander zwei Ad-hoc-Verfahren zum Einsatz. Bei der Probability-Imputation-Methode werden fehlende Datenwerte durch den Mittelwert der vorhandenen Werte aufgefüllt, wohingegen fehlende Dateneinträge beim Indikatorvariablen-Verfahren auf den Wert Null gesetzt und durch die zusätzliche Einführung einer binären Dummy-Variablen gekennzeichnet werden. Beide Verfahren können Ergebnisverzerrungen hervorrufen, wobei speziell beim Indikatorvariablen-Verfahren davon auszugehen ist, dass die größten Verzerrungen durch Dummy-Variablen hervorgerufen werden, die eine große Anzahl von fehlenden Werten kennzeichnen. Deshalb wurde im Zusammenhang mit dem Indikatorvariablen-Verfahren ein zweites Modell betrachtet, von welchem a priori alle Merkmale mit mehr als 200 fehlenden Werten ausgeschlossen wurden.

Für alle drei Modelle wurden Rückwärtsauswahlprozeduren durchgeführt und beim Vergleich der finalen Modelle sowohl Gemeinsamkeiten als auch Unterschiede festgestellt. Obwohl kein Widerspruch, in dem Sinne, dass ein Merkmal in einem Modell das Brustkrebsrisiko verringert und in einem anderen das Brustkrebsrisiko erhöht, beobachtet wurde, zeigt ein Vergleich der von Modell zu Modell variierenden Odds-Ratio-Schätzwerte, dass bedingt durch die Miteinbeziehung fehlender Werte eine zuverlässige Quantifizierung der Merkmalseinflüsse nicht möglich ist. Folglich können Variablen in Bezug auf ihre Relevanz für die Brustkrebskrankheit nur in Abhängigkeit davon, in welchen der drei Modelle ihnen eine Bedeutung zukam, in drei Gruppen aufgeteilt werden.
Am sichersten erscheint der Einfluss von Merkmalen, denen sowohl im Probability-Imputation-Modell als auch im Indikatorvariablen-Modell bei Berücksichtigung aller Merkmale eine Bedeutung zukam. Zu diesen gehören neben dem Lebensalter die Merkmale „Anzahl Lebendgeburten“, „Schwangerschaftsabbruch“, „Unregelmäßige Periode“ und „Späte Menarche“. In beiden Modellen gilt, dass das Brustkrebsrisiko mit jeder Lebendgeburt sinkt, und dass der Abbruch einer Schwangerschaft, das nicht regelmäßige Eintreten der Periode und eine späte Menarche (>12 Jahre) das Erkrankungsrisiko erhöhen. Das Ergebnis bezüglich

des Merkmals „Späte Menarche“, steht damit im Widerspruch zu den Ergebnissen anderer Studien, denen zufolge – ganz gegenteilig - eine frühe Menarche als Risikofaktor anzusehen ist (vgl. zum Beispiel Kelsey und Kelsey et al. 1993).
Ebenfalls deutet vieles darauf hin, dass auch die Faktoren „Fehlgeburt“, „Menopause-Status“ und „Verlust des Ehemannes“ von Bedeutung für das Brustkrebsrisiko sind. Allen dreien kommt im Probability-Imputation-Modell und im Indikatorvariablen-Modell nach Vernachlässigung der besonders lückenhaften Variablen (u.a. „Späte Menarche“ und „Unregelmäßige Periode“) ein signifikanter Erklärungswert zu. Die Schätzwerte zeigen, dass sowohl Frauen, die ihren Ehemann (durch Versterben oder Scheidung) verloren haben, als auch Frauen, die mindestens eine Fehlgeburt erlitten haben, brustkrebsgefährdet sind. Die Schätzwerte der kategorialen Variablen „Menopause-Status“ zeigen, dass in beiden Modellen ausschließlich Frauen, bei denen die Menopause auf regulärem Wege eingetreten ist, signifikant weniger brustkrebsgefährdet sind als Frauen, bei denen die Menopause operativ (vorwiegend durch Hysterektomien) herbeigeführt wurde.
Als eher unsicher angesehen werden müssen die Bedeutungen der Merkmale „Alter erste Schwangerschaft“ und „Familiäres Einkommen“. Nur im Indikatorvariablen-Modell, bei welchem a priori alle Merkmale mit mehr als 200 fehlenden Merkmalswerten ausgeschlossen wurden, zeigt sich, dass das Brustkrebsrisiko signifikant mit dem Alter der ersten vollständigen Schwangerschaft steigt und ausschließlich im Probability-Imputation-Modell steigt das Risiko mit dem familiären Einkommen. Obwohl die Bedeutung beider Faktoren damit fragwürdig erscheint, kann Kelsey und Kelsey et al. (1993) entnommen werden, dass beide Merkmale in anderen Studien als Risikofaktoren für Brustkrebs identifiziert wurden.

Im Anschluss wurden noch zwei separate Modelle für die Frauen vor und nach der Menopause generiert, wobei ausschließlich Merkmale mit wenigen fehlenden Werten betrachtet wurden, so dass auf eine verzerrende Miteinbeziehung fehlender Werte verzichtet werden konnte. Aufgrund ihrer Unvollständigkeit mussten dabei allerdings auch Merkmale („Einkommensklasse“, „Regelmäßigkeit der Periode“ und „Späte Menarche“) ausgeschlossen werden, die zuvor in Bezug auf die gesamte Studienpopulation als bedeutsam eingestuft wurden (siehe oben). Diese sind folglich im Rahmen der Untersuchungen für Frauen vor und nach der Menopause als unkontrollierte Störgrößen anzusehen.

Für die Frauen der Prae-Menopause konnten von den nur wenig unvollständig erhobenen Merkmalen neben dem Lebensalter lediglich die Faktoren „Schwangerschaftsprobleme“ und

„Alter erste Schwangerschaft“ als Risikofaktoren identifiziert werden. Die Schätzwerte zeigen, dass die Durchführung eines Schwangerschaftsabbruchs und/ oder das Erleiden einer Fehlgeburt das Risiko schon vor der Menopause an Brustkrebs zu erkranken erhöht. Darüber stellt die Geburt eines Kindes einen Risikofaktor dar, wobei die Erkrankungschance insbesondere mit dem Alter zum Zeitpunkt der ersten Schwangerschaft ansteigt. Die anschließende Überprüfung der Modellanpassung zeigte, dass das Modell die Daten verhältnismäßig gut beschreibt. Lediglich bei der Regressionsdiagnostik fiel auf, dass einige wenige, brustkrebserkrankte Frauen, die dafür untypische Werte aufweisen, vom Modell schlecht beschrieben werden und deshalb einen entsprechend großen Einfluss auf die Modellparameterschätzungen nehmen. Darüber hinaus wurden im Rahmen der Regressionsdiagnostik einige Frauen mit untypisch hohen Lebensaltern (für Frauen vor der Menopause) identifiziert. Da diese jedoch keinen großen Einfluss auf die Parameterschätzwerte nehmen, wurde ohne substanzwissenschaftliche Rückfrage entschieden diese nicht aus dem Modell auszuschließen.

Im finalen Modell für die Frauen der Post-Menopause ergaben sich bei Ausschluss aller Frauen mit fehlenden Merkmalswerten, die folgenden Beziehungen.
Mit jeder Lebendgeburt sinkt das Brustkrebsrisiko, wobei insbesondere schon eine hinreichend frühe erste Geburt eines einzigen Kindes das Brustkrebsrisiko senkt. Auf der anderen Seite steigt das Brustkrebsrisiko mit dem Alter zum Zeitpunkt der ersten Schwangerschaft, so dass Frauen mit einer späten ersten Schwangerschaft sogar brustkrebsgefährdeter sind als kinderlose Frauen. Darüber hinaus kommt es jeweils zu einer Erhöhung der Erkrankungschance, wenn für Frauen gilt, dass sie mindestens eine Fehlgeburt erlitten haben, dass sie die Menopause auf operativem Wege (Hysterektomie) erreicht haben, oder dass sie ihren Ehemann verloren haben. Zu einem gewissen Grade von statistischer Signifikanz ist zudem, dass Frauen, die an einer Nichtbrustkrebs-Tumorkrankheit leiden, weniger brustkrebsgefährdet sind.
Die anschließend durchgeführte Überprüfung der Modellanpassung zeigte auch in dieser Subpopulation, dass das Modell die Daten hinreichend gut beschreibt. Bei der Regressionsdiagnostik wurden ausschließlich Frauen identifiziert, die an Brustkrebs erkrankt sind, dafür aber untypische Merkmalswerte aufweisen. Ein großer Einfluss auf die Parameterschätzung geht von diesen Studienteilnehmerinnen allerdings nicht aus.

Sowohl für Frauen vor als auch Frauen nach der Menopause sind somit neben dem Lebensalter auch Fortpflanzungsmerkmale von Bedeutung. In beiden Subpopulationen stellen eine späte erste Schwangerschaft und Schwangerschaftsprobleme (Fehlgeburten und/oder Schwanger-

schaftsabbrüche) Faktoren dar, die das Brustkrebsrisiko erhöhen. Für die Frauen der Post-Menopause gilt zudem, dass das Erkrankungsrisiko mit der Anzahl Lebendgeburten sinkt.
Ein wesentlicher Unterschied besteht jedoch darin, dass für Frauen der Prae-Menopause jede vollständige Schwangerschaft bereits als solche einen deutlichen Risikofaktor darstellt, wohingegen für die Frauen der Post-Menopause ausschließlich späte erste Schwangerschaften das Brustkrebsrisiko erhöhen. Warum sich die Geburt eines Kindes für Frauen der Prae-Menopause als Risikofaktor präsentiert, kann inhaltlich nicht gedeutet werden. Die anderen beiden Faktoren („Menopause operativ erreicht“ und „Verlust des Ehemannes“), die in der Subpopulation Post-Menopause von Bedeutung sind, stellen für die Frauen vor der Menopause keine inhaltlich sinnvollen Expositionen dar.

Als Fazit kann festgehalten werden, dass im Rahmen dieser Arbeit trotz der systematischen Unvollständigkeit des Datenmaterials Risikofaktoren für Brustkrebs identifiziert werden konnten. Eine zuverlässige Quantifizierung der Bedeutung dieser Faktoren für das Brustkrebsrisiko war hierbei nicht möglich, da aufgrund der Fülle des Datenmaterials fehlende Merkmalswerte nur durch den Einsatz zweier Ad-hoc-Verfahren, die zu nicht abschätzbaren Ergebnisverzerrungen führen, miteinbezogen werden konnten. Die Gemeinsamkeiten der Ergebnisse, die sich durch Einsatz der beiden Verfahren im Umgang mit fehlenden Werten ergaben, und der Ergebnisse der beiden Modelle für Frauen vor und nach der Menopause, bei denen unvollständige Merkmale als Störgrößen akzeptiert wurden, deuten jedoch darauf hin, dass die Ergebnisverzerrungen in einem vertretbaren Rahmen liegen. Speziell die Modelle, die für die Frauen vor und nach der Menopause generiert wurden, bestätigen in diesem Zusammenhang, dass die Bedeutung der meisten Variablen nicht ausschließlich auf die Miteinbeziehung (im Sinne von Manipulation) fehlender Werte zurückzuführen ist. Lediglich in Bezug auf Merkmale, die aufgrund ihrer Unvollständigkeit in den separaten Modellen unberücksichtigt blieben („Einkommensklasse“, „Regelmäßigkeit der Periode“ und „Späte Menarche“), liefern die Ergebnisse des Prae- und Post-Menopause-Modells keinen zusätzlichen Informationsgewinn.

Speziell im Hinblick auf weiterführende Brustkrebsstudien, sollten deshalb in Anbetracht der Ergebnisse dieser Arbeit die folgenden Merkmale als nicht-familiäre Einflussgrößen berücksichtigt oder zumindest als potentielle Störgrößen kontrolliert werden:

- Lebensalter
- Menopause-Status

- Regelmäßigkeit des Periodenzyklus
- Menarche-Alter
- Anzahl Lebendgeburten
- Alter erste Schwangerschaft
- Fehlgeburten
- Schwangerschaftsabbrüche
- Familienstand
- Familiäres Einkommen
- Angaben zu Nichtbrustkrebs-Tumorleiden

Als problematisch in Bezug auf die statistische Auswertung stellte sich heraus, dass lediglich Daten von Familien zur Verfügung gestellt wurden, in denen es mindestens einen Brustkrebsfall gegeben hat, so dass in Anbetracht des Fehlens von Kontrollfamilien-Daten die familiären Abhängigkeiten zwischen den Individuen im Rahmen der logistischen Modelle nicht berücksichtigt werden konnten. Da die Ergebnisse anderer Studien zeigen, dass genetische Dispositionen Einfluss auf das Brustkrebsrisiko nehmen (vgl. Kelsey et al. 1993), stellten diese im Rahmen der durchgeführten Untersuchungen unkontrollierbare Störgrößen dar. In einer weiterführenden Untersuchung wurden deshalb ausschließlich die sicheren BCA-Indexfälle der 253 Familien betrachtet und versucht, ihre Brustkrebserkrankungsalter mit Hilfe eines Proportional-Hazards-Modells (vgl. Klein und Moeschberger 1997) zu ihren Merkmalsausprägungen in Beziehung zu setzen. Hierbei zeigte sich jedoch, dass aufgrund der unterschiedlichen Geburtsdaten und der Korrelation zwischen den Lebens- und Erkrankungsaltern der Indexfälle bei dieser Untersuchungsform häufig nur Scheinassoziationen zwischen den Einflussvariablen und der Zielvariablen „Erkrankungsalter“ resultieren. Nach Anpassung eines Proportional-Hazards-Modells an die unproblematischen Variablen, zeigten graphische Untersuchungen, dass die für diese Modellklasse notwendige Annahme, proportionaler Hazard-Funktionen, nicht gehalten werden kann.

Abschließend soll noch einmal angemerkt werden, dass als Datenmaterial ausschließlich epidemiologische Angaben über afroamerikanischer Frauen zur Verfügung gestellt wurden, so dass auf Grundlage dieser Arbeit nicht entscheidbar ist, ob für Frauen anderer ethnischer Abstammung dieselben Risikofaktoren von Bedeutung für die Brustkrebskrankheit sind.

Anhang 1: Überlegungen zur naiven (–1/0/1)-Kodierung

In diesem ersten Anhang 1 werden theoretische Überlegungen dazu angestellt, warum die in Unterkapitel 5.3.1 vorgestellte (-1/0/1)-Kodierung des Merkmals „Fettleibigkeitsstatus“ kein geeignetes Verfahren im Umgang mit fehlenden Werten darstellt.
Diese Überlegungen zeigen, dass aller Voraussicht nach, die ungleiche Verteilung der fehlenden Werte in Fall- und Kontrollgruppe in Verbindung mit der unterschiedlichen Expositionshäufigkeit in den Daten, zu der Ergebnisverzerrung führt.

Zunächst werden die Erkenntnisse zusammengefasst, die sich aus den zuvor durchgeführten Analysen ergaben.

Da fehlende Werte in der Kontrollgruppe häufiger auftreten als in der Fallgruppe ist die Merkmalsausprägung „fehlender Wert“ (Y=0) mit einem hohen Brustkrebsrisiko assoziiert. Die Untersuchungen, basierend auf dem vollständig vorhandenen Datenmaterial, zeigten zudem, dass sich für die beiden regulären Ausprägungen (Y=-1 bzw. Y=1) die Erkrankungsrisiken nicht (signifikant) unterscheiden. Grundsätzlich besteht also kein Unterschied zwischen den Erkrankungswahrscheinlichkeiten der beiden Gruppen Exponierter und Nichteexponierter, wohl aber zwischen den Erkrankungswahrscheinlichkeiten dieser beiden Gruppen und der Gruppe von Studienteilnehmerinnen mit fehlenden Werten. Die Gruppe „fehlender Wert“ weist eine deutlich geringere Erkrankungswahrscheinlichkeit auf.
Berücksichtigt man nun, dass die, zu maximierende, Likelihood-Funktion, in Abhängigkeit vom Parametervektor β die Wahrscheinlichkeit für die Realisierung der beobachteten Daten beschreibt, ergeben sich bei einer Wahl von $\beta(Y)$ nahe Null, im Modell für alle drei Gruppen (Exponierte, Nichtexponierte und „fehlender Wert“) ungefähr dieselben Grunderkrankungswahrscheinlichkeiten $\exp\{\beta_0\}$, welche nur noch in Abhängigkeit vom Lebensalter modifiziert werden. Das bedeutet, dass die Wahl $\beta(Y)\cong 0$ in Abhängigkeit von der Größe des Parameters β_0 zwangsläufig entweder für die 704 Frauen mit bekanntem Expositionsstatus oder für die 431 Frauen mit fehlendem Y-Eintrag zu einer schlecht angepassten Grunderkrankungschance führt.

Auf der anderen Seite kann bei geeigneter Wahl von $\beta(Y)<0$ und β_0 erreicht werden, dass sich ausschließlich für die 237 exponierten Frauen eine schlecht angepasste Erkrankungswahrscheinlichkeit ergibt.

Da für das Modell die folgenden Odds-Ratio-Beziehungen gelten:

Odds („fehlender Wert“) = Odds(Y=0) = $\exp\{\beta_0\}$

Odds („nicht fettleibig“) = Odds(Y=-1) = $\exp\{\beta_0\} \cdot \exp\{-\beta(Y)\}$ und

Odds (“fettleibig”) = Odds(Y=1) = $\exp\{\beta_0\} \cdot \exp\{\beta(Y)\}$,

kann zunächst β_0 so gewählt werden, dass $\exp\{\beta_0\}$ die Grunderkrankungschance der Gruppe „fehlender Wert“ gut wiedergibt. Anschließend kann in Abhängigkeit von β_0 der Parameter $\beta(Y)$ so gewählt werden, dass $\exp\{\beta_0\}\cdot\exp\{-\beta(Y)\}$ die Grunderkrankungschance der Nicht-exponierten gut beschreibt. Diese Parameterwahl führt dazu, dass für insgesamt 898 Frauen gut angepasste Erkrankungschancen vorliegen. Darüber hinaus legen die beiden Parameter β_0 und $\beta(Y)$ allerdings auch die Erkrankungschance für die 237 exponierten Studienteilnehmerinnen fest. Es gilt die Beziehung:

Odds(“fettleibig”) $=\exp\{\beta_0\}\cdot\exp\{\beta(Y)\}$ =OR(Y=0,Y=-1)^2 · Odds(“nicht fettleibig”).

Da die Grunderkrankungschance Nichtexponierter (Y=-1) in den vorliegenden Daten größer ist als die Chance von Frauen mit fehlenden Merkmalswerten (Y=0), ist das Odds-Ratio OR:=OR(Y=0,Y=-1) kleiner als Eins.
Folglich resultiert für die Gruppe Exponierter der Größe 237 eine zu geringe Erkrankungschance: $OR^2\cdot$Odds(„nicht fettleibig“).

Die realisierten Maximum-Likelihood-Schätzwerte (vgl. Unterkapitel 5.3.1) zeigen, dass trotz dieser zu geringen Erkrankungschance für die 237 Exponierten, diese Parameterwahl angesichts der guten Anpassung an die 898 Frauen offensichtlich noch immer zu einen größeren Likelihood-Wert führt als ein $\beta(Y)$-Wert nahe bei Null.

Anhang 2: Theoretische Überlegungen zum (unbedingten) Probability-Imputation-Verfahren

Nachdem in Anhang 1 Überlegungen dazu angestellt wurden, warum das (-1/0/1)-Kodierungsschema des Merkmals „Fettleibigkeitsstatus“ zu verzerrten Parameterschätzungen führt, wird in diesem zweiten Anhang kurz erläutert, warum das Verfahren der Probability-Imputation besser geeignet ist.

Im Hinblick auf die zu maximierende Likelihood-Funktion ergeben sich im Gegensatz zur naiven (–1/0/1)-Codierung nach einer Probability-Imputation bzgl. des dichotomen Merkmals „Fettleibigkeitsstatus“ die folgenden Odds-Beziehungen:

Odds („fehlender Wert“) = Odds$\{I(Y)_2=p\}$ = exp$\{\beta_0\}\cdot$ exp$\{p\cdot\beta(Y)\}$

Odds („nicht fettleibig“) = Odds$\{I(Y)_2=0\}$ = exp$\{\beta_0\}$ und

Odds (“fettleibig”) = Odds$\{I(Y)_2=1\}$ = exp$\{\beta_0\}\cdot$ exp$\{\beta(Y)\}$,

Wie bei der (–1/0/1)-Codierung gewährleistet eine geeignete Wahl von β_0 und $\beta(Y)$, dass die Grunderkrankungschancen der Nichtexponierten und Frauen mit fehlenden Merkmalswerten gut durch die Ausdrücke exp$\{\beta_0\}$ bzw. exp$\{\beta_0\}\cdot$ exp$\{p\cdot\beta(Y)\}$beschrieben werden.
Durch entsprechende Parameterwahl können insbesondere dieselben Erkrankungschancen und damit auch dasselbe Odds-Ratio OR:=OR$\{I(Y)_2=p, I(Y)_2=0\}<1$ wie bei der (–1/0/1)-Codierung erreicht werden.

Für die Erkrankungschance der 231 Exponierten gilt nun allerdings nicht wie bei der (–1/0/1)-Kodierung die vom Anteil Exponierter p unabhängige Beziehung:

$$\text{Odds}(\text{„fettleibig“}) = \text{OR}^2\cdot\text{Odds}(\text{„nicht fettleibig“}),$$

sondern die folgende, von p abhängige, Beziehung:

$$\text{Odds}(\text{„fettleibig“}) = \exp\{\beta_0\}\cdot\exp\{\beta(Y)\} = \text{OR}^{1/p}\cdot\text{Odds}(\text{“nicht fettleibig”}).$$

<u>Hierbei gilt</u>:

1) OR<1, denn in den vorliegenden Daten ist die Grunderkrankungschance der Nichtexponierten größer als die Chance der Frauen mit fehlenden Merkmalswerten,

2) $p<0,5$, denn andernfalls wäre die Gruppe der Exponierten nicht kleiner als die Gruppe der Nichtexponierten,

so dass folgt: $OR^2>OR^{1/p}$.

Das heißt, bei dem Verfahren der Probability-Imputation führt eine Wahl von β_0 und $\beta(Y)$, die gute Grunderkrankungschancen für die beiden größten Gruppen (Nichtexponierte und Gruppe „fehlender Wert“) gewährleistet in Abhängigkeit von $p<0,5$ zu noch ungeeigneteren bzw. geringeren Erkrankungschancen als die (–1/0/1)-Kodierung.
Die realisierten Schätzwerte (vgl. Unterkapitel 5.3.1) zeigen, dass diese noch ungünstigeren Erkrankungschancen für die Exponierten offensichtlich dazu führen, dass nach der (unbedingten) Probability-Imputation eine Wahl von $\beta(Y)$ nahe bei Null zu einem größeren Likelihood-Wert führt als eine Modellparameterwahl, die hinsichtlich der beiden größten Gruppen eine gute Anpassung gewährleistet und dabei äußerst ungünstige Erkrankungschancen der Exponierten toleriert.

Anhang 3: Regressionsdiagnostik

In diesem Anhang werden die beiden Modelle aus Unterkapitel 5.3.8 jeweils einer Regressionsdiagnose unterzogen. In beiden Fällen gilt es herauszufinden, ob einzelne Beobachtungen besonderen Einfluss auf die Parameterschätzungen nehmen und/oder, ob bestimmte Beobachtungen besonders schlecht vom Modell beschrieben werden.

Nur im Rahmen der Untersuchung des Prae-Menopause-Modells (vgl. A.3.1) wird noch einmal etwas ausführlicher auf die Methoden der Regressionsdiagnose (vgl. 4.3.7) eingegangen. Im darauf folgenden Unterkapitel A.3.2, das heißt bei der Regressionsdiagnose des Post-Menopause-Modells, kommen dieselben Diagnosegrößen zum Einsatz, so dass diese als bekannt vorausgesetzt werden.

A.3.1 Regressionsdiagnose – Prae-Menopause-Modell

Das zuerst zu untersuchende Modell, welches an die 355 Frauen der Prae-Menopause mit vollständigen Merkmalseinträgen angepasst wurde, lautet:

$$Odds\{P(D=1 \mid X)\} = \exp\{-9.202 + (0.270 \cdot X_1 - 0.002 \cdot X_1^2) + 0{,}603 \cdot X_2 + 0.0512 \cdot X_3\},$$

mit $X=(X_1, X_2, X_3)^T$, wobei X_1 das Lebensalter und X_2 das Alter, zum Zeitpunkt der ersten vollständigen Schwangerschaft, beschreibt. Die Indikatorvariable X_3 kennzeichnet, ob es Schwangerschaftsprobleme (Fehlgeburten und/oder Schwangerschaftsabbrüche) gegeben hat.

Als Regressionsdiagnostiken dienen neben den Leverage-Werten zwei der drei in Unterkapitel 4.3.7 vorgestellten Diagnosemaße. Die dritte Diagnosegröße $\Delta D_{(.)}$ wird nicht benötigt, da sie sich im vorliegenden Fall äquivalent zur Diagnosegröße $\Delta X^2_{(.)}$ verhält, und ihre Betrachtung daher keinen zusätzlichen Informationsgewinn einbringt. Im Folgenden wird kurz wiederholt, welche Aussagen die Maße im Hinblick auf eine Regressionsdiagnose erlauben.

Unter Zuhilfenahme der Ausdrücke

$$\Delta\hat{\beta}_{(-j)} \approx (r_j)^2 \cdot \frac{h_j}{(1-h_j)^2} = \frac{(d_j - \hat{p}_j)^2}{\hat{p}_j \cdot (1-\hat{p}_j)} \cdot \frac{h_j}{(1-h_j)^2}$$

kann zunächst für jede Frau (bzw. Beobachtung) beurteilt werden, inwieweit die bei ihr vorliegende Kombination von Einflussvariablen X_1, X_2, X_3 und Brustkrebszustand D die Schätzung des Parametervektors beeinflusst. Große Werte weisen darauf hin, dass sich die Berück-

sichtigung der dazugehörigen Studienteilnehmerin in besonderer Weise auf die Parameterschätzungen auswirkt.

Die Kenngrößen ergeben sich jeweils als Produkt aus dem quadrierten Pearson-Residuum r_j^2 und einer streng monoton steigenden Funktion des dazugehörigen Leverage-Wertes h_j (vgl. 4.3.7). Ursachen für die besondere Beeinflussung können also sein, untypische Kombinationen von Brustkrebszustand und Ausprägungen der Einflussvariablen und/oder für sich betrachtete untypische Ausprägungen der Einflussvariablen.

Die Sensitivität der Pearson-Statistik kann unter Zuhilfenahme der Kenngrößen

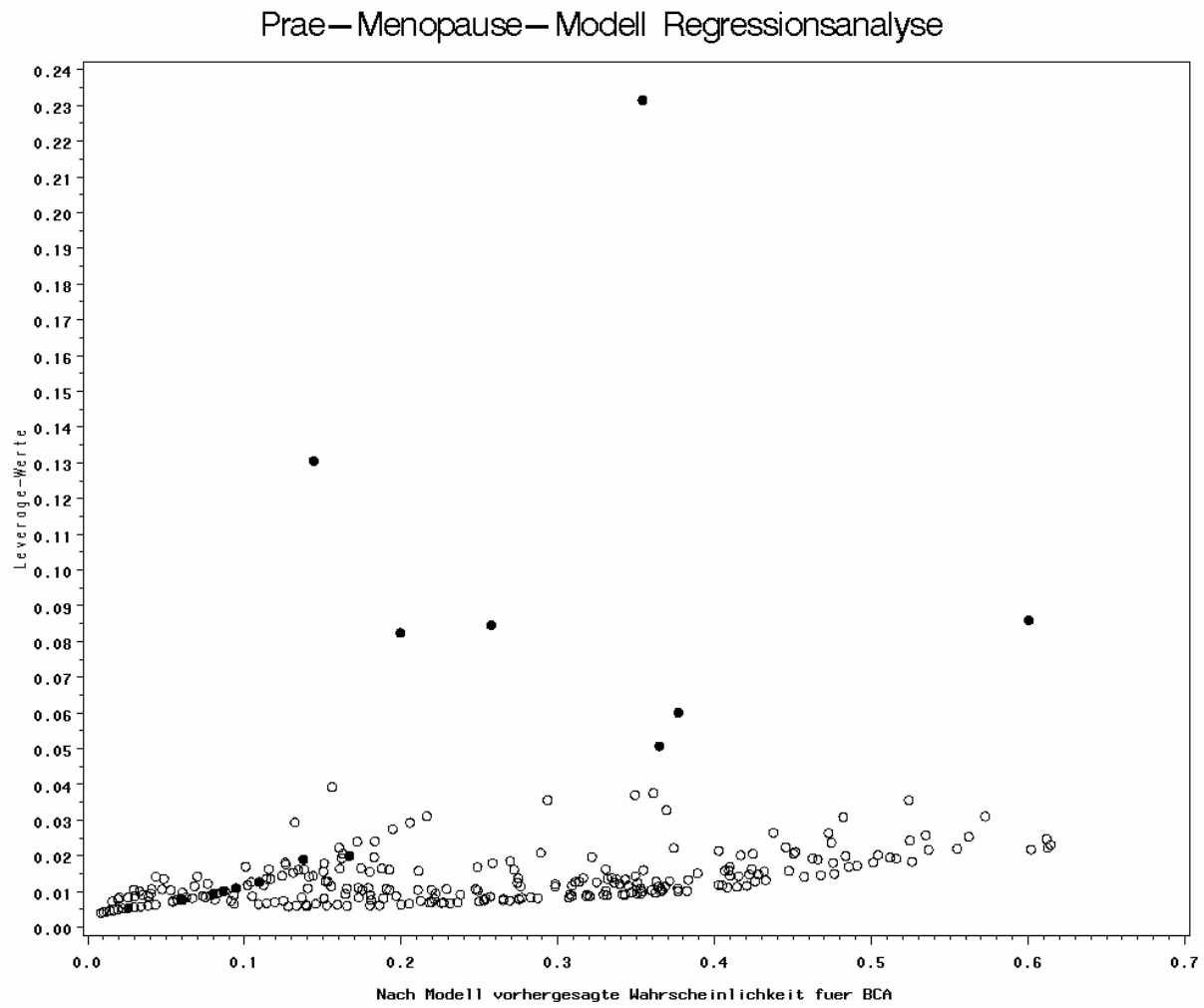
$$\Delta X^2_{(-j)} \approx (r_j)^2 \cdot \frac{1}{1-h_j} = \frac{(d_j - \hat{p}_j)^2}{\hat{p}_j \cdot (1-\hat{p}_j)} \cdot \frac{1}{1-h_j}$$

beurteilt werden. Für j=1,…,355 approximieren diese Größen die Veränderung, die die Pearson-Statistik unterliegt, wenn die j-te Studienteilnehmerin (bzw. Beobachtung) von der Analyse ausgeschlossen wird. Ein großer Wert weist auch hier darauf hin, dass bei der dazugehörigen Frau eine große Diskrepanz zwischen dem nach Modell zu erwartenden und beobachteten Krankheitszustand vorliegt, kann aber zu einem gewissen Grade auch dadurch bedingt sein, dass lediglich die Ausprägungen der Einflussvariablen sehr untypische Werte aufweisen. Obiger Formel kann allerdings entnommen werden, dass die Leverage-Werte, als Maß dafür, wie untypisch die Merkmalsausprägungen sind, bei diesem Diagnosemaß von geringer Bedeutung sind.

1) Untersuchung der Leverage-Werte

In einem ersten Schritt der Regressionsdiagnose werden die Leverage-Werte untersucht. Diese variieren im vorliegenden Fall zwischen 0,004 (Minimum) und 0,2315 (Maximum) und liegen durchschnittlich bei 0,014. Ein Streudiagramm, in welchem die Leverage-Werte gegen die geschätzten Erkrankungswahrscheinlichkeiten abgetragen werden (Grafik A.3.1), zeigt, dass nur 7 Leverage-Werte (ausgefüllte Punkte) nach oben abweichen, wohingegen alle Anderen deutlich näher (in einem Streifen) beieinander liegen.

<u>Grafik A.3.1</u>: Streudiagramm: Leverage-Werte gegen geschätzte Wahrscheinlichkeiten



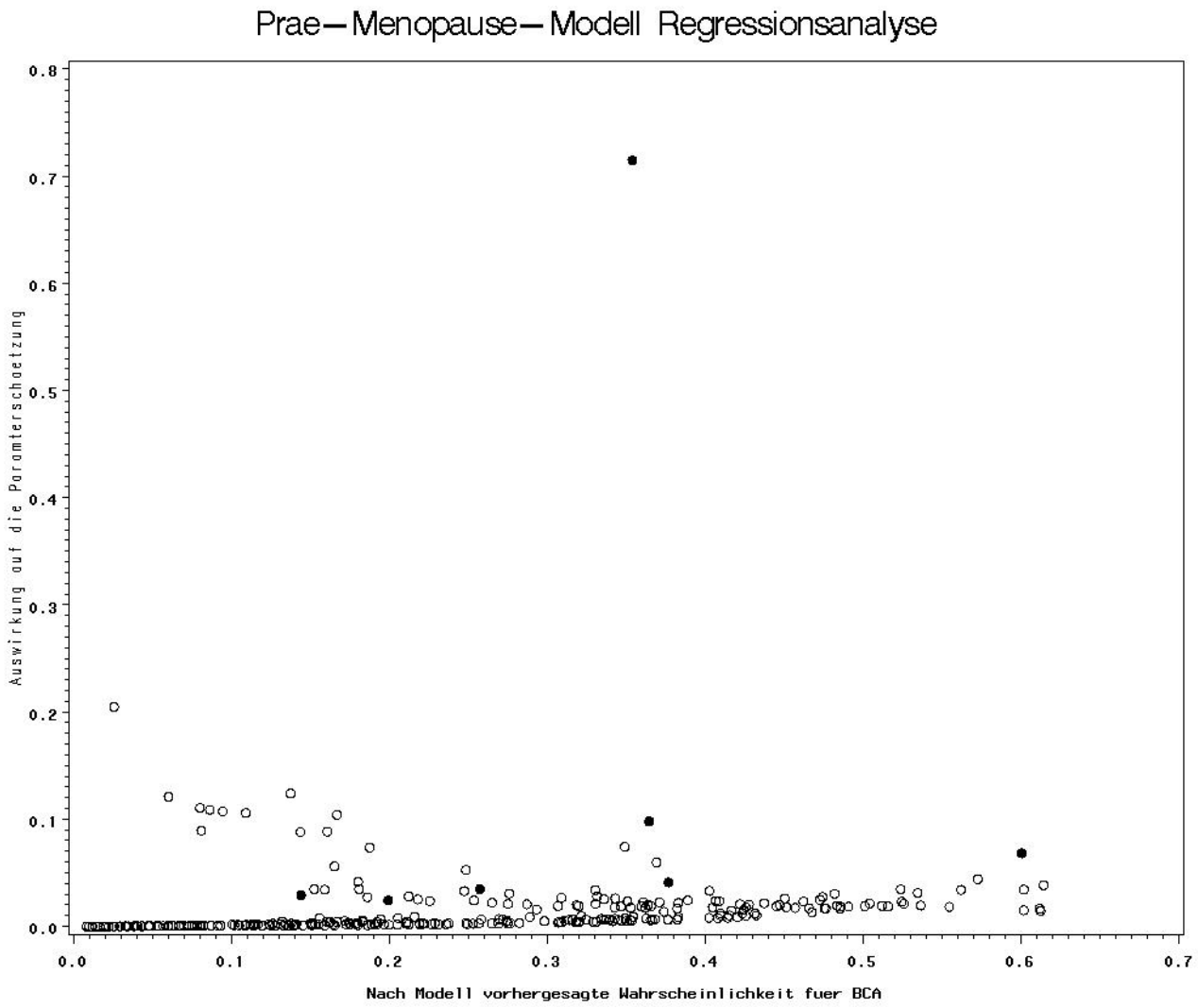
Weiterführende Untersuchungen zeigen, dass die zu den 7 großen Leverage-Werte gehörigen Studienteilnehmerinnen ein überdurchschnittlich hohes und damit für die Studienpopulation untypisches Lebensalter aufweisen. Während der Altersdurchschnitt der gesamten Prae-Menopause-Population bei 38,6 Jahren liegt, weisen diese 7 Frauen ein durchschnittliches Lebensalter von 72,1 Jahren auf. Die Ursache dafür, dass bei diesen Frauen trotz ihres hohen Lebensalters die Menopause noch nicht eingetreten ist, kann sowohl in einer biologischen Besonderheit (Abnormalität) als auch in einer fehlerhaften Datenerhebung lokalisiert sein.

Davon ausgehend, dass möglicherweise speziell diese 7 Studienteilnehmerinnen einen besonderen Einfluss auf die Parameterschätzungen nehmen, werden sie im nächsten Abschnitt, bei der graphischen Untersuchung der Sensitivität der Parameterschätzungen ($\Delta\hat{\beta}_{(.)}$-Diagnose), besonders hervorgehoben.

2) Untersuchung der Sensitivität der Parameterschätzung

Zur Identifikation der Studienteilnehmerinnen, deren Kombination von Einflussvariablen-Ausprägungen und Brustkrebszustand die Parameterschätzung stark beeinflusst, werden die $\Delta\hat{\beta}_{(.)}$ in einem Streudiagramm gegen die geschätzten Erkrankungswahrscheinlichkeiten $\hat{p}$ abgetragen. Die 7 Frauen, bei denen aufgrund ihres untypisch hohen Lebensalters besonders große Leverage-Werten resultieren, werden in dem Diagramm durch ausgefüllte Punkte gekennzeichnet. Das Streudiagramm (vgl. Grafik A.3.2) zeigt allerdings, dass es nicht ausschließlich die 7 Frauen mit großen Leverage-Werten bzw. überaus hohen Lebensaltern sind, die die Parameterschätzung stark beeinflussen.

Grafik A.3.2: Streudiagramm zur Identifikation einflussreicher Beobachtungen



Am auffälligsten ist der maximale Wert 0,71 des Diagnosemaßes $\Delta\hat{\beta}_{(.)}$, welcher bei der Frau mit dem größten Leverage-Wert von 0,2315 auftritt. Diese Studienteilnehmerin ist 78 Jahre alt und wurde erst im Alter von 38 Jahren zum ersten Mal (vollständig) schwanger. Schwangerschaftsprobleme gab es bei ihr nicht. Nach dem geschätzten Modell ergibt sich für diese brustkrebserkrankte Frau eine Erkrankungswahrscheinlichkeit von 0,354. Da somit keine allzu große Diskrepanz zwischen dem geschätzten und beobachteten Krankheitszustand vorliegt, ist der Einfluss dieser Beobachtung auf den geschätzten Parametervektor in erster Linie auf den überaus große Leverage-Wert zurückzuführen. Darüber hinaus kann dem Streudiagramm entnommen werden, dass im Bereich geringer geschätzter Erkrankungswahrscheinlichkeiten (<0,2) eine Häufung mittelgroßer $\Delta\hat{\beta}_{(.)}$-Werte vorliegt. Im Folgenden werden deshalb die 8 Frauen näher untersucht, bei denen das Diagnosemaß $\Delta\hat{\beta}_{(.)}$ über 0,1 liegt und die geschätzte Erkrankungswahrscheinlichkeit unter 0,2 liegt.

Für diese 8 Frauen gilt, dass sie brustkrebserkrankt sind, obwohl ihnen das geschätzte Modell nur eine geringe Erkrankungswahrscheinlichkeit unterstellt. Große Leverage-Werte liegen nicht vor. Entsprechend ist anzunehmen, dass der Einfluss dieser Beobachtungen primär darauf zurückzuführen ist, dass diese Frauen an Brustkrebs erkrankt sind, obwohl sich ihre Merkmalsausprägungen grundliegend von denen der anderen brustkrebserkrankten Frauen unterscheiden. Bemerkenswert ist in diesem Zusammenhang vor allen Dingen, dass nur eine dieser 8 Studienteilnehmerinnen schon eine vollständige Schwangerschaft hinter sich gebracht hat und, dass ebenfalls nur bei einer Schwangerschaftsprobleme aufgetreten sind.

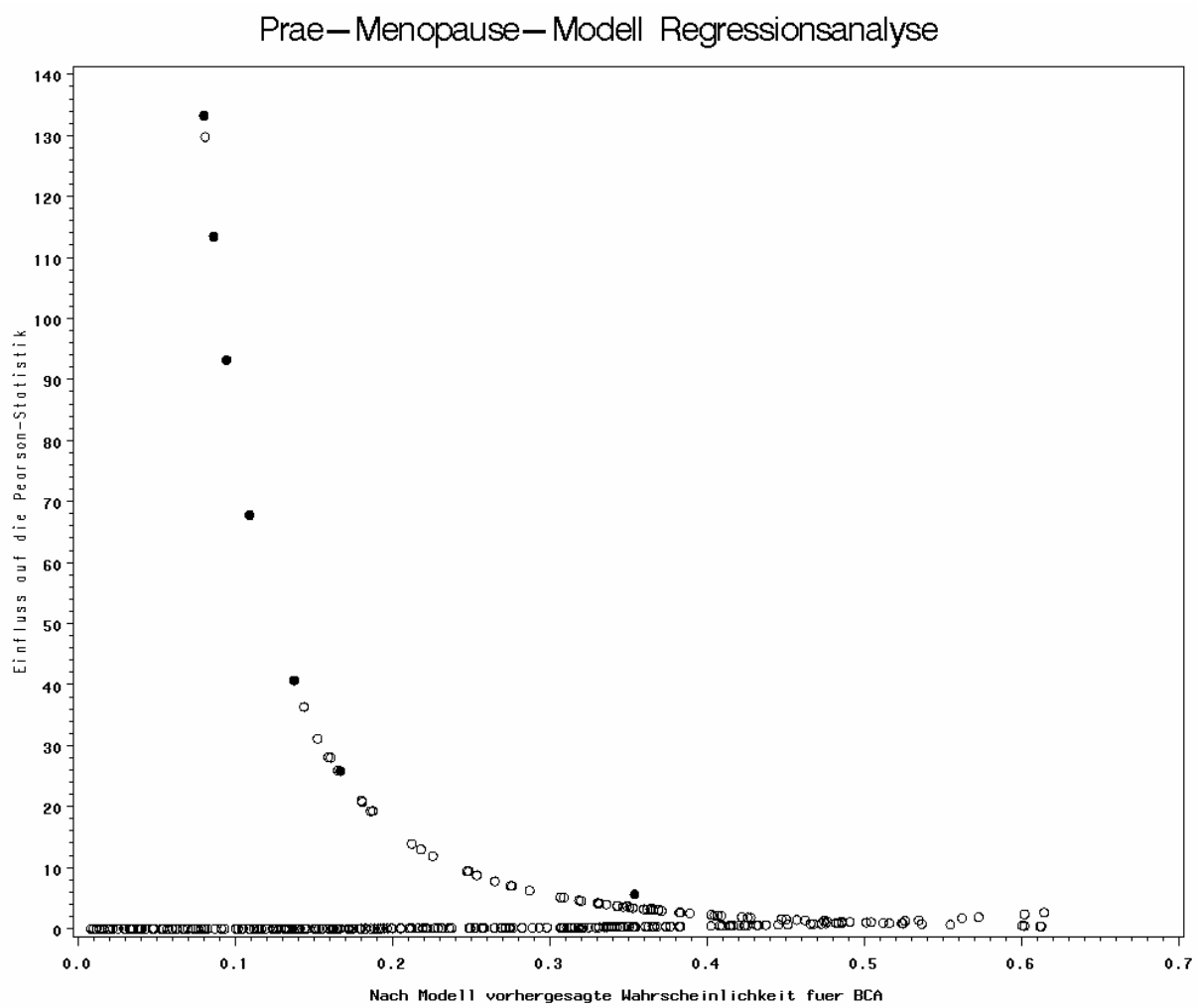
3) Untersuchung der Sensitivität der Pearson-Statistik

Abschließend wird graphisch untersucht, inwieweit die einzelnen Studienteilnehmerinnen Einfluss auf die Pearson-Statistik nehmen. Dazu werden die Diagnosemaße $\Delta X^2_{(.)}$ einem Streudiagramm gegen die geschätzten Erkrankungswahrscheinlichkeiten abgetragen. Die ausgefüllten Punkte kennzeichnen die Frauen, bei denen ein Diagnosemaß $\Delta\hat{\beta}_{(.)}$ größer 0.1 beobachtet wurde. Da sich für zwei Frauen deutlich größere Diagnosemaß $\Delta X^2_{(.)}$ ergeben als für die Anderen, werden die beiden dazugehörigen Punkte bei Erstellung des Streudiagramms - aus Gründen der Übersichtlichkeit – ausgeschlossen.

Die Koordinaten (x;y) der vernachlässigten Punkte sind (0.026; 1416,99) und (0.061; 244,80) und beide hätten ausgefüllt dargestellt werden müssen (siehe oben). Das Streudiagramm (Grafik A.3.3) zeigt, dass große Diagnosemaße vorwiegend im Bereich kleiner geschätzter Er-

krankungswahrscheinlichkeiten zu finden sind. Insbesondere befinden sich in diesem Bereich 6 der 7 sichtbaren ausgefüllten Punkte. Diese 6 Punkte gehören zu den bereits erwähnten Frauen, bei denen zwar kein großer Leverage-Wert aber ein großes Diagnosemaße $\Delta\hat{\beta}_{(.)}$ vorliegt. Genau wie die beiden vernachlässigten Frauen leiden sie an Brustkrebs, weisen allerdings dafür untypische Merkmalswerte auf.

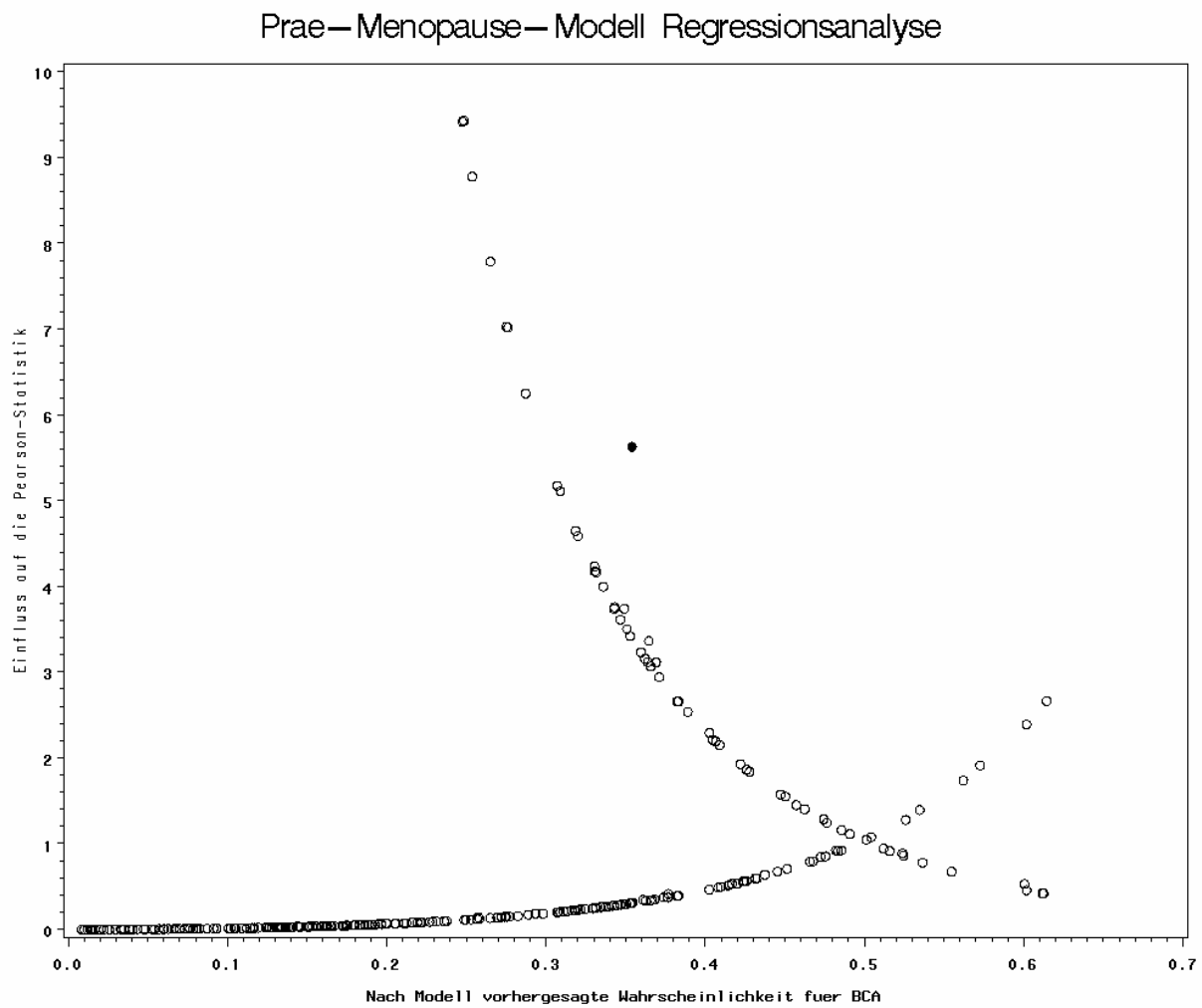
<u>Grafik A.3.3</u>: Streudiagramm: Erkrankungswahrscheinlichkeiten gegen $\Delta X^2_{(.)}$



Der im Streudiagramm mittig gelegene, ausgefüllte Punkt kennzeichnet die Studienteilnehmerin mit dem maximalen Leverage-Wert von 0,2315. Obwohl diese Beobachtung einen enormen Einfluss auf die Parameterschätzung nimmt ($\Delta\hat{\beta}_{(.)}$=0,71), beeinflusst sie die Pearson-Statistik nur unwesentlich.

Erst wenn ausschließlich die Punkte der 333 Frauen, bei denen das Diagnosemaß $\Delta X^2_{(.)}$ kleiner als 10 ist, in einem Diagramm dargestellt werden, weist das Streudiagramm (vgl. Grafik A.3.4) das typische Erscheinungsbild zweier gekreuzter Parabel-Äste auf.

<u>Grafik A.3.4</u>: Ausschnitt aus obigem Streudiagramm (Grafik A.3.3)



Der Parabel-Ast, der sich von links oben nach rechts unten durch das Streudiagramm zieht, korrespondiert zu den brustkrebserkrankten Frauen, wohingegen der zweite Ast (von links unten nach rechts oben) von den Punkten der nichtbrustkrebserkrankten Frauen gebildet wird. Da die meisten Leverage-Werte nur unwesentlich vom Mittelwert abweichen, so dass auch der Faktor $(1\text{-}h_j)^{-1}$ im Diagnosemaß $\Delta X^2_{(.)}$ näherungsweise konstant den Wert c aufweist, gelten annähernd die folgenden Beziehungen:

$$\Delta X^2_{(-j)} \approx \frac{(1-\hat{p}_j)^2}{\hat{p}_j \cdot (1-\hat{p}_j)} \cdot c \quad \text{(für die brustkrebserkrankten Frauen) und}$$

$$\Delta X^2_{(-j)} \approx \frac{\hat{p}_j}{1-\hat{p}_j} \cdot c \quad \text{(für die nichtbrustkrebserkrankten Frauen).}$$

Neben einigen unwesentlichen Abweichungen führt ausschließlich der Leverage-Wert von 0,2315 (ausgefüllter Punkt) zu einer deutlichen Abweichung von obiger funktionaler Beziehung.

Zusammenfassung und Auswertung der Ergebnisse der Regressionsdiagnose

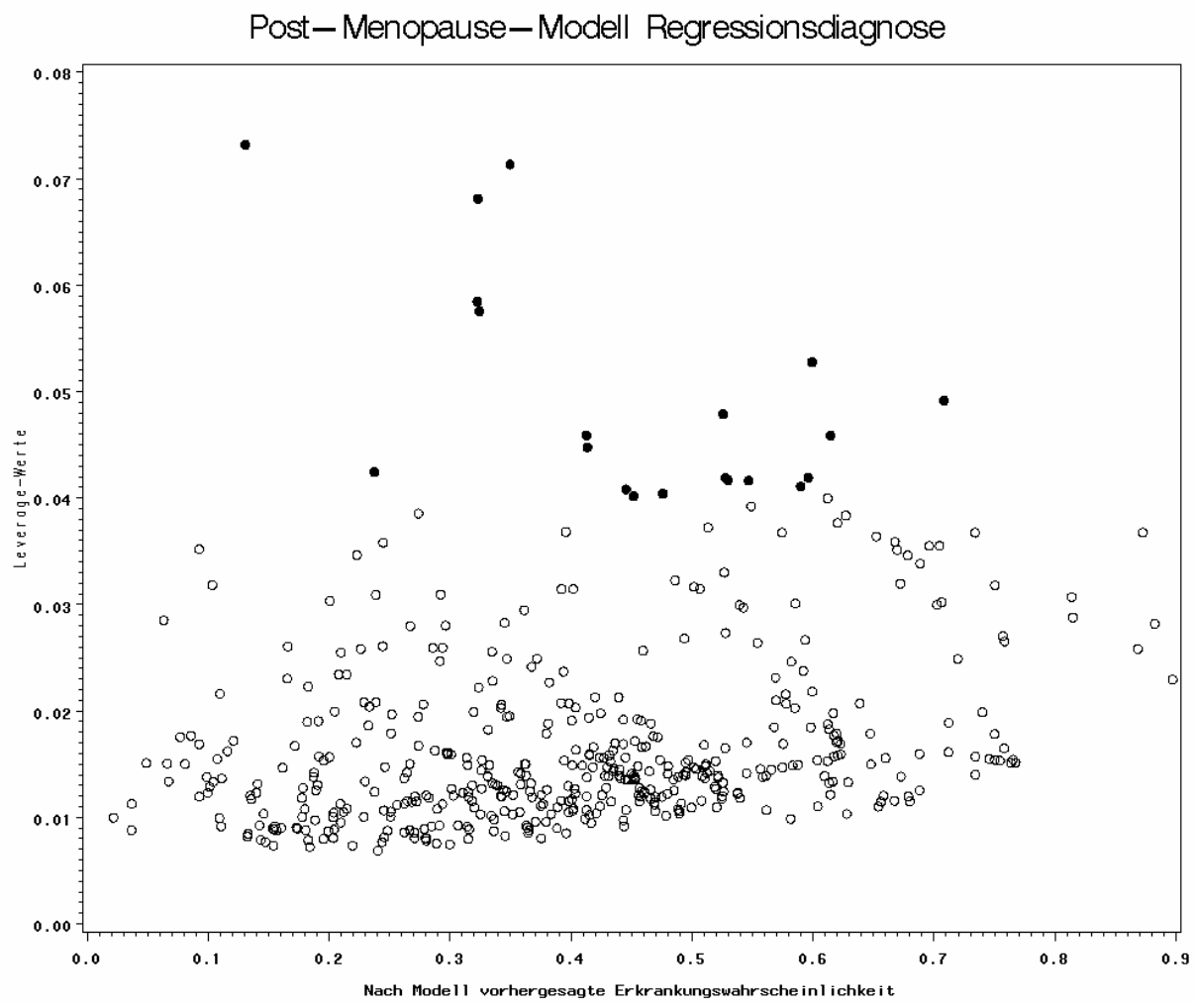
Große Leverage-Werte ergeben sich ausschließlich für die 7 überdurchschnittlich alten Frauen der Prae-Menopause. Die Diagnosemaße $\Delta\hat{\beta}_{(.)}$ zeigen jedoch, dass nur eine dieser Frauen einen unverhältnismäßig großen Einfluss auf die Parameterschätzung nimmt, so dass eine Berücksichtigung der überdurchschnittlich alten Studienteilnehmerinnen – unabhängig von der biologischen Plausibilität ihres Lebensalters - kein besonders großes Problem darstellt.
Problematischer hingegen erscheint, dass einige brustkrebserkrankte Frauen untypische Merkmalswerte (für Brustkrebs) aufweisen und daher sowohl Einfluss auf die Parameterschätzungen nehmen (große $\Delta\hat{\beta}_{(.)}$–Werte), als auch die Anpassung des Modells verschlechtern (große $\Delta X^2_{(.)}$-Werte). Allerdings stellen ihre Merkmalswerte - als solche betrachtet - keine substanzwissenschaftlichen Besonderheiten bzw. Abnormalitäten dar. Das heißt, sowohl die damit verbundene Verschlechterung der Modellanpassung als auch die resultierende Auswirkung auf die Parameterschätzung ist ausschließlich darauf zurückzuführen, dass trotz dafür untypischer Merkmalswerte Brustkrebskrankheiten aufgetreten sind. Da somit an der biologischen Plausibilität kein Zweifel besteht, kann eine Vernachlässigung dieser Frauen – zwecks Modellverbesserung - aus substanzwissenschaftlicher Sicht nicht gerechtfertigt werden. Es ist zu akzeptieren, dass auch Frauen, denen im Modell eine geringe Erkrankungswahrscheinlichkeit zukommt, eine Brustkrebskrankheit erleiden können.

Zusammenfassend lässt sich deshalb festhalten, dass die Ergebnisse der durchgeführten Regressionsdiagnosen nur eine geringfügige Unzulänglichkeit des Prae-Menopause-Modells aufzeigen. Lediglich bei einigen wenigen Frauen liegen Brustkrebserkrankungen vor, obwohl sie dafür untypische Werte aufweisen, so dass ihnen im Modell kleine Erkrankungswahrscheinlichkeiten zukommen. Der Einfluss, den speziell diese Beobachtungen auf den Parametervektor nehmen, ist in Anbetracht der Plausibilität der Beobachtungen zu akzeptieren.

A.3.2 Regressionsdiagnose – Post-Menopause-Modell

In diesem Unterkapitel wird eine Regressionsdiagnose für das Post-Menopause-Modell durchgeführt. Zunächst ist auch hier von Interesse, ob es Frauen gibt, bei denen untypische Ausprägungskombinationen vorliegen, so dass von ihnen ein besonderen Einfluss auf die Parameterschätzung ausgeht. In einem ersten Schritt werden deshalb die Leverage-Werte der 518 Studienteilnehmerinnen (bzw. Beobachtungen) in einem Streudiagramm gegen die geschätzten Erkrankungswahrscheinlichkeiten abgetragen. In dem Streudiagramm (vgl. Grafik A.3.5) wurden die 20 größten Leverage-Werte durch ausgefüllte Punkte markiert.

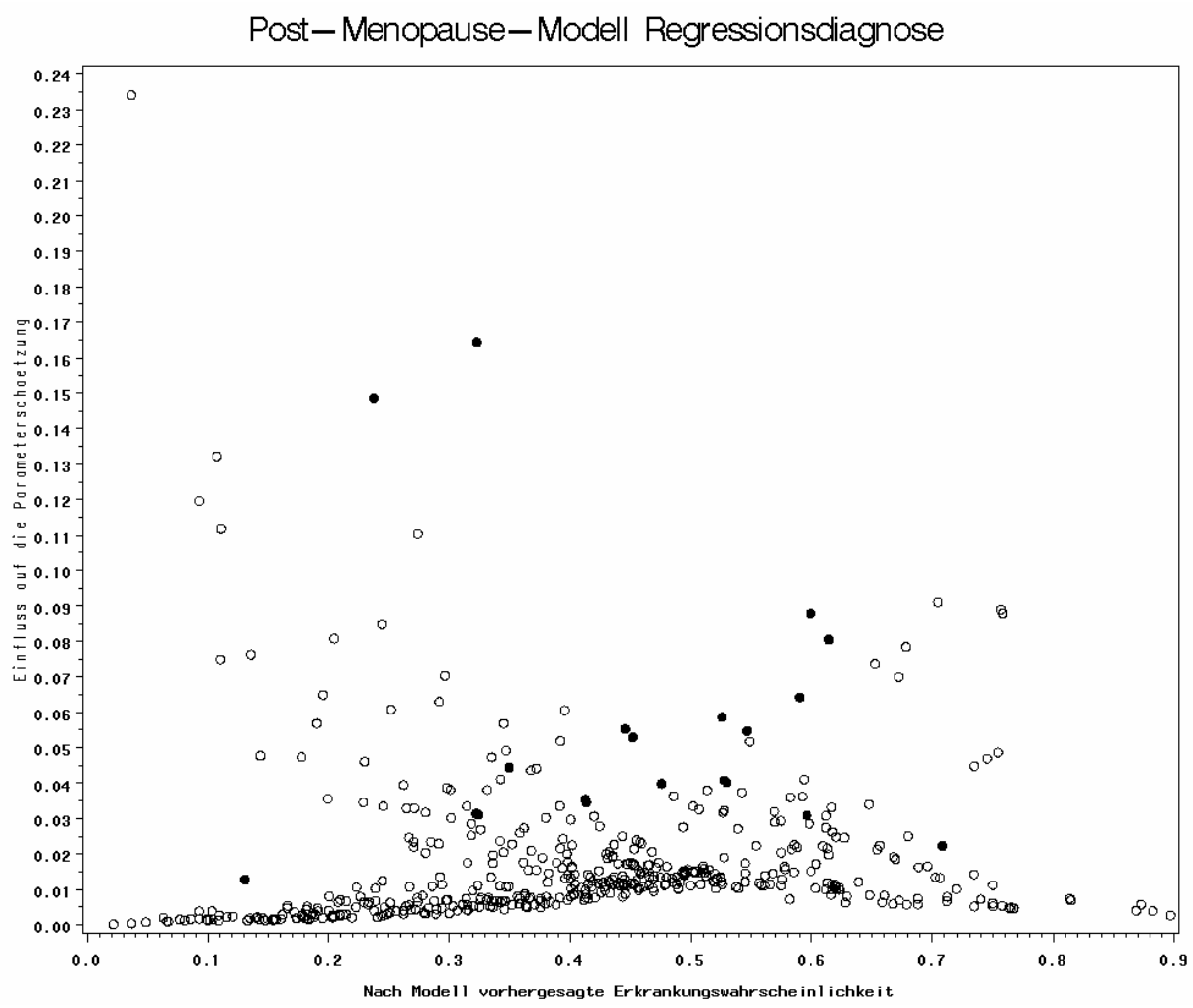
Grafik A.3.5: Leverage-Werte des Post-Menopause-Modells



Dem Streudiagramm kann entnommen werden, dass die Größen der Leverage-Werte nur unwesentlich mit den geschätzten Wahrscheinlichkeiten variieren. Die Einflussvariablen-Ausprägungen der Frauen mit den 20 größten Leverage-Werten (>0,04), werden im Folgenden genauer untersucht, um zu klären, ob diese eine Gemeinsamkeit aufweisen. Ein Blick auf

die Einflussvariablen-Ausprägungen dieser 20 Studienteilnehmerinnen zeigt allerdings, dass keine bestimmte Systematik zugrunde liegt. Offensichtlich führen bereits extreme Ausprägungen in einer stetigen Variablen („Lebensalter“ oder „Alter erste Schwangerschaft“) und auch das gleichzeitige Vorliegen mehrerer Expositionen zu großen Leverage-Werten. Der größte Leverage-Wert von 0,073 liegt zum Beispiel bei einer 101-jährigen Frau vor, deren Einflussvariablen-Ausprägungen ansonsten keine Besonderheit darstellen. Der zweitgrößte Leverage-Wert von 0,071 ergibt sich für eine 67-jährige Frau, die in ihrem Leben keine vollständige Schwangerschaft hinter sich gebracht hat. Obwohl die Einflussvariablen-Kombination bei keiner der 20 Studienteilnehmerinnen fragwürdig bzw. unrealistisch erscheint, stellt sich die Frage, ob die Parameterschätzung speziell von diesen 20 Frauen in besonderem Maße beeinflusst wird. Daher werden die zu diesen Beobachtungen (bzw. Frauen) gehörigen Punkte zwecks Wiedererkennung in dem Streudiagramm der $\Delta\hat{\beta}_{(.)}$ Diagnosegrößen gegen die geschätzten Erkrankungswahrscheinlichkeiten $\hat{p}$ als ausgefüllt dargestellt.

<u>Grafik A.3.6</u>: Streudiagramm zur Identifikation einflussreicher Beobachtungen

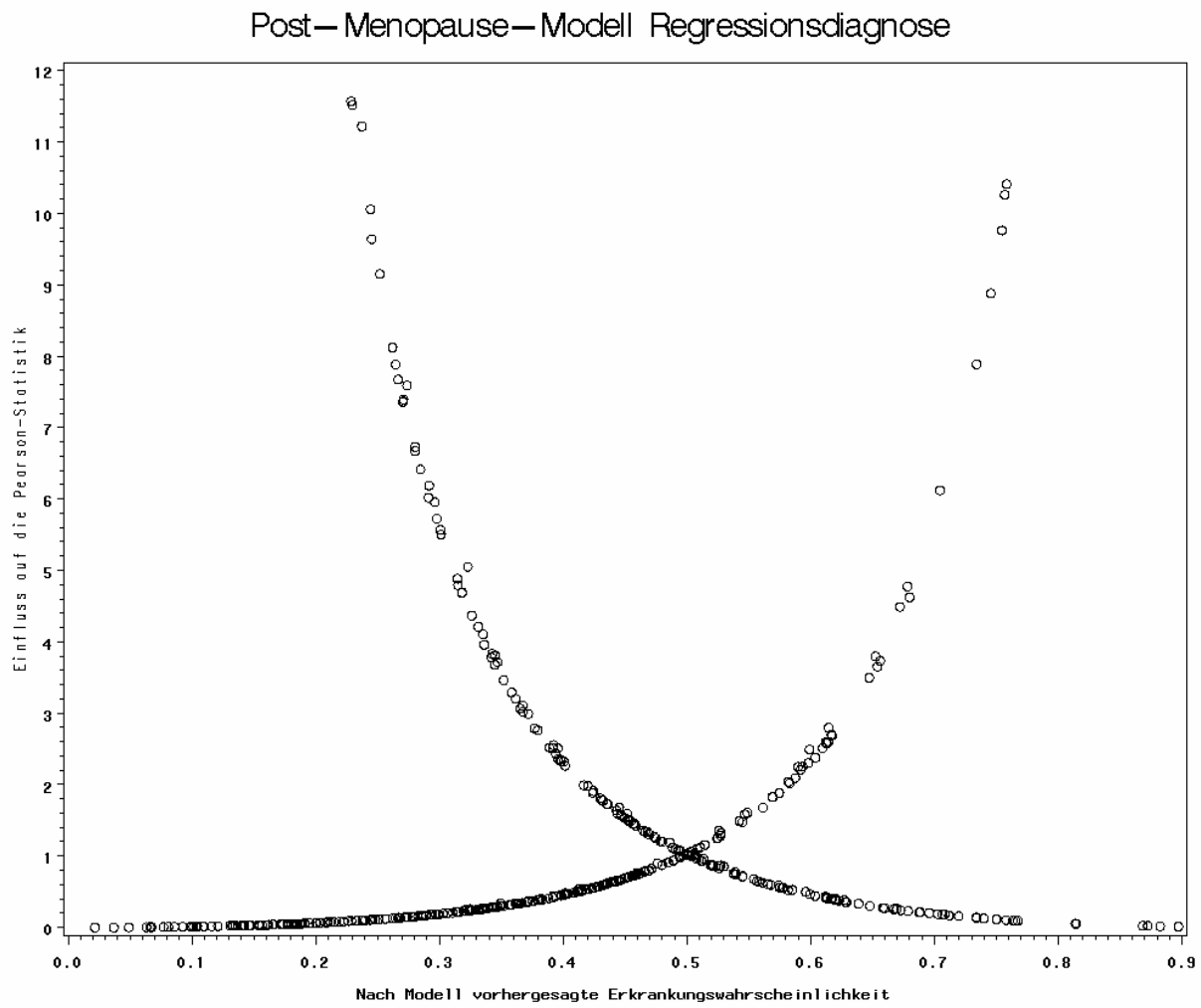


Das Streudiagramm Grafik A.3.5 zeigt, dass nicht nur die Beobachtungen mit den großen Leverage-Werten (ausgefüllte Punkte) einen überdurchschnittlich großen Einfluss auf die Parameterschätzung nehmen. Bemerkenswert in vor allen Dingen, dass der größte Einfluss auf die Parameterschätzung ausschließlich von Frauen mit kleinen geschätzten Erkrankungswahrscheinlichkeiten ausgeht. Eine genauere Untersuchung der 7 Frauen mit den größten $\Delta\hat{\beta}_{(.)}$ – Werten zeigt, dass es sich ausschließlich um Brustkrebsfälle handelt, denen aufgrund ihrer dafür untypischen Merkmalsausprägungen im Modell geringe Erkrankungswahrscheinlichkeiten zukommen. Ein bestimmter Zusammenhang ist dabei auch hier nicht zu erkennen. Den extremsten Einfluss auf die Parameterschätzung ($\Delta\hat{\beta}_{(.)}$=0,234) nimmt zum Beispiel eine 38-jährige Studienteilnehmerin, die bereits im Alter von 16 Jahren zum ersten Mal schwanger wurde, insgesamt 7 lebendige Kinder geboren hat, und bei der ansonsten kein Risikofaktor für Brustkrebs (Schwangerschaftsabbruch, operativ herbeigeführte Menopause oder Verlust des Ehemannes) vorliegt. Der zweit- und drittgrößte Einfluss geht von brustkrebserkrankten Frauen aus, für die sich im Modell zwar höhere Erkrankungswahrscheinlichkeiten ergeben, bei denen dafür aber aufgrund untypischer Merkmalswerte-Kombinationen große Leverage-Werte resultieren. Bemerkenswert ist zudem, dass die 101-jährige Studienteilnehmerin mit dem größten Leverage-Wert von 0,073 – (ausgefüllter) Koordinatenpunkt (0.130;0.013) - einen eher geringen Einfluss auf die Parameterschätzung nimmt.
Insgesamt besteht somit kein Grund zu der Annahme, dass einzelne Beobachtungen mit bestimmten Ausprägungen der Einflussvariablen einen besonderen Einfluss auf die Parameterschätzung nehmen.

Abschließend wird noch untersucht, ob bestimmte Beobachtungen besonders schlecht vom Modell beschrieben werden und daher die globale Modellanpassung beeinträchtigen.
Zur Identifikation solcher Beobachtungen können die Diagnosemaße $\Delta X^2_{(.)}$ in einem Streudiagramm gegen die geschätzten Erkrankungswahrscheinlichkeiten abgetragen werden. Grafik A.3.7 zeigt einen Ausschnitt aus diesem Diagramm, in welchem die Punkte von 12 Beobachtungen (bzw. Studienteilnehmerinnen) aufgrund zu großer y-Koordinaten ($\Delta X^2_{(.)}$-Werte) ignoriert wurden. Nach Ausschluss dieser 12 Frauen weist das Streudiagramm das typische Erscheinungsbild zweier gekreuzter Parabel-Äste auf.
Eine genauere Untersuchung der 12 unberücksichtigten Frauen zeigt, dass es sich um Brustkrebsfälle handelt, denen im Modell aufgrund dafür untypischer Merkmalswerte eine geringe Erkrankungswahrscheinlichkeit zukommt. Keine der 12 Frauen gehört zu den 20 Studienteil-

nehmerinnen mit den größten Leverage-Werten (>0.04) und nur 4 der 12 Frauen waren bereits durch große $\Delta\hat{\beta}_{(.)}$ –Werte (>0.1) aufgefallen.

<u>Grafik A.3.2</u>: Streudiagramm: Erkrankungswahrscheinlichkeiten gegen $\Delta X^2_{(.)}$



Diese Gesichtspunkte und eine eingehendere Betrachtung der Merkmalsausprägungen zeigt, dass es sich um Studienteilnehmerinnen handelt, deren Merkmalswerte zwar für eine Brustkrebserkrankungen untypisch sind, in der Gesamtpopulation jedoch keine Besonderheit darstellen. Die $\Delta X^2_{(.)}$-Werte ergeben sich aufgrund der geringen Leverage-Werte im wesentlichen in Abhängigkeit von den („zu gering“) geschätzten Erkrankungswahrscheinlichkeiten. Sie liegen für 5 Frauen im Bereich von 12 bis 20, für 6 Frauen im Bereich von 35 bis 100 und für eine Frau ergibt sich sogar ein $\Delta X^2_{(.)}$-Wert von 691. Bei Letzterer wurde bereits der extremste $\Delta\hat{\beta}_{(.)}$ –Wert von 0,23 beobachtet. Es handelt es sich um die 38-jährige Frau, die be-

reits im Alter von 16 bereits zum ersten Mal schwanger wurde, insgesamt 7 Kinder zur Welt gebracht hat und, bei der auch ansonsten kein Risikofaktor für Brustkrebs vorliegt. Für diese ergibt sich trotz ihrer Brustkrebserkrankung im Modell nur eine äußerst geringe Erkrankungswahrscheinlichkeit von 0,037, was zu einem unverhältnismäßig großen Pearson-Residuum führt.

Zusammenfassung der Ergebnisse der Regressions-Diagnose (Post-Menopause)

Zusammenfassend kann festgehalten werden, dass bei der Regressionsdiagnose des Post-Menopause-Modells keine besonderen Auffälligkeiten beobachtet wurden.

Große Leverage-Werte ergeben sich für verschiedene Kombinationen von Ausprägungen der Einflussvariablen, wobei schon einzelne extreme Ausprägungen und/oder das Vorliegen mehrerer Expositionen zu großen Werten führen. Sowohl die größte Beeinflussung der Parameterschätzwerte als auch die größte Beeinträchtigung der Modellanpassung (Vergrößerung der Pearson-Quadratsumme) geht von Frauen aus, bei denen trotz dafür untypischer Merkmalswerte Brustkrebserkrankungen aufgetreten sind. Beobachtungen bzw. Studienteilnehmerinnen mit großen Leverage-Werten bzw. mit – bezogen auf die Gesamtpopulation - untypischen Merkmalswerten tragen im Vergleich dazu nur unwesentlich zur Vergrößerung der Pearson-Residualsumme und/oder Beeinflussung der Parameterschätzwerte bei.

Anhang 4: Datenmaterial

In diesem vierten Anhang wird stichwortartig beschrieben, welche ausführlicheren epidemiologischen Informationen von 1198 Studienteilnehmern zur Verfügung gestellt wurden. In der folgenden Aufzählung wird jedes erhobene Merkmal namentlich genannt, wobei sofern notwendig in Klammern in Abhängigkeit vom Messniveau noch Angaben zur verwendeten Maßeinheit oder Angaben zur Klassifizierung folgen. Ausführlichere Beschreibungen zu den (mutmaßlich für die Brustkrebskrankheit) relevanten Merkmalen können Unterkapitel 5.3 entnommen werden.

1) **Geschlecht**
   (männlich, weiblich)
2) **Geburtsdatum**
   2.2) Todesdatum
   (sofern Studienteilnehmer zum Zeitpunkt des Studienbeginns bereits verstorben war)
2) **Familienstand**
   (ledig, verheiratet, geschieden oder verwitwet)
3) **Fettleibigkeitsleiden**
   (Ja/ Nein)
4) **Schulbildungsjahre**
   (1-12 entsprechen den absolvierten Schuljahren an einer Primary- bzw. Secondary-School; 13-16 korrespondieren zu zusätzlichen Highschool bzw. Universitätsjahren; 17 entspricht einer akademischen Ausbildung, die über 16 Jahre gedauert hat („postgraduate education“))
5) **Familiäres Jahreseinkommen**
   (unterteilt in 6 Einkommensklassen)
6) **Körpergröße**
   (in [Zoll=2,54cm])
7) **Körpergewicht**
   (in [Pfund=453,6g])
8) **Anzahl der Schwangerschaften,**
   (die zu Lebend- oder Totgeburten geführt haben)
   8.2) Alter zum Zeitpunkt der ersten Schwangerschaft
   (ausschließlich Schwangerschaften, die zu Lebend- oder Totgeburt geführt haben)
   8.3) Alter zum Zeitpunkt der letzten Schwangerschaft
   (ausschließlich Schwangerschaften, die zu Lebend- oder Totgeburt geführt haben)

8.4) Durchschnittliche Anzahl Jahre zwischen zwei Schwangerschaften

(ausschließlich Schwangerschaften, die zu Lebend- oder Totgeburt geführt haben)

9) Anzahl Lebendgeburten

10) Anzahl Totgeburten

11) Anzahl Schwangerschaftsabbrüche

12) Anzahl Fehlgeburten

13) Anzahl anderer Schwangerschaftsergebnisse

14) Anzahl Söhne

15) Anzahl Töchter

16) Anzahl Kinder, die brustgestillt wurden

17) Anzahl Monate, in denen Kinder brustgestillt wurden

18) Probleme beim Stillen

(Ja/ Nein)

19) Menarche-Alter

(in Lebensjahren)

20) Regelmäßigkeit der Periode

(regelmäßig, gelegentlich unregelmäßig, fast immer unregelmäßig, immer unregelmäßig)

21) Durchschnittliche Periodendauer

(in Tagen)

22) Menopause-Status

(Prae-Menopause, Klimakterium, Post-Menopause (biologisch erreicht) oder Post-Menopause (operativ herbeigeführt)

22.2) Menopause-Alter

(in Lebensjahren)

23) Hysterektomie-Operation durchgeführt

(Ja/ Nein)

23.2) Alter zum Zeitpunkt der Hysterektomie-Operation

(in Lebensjahren)

24) Anzahl operativ entfernter Eierstöcke

(0, 1 oder 2)

24.2) Alter zum Zeitpunkt der (ersten) Ovar-Entfernung

(in Lebensjahren)

25) Anti-Baby-Pille eingenommen

(Ja/ Nein)

25.2) Alter, als zum ersten Mal Anti-Baby-Pille eingenommen

(in Lebensjahren)

25.3) Anzahl Jahre, in denen die Anti-Baby-Pille eingenommen wurde

25.4) Zeitspanne vor der ersten Geburt, in der Anti-Baby-Pille genommen wurde

(in Jahren)

26) weibliche Hormone eingenommen

(Ja/ Nein)

26.2) Alter, als zum ersten Mal weibliche Hormone eingenommen wurden

(in Lebensjahren)

26.3) Anzahl Jahre, in denen weibliche Hormone eingenommen wurden

27) männliche Hormone eingenommen

(Ja/ Nein)

27.2) Alter, als zum ersten Mal männliche Hormone eingenommen wurden

(in Lebensjahren)

27.3) Anzahl Jahre, in denen männliche Hormone eingenommen wurden

28) „NME“-Hormone eingenommen

(Ja/ Nein)

28.2) Alter, als zum ersten Mal „NME“-Hormone eingenommen wurden

(in Lebensjahren)

28.3) Anzahl Jahre, in denen „NME“-Hormone eingenommen wurden

29) andere Hormone eingenommen

(Ja/ Nein)

29.2) Anzahl Jahre, in denen andere Hormone eingenommen wurden

30) Wurde eine Tubensterilisation durchgeführt

(Ja/ Nein)

30.1) Alter zum Zeitpunkt der Tubensterilisation

(in Lebensjahren)

31) Rauchverhalten

(Raucher/ Nichtraucher)

31.2) Alter, als mit dem Rauchen begonnen wurde

(in Lebensjahren)

31.3) Zeitspanne, in der Zigaretten konsumiert wurden

(in Jahren)

31.4) durchschnittlicher Zigarettenkonsum

(Anzahl Zigaretten pro Tag)

32) Regelmäßiger Alkoholkonsum

(Ja/ Nein)

32.2) Zeitspanne, in der regelmäßig getrunken wurde

(in Jahren)

32.3) durchschnittliche getrunkene Biermenge (Gläser [200 ml]pro Tag)

32.4) durchschnittliche getrunkene Weinmenge (Gläser [30 ml] pro Tag)

32.4) durchschnittliche getrunkene Likörmenge (auf 30ml genau)

33) Brustkrebsfälle in der Familie

33.1) Mutter (Ja/ Nein)

33.2) Vater (Ja/ Nein)

33.3) Schwestern (Anzahl)

33.4) Brüder (Anzahl)

33.5) Halbschwestern (Anzahl)

33.6) Halbbrüder (Anzahl)

33.7) Töchter (Anzahl)

33.8) Söhne (Anzahl)

33.9) Großmutter (mütterlicherseits) (Ja/ Nein)

33.10) Großvater (mütterlicherseits) (Ja/ Nein)

33.11) Großmutter (väterlicherseits) (Ja/ Nein)

33.12) Großvater (väterlicherseits) (Ja/ Nein)

34) Taillenumfang

(in [Zoll=2,54cm])

35) Hüftumfang

(in [Zoll=2,54cm])

Anhang 5: Verzerrungen beim Indikatorvariablen-Verfahren

In diesem fünften Anhang wird anhand theoretischer Überlegungen motiviert, dass im Rahmen multipler logistischer Regressionsmodelle das Indikatorvariablen-Verfahren im Umgang mit fehlenden Werten (vgl. 4.3.9) bei Vorliegen eines Missing-Randomly-At-Outcome-Mechanismus durchaus zu Ergebnisverzerrungen führen kann.

Bei Vorliegen eines Missing-Randomly-At-Outcome-Mechanismus ist zu erwarten, dass der zusätzlichen Indikatorvariablen für fehlende Werte ein inhaltlich nicht interpretierbarer Erklärungswert für die Zielvariable zukommt. Dieser Erklärungswert resultiert daraus, dass die zusätzliche Indikatorvariable fehlende Werte kennzeichnet, die ihrerseits wiederum mit einer bestimmten Ausprägung der Zielvariablen assoziiert sind, und stellt somit ein Artefakt der Datenstruktur dar.

Hinsichtlich der Schätzung von Chancenverhältnissen zwischen den regulären Kategorien bzw. Werten des zugehörigen Merkmals X stellt dieser Erklärungswert nicht unmittelbar ein Problem dar. Bedingt durch die Indikatorvariable „fehlender Wert“ I=I(X) wird der Datensatz gewissermaßen in zwei Teile aufgespaltet. Die Frauen mit fehlenden Merkmalswerten bilden den ersten Unterdatensatz (I=1) und alle Frauen mit regulären Merkmalsausprägungen den zweiten Unterdatensatz (I=0). Lediglich der zusätzliche Parameter β_I charakterisiert das Chancenverhältnis zwischen der Kategorie „fehlender Wert“ (I=1) und einer Referenzausprägung (I=0,X=0) und stellt somit die einzige Verbindung zwischen den beiden Datenteilen her. Der oder die anderen Parameter, die die Chancenverhältnisse zwischen den regulären Ausprägungen von X charakterisieren, ergeben sich ausschließlich aus dem zweiten Datenteil und repräsentieren daher, die tatsächlich beobachteten Zusammenhänge zwischen den regulär beobachteten Merkmalsausprägungen und der Zielvariablen.

In Bezug auf die Schätzung der Einflüsse anderer Merkmale $Y_1,\ldots,Y_m$ stellt die zusätzliche Indikatorvariable I=I(X) im logistischen Regressionsmodell hingegen schon ein Problem dar. Dieses ergibt sich daraus, dass im Modell nicht zwischen den regulären Einflussvariablen $Y_1,\ldots,Y_n$ und der Kunstvariable I, deren Erklärungswert für die Zielvariable nur ein Artefakt des Missing Randomly at Outcome-Mechanismus darstellt, unterschieden wird. Diese Einflussüberlagerung hat zur Konsequenz, dass bei simultaner Betrachtung aller Variablen nicht ausgeschlossen werden kann, dass der tatsächlich vorliegende Einfluss der Variablen $Y_1,\ldots,Y_n$ durch den inhaltlich nicht interpretierbaren Erklärungswert von I verzerrt wird. Bedingt durch die Verzerrung der anderen Parameter kann es somit zumindest indirekt auch zu einer Verzerrung des Parameters β_X kommen.

LITERATURVERZEICHNIS

Bleich, S. , Kropp, S. und Stipriaan, H. (1995): Kurzlehrbuch: Allgemeine Pathologie. Schattauer, Stuttgart.

Booney, G.E., Chen, G., Demanais, F.M., Kissling, G., Laing, A.E. und Williams, R. (1993): Breast cancer risk factors in African-American women: Howard University Tumor Registry Experience.
Journal of then National Medical Association; 85.

Bosch, K. (1995): Elementare Einführung in die Wahrscheinlichkeitstheorie.
Vieweg Studium, Berlin.

Breslow, N.E. und Day, N.E. (1980): Statistical methods in cancer research, Vol.1: The analysis of case-control studies.
IARC Scientific Publications No. 32, Lyon.

Büning, H. und Trenkler, G. (1994): Nichtparametrische statistische Methoden.
Walter de Gryter, Berlin.

Cochran, W.G. (1954): Some methods für strengthening the common $\chi 2$ tests.
Biometrics, 10.

Cohen, B.H., Beaty, T.H. und Khoury, M.J. (1993): Fundamentals of genetic epidemiology.
Oxford University Press, New York.

Fahrmeir, L., Hamerle, A. und Tutz, G. (1995): Multivariate statistische Verfahren.
Walter de Gryter.

Greskötter, K.-R. (1996): Pathologie/ Klinische Medizin systematisch. Band 2.
Uni-Med Verlag, Lorch/ Württemberg.

Grundmann, E. (1994): Einführung in die allgemeine Pathologie.
Urban & Fischer, München.

Hartung, J. (1995): Statistik.
Oldenbourg.

Hosmer, D.W., Lemeshow, S. (1980): A goodness-of-fit test for the multiple logistic regression model.
Communications in Statistics; A10.

Hosmer, D.W., Lemeshow, S. (1989): Applied logistic regression
John Wiley & Sons, New York.

Ibrahim, J.G. (1990): Incomplete data in generalized linear models.
Journal of the American Statistical Association, 85.

Kale B.K. (1962): On the solution of likelihood equatations by iteration processes.
Biometrika; 15.

Kelsey, J.L. (1993): Breast cancer epidemiology: Summary and future directions.
Epidemiologic Reviews; 15(1).

Kelsey, J.L. und Horn-Ross P.L (1993): Breast cancer: Magnitude of the problem and descriptive epidemiology.
Epidemiologic Reviews; 15(1).

Klein, J.P. und Moeschberger, M.L. (1997): Survival analysis.
Springer, New York.

Kleinbaum, D.G. (1998): Logistic regression.
Springer, New York.

Kleinbaum, D.G., Kupper, L.L. und Muller, K.E.: (1982): Epidemiologic Research: Principles and Quantitative Methods.
Van Nostrand Reinhold, New York.

Kreienbrock, L. und Schach S., (2000): Epidemiologische Methoden.
Spektrum, Heidelberg.

Künzi, H.P., Krelle, W. und Randow, R. (1962): Nichtlineare Programmierung.
Springer, Berlin.

Little, R.J.A., Rubin, D.B. (1987): Statistical analysis with missing data.
John Wiley & Sons, New York.

Pepe, M.S. und Fleming, T.R. (1991): A nonparametric method for dealing with mis-measured covariate data.
Journal of the American Statistical Association, 86.

Pregibon, D. (1981): Logistic regression diagnostics.
Annals of Statistics; 9.

Pschyrembel (1998): Klinisches Wörterbuch.
Walter de Gryter, Berlin.

Rothmann, K.J. und Greenland, S. (1998): Modern Epidemiology.
Lippincott-Raven, Philadelphia.

Rubin, D.B. (1976): Inference and missing data.
Biometrika, 63.

Schoenfeld, D.A. (1982): Logistic Models. In Mikey, V. und Stanley, K.E.: Statistics in medical research.
Wiley and Sons, New York.

Schumacher, M. und Vach, W. (1993): Logistic regression with incompletely observed categorical covariates.
Biometrika, 80.

Vach, W. und Blettner, M. (1991): Biased estimation of the odds ratio in case-control studies due to the use of ad-hoc methods of correcting for missing values in confounding variables.
American Journal of Epidemiology, 134.

Witting, H. und Nölle, G. (1970): Angewandte mathematische Statistik.
Teubner, Stuttgart.

Yarnold, J.K. (1970): The minimum expectation in χ^2 goodness of fit tests and accuracy of approximations of the null distributions.
Journal of the American Statistical Association, 65.

Yuen Fung, K. und Wrobel, A. (1989): The treatment of missing values in logistic regression.
Biometrical Journal, 31.

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig bzw. nur unter Zuhilfenahme der im Literaturverzeichnis genannten Quellen erstellt habe.

Marco Grzegorczyk, Februar 2003