ROC-Methodologie

Johannes Bracher

Seminararbeit zu
„Aktuelle Anwendungsgebiete der Statistik“

Dozent: Prof. Dr. Augustin

Inhaltsverzeichnis

Kapitel 1

Klassifikationsverfahren

In den verschiedensten Bereichen stößt man auf das Problem, Untersuchungseinheiten von bestimmten gegebenen Informationen bzw. Kovariablen ausgehend zwei oder mehr Gruppen zuordnen zu müssen: Rating-Verfahren von Banken sollen kreditwürdige Kunden von nicht kreditwürdigen unterscheiden, also eine möglichst gute Prognose abgeben (Henking et al., 2006). Diagnostische Tests in der Medizin zielen darauf ab, zwischen Kranken und Gesunden zu differenzieren. Die Ausführungen in diesem Kapitel orientieren sich an diesem Anwendungsbereich und bauen im Wesentlichen auf Wehberg et al. (2007) auf.

Betrachtet wird ein skalarer Wert x_i (*Score*). Dieser ist eine Funktion der Kovariablen, die bei der Untersuchung erhoben wurden, etwa eine Linearkombination. Von ihm ausgehend sollen Rückschlüsse auf den Status S_i einer Person i gezogen werden. Damit ein Verfahren verwertbare Ergebnisse liefern kann muss sich die Verteilung von x in den beiden Gruppen Kranke ($S = 1$) und Gesunde ($S = 0$) unterscheiden.

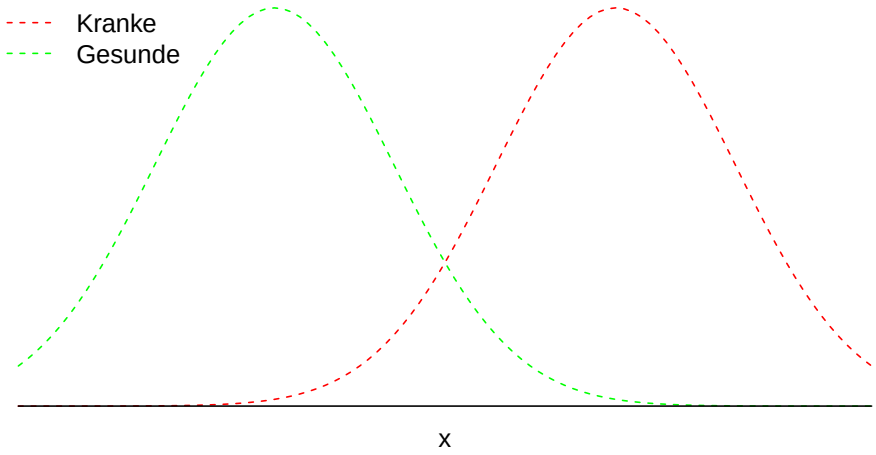

Abbildung 1.1: Dichten in zwei Gruppen unter Normalverteilungsannahme.

Um eine Diagnose R abgeben zu können, bedarf es noch einer Zuordnungsregel. Man legt hierfür einen Schwellenwert oder Cutpoint c für x fest, ab dem ein Patient als krank

($R = 1$) diagnostiziert wird. Im Gegensatz zu vielen Verfahren aus dem Kreditrating, wo hohe Scorewerte für hohe Kreditwürdigkeit ($R = 0$) sprechen und Niedrige auf einen Ausfall hindeuten ($R = 1$), gilt im Folgenden also stets:

$$\begin{aligned} x_i > c &\iff R_i = 1 \\ x_i \leq c &\iff R_i = 0 \end{aligned} \tag{1.1}$$

Wenn es Überlappungen der Verteilungen von x in den beiden Gruppen gibt, x die Gesunden ($S = 0$) und Kranken ($S = 1$) also nicht vollständig trennen kann, kommt es zwangsläufig zu Fehlklassifikationen. Man unterscheidet:

- **Fehler erster Art:** Negatives Testergebnis, obwohl die Person krank ist, also $x \leq c$ und damit $R = 0$, obwohl $S = 1$
- **Fehler zweiter Art:** Positives Testergebnis, obwohl die Person gesund ist, also $x > c$ und damit $R = 1$, obwohl $S = 0$

Die Wahrscheinlichkeiten korrekter Klassifikationen bedingt auf den tatsächlichen Status werden verwendet, um die Qualität eines Diagnoseverfahrens mit festem Cutpoint zu beschreiben:

- **Sensitivität:** Wahrscheinlichkeit, eine kranke Person korrekt als krank zu diagnostizieren: $P(x > c|S = 1)$. Entspricht der Vermeidung des Fehlers erster Art.
- **Spezifität:** Wahrscheinlichkeit, eine gesunde Person korrekt als gesund zu diagnostizieren: $P(x \leq c|S = 0)$. Entspricht der Vermeidung des Fehlers zweiter Art.

Es ist anzumerken, dass es im medizinischen Bereich, zur Berechnung von Sensitivität und Spezifität meist eines zweiten, sehr guten Diagnosemechanismus bedarf, der dazu dient, den tatsächlichen Status festzustellen. Man bezeichnet dieses Verfahren dann auch als Goldstandard. Wie leicht nachzuvollziehen ist, gibt es einen Tradeoff zwischen Sensitivität und Spezifität: Niedrige Schwellenwerte erreichen hohe Sensitivität zum Preis niedriger Spezifität, bei hohen Schwellenwerten ist es umgekehrt:

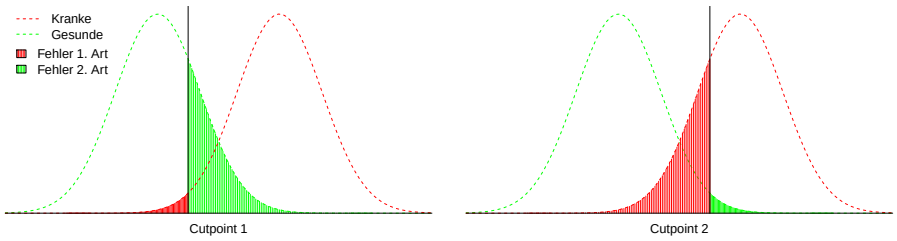

Abbildung 1.2: Fehler erster und zweiter Art bei verschiedenen Cutpoints.

Um dies zu veranschaulichen und Testverfahren unabhängig vom gewählten Cutpoint bezüglich ihrer Trennschärfe zu vergleichen, verwendet man CAP- oder ROC-Kurven.

Kapitel 2

Grafische Darstellungsformen der Trennschärfe

2.1 CAP-Kurven

Es existieren verschiedene grafische Darstellungsweisen, die die Trennschärfe eines Diagnoseverfahrens veranschaulichen sollen. Eine vor allem in den Wirtschaftswissenschaften beliebte Variante ist die CAP-Kurve *(Cumulative Accuracy Profile)*, auch Lorenzkurve genannt (Reitz, 2011; Henking et al., 2006; Krämer and Bücker, 2009). Hier wird die Wahrscheinlichkeit eines positiven Testresultats in der gesamten Population gegen die in der Subpopulation der Kranken (bzw. in der wirtschaftswissenschaftlichen Anwendung: nicht Kreditwürdigen) angetragen. Die Kurve besteht also aus den Punkten:

$$(P(R=1), P(R=1|S=1)) = (P(x>c), P(x>c|S=1)) \tag{2.1}$$

Selbstverständlich ist die zugrundeliegende Verteilung nicht bekannt. Eine Möglichkeit wäre nun eine parametrische Schätzung (vgl. hierzu Kap. 2.2). Gebräuchlicher ist es allerdings, die Kurve mithilfe der empirischen Verteilungsfunktionen anzunähern. Zur Veranschaulichung betrachten wir folgendes Beispiel: Es liegen Daten zu je fünf Kranken und fünf Gesunden vor. Zum Vergleich stehen drei verschiedene Merkmale oder Scores x, y, z, deren Trennschärfe betrachtet werden soll.

Status	0	0	0	0	0	1	1	1	1	1
w	1	2	2	3	4	6	7	7	8	10
x	1	7	8	4	5	5	6	7	9	10
y	1	2	3	4	5	1	2	3	4	5

Man erkennt: Das Merkmal w trennt die Gesunden und Kranken perfekt. Bei x werden die Kranken tendenziell höher eingestuft als die Gesunden. Das Merkmal hat also eine gewisse diagnostische Kraft, auch wenn einige Kranke niedrigere Werte erzielen als einige Gesunde. Die Diagnosen mit y sind im Grunde wertlos - die empirischen Verteilungen in

den beiden Gruppen sind hier identisch. Bei den CAP-Kurven ergibt sich dann das folgende Bild:

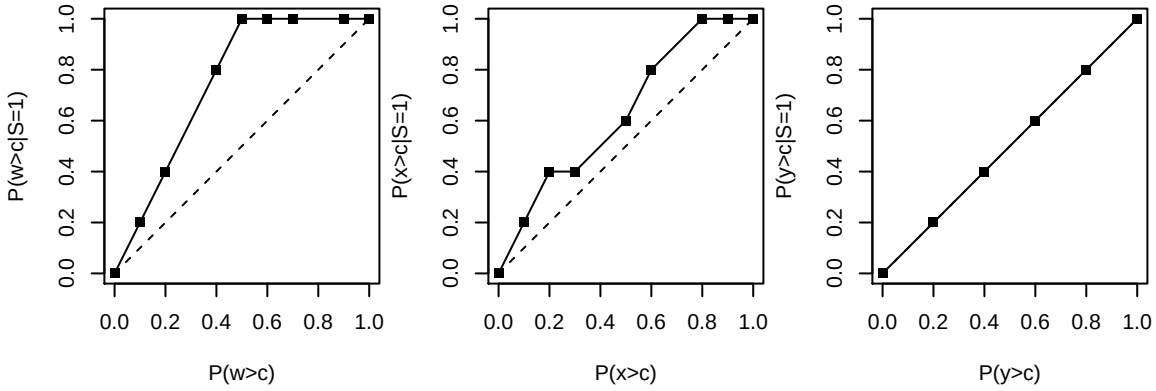

Abbildung 2.1: CAP-Kurven von Diagnoseverfahren mit unterschiedlicher Trennschärfe

Die Kurve für w, das perfekt trennscharf ist, hat im unteren Bereich als Steigung den Kehrwert des Anteils der Kranken im Datensatz. Die Kurve von z verläuft auf der Winkelhalbierenden, was widerspiegelt, dass der Anteil der als krank diagnostizierten unter den tatsächlich Kranken genauso groß ist wie in der gesamten Stichprobe. Die Kurve von x verläuft deutlich über der Diagonalen, was daran liegt, dass z bei den Kranken tendenziell höhere Werte annimmt als in der gesamten Stichprobe. Es wird also deutlich, dass ein Diagnoseverfahren desto besser ist, je weiter sich die Kurve von der Diagonalen wegbeugt. Allerdings ist zu beachten, dass der Verlauf der CAP auch immer vom Anteil der Kranken in der Stichprobe abhängt, Vergleiche von Kurven, die auf unterschiedlichen Stichproben beruhen also sehr problematisch sind (vgl. Reitz (2011), 282).

2.2 ROC-Kurven

Eine Alternative zur CAP-Kurve stellt die ROC-Kurve *(Receiver Operating Characteristics)* dar (Wehberg et al., 2007; Henking et al., 2006). Hier werden die Wahrscheinlichkeiten eines positiven Testergebnisses in den Teilpopulationen der Gesunden und der Kranken gegeneinander angetragen, also 1-Spezifität und Sensitivität. Die Kurve besteht dementsprechend aus den Punkten:

$$(P(R = 1|S = 0), P(R = 1|S = 1)) = (P(x > c|S = 0), P(x > c|S = 1)) \tag{2.2}$$

Obwohl sich dies formal wesentlich vom Prinzip der CAP-Kurve unterscheidet, ergibt sich qualitativ ein ähnlicher Verlauf (Henking et al., 2006, 219). Dies wird auch deutlich, wenn man die ROC-Kurven der Beispieldaten aus Kapitel 2.1 zeichnet:

Die ROC-Kurve eines vollkommen trennscharfen Diagnoseverfahrens steigt sofort auf (0,1), schließlich gibt es einen Cutpoint, für den alle Kranken und kein einziger Gesunder als krank diagnostiziert werden. Beim uninformativen Diagnoseverfahren y ergibt sich auch

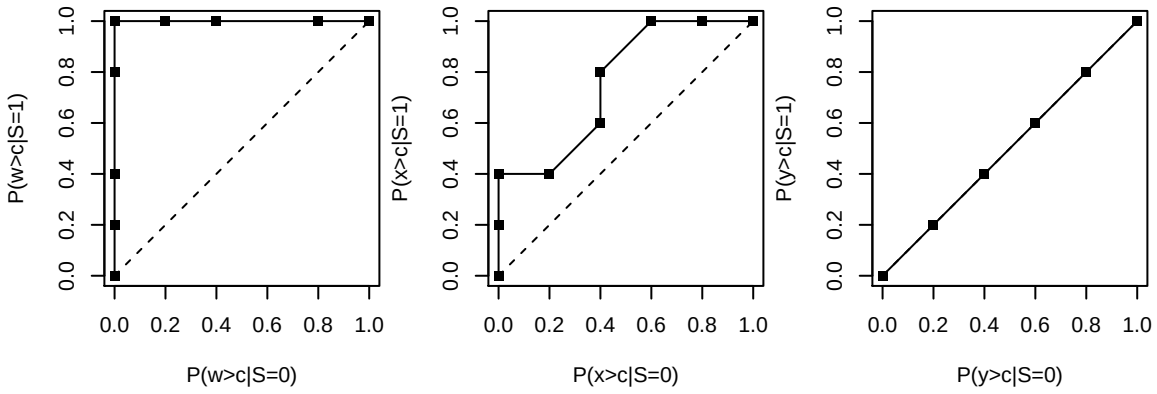

Abbildung 2.2: ROC-Kurven von Diagnoseverfahren mit unterschiedlicher Trennschärfe

bei der ROC-Kurve die Winkelhalbierende. Allgemein gilt auch hier, dass sich die Kurven trennschärferer Tests weiter von der Winkelhalbierenden wegbeugen.
Hier wurde wieder eine direkte Schätzung der Kurve mittels der empirischen Verteilungsfunktionen durchgeführt. Dies ist vor allem deshalb ein beliebtes Vorgehen, weil sich eine interessante Interpretation für die Fläche unter der Kurve ergibt: Wählt man zufällig je ein Individuum aus den Gesunden und eines aus den Kranken aus, so entspricht die Fläche genau der Wahrscheinlichkeit, richtig zuzuordnen, welcher der beiden krank und welcher gesund ist (Wehberg et al. (2007, 335), vgl. auch Kapitel 3). Der Wertebereich [0, 1], der sich aus der grafischen Darstellung ergibt steht also im Einklang mit der stochastischen Interpretation. In der Praxis werden freilich nur Werte zwischen 0.5 und 1 auftreten, geringere Werte bedeuten schlicht, dass der Test falsch herum gepolt ist. Umpolen liefert dann einen Wert aus [0.5,1].

Eine Alternative zur direkten Schätzung der ROC-Kurve ist eine parametrische Schätzung unter der Annahme, dass die Scorewerte $x_{(1)}$ der Kranken und $x_{(0)}$ der Gesunden jeweils normalverteilt sind:

$$\begin{aligned} x_G &\sim N(\mu_G, \sigma_G) \\ x_K &\sim N(\mu_K, \sigma_K) \end{aligned} \tag{2.3}$$

Möglich wäre es etwa, mittels ML-Schätzung die Verteilungsfunktionen in den beiden Subpopulationen zu schätzen. Mit deren Hilfe könnte dann eine glatte ROC-Kurve gezeichnet werden. Ein Vorgehen, das mit weniger zu schätzenden Größen auskommt stammt von Dorfman und Alf (Dorfman and Alf, 1968). Die 1-Spezifität wird dabei unter Ausnutzung der Normalverteilungsannahme mit $Z := (c - \mu_G)/\sigma_G$ wie folgt umgeformt:

$$P(x > c|S = 0) = 1 - P(x \leq c|S = 0) = 1 - \Phi\left(\frac{c - \mu_G}{\sigma_G}\right) = 1 - \Phi(Z) = \Phi(-Z) \tag{2.4}$$

Mit $a := (\mu_K - \mu_G)/\sigma_K$ und $b := \sigma_G/\sigma_K$ kann in ähnlicher Weise die Sensitivität umgeformt werden:

$$\begin{aligned} P(x > c|S = 1) = 1 - P(x \leq c|S = 1) = 1 - \Phi\left(\frac{c - \mu_K}{\sigma_K}\right) \\ = 1 - \Phi\left(\frac{c - \mu_G}{\sigma_G}\frac{\sigma_G}{\sigma_K} - \frac{\mu_K - \mu_G}{\sigma_K}\right) = 1 - \Phi(Zb - a) = \Phi(a - bZ) \end{aligned} \tag{2.5}$$

Es genügt also, die beiden Parameter a und b zu schätzen, wofür Dorfman und Alf einen ML-Ansatz vorstellen. Dann können durch Einsetzen verschiedener Werte für Z beliebig viele Punkte der ROC-Kurve berechnet werden. Eine Zuordnung zu einem bestimmten Cutpoint ist allerdings nicht unmittelbar möglich.

Vorteil dieses Verfahrens ist, dass eine glatte Kurve erzeugt werden kann und relativ einfach Aussagen über die Güte der Schätzung getroffen werden können. Es bieten sich also einfachere Vorgehensweisen für Konfidenzbänder o.ä. an als beim nicht-parametrischen Ansatz. Nachteilig ist allerdings, dass eine Verteilungsannahme getroffen werden muss, sodass das Verfahren nicht auf jeden Datensatz gleichermaßen anwendbar ist. Es ist zu beachten, dass die einzelnen geschätzten Werte für Sensitivität und Spezifität, also die Punkte, die sich bei der direkten Schätzung für die ROC-Kurve ergeben, nicht auf der mit dem parametrischen Ansatz geschätzten Kurve liegen müssen (Wehberg et al., 2007, 333). In der Regel ergibt sich beim parametrischen Ansatz eine kleinere Fläche unter der Kurve als beim nicht-parametrischen (ebd., 335).

Auch bei ROC-Kurven gilt, dass den zu vergleichenden Kurven die selben Stichproben zugrundeliegen sollten. Zwar ist der Verlauf der Kurve nicht wie bei der CAP-Kurve vom Anteil der Kranken in der Population abhängig (Krämer and Bücker, 2009, 15), doch können sich Stichproben auch in ihrer Homogenität bezüglich der Kovariablen unterscheiden. Bei stark heterogenen Stichproben ist es selbstverständlich leichter, eine gute Trennschärfe zu erreichen. Unterscheiden sich die Individuen hingegen kaum, so kann auch das beste Diagnoseverfahren wenig ausrichten. Insbesondere ist auch darauf zu achten, dass bei den zu vergleichenden ROC-Kurven die selbe Krankheitsdefinition bzw. der selbe Goldstandard verwendet wird (vgl. Henking et al. (2006), 227).

Ein allgemeines Problem beim Vergleich von Diagnoseverfahren anhand von derartigen Kurven, sowohl ROC- als auch CAP-Kurven, ist das Folgende: Es kann vorkommen, dass sich die Kurven zweier zu vergleichender Kurven ein- oder mehrmals schneiden. Inhaltlich bedeutet dies, dass eines der Verfahren in einem bestimmten Teil des Definitionsbereichs von c besser, in einem anderen aber schlechter ist als das andere. Es ist dann nicht ohne Festlegung weiterer Kriterien möglich, zu entscheiden, welches Verfahren vorzuziehen ist. Um ein allgemein anwendbares Kriterium zum Vergleich zu bekommen, wurden verschiedene aus den Kurven ableitbare skalare Größen vorgeschlagen.

Kapitel 3

Skalare Kenngrößen zum Vergleich von Diagnoseverfahren

3.1 Area under the Curve

In Bezug auf die ROC-Kurve stellt die Fläche unter der Kurve eine sehr verbreitete Kennzahl dar. Man bezeichnet diese als AUC (*Area under the Curve*). Wie bereits beschrieben, ist ein Diagnoseverfahren umso besser, je weiter sich seine ROC-Kurve von der Diagonalen wegbiegt. Insofern ist es schon intuitiv sinnvoll, die Fläche, die sich unter der Kurve zu betrachten. Doch lässt sich diese Vorgehensweise auch theoretisch rechtfertigen: Green und Swets zeigen, dass die Fläche unter der Kurve genau der Wahrscheinlichkeit entspricht, zwischen einem zufällig ausgewählten Gesunden und einem zufällig ausgewählten Kranken korrekt den Kranken auszumachen. Die Autoren sprechen dabei von einem *forced choice task*, es muss also eine Entscheidung getroffen werden, auch wenn kein Unterschied ausgemacht wurde. Sie liefern dafür einen formalen Beweis (Green and Swets, 1966), hier soll nur eine etwas vereinfachte Veranschaulichung erfolgen.

Es wird davon ausgegangen, dass der Scorewert nur endlich viele Ausprägungen annimmt. Selbst wenn das Merkmal in Wirklichkeit stetig ist, ist dies der Fall, wenn wir nur auf die Stichprobe Bezug nehmen, die der geschätzten ROC-Kurve zugrundeliegt. Schließlich ist die Stichprobengröße begrenzt. Wir betrachten also die Wahrscheinlichkeit, mit der ein zufällig aus der Stichprobe ausgewähltes Paar eines Gesunden G und eines Kranken K korrekt zugeordnet wird. Diese Wahrscheinlichkeit sei im Folgenden mit Θ bezeichnet. Da einzelne Werte von x dann tatsächlich von 0 unterschiedliche Ziehungsswahrscheinlichkeiten haben, lässt sich besagte Wahrscheinlichkeit folgendermaßen darstellen:

$$\Theta = \sum_{i=1}^{l} \left(P(x_G = x_{(i)}) \cdot P(x_K > x_{(i)}) + 0.5 \cdot P(x_G = x_{(i)}) \cdot P(x_K = x_{(i)}) \right) \tag{3.1}$$

Die $x_{(i)}$ bezeichnen dabei diejenigen Werte von x, die in der Stichprobe vorkommen (in aufsteigender Ordnung). l bezeichnet also die Anzahl der unterschiedlichen Ausprägungen

von x. Der hintere Term $0.5 \cdot P(x_G = x_{(i)}) \cdot P(x_K = x_{(i)})$ reflektiert, dass bei gleichem Scorewert randomisiert werden muss, um im Sinne des *forced choice tasks* eine Entscheidung zu treffen.

Betrachtet man nun eine nach dem nicht-parametrischen Verfahren erzeugte ROC-Kurve, so kann leicht gezeigt werden, dass diese Wahrscheinlichkeit genau der Fläche unterhalb der Kurve entspricht. Zu Veranschaulichung werden wieder die Daten aus Kapitel 2 mit Score x verwendet.

Status	0	0	0	0	0	1	1	1	1	1
x	1	7	8	4	5	5	6	7	9	10

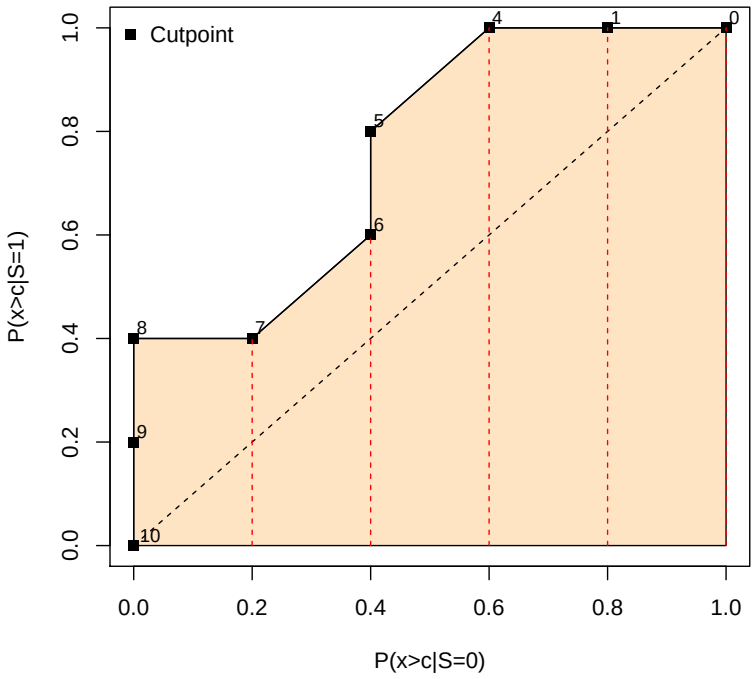

Abbildung 3.1: ROC-Kurve mit Aufteilung in Trapeze

Wie leicht ersichtlich ist, besteht die ROC-Kurve aus $(k + 1)$ Punkten und den Verbindungen zwischen ihnen: Jeder Auspräqung von x entspricht ein Punkt, hinzu kommt der Punkt (1,1) welcher einem beliebigen Wert unterhalb des Minimums von x entspricht. Durch Hinzufügen der Lote von diesen Punkten auf die Abszisse lässt sich die Fläche unter der Kurve in k Trapeze zerlegen. Da manche Werte von x nur bei Kranken vorkommen, sind einige dieser Trapeze entartet und haben die Breite 0. Allgemein lässt sich die Breite desjenigen Trapezes, das sich rechts vom dem Scorewert $x_{(i)}$ entsprechenden Punkt befindet

folgendermaßen ausdrücken:

$$P(x > x_{i-1}|S = 0) - P(x > x_i|S = 0) = P(x_G > x_{i-1}) - P(x_G > x_i) = P(x_G = x_i) \quad (3.2)$$

Für die Höhe auf der linken Seite gilt:

$$P(x > x_i|S = 1) = P(x_K > x_i) \quad (3.3)$$

Entsprechend für die rechte Seite:

$$P(x > x_{i-1}|S = 1) = P(x_K > x_{i-1}) \quad (3.4)$$

Es ist anzumerken, dass diese beiden Höhen für den Fall, dass die zugrundeliegende Variable stetig ist fast sicher gleich sind, da dann keine Bindungen zu erwarten sind. Für den Fall kategorialer Daten ist die Unterscheidung allerdings zu beachten. In jedem Fall lässt sich die Fläche unter der Kurve unter Verwendung der Flächenformel für Trapeze schreiben als:

$$\begin{aligned} AUC &= \sum_{i=1}^{l} \left(0.5 \cdot \left(P(x_K > x_{(i)}) + P(x_K > x_{(i-1)})\right) \cdot P(x_G = x_{(i)})\right) \\ &= \sum_{i=1}^{l} \left(\left(P(x_K > x_{(i)}) + 0.5 \cdot (P(x_K > x_{(i-1)}) - P(x_K > x_{(i)}))\right) \cdot P(x_G = x_{(i)})\right) \\ &= \sum_{i=1}^{l} \left(P(x_G = x_{(i)}) \cdot P(x_K > x_{(i)}) + 0.5 \cdot P(x_G = x_{(i)}) \cdot P(x_K = x_{(i)})\right) = \Theta \end{aligned} \quad (3.5)$$

Dies entspricht also genau der Wahrscheinlichkeit einer richtigen Klassifikation im beschriebenen Szenario. Dass die Interpretation in Bezug auf die zugrundeliegende Stichprobe zulässig ist, kann also mit einfachen geometrischen Überlegungen veranschaulicht werden.

Interessanterweise sind die Eigenschaften dieser Größe unabhängig von ihrem Auftreten in ROC-Analysen schon detailliert untersucht: Sie ist abgesehen von der Normierung auf [0,1] identisch mit der Mann-Whitney-Statistik (Hanley and McNeil, 1982, 31). Diese berechnet sich folgendermaßen (Neuhäuser, 2011):

$$\begin{aligned} U &= \sum_{i=1}^{n_k} \sum_{j=1}^{n_g} S(x_{K_i}, x_{G_j}) \\ S(a, b) &= \begin{cases} 1 \text{ für } a > b \\ 0.5 \text{ für } a = b \\ 0 \text{ für } a > b \end{cases} \end{aligned} \quad (3.6)$$

Für diese Teststatistik werden also alle möglichen Paarvergleiche zwischen Kranken und Gesunden betrachtet und die Zahl derjenigen festgehalten, in denen der Kranke einen höheren Score erhält wird als der Gesunde. Die im Fall von Bindungen verwendete Gewichtung von 0.5 ist wieder so zu interpretieren, dass hier eine Randomisierung nötig ist. Normiert man die Teststatistik mithilfe von n_G und n_K, die die Zahl der Individuen in den Subpopulationen bezeichnen, so kann mit einigen kleinen Umstellungen die Übereinstimmung mit den obigen Größen deutlich gemacht werden:

$$\begin{aligned} U^* &= \frac{1}{n_G n_K} \sum_{i=1}^{n_K} \sum_{j=1}^{n_G} S(x_{K_i}, x_{G_j}) \\ &= \sum_{j=1}^{n_G} \frac{1}{n_G} \underbrace{\frac{1}{n_K} \sum_{i=1}^{n_K} S(x_{K_i}, x_{G_j})}_{P(x_K > x_{G_j}) + 0.5 \cdot P(x_K = x_{G_j})} \\ &= \sum_{j=1}^{n_G} \left(\frac{1}{n_G} \left(P(x_K > x_{G_j}) + 0.5 \cdot P(x_K = x_{G_j}) \right) \right) = \Theta \end{aligned} \tag{3.7}$$

Dies ist offensichtlich identisch mit (1.1) und (1.5), lediglich die Indizierung ist anders gehandhabt (es werden nicht die Levels von x mit Gewichtung nach ihrer Wahrscheinlichkeit durchlaufen, sondern einfach die Untersuchungseinheiten in der Stichprobe).
Beim Mann-Whitney-Test handelt es sich um einen Homogenitätstest. Unter der Annahme, dass die Verteilungsfunktion von x in beiden Subpopulationen die selbe Form aufweist, jedoch möglicherweise um einen Betrag verschoben ist, kann die Nullhypothese gleicher Verteilungen getestet werden, die Nullhypothese wird dabei für sehr große und sehr kleine Werte abgelehnt. Die Fläche unter der ROC-Kurve kann dementsprechend als ein Maß dafür angesehen werden, wie plausibel die Annahme ist, dass x in beiden Subpopulationen identisch verteilt ist und damit keinerlei Aussagekraft besitzt. Für die Verteilung unter H_0 gibt es Tabellierungen. In statistischen Programmpaketen ist häufig die Verteilung für die Teststatistik des Wilcoxon-Rangsummentest implementiert (z.B. im R-Paket textitstats mit den Befehlen `dwilcox, pwilcox, qwilcox`). Diese baut zwar nicht auf Paarvergleichen, sondern auf den Rängen in einer gepoolten Stichprobe auf, ist aber äquivalent zur Mann-Whitney-Statistik und kann leicht aus ihr berechnet werden (Neuhäuser, 2011):

$$W = U + \frac{n_K(n_K + 1)}{2} \tag{3.8}$$

Da die stochastischen Eigenschaften von W auch für andere Szenarien als H_0 bekannt sind, sind hieran anschließend weitere Überlegungen möglich. Hanley und McNeill befassen sich etwa mit der Konzeption von Stichproben (Hanley and McNeil, 1982, 32) und verweisen auf Metz and Kronman (1980), die ein Verfahren vorstellen, mit dem untersucht werden kann, ob die AUC einer Kurve signifikant größer ist als die einer anderen.

In Bezug auf CAP-Kurven ist die Verwendung der Fläche unter der Kurve nicht gebräuchlich. Es ergibt sich auch keine ähnlich einfache Interpretation wie bei den ROC-Kurven.

3.2 Gini-Koeffizient

Eine weitere Maßzahl stellt der sogenannte Gini-Koeffizient dar (Henking et al., 2006). Man betrachtet dabei die Fläche, die zwischen der Diagonalen und der ROC- bzw. CAP-Kurve eingeschlossen ist. Diese wird dann zu derjenigen Fläche ins Verhältnis gesetzt, die die Kurve eines optimalen Verfahrens und die Diagonale einschließen. Dieser Quatient wird als Gini–Koeffizient bezeichnet. Mit den bereits bekannten Beispieldaten ergibt sich die folgende Veranschaulichung. Die Fläche zwischen Kurve und Diagonale ist dunkler eingefärbt. Die bei der idealen Kurve noch hinzukommende Fläche ist heller gekennzeichnet.

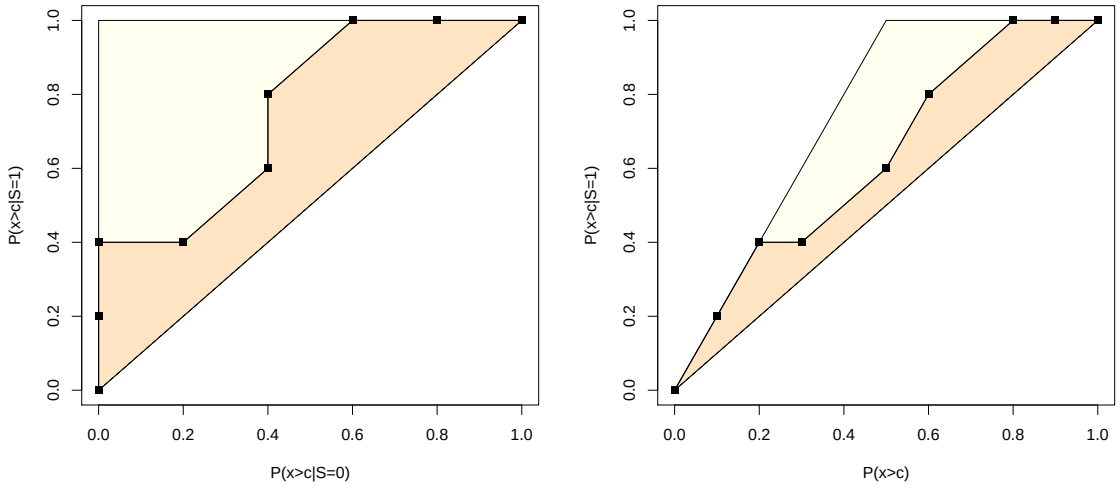

Abbildung 3.2: Veranschaulichung des GINI-Koeffizienten für ROC-Kurve (links) und CAP-Kurve (rechts)

Da die Fläche zwischen optimaler Kurve und Diagonale bei einer ROC-Kurve immer 0.5 beträgt, ergibt sich ein einfacher Zusammenhang mit der AUC:

$$\text{Gini} = 2 \cdot \text{AUC} - 1 \tag{3.9}$$

Interessanterweise lässt sich zeigen, dass es keinen Unterschied macht, ob man den Gini-Koeffizienten aus der ROC- oder der CAP-Kurve ableitet (Henking et al., 2006, 221). Die flacheren Verläufe der CAP-Kurve und der idealen CAP-Kurve gleichen sich also genau aus. Das wiederum bedeutet, dass auch aus einer CAP-Kurve die AUC der zugehörigen ROC-Kurve leicht abzuleiten ist. Da es sich bei (3.9) um eine monotone Transformation handelt,

können (qualitative) Vergleiche bezüglich der in 3.1 beschriebenen Wahrscheinlichkeit θ also auch anhand von CAP-Kurven und den dort vorkommenden Flächen angestellt werden.

3.3 ROC-gap

Im wirtschaftswissenschaftlichen Kontext, etwa in Bezug auf Kreditratings, findet außerdem der ROC-gap Verwendung. Diese Größe bezeichnet den maximalen vertikalen Abstand zwischen der ROC-Kurve und der Diagonalen.

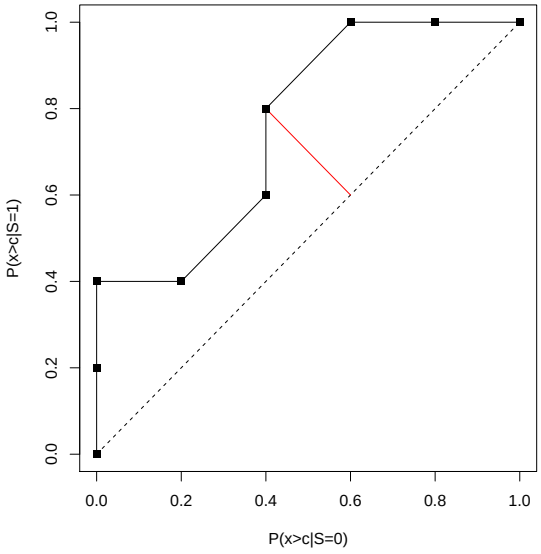

Abbildung 3.3: ROC-gap

Der ROC-gap ist eine monotone Funktion des Kolmogoroff-Smirnoff-Tests (Krämer and Bücker, 2009, 11). Mit diesem kann getestet werden, ob zwei Stichproben, also hier die Teilstichproben der Kranken und Gesunden der selben Verteilung folgen (Sachs and Hedderich, 2006). Es wird also das selbe getestet wie beim Mann-Whitney Test, jedoch entfällt die Annahme gleichartiger Verteilungen. Als Teststatistik dient der maximale Abstand zwischen den empirischen Verteilungsfunktionen in den beiden Teilstichproben.

Kapitel 4

Ein Verfahren zur Behandlung systematisch fehlender Verifizierung

4.1 Das Problem unvollständiger Verifizierung

Bisher wurde implizit stets davon ausgegangen, dass für alle Untersuchungseinheiten der wahre Status vorliegt. Dies ist in der Praxis allerdings in den wenigsten Fällen gegeben: Bei der Kreditvergabe etwa liegt es in der Natur der Sache, dass man von denjenigen, die einen schlechten Scoringwert erzielen, nicht erfahren wird, ob sie den Kredit zurückgezahlt hätten. Schließlich werden die Betreffenden dann gar keinen Kredit erhalten. Wenn das bisher verwendete Verfahren etwas taugt, ist davon auszugehen, dass dadurch der Anteil der ausgefallenen Kredite in den Daten, bei denen die Verifizierung vorliegt, geringer ist als in der gesamten Population. Entsprechend ist vor allem die CAP-Kurve mit Vorsicht zu interpretieren.

Im medizinischen Bereich werden Personen, bei denen das zu untersuchende Diagnoseverfahren keinerlei Hinweis auf eine Erkrankung geliefert hat, in vielen Fällen nicht weiter untersucht. Grund hierfür ist, dass die Anwendung des Goldstandards oft mit hohen Kosten oder Nebenwirkungen verbunden ist. Vor allem bei seltenen Krankheiten kann es also passieren, dass ein großer Teil der Untersuchungseinheiten keiner Verifizierung unterzogen wird.

Berechnet man mit den dann vorliegenden Daten Sensitivität und Spezifität, so erhält man im Grunde nur die jeweiligen Größen bedingt darauf, dass das Untersuchungsergebnis in einem bestimmten Bereich liegt. In den Wirtschaftswissenschaften fasst man Verfahren, die hier Abhilfe schaffen sollen unter dem Begriff *reject inference* (Krämer and Bücker, 2009, 13) zusammen. Im Folgenden soll allerdings kein Überblick über dieses umfangreiche Gebiet gegeben werden, sondern lediglich ein Verfahren aus dem medizinischen Bereich herausgegriffen werden. Anhand einer einfachen Simulation sollen außerdem noch etwaige Probleme herausgearbeitet werden.

4.2 Lloyd/Frommer: Anwendung logistischer Regression zur Schätzung der Test-Performance bei unvollständiger Verifizierung

Lloyd and Frommer (2008) befassen sich mit einem Screening-Test, der dazu dienen soll, Personen mit erhöhtem Darmkrebsrisiko auszumachen. Für den Test wird an sechs aufeinanderfolgenden Tagen festgestellt, ob sich im Stuhl der Untersuchten okkultes Blut befindet. Dieses stellt eines der Hauptsymptome bei Darmkrebs dar. Es ist zu beachten, dass das Blut im Stuhl auch bei Polypen, also gutartigen Geschwülsten im Darm, auftritt. Diese haben mit dem Krebs an sich nichts zu tun, bringen aber ähnliche Symptome mit sich und können deshalb quasi nebenbei mitdiagnostiziert werden. Es gibt also im vorliegenden Fall drei mögliche Status: Gesund, Polypen und Krank. Die Herausforderung bei der Schätzung der Performance des Tests besteht nun darin, dass nicht bei allen Probanden der tatsächliche Status festgestellt wurde: Wurde eine Person an allen sechs Tagen negativ getestet, so wurde keine weitere Untersuchung vorgenommen, da keinerlei Hinweis auf eine Erkrankung vorlag.

4.2.1 Das Verfahren

Das auf einem multinomialen logistischen Modell beruhende Verfahren, das Lloyd und Frommer vorschlagen, um dennoch die ROC-Kurve schätzen zu können soll im Folgenden in leicht vereinfachter Form dargestellt werden. Da die Begriffe Sensitivität und Spezifität und damit auch die herkömmliche ROC-Kurve bei drei Responsekategorien nicht recht passen, wird eine Übertragung des Verfahrens auf einen gewöhnlichen diagnostischen Test vorgestellt, bei dem zwischen Gesunden ($S = 0$) und Kranken ($S = 1$) diskriminiert werden soll. Entsprechend wir auch kein multinomiales sondern ein binomiales logistisches Modell verwendet.

Die Darstellung erfolgt anhand eines computergenerierten Beispieldatensatzes mit $n = 10000$ Untersuchungseinheiten. Ausfälle aufgrund von Abbrechern o.ä. werden nicht berücksichtigt. Um den Datensatz einfacher darstellen zu können, werden nur vier binäre Kovariablen verwendet. Der wahre Zusammenhang zwischen den Kovariablen und dem Status ist festgelegt als:

$$\begin{aligned} P(S = 1|V1, V2, V3, V4) &:= \frac{exp(Z)}{1 + exp(Z)} \\ Z &:= -2.5 + 1.2 \cdot V1 + 1.4 \cdot V2 + 1.5 \cdot V3 + 1.6 \cdot V4 \end{aligned} \tag{4.1}$$

Für die Trefferwahrschheinlichkeiten bei den binären Kovariablen gilt:

$$\begin{aligned} P(V1 = 1) &:= 0.15 \\ P(V2 = 1) = P(V3 = 1) &:= 0.1 \\ P(V4 &:= 0.2 \end{aligned} \tag{4.2}$$

Für diejenigen, bei denen mindestens eine Kovariable den Wert 1 angenommen hat, liegt zusätzlich der tatsächliche Status vor, bei den Restlichen fehlt er. Zusammengefasst sieht der simulierte Datensatz folgendermaßen aus:

L	V1	V2	V3	V4	Score	Gesunde	Kranke	Gesamt
1	0	0	0	0	0	NA	NA	5532
2	0	0	0	1	1	931	412	1343
3	0	0	1	0	1	454	199	653
4	0	0	1	1	2	52	95	147
5	0	1	0	0	1	472	150	622
6	0	1	0	1	2	64	93	157
7	0	1	1	0	2	27	34	61
8	0	1	1	1	3	1	14	15
9	1	0	0	0	1	775	221	996
10	1	0	0	1	2	102	116	218
11	1	0	1	0	2	45	48	93
12	1	0	1	1	3	1	24	25
13	1	1	0	0	2	53	44	97
14	1	1	0	1	3	5	19	24
15	1	1	1	0	3	1	11	12
16	1	1	1	1	4	0	5	5

Die zusätzlich eingeführte Variable L stellt lediglich eine Rekodierung des Ergebnisses dar, um leichter auf die unterschiedlichen Kombinationen Bezug nehmen zu können. Der einer Person i durch das zu betrachtende Diagnoseverfahren zugeordnete Score x_i sei schlicht die Anzahl der positiv ausgefallenen binären Tests:

$$x_i := V1_i + V2_i + V3_i + V4_i \tag{4.3}$$

Um die ROC-Kurve schätzen zu können, benötigt man Schätzungen von $P(x > c|S = 0)$ und $P(x > c|S = 1)$, also der auf S bedingten Verteilung von x. Bei vollständiger Verifizierung kann man hierfür die empirischen Verteilungen verwenden (Kapitel 2). Dies ist hier allerdings nicht ohne weiteres möglich, schließlich liegt nur für etwa die Hälfte der Untersuchten der echte Status vor. Zudem ist offensichtlich, dass sich diejenigen, bei denen der Status nicht festgestellt wurde, systematisch von den Restlichen unterscheiden, zumindest wenn man davon ausgeht, dass das Diagnoseverfahren nicht völlig nutzlos ist. Lloyd und Frommer schlagen nun gewissermaßen einen kleinen Umweg vor: Im ersten Schritt wird ein logistisches Regressionsmodell für $P(S|V1, V2, V3, V4)$ angepasst. (Die Autoren passen natürlich in ihrem Artikel das Modell etwas sorgfältiger an die Daten an, hier wird von dieser Vereinfachung ausgegangen). Auf diese Weise kann mehr Information mit einbezogen werden, als bei einer Modellierung mit der zusammengefassten Größe x als einziger Kovariable, was eine bessere Schätzung ermöglicht. Im nächsten Schritt wird aus den gefitteten Wahrscheinlichkeiten $\hat{P}(S|V1, V2, V3, V4)$ eine Schätzung von $P(S|L)$ abgeleitet. Es wird also nicht nur für $P(S|L = 1)$, wo die Verifizierung fehlt, sondern überall die mit dem Mo-

dell geschätzte Wahrscheinlichkeit weiterverwendet. Für den Beispieldatensatz lässt sich $\hat{P}(S|L)$ dann in Gegenüberstellung zu den relativen Häufigkeiten wie folgt darstellen:

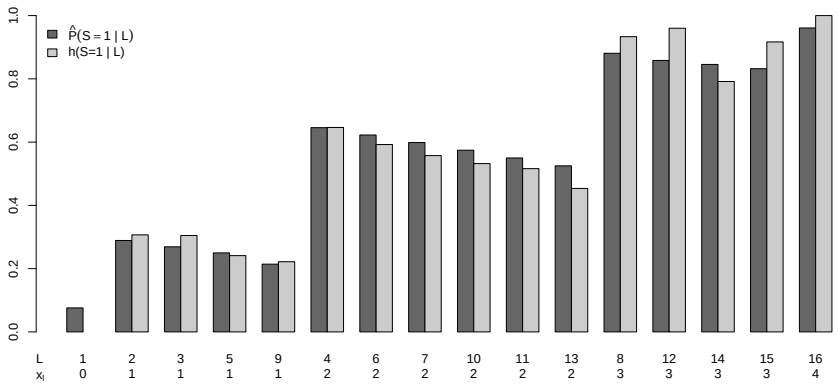

Abbildung 4.1: Mithilfe des logistischen Modells geschätzte $P(\hat{S = 1}|L)$ und zugehörige relative Häufigkeiten

Zusätzlich wird die Randverteilung von L direkt durch die empirische Verteilung geschätzt (hier wird also davon ausgegangen, dass eine einfache Zufallsstichprobe vorliegt):

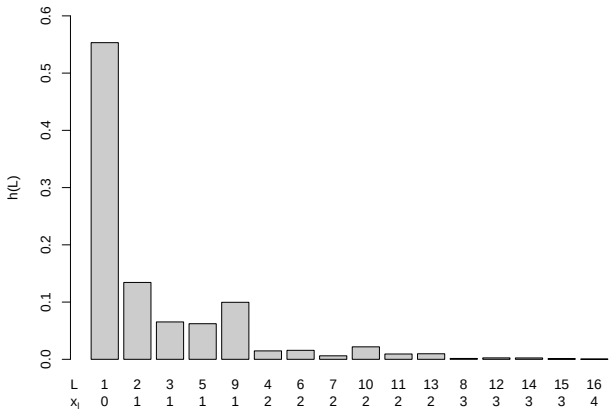

Abbildung 4.2: Relative Häufigkeiten der Ergebnisse L

Aus diesen beiden Schätzungen wird dann mittels des Satzes von Bayes eine Schätzung für $P(L|S)$ errechnet:

$$\hat{P}(L = l|S = s) = \frac{\hat{P}(S = s|L = l)\hat{P}(L = l)}{\sum_{i=1}^{16} \hat{P}(S = s|L = i)\hat{P}(L = i)} \tag{4.4}$$

Es ergibt sich dann das folgende Bild für die geschätzten Verteilungen in den beiden Subpopulationen:

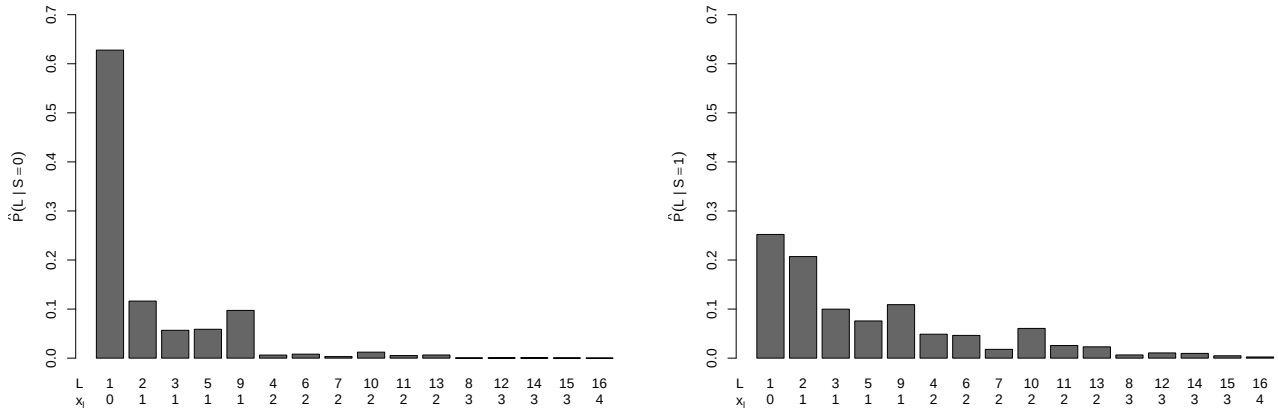

Abbildung 4.3: Geschätzte bedingte Verteilung von L in den Gruppen der Gesunden (links) und Kranken (rechts)

Diese implizieren nun die zum Zeichnen der ROC-Kurve nötigen $\hat{P}(x|S)$:

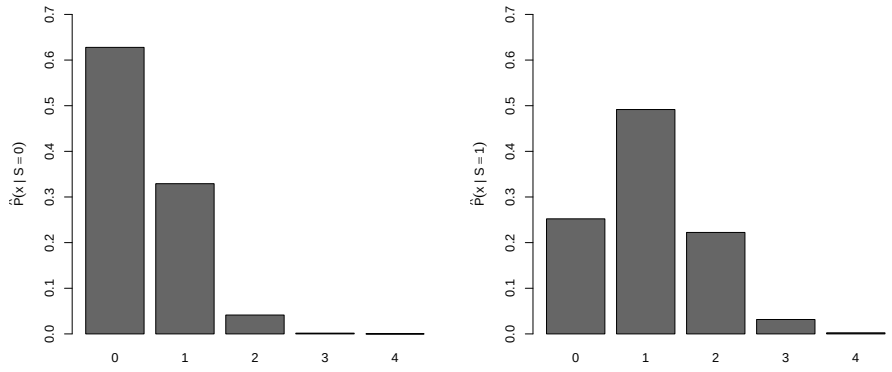

Abbildung 4.4: Geschätzte bedingte Verteilung des Scores x in den Gruppen der Gesunden (links) und Kranken (rechts)

Nun kann eine ROC-Kurve gezeichnet werden. Zusätzlich wird die wahre ROC Kurve, die auf den in der Simulation festgelegten echten Verteilungen beruht eingezeichnet. Ebenfalls wird diejenige ROC-Kurve ergänzt, die sich ergibt, wenn man die Untersuchungseinheiten ohne Verifizierung einfach weglässt.

Man sieht, dass die ROC-Kurve, bei der die Untersuchungseinheiten ohne Verifizierung einfach weggelassen wurden, deutlich unter der echten liegt. Dies verwundert nicht, wenn

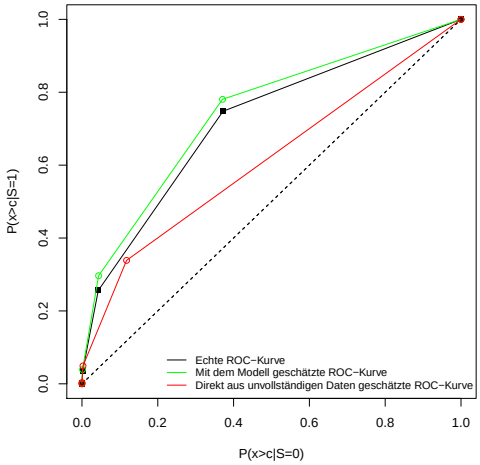

Abbildung 4.5: Geschätzte, echte und direkt aus unvollständigen Daten geschätzte ROC-Kurve

man bedenkt, dass hier eine vergleichsweise homogene Teilstichprobe verwendet wurde, in der die Unterscheidung Kranker und Gesunder notwendigerweise schwieriger ist. Die ROC-Kurve, die mit dem Verfahren nach Lloyd und Frommer erstellt wurde, stellt in diesem Fall hingegen eine sehr gute Näherung dar. Allerdings entsprach der Zusammenhang zwischen Kovariablen und Status im simulierten Datensatz genau der Modellgleichung des verwendeten logistischen Regressionsmodells, es kam nur darauf an, die Parameter richtig zu schätzen. Das gute Ergebnis überrascht also angesichts des recht hohen Stichprobenumfangs und nur vier Kovariablen nicht. Im Folgenden soll deshalb noch für verschiedene andere Szenarien die Leistungsfähigkeit des Verfahrens überprüft werden.

4.2.2 Simulation zur Leistungsfähigkeit in verschiedenen Szenarien

Einfluss der Stichprobengröße Als erstes soll betrachtet werden, wie sich die Genauigkeit der Schätzung bei Vergröserung des Stichprobenumfangs verbessert. Als Beispieldatensatz wird ein Datensatz mit 6 unabhängigen binären Teiltests als Kovariablen simuliert. Die Trefferwahrscheinlichkeiten seien:

$$\begin{aligned} P(V1 = 1) = P(V4 = 1) &:= 0.1 \\ P(V2 = 1) = P(V5 = 1) &:= 0.15 \\ P(V3 = 1) = P(V6 = 1) &:= 0.2 \end{aligned} \tag{4.5}$$

Der echte Zusammenhang zwischen Status und den Ergebnissen der binären Tests sei:

$$P(S = 1|V1, ..., V8) := \frac{exp(Z)}{1 + exp(Z)} \qquad (4.6)$$
$$Z := -2 + 0.8 \cdot (V1 + V5) + 0.7 \cdot (V2 + V6) + 0.6 \cdot V3 + 0.5 \cdot V4$$

Der Zusammenhang entspricht also genau den Annahmen des im Verfahren verwendeten Logit-Modells. Der Score, der einer Untersuchungseinheit zugewiesen wird sei wieder einfach die Zahl der positiven Teiltests. Nun wird das Verfahren für die Stichprobenumfänge 100, 1000 und 10000 jeweils 50 mal angewendet. Allgemeine Aussagen sind aus dieser einfachen Simulation natürlich nicht ableitbar, schließlich hängt die Genauigkeit der Schätzung unter anderem auch davon ab, wie groß der Anteil der nicht Verifizierten ist. Im Beispiel ergibt sich hier ein Erwartungswert von 0.37, was im Vergleich mit der Darmkrebs-Studie (0.93) ein recht geringer Anteil ist.

Am simulierten Beispiel lassen sich einige Tendenzen ablesen: Das Verfahren ist für kleine Stichprobenumfänge nicht geeignet, die Varianz der Schätzung ist sehr groß. Bei wachsendem Stichprobenumfang verbessert sich die Schätzung, bei n=10000 sind im Beispiel kaum Abweichungen von der echten Kurve zu erkennen. Allerdings ist immer im Hinterkopf zu behalten, dass hier die Modellannahmen genau erfüllt sind. In der Praxis ist das nicht unbedingt zu erwarten und angesichts der fehlenden Daten auch schwer zu überprüfen. Grafisch stellen sich die Ergebnisse für das Beispiel wie folgt dar:

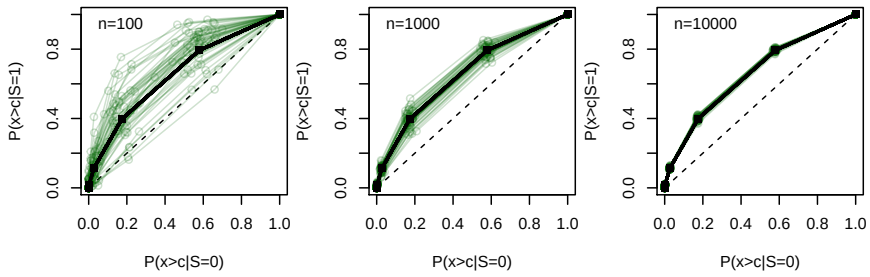

Abbildung 4.6: Geschätzte ROC-Kurven bei Stichprobenumfängen von 100, 1000 und 10000

Fehlende Kovariablen Im Folgenden soll für die selben Daten wie oben (n=1000 bzw. n=10000) betrachtet werden, wie sich das Fehlen von Kovariablen auswirkt. Der datengenerierende Prozess ist der selbe wie bisher und bezieht 6 Kovariablen ein. Jedoch werden nur 3, dann 4, dann 5 der Kovariablen in der Analyse berücksichtigt. Als Score wird jeweils die Zahl der positiven unter diesen Tests betrachtet. Es handelt sich also jedes Mal um ein anderes Diagnoseverfahren, entsprechend unterscheiden sich die echten ROC-Kurven in den drei Szenarien. Die Verifizierung wird nur bei denjenigen Untersuchungseinheiten als bekannt angenommen, bei denen einer der einbezogenen Tests positiv ausgefallen ist.

Auch beim Logit-Modell werden selbstverständlich nur die jeweils zur Verfügung stehenden Variablen verwendet. Bei einem Stichprobenumfang von 1000 bzw. 10000 ergibt sich folgendes Bild:

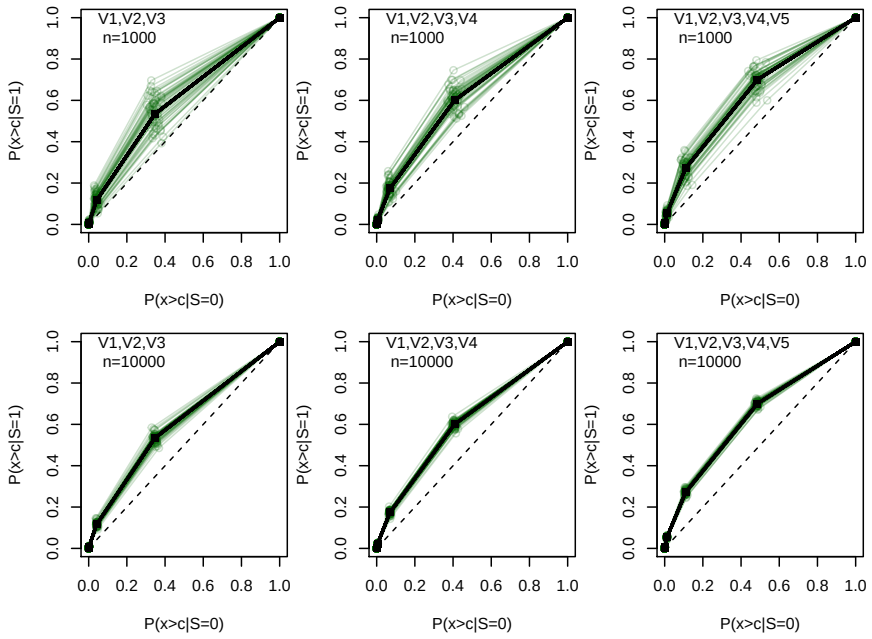

Abbildung 4.7: Geschätzte ROC-Kurven bei 3, 4 bzw. 5 berücksichtigten Kovariablen

Auch wenn Einflussgrößen fehlen, funktioniert das Verfahren für wirklich große Stichprobenumfänge noch gut. Dass es für geringere Stichprobenumfänge nicht mehr so gut klappt, hängt natürllich auch damit zusammen, dass bei noch mehr Fällen die Verifizierung fehlt und somit noch weniger vollständige Beobachtungen vorliegen.
Dass das Fehlen von Variablen, selbst wenn sie keine Confounder sind, nicht zu einer verzerrten Schätzung führt war keineswegs selbstverständlich. Schließlich kommt es ja dennoch bei der Schätzung des Intercepts zu Verzerrungen, sofern dieser nicht null ist. Und gerade der Intercept ist entscheidend bei der Extrapolation auf $P(S|L = 1) = P(S|V1 = ... = V6 = 0)$. Die Vermutung, dass sich hier Probleme ergeben könnten hat sich allerdings nicht bestätigt.

Nicht auszumachende Interaktion Eine Annahme, die zentral für das Funktionieren des Verfahrens ist, ist, dass von den Individuen mit mindestens einem positiven Test auf die ohne positive Tests geschlossen werden kann. Auf die Darmkrebs-Studie bezogen heißt das, dass der Einfluss eines bestimmten positiven Tests unabhängig davon ist, ob es noch andere positiv ausgefallene Tests gibt oder nicht. Es wäre nun aber denkbar, dass das Vorliegen mindestens eines positiven Tests schon auf ein erhöhtes Risiko hindeutet und es gar nicht so sehr darauf ankommt, um welchen Test es sich handelt. Gerade bei den ja im Grunde identischen Tests in der Darmkrebs-Studie, die nur an verschiedenen Tagen durchgeführt

wurden, erscheint dies durchaus plausibel. Weitere positive Tests lassen dann zwar trotzdem auf ein noch höheres Risiko schließen, wirken sich aber nicht mehr so drastisch aus. Für die Simulation dieses Szenarios wird der wahre Zusammenhang zwischen den binären Tests und dem Status wie folgt festgelegt:

$$\begin{aligned} P(S=1|V1,...,V8) &:= \frac{exp(Z)}{1+exp(Z)} \\ Z &:= -2 + 0.8 \cdot (V1+V5) + 0.7 \cdot (V2+V6) + 0.6 \cdot V3 + 0.5 \cdot V4 - 3 \cdot VZ \\ VZ &:= (1-V1) \cdot (1-V2) \cdot (1-V3) \cdot (1-V4) \cdot (1-V5) \cdot (1-V6) \end{aligned} \tag{4.7}$$

Der Interaktionsterm VZ ist also genau so gestaltet, dass er 1 wird, wenn alle Tests negativ sind und ansonsten den Wert 0 annimmt. Er führt dazu, dass die Krankheitswahrscheinlichkeit bei Personen ohne positiven Teiltest deutlich sinkt. Wieder soll die Anzahl der positiven binären Tests als Score verwendet werden. Es ergibt sich das folgende Bild:

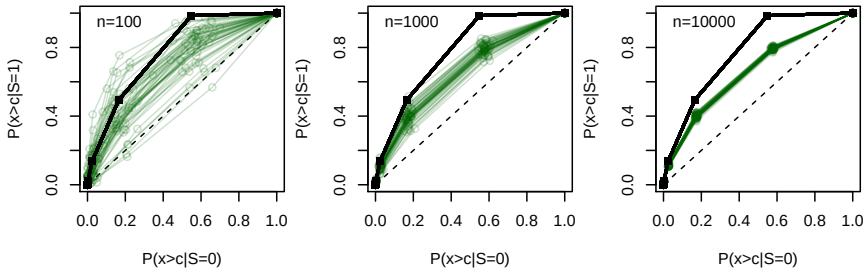

Abbildung 4.8: Geschätzte ROC-Kurven bei Stichprobenumfängen von 100, 1000 und 10000, falsche Modellgleichung

Dass das Verfahren hier nicht richtig funktionieren kann, liegt auf der Hand: Die Daten, die zur Verfügung stehen folgen den selben Verteilungen wie im ersten Beispiel, jedoch ist die wahre ROC-Kurve eine andere, da der Anteil der Kranken unter den nicht Verifizierten geringer ist. Problematisch ist dieser Punkt vor allem deshalb, weil die in der wahren Modellgleichung vorkommende Interaktion selbst dann, wenn man sie vermutet angesichts der Datenlage nicht modellierbar ist. Schließlich liegt in dem Teil des Datensatzes, in dem die Verifizierung verfügbar ist, keine Varianz bei VZ vor.

Weitere Szenarien, die noch zu überprüfen wären, sind etwa das Weglassen von Confoundern oder die falsche Spezifikation der Linkfunktion. Auch der Einfluss der Zahl der im Modell zu schätzenden Parameter wäre noch eine Untersuchung wert.

Literaturverzeichnis

Dorfman, D. D. and Alf, E. (1968). Maximum likelihood estimation of parameters of signal detection theory - a direct solution. *Psychometrika*, 33(1):117–124.

Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. Wiley, New York.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

Henking, A., Bluhm, C., and Fahrmeir, L. (2006). *Kreditrisikomessung. Statistische Grundlagen und Modellierung*. Springer, Berlin.

Krämer, W. and Bücker, M. (2009). Statistischer Vergleich von Kreditausfallprognosen.

Lloyd, C. J. and Frommer, D. J. (2008). An application of multinomial logistic regression to estimating performance of a multiple-screening test with incomplete verification. *Applied Statistics*, 57:89–102.

Metz, C. E. and Kronman, H. B. (1980). Statistical significance tests for binormal roc curves. *Journal of Mathematical Psychology*, 22(3):218–243.

Neuhäuser, M. (2011). Wilcoxon–mann–whitney test. In Lovric, M., editor, *International Encyclopedia of Statistical Science*, pages 1656–1658. Springer, Berlin.

Reitz, S. (2011). *Mathematik in der modernen Finanzwelt*. Vieweg+Teubner, Wiesbaden.

Sachs, L. and Hedderich, J. (2006). *Angewandte Statistik: Methodensammlung mit R*. Springer, Berlin.

Wehberg, S., Sauerbrei, W., and Schumacher, M. (2007). Diagnosestudien: Wertigkeit der sonographie bei der differenzierung von gut- und bösartigen brusttumoren bei patientinnen mit klinischen symptomen. In Schumacher, M. and Schulgen, G., editors, *Methodik klinischer Studien*, pages 319–340. Springer, Berlin.