

Kreditrisikomessung ist keine Geheimwissenschaft

vereinbarung (Basel II) werden Kreditinstitute alle Kunden einem Rating unterziehen. Im Rahmen der IRB-Ansätze ist dafür die jeweilige Ausfallwahrscheinlichkeit der Forderung zu bestimmen, für deren Schätzung im Wesentlichen logistische Regressionen verwendet werden.

Basel II: Logistische Regression als das Herz einer Rating-Maschine

it der Revision der Eigenkapitalübereinkunft aus dem Jahr 1988, kurz Basel II genannt, verfolgt der Baseler Ausschuss für Bankenaufsicht das Ziel, ein flexibles risikosensitives System für die Quantifizierung von Kreditrisiken zu schaffen. Insbesondere durch die Verwendung bankintern bestimmter Ratings soll das Ausmaß des eingegangenen Kreditrisikos bei einem Geschäft und damit die Höhe des für dieses Risiko als Puffer vorzuhaltenden Eigenkapitals ermittelt werden. Je schlechter das Rating eines Kunden ist, desto mehr Eigenkapital muss ein Institut für einen Kredit an diesen Kunden vorhalten. Basel II stellt die Ermittlung eines risikosensitiven Bonitätsgewichts für eine Forderung oder eine Geschäftsart in den Mittelpunkt der Kreditrisikoquantifizierung. Wählt ein Institut den internen Rating-Ansatz (Internal Ratings-Based Approach, IRB) als Alternative zum obligatorischen Standardansatz, hat es die Wahl zwischen dem Basisverfahren und dem fortgeschrittenen Verfahren.¹ Im IRB-Ansatz wird das Bonitätsgewicht insbesondere von dem Parameter "Ausfallwahrscheinlichkeit (Probability of Default; PD)" und einer vom Baseler Ausschuss vorgegebenen Risikogewichtsfunktion bestimmt (IRB-Formel).

Wie lassen sich ökonomische Zusammenhänge durch Gleichungen erfassen?

Die Institute prognostizieren im Rahmen des IRB-Ansatzes die Ausfallwahrscheinlichkeit PD eines Kredites mit Hilfe von statistischen Modellen. Ein sehr gebräuchliches statistisches Verfahren ist hierbei die logistische Regression. Eine übliche Fragestellung in der Ökono-

mie bezieht sich auf die Darstellung des Zusammenhanges zwischen zwei oder mehr Variablen. Man geht beispielsweise davon aus, dass ein Unternehmenskredit umso ausfallbedrohter ist, je höher die Verschuldung des Kreditnehmers ist. Umgekehrt sinkt die Wahrscheinlichkeit eines Ausfalles, wenn das Unternehmen nur mit Eigenkapital finanziert ist. Eine Vielzahl weiterer Einflussfaktoren (Risiko-Treiber) ist denkbar. Ausgangspunkt der Analyse ist daher die Überlegung, wie man eine Gruppe von potenziellen Risiko-Treibern untereinander gewichten soll, damit man mit ihnen eine optimale Prognose der Ausfallwahrscheinlichkeit abgegeben kann. Hier weiß die Statistik Rat. Mit Hilfe der Regressionsanalyse kann man eine abhängige (endogene) Variable Y mit einer Anzahl von unabhängigen (exogenen), also bereits bekannten, quantitativ messbaren Variablen $X_1, X_2,..., X_l$ erklären. Jede einzelne der exogenen Variablen geht mit dem Gewicht b_i in die Prognose ein, wobei der Index į die j-te Einflussvariable bezeichnet. Drückt man diesen Zusammenhang als Gleichung aus, so erhält man:

(1)
$$Y = a + b_1 \cdot X_1 + b_2 \cdot X_2 + ... + b_1 \cdot X_1$$
.

Der Wert a bezeichnet eine Konstante, die wahlweise in die Gleichung mit aufgenommen werden kann. Im Folgenden wird angenommen, dass nur eine einzige Einflussvariable existiert, diese wird mit X bezeichnet. Gleichung (1) vereinfacht sich dann zu

(2)
$$Y_i = a + b \cdot X_i$$
,

wobei der Index i den i-ten Schuldner bezeichnet. Somit ist Y, die Ausprägung der Variablen Y bei Schuldner i, analog interpretiert man X_i und u_i .

Ein zusätzlicher Term u_i wird eingeführt, der so genannte "Störterm", der die individuellen Abweichungen von der Gleichung (2)

einfangen soll. Selbst wenn Y maßgeblich von X verursacht wird, so wird man in der Praxis, d. h. bei der Betrachtung einzelner Kreditnehmer, immer Abweichungen feststellen. Insgesamt gilt also:

(3)
$$Y_i = a + b \cdot X_i + u_i$$
.

Über den Störterm nimmt man an, dass er im Durchschnitt den Wert Null annimmt, so dass sich, über alle Kreditnehmer hinweg gesehen, die positiven und negativen Abweichungen gerade aufheben. Ferner wird seine Varianz als konstant, also nicht schuldnerspezifisch, angenommen. Dieser Ansatz ist unter dem Namen "lineare Regression" bekannt,

Wie werden die Parameter a und b einer Regressionsgleichung ermittelt?

Nun stellt sich die Frage, wie man die Werte der Parameter a und b in Gleichung (3) ermittelt? In der Statistik gibt es hierauf nur eine Antwort: Sie müssen geschätzt werden! Hierzu wird eine Stichprobe von n Kreditnehmern benötigt, bei denen die Werte der Variablen Y und X bereits bekannt sind. Es werden dann genau die Werte von a und b gesucht, bei denen die Gleichung

$$(4) \quad \hat{Y}_i = \hat{a} + \hat{b} \cdot X_i$$

innerhalb der Stichprobe am besten erfüllt wird. Die Frage, wann dieses Kriterium erfüllt ist, wird von den verschiedenen Schätzverfahren unterschiedlich beantwortet (Box 1). Der Störterm bildet gerade die Abweichung der Schätzung von der Realität ab. Da die Parameter \hat{a} und \hat{b} auf Grundlage einer Stichprobe ermittelt werden, handelt es sich bei ihnen um Schätzwerte. Dies wird durch das "Dach" über einem Parameter angedeutet. Die Werte der exogenen Variable sind hingegen bekannt. Insgesamt folgt hieraus, dass auch der Wert von Yi nur geschätzt werden kann, so dass in Gleichung (4)

Vgl. Boos/Schulte-Mattler (2001), Basel II: Externes und internes Rating, in: Die Bank 5/2001, S. 346-354.

auch diese Variable von einem Dach geziert wird.

Wie prognostiziert die logistische Regression Kreditausfälle?

Bei der Konstruktion eines Rating-Systems möchte man insbesondere einen Zusammenhang zwischen dem Ausfallstatus eines Kreditnehmers und seinen potenziellen Risiko-Faktoren herstellen. Solche Faktoren ergeben sich insbesondere aus der Vermögens- und Kapitalstruktur (wie Eigenkapitalquote und Kapitalrückflussquote) und aus der Liquiditäts- und Ertragslage (wie Verschuldungsgrad und Eigenkapitalrendite) einer Unternehmung. Man steht jedoch vor der besonderen Schwierigkeit, dass die endogene Variable Y nur zwei Werte annehmen kann: Ausfall und Nicht-Ausfall des Kreditnehmers. Daher codiert man diese beiden Fälle mit

(5)
$$Y_i = \begin{cases} 1 & \text{falls der Kreditnehmer } i \\ \text{ausfällt} \\ 0 & \text{falls der Kreditnehmer } i \\ \text{nicht ausfällt} \end{cases}$$

Eine optimal Prognosegleichung für Y müsste also ebenfalls nur die Werte Null und Eins prognostizieren. Dies ist nicht ohne weiteres machbar, daher begnügt man sich damit, dass der Wert 0 nicht unterund der Wert 1 nicht überschritten wird. Um dies zu erreichen, wird die "logistische Regression" benutzt:

(6)
$$\hat{Y}_i = \frac{1}{1 + e^{-(\hat{a} + \hat{b} \cdot X_i)}}$$

Auch bei diesem Verfahren wird der Ausfallstatus Y mit Hilfe der Beziehung $\hat{a} + \hat{b} \cdot X$ erklärt, allerdings wird das Ergebnis dieser Gleichung durch den Bruch auf den Bereich von Null bis Eins normiert. Dieser Bruch entspricht der logistischen Verteilungsfunktion, daher auch der Name "logistische Regression". Grundsätzlich können verschie-

Tabelle 1: Änderung der PD bei Erhöhung von X (mit a = 0, b = 1).		
Х	PD	Zunahme der PD
-3	4,74%	-
-2	11,92%	7,18%
-1	26,89%	14,97%
0	50,00%	23,11%
1	73,11%	23,11%
2	88,08%	14,97%
3	95,26%	7,18%

dene Verteilungsfunktionen verwendet werden. So verwendet das in Box 4 beschriebene Probit-Modell die Standardnormalverteilung. Außerdem scheint es zunächst, dass bei der logistischen Regression kein Störterm berücksichtigt wird. Box 2 zeigt, dass dies jedoch nicht der Fall ist. Die Parameter a und b werden mit dem ML-Verfahren (siehe Box 1) geschätzt.

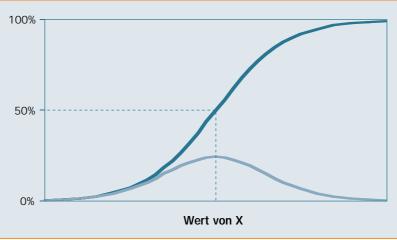
Im Gegensatz zu Y wird \hat{Y}_i auch Werte zwischen Null und Eins annehmen. Man interpretiert diese Werte als Ausfallwahrscheinlichkeiten: $\hat{Y}_i = \hat{P}D(Y_i = 1) = \hat{P}D_i$. Um auf den tatsächlichen Ausfallzustand zu schließen, definiert man häufig "Cut-Off-Werte": Wenn \hat{Y}_i diesen Wert übersteigt, wird ein Ausfall prognostiziert (man nimmt an, dass dann $Y_i = 1$ ist), andernfalls geht man vom Nicht-Ausfall aus. Al-

ternativ könnte man auf Basis von \hat{Y}_i Rating-Klassen bilden: Wenn \hat{Y}_i beispielsweise zwischen 0 und 0,001 liegt, geht man von einem "sicheren" Kredit aus, Werte zwischen 0,001 und 0,002 deuten eine geringe Ausfallwahrscheinlichkeit an etc. Jeder Klasse würde dann ihre Mittlere Ausfallwahrscheinlichkeit zugewiesen.

Wie werden die Parameter a und b einer Regressionsgleichung interpretiert?

Es ist offensichtlich, dass das Vorzeichen von b wie gewöhnlich interpretiert werden kann: Ist es positiv, so führt ein Anstieg von X zu einem Anstieg der Ausfallwahrscheinlichkeit. Wie sieht es aber aus, wenn man X um Eins erhöht? Macht es einen Unterschied, von welchem Niveau aus man dies tut?





Wirkt sich ein Anstieg von zwei auf drei genauso stark auf die Ausfallwahrscheinlichkeit aus wie ein Anstieg von drei auf vier? Die Antwort lautet: Nein! Die Interpretation der Werte der Parameter a und b erfordert nämlich einen weiteren Gedankenschritt: Die Berücksichtigung der Nichtlinearität. Ein einfaches Beispiel mit a=0 und b=1 mach dies deutlich. Die Ergebnisse sind Tabelle 1 angeführt, Abbildung 1 stellt sie grafisch dar. Der mathematisch interessierte Leser kann diesen Effekt ferner anhand der Ausführungen in Box 3 formal nachvollziehen.

Die schwarze Linie zeigt den Wert der PD bei verschiedenen Werten der Gleichung $a + b \cdot X$. Die graue Linie zeigt den Anstieg der PD (marginale PD), der mit dem Anstieg von X verbunden ist. Man erkennt deutlich, dass dieser marginale Anstieg zunächst selber ansteigt um anschließend wieder schwächer zu werden. Das Maximum der marginalen PD liegt genau dort, wo die PD selber den Wert 0,5 (oder 50 %) annimmt.

Wie erstellt man eine PD-Prognose?

Wenn eine logistische Regression geschätzt wurde und auch die Parameterwerte auf ihre Signifikanz und Plausibilität hin überprüft wurden, kann die so erhaltene Gleichung zur Prognose der PD eines neuen Kunden herangezogen werden. Hierzu setzt man den entsprechenden Wert der Variable X in die Gleichung (6) ein und erhält die geschätzte PD.

Worin besteht das Datenproblem?

Aus den bisherigen Ausführungen wird ersichtlich, dass der Aussagegehalt einer logistischen Regression mit der Qualität der Daten steht und fällt. Diese Qualität hat mehrere Aspekte:

Umfang des Datensatzes: Bei der Verwendung von Stichproben möchte man auf die zu Grunde liegende Grundgesamtheit schließen. Hat man nur geringe Stichprobenumfänge zur Verfügung, so besteht die Gefahr, dass einzelne Ausreißer übermäßig stark ins Gewicht fallen und die Stichprobe die Heterogenität der Grundgesamtheit überzeichnet. Dies gilt selbstverständlich auch umge-

Verteilung zwischen ausgefallenen und nicht-ausgefallenen Kreditnehmern (Defaulter und Non-Defaulter): Gerade beim Rating von Kreditnehmern sind Daten zu ausgefallenen Kreditnehmern von besonderem Interesse, so dass auch sie in ausreichender Anzahl vorliegen müssen. Faustregeln besagen, dass sie mindestens 5 % der Stichprobe ausmachen sollten, oftmals wird aber versucht, einen weitaus höheren Anteil zu erreichen: So kann man die speziellen Eigenschaften der Defaulter besser erfassen. Eine nachträgliche Anpassung der Schätzung an die tatsächliche Ausfallrate in der Grundgesamtheit ist problemlos möglich (Schichtung der Stichprobe).

Vollständigkeit: Die Daten eines Kreditnehmers können nur dann statistisch analysiert werden, wenn sie vollständig sind. Dies mag trivial klingen, ist aber ein häufig anzutreffendes Problem der angewandten Statistik.

Gleichartigkeit: Die verwendeten Daten müssen in ihrer Art identisch sein. So treten Probleme auf, wenn das Ereignis "Kreditausfall" von verschiedenen Sachbearbeitern unterschiedlich definiert wird. In diesem Zusammenhang sei auch die Baseler Ausfalldefinition aus Ziffer 414 des Dritten Konsultationspapieres genannt: Ein so "weiches" Kriterium wurde bisher von den wenigsten Banken verwendet, so dass die momentan entwickel-

ten Modelle die zukünftigen Baselkonformen Ausfallraten möglicherweise unterschätzen werden. Weitere Probleme treten auf, wenn Daten in verschiedenen Einheiten erfasst werden. In diesem Fall sind die Koeffizienten nicht verzerrungsfrei schätzbar.

Zeitnahe Verfügbarkeit: Die Variablen, mit denen man die Ausfallwahrscheinlichkeit prognostizieren will, müssen vor dem Ausfallereignis bekannt sein. Probleme bereiten beispielsweise Bilanzdaten. Für die Prognose der Ein-Jahres-PD des Jahres 2003 müssten die Risikotreiber aus den Bilanzdaten aus 2002 zum 1. Januar 2003 gebildet werden und in Gleichung (6) eingesetzt werden. Dann könnte man die PD für 2003 schätzen. Tatsächlich liegen diese Daten oft erst in der Jahresmitte vor, so dass die dann geschätzte PD sich nur noch auf eine halbes Jahr bezieht. Verwendet man statt dessen Daten aus 2001, so würde man im Nachhinein die PD für 2002 schätzen. Daher ist bei Auswahl und Erhebung der Daten sehr genau auf ihre zeitliche Verfügbarkeit zu achten.

Welches Fazit lässt sich ziehen?

Bei den statistischen Methoden, die in der Kreditrisikomessung eingesetzt werden, handelt es sich nicht um "Geheimwissenschaften". Mit ihnen können zuverlässige Risiko-Einschätzungen vorgenommen werden, die den Vorteil einer hohen Transparenz aufweisen. Der mechanische Charakter dieser Verfahren sollte aber stets bedacht werden, so dass auch eine PD-Schätzung mittels logistischer Regressionen interpretiert werden muss. Die Qualität eines Rating-Verfahrens, das auf statistischen Methoden basiert, hängt wesentlich von der Qualität der verwendeten Daten ab.

Box 1: Schätzverfahren

Ein Schätzverfahren ist eine Methode, um Schätzungen für die Parameter a und b zu erhalten. In der Statistik haben sich zwei Schätzverfahren durchgesetzt:

- OLS-Verfahren: Das Ordinary-Least-Squares-Verfahren wählt \hat{a} und \hat{b} so, dass innerhalb der Stichprobe die Summe der quadrierten individuellen Abweichungen der Prognose von der Realität minimal ist. Die Idee ist offensichtlich, dass die Werte von a und b umso besser geschätzt werden, je geringer die resultierenden Störterme sind.
- ML-Verfahren: Das Maximum-Likelihood-Verfahren wählt \hat{b} so, dass die Wahrscheinlichkeit, genau die vorliegende Stichprobe zu erhalten, maximal ist. Hier überlegt man also, woher die Stichprobe stammen könnte.

Box 2: Wo ist der Störterm u?

Um die Rolle und die Erscheinungsform des Störterms einer logistischen Regression darzustellen, muss etwas weiter ausgeholt werden. Man nimmt zunächst an, dass der Ausfallstatus eines Kunden von einer nicht beobachtbaren Größe Y* abhängt, einer Art "Bonitätsindex". Wenn Y^* einen Wert kleiner oder gleich Null annimmt ($Y^* \le 0$), so fällt der Kunde aus und die beobachtbare Variable "Ausfallzustand", als Y bezeichnet, nimmt für diesen Kunden den Wert Eins an $(Y_i = 1)$. Gilt hingegen $Y_i^* > 0$, so kann Kunde i seinen Kredit weiterhin bedienen ($Y_i = 0$). Offenbar gilt für die Ausfallwahrscheinlichkeit (PD) von Kunde i:

$$PD_i = P(Y_i = 1) = P(Y_i^* \le 0)$$

Wird weiterhin angenommen, dass der Bonitätsindex Y^* linear von der Variablen X und dem Störterm u abhängt: $Y_i^* = c + d \cdot X_i$ + u_i . Für die PD gilt somit:

$$PD_i = P(Y_i^* \le 0) = P(c + d \cdot X_i + u_i \le 0) = P(u_i \le -c - d \cdot X_i)$$

Der Störterm u ist annahmegemäß eine logistisch verteilte Zufallsvariable. Wenn F(...) die Verteilungsfunktion der logistischen Verteilung bezeichnet, so gilt:

$$PD_i = F(-c - d \cdot X_i)$$

Mit a = -c und b = -d erhält man die bereits bekannt Formulierung der logistischen Regression

$$PD_i = F(a + b \cdot X_i) = \frac{1}{1 + e^{-(\hat{a} + b \cdot X_i)}}$$

Man erkennt, dass der Störterm u nur scheinbar wegfällt. Er ist nach wie vor durch die Verteilungsfunktion präsent und übt einen maßgeblichen Einfluss auf die PD aus.

Box 3: Die Niveauabhängigkeit des Einfluss von X auf die PD

Um den Effekt einer Erhöhung des unabhängigen Risikotreibers X auf die Ausfallwahrscheinlichkeit PD zu ermitteln, muss die erste Ableitung der PD-Funktion nach X gebildet werden. Es gilt offenbar:

$$\frac{\partial PD}{\partial X} = b \cdot \frac{e^{-(a+b \cdot X)}}{(1+e^{-(a+b \cdot X)})^2} = b \cdot PD \cdot (1-PD)$$

Das Vorzeichen des gesamten Ausdrucks hängt nur von b ab, da die anderen beiden Faktoren mit PD und 1 – PD immer positiv sind. Die Wirkungsrichtung von X kann also am Vorzeichen von b festgemacht werden. Man erkennt aber auch, dass der Einfluss von X auf die PD von X selber abhängt, der marginale Einfluss ist also nicht konstant. Querverweis: Im Falle der linearen Regression wäre der marginale Einfluss von X auf Y konstant und entspräche b.

Box 4: Das Probit-Modell

Das Probit-Modell verwendet die Standardnormalverteilung zur Ermittlung der PD. Es gilt:

$$PD_i = F_N (a + b \cdot X_i)$$

wobei F_N die Verteilungsfunktion der Standardnormalverteilung bezeichnet. So wird die Annahme des logistisch verteilten Störtermes durch die Annahme eines standardnormalverteilten Störtermes ersetzt. Ein wesentlicher Unterschied zwischen diesen Verteilungen ist, dass die logistische Verteilung mehr Wahrscheinlichkeitsmasse auf den Flanken aufweist, so dass sehr extreme Ausreißer eine höhere Wahrscheinlichkeit als bei einer Normalverteilung besitzen. Der marginale Einfluss von X auf Y ist auch im Probit-Modell nicht konstant, sondern abhängig vom Niveau von X. Es gilt

$$\frac{\partial PD}{\partial X} = b \cdot f_N \left(a + b \cdot X \right)$$

wobei f_N (...) die Dichtefunktion der Standardnormalverteilung bezeichnet. Auch im Probit-Modell weist der marginale Effekt einen glockenförmigen Verlauf in Abhängigkeit von X auf und ist in seiner Wirkungsrichtung nur vom Vorzeichen von b abhängig.

LITERATUREMPFEHLUNGEN

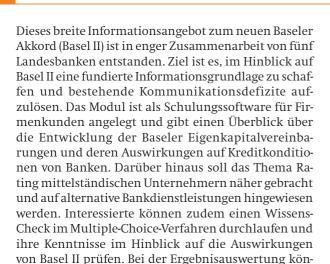
- Kaiser, U.; A. Szczesny (2000), Einfache ökonometrische Verfahren für die Kreditrisikomessung: Logit- und Probit-Modelle, Johann Wolfgang Goethe-Universität Frankfurt am Main, Fachbereich Wirtschaftswissenschaften, Working Paper No. 61, Dezember 2000. Ein Übersichtsartikel, der auch die mathematischen Aspekte des Themas ein wenig höher gewichtet, ohne dabei den Praxisbezug zu verlieren.
- Pindyck, R. S.; D. L. Rubinfeld (2000), Econometric Models and
- Economic Forecasts, 5. Auflage, Boston (McGraw Hill) 2000. Der Leser findet in diesem gängigen Lehrbuch der Ökonometrie im Kapitel 4 eine vertiefte und verallgemeinerte Darstellung der linearen Regression in Vektorenschreibweise. In Kapitel 11 findet sich eine gut lesbare Einführung in die logistische Regression.
- King, G.; L. Zeng (2001), "Logistic Regression in Rare Events Data", in: Political Analysis 2001 (9), No. 2, S. 137-163. In diesem Beitrag findet sich näheres zum Problem der "Schichtung" der Stichprobe.
- · Hosmer, D. W.; S. Lemeshow (2000), Applied Logistic Regression, 2. Auflage, New York (John Wiley & Sons) 2000. In diesem Buch wird auch die logistische Regression ausführlich in Vektorschreibweise dargestellt. Im Kap. 6 wird ferner auf das Problem der "Schichtung" der Stichprobe eingegangen.

Prof. Dr. Hermann Schulte-Mattler lehrt Finanzwirtschaft an der FH Dortmund. Ulrich Daun

ist Mitarbeiter im Rating-Gemeinschaftsprojekt des Bank-Verlags, Köln, und dort für Statistik zu-

Neues

Informationsportal zum Thema Basel II





nen die richtigen Antworten bei Bedarf in einer kommentierten Langversion eingesehen werden. Die Internet-Seite hält auch anschauliche Texte über die Konditionengestaltung bei der Kreditvergabe, Finanzierungsalternativen zum klassischen Bankkredit sowie einige Buch- und Link-Tipps, ein Glossar und eine Sammlung häufig gestellter Fragen vor. Informativ gestaltet sich der Abschnitt zum Themenkomplex Rating. Neben den verschiedenen Basel-II-Ansätzen liegt ein Schwerpunkt auf dem internen Rating. Positiv fällt auf, dass den Adressaten auch mögliche Nachteile eines schlechten Ratings vor Augen geführt werden. Die Texte sind durchgehend mit Links unterlegt, die einzelne Fachbegriffe über das Glossar leicht aufzulösen.

Link-Tipp: Kunden-Modul "Basel II"

> http://basel2.sachsenlb.de