



Koeffizienten der Logitanalyse

Kurt Holm

Almo Statistik-System
www.almo-statistik.de
holm@almo-statistik.de
kurt.holm@jku.at

Autor: em. Prof. Dr. Kurt Holm, Universität Linz, Österreich

Siehe auch das Almo-Dokument Nr. 9 "Logitanalyse".

Im Text wird häufig auf das Dokument **P0** Bezug genommen. Dabei handelt es sich um das Almo-Dokument "Arbeiten mit Almo.PDF" (Dokument 0).

Weitere Almo-Dokumente

Die folgenden Dokumente können alle kostenlos von der Handbuchseite in <http://www.almo-statistik.de> heruntergeladen werden.

0. Arbeiten_mit_Almo.PDF (1 MB)
- 1a. Eindimensionale Tabellierung.PDF (1.8 MB)
- 1b. Zwei- und drei-dimensionale Tabellierung.PDF (1.1 MB)
2. Beliebig-dimensionale Tabellierung.PDF (1.7 MB)
3. Nicht-parametrische Verfahren.PDF (0.9 MB)
4. Kanonische Analysen.PDF (1.8 MB)
Diskriminanzanalyse.PDF (1.8 MB)
enthält: Kanonische Korrelation, Diskriminanzanalyse, bivariate Korrespondenzanalyse, optimale Skalierung
5. Korrelation.PDF (1.4 MB)
6. Allgemeine multiple Korrespondenzanalyse.PDF (1.5 MB)
7. Allgemeines ordinales Rasch-Modell.PDF (0.6 MB)
- 7a. Wie man mit Almo ein Rasch-Modell rechnet.PDF (0.2 MB)
8. Tests auf Mittelwertsdifferenz, t-Test.PDF (1,6 MB)
9. Logitanalyse.pdf (1,2MB) enthält Logit- und Probitanalyse
10. Koeffizienten der Logitanalyse.PDF (0,06 MB)
11. Daten-Fusion.PDF (1,1 MB)
12. Daten-Imputation.PDF (1,3 MB)
13. ALM Allgemeines Lineares Modell.PDF (2.3 MB)
- 13a. ALM Allgemeines Lineares Modell II.PDF (2.7 MB)
14. Ereignisanalyse: Sterbetafel-Methode, Kaplan-Meier-Schätzer, Cox-Regression.PDF (1,5 MB)
15. Faktorenanalyse.PDF (1,6 MB)
16. Konfirmatorische Faktorenanalyse.PDF (0,3 MB)
17. Clusteranalyse.PDF (3 MB)
18. Pisa 2012 Almo-Daten und Analyse-Programme.PDF (17 KB)
19. Guttman- und Mokken-Skalierung.PFD (0.8 MB)
20. Latent Structure Analysis.PDF (1 MB)
21. Statistische Algorithmen in C (80 KB)
22. Conjoint-Analyse (PDF 0,8 MB)
23. Ausreisser entdecken (PDF 170 KB)
24. Statistische Datenanalyse Teil I, Data Mining I
25. Statistische Datenanalyse Teil II, Data Mining II
26. Statistische Datenanalyse Teil III, Arbeiten mit Almo-Datenanalyse-System
27. Mehrfachantworten, Tabellierung von Fragen mit Mehrfachantworten (0.8 MB)
28. Metrische multidimensionale Skalierung (MDS) (0,4 MB)
29. Metrisches multidimensionales Unfolding (MDU) (0,6 MB)
30. Nicht-metrische multidimensionale Skalierung (MDS) (0,5 MB)
31. Pfadanalyse (0,7 MB)
32. Datei-Operationen mit Almo (1,1 MB)
33. Wählerstromanalyse und Wahlhochrechnung.PDF (1,6 MB)
34. Soziometrie. Auswertung soziometrischer Daten (0,5 MB)

Kurt Holm

Koeffizienten der Logitanalyse

Eine häufig gestellte Frage lautet:

Wie sind die im Rahmen der Logit-Analyse errechneten Regressionskoeffizienten und Risiko-Koeffizienten zu interpretieren ?

Vorweg ist folgendes zu sagen:

Anstelle des Begriffs "Risiko-Koeffizient", den wir hier verwenden, wird in der Literatur auch der Begriff "Effekt-Koeffizient" gebraucht (so bei D. Urban, 1993).

Betrachten wir ein Beispiel:

Kunden kaufen auf Kredit. Werden sie ihren Kreditrückzahlen ?

Die Variablen für unser Beispiel sollen folgende sein:

Die Zielvariable ist Kredit-Rückzahlung: nein,
ja

Zur Erklärung der Zielvariablen werden folgende Variable eingeführt:

Die unabhängigen nominalen Variablen sind
Wohnort: Stadt
Land
Hausbesitz: kein Haus
hat Haus

Die unabhängigen quantitativen Variablen sind
Einkommen
Rückzahlungsrate
Kredit-Laufzeit

Almo liefert folgende Ausgabe (verkürzt).

Ergebnisse fuer 2. Auspraegung "ja" der abhaengigen Variablen V10 Rückzahl
(als Referenz wird die 1. Auspraegung "nein" verwendet)

unabhaengige Variable	Regress. koeff.β	Risiko ep̄(β)	relatives Risiko	Signifikanz (1-p)*100	partielle Korrelation
A1 Wohnort: Stadt	-0.43493	0.64731	-35.26902	100.00	-0.13168
A2 Wohnort: Land	0.43493	1.54486	54.48553	100.00	0.13168
B1 Hausbesi:kein Hau	-0.74569	0.47440	-52.55955	100.00	-0.15825
B2 Hausbesi:hat Haus	0.74569	2.10791	110.79059	100.00	0.15825
V4 Einkommen	0.68943	1.99257	99.25744	100.00	0.25486
V7 Rueckrate	-0.00077	0.99923	-0.07689	100.00	-0.25619
V8 Laufzeit	0.04562	1.04667	4.66727	99.22	0.06526

Betrachten wir die beiden Regressionskoeffizienten für den Wohnort

A1 Wohnort: Stadt -0.43493
A2 Wohnort: Land 0.43493

Das Logit-Modell lautet

$$P_1 = \frac{1}{1 + e^{-(c + a(i) + b(j) + \beta_1 \cdot E + \beta_2 \cdot R + \beta_3 \cdot L_j)}}$$

Diese Gleichung kann so umgewandelt werden, daß auf der rechten Seite ein linearer Ausdruck steht

$$(1) \quad \ln(p1/p2) = c + a(i) + b(j) + \beta1 \cdot E + \beta2 \cdot R + \beta3 \cdot L$$

p1=Wahrscheinlichkeit für Kreditkauf: ja
p2=Wahrscheinlichkeit für Kreditkauf: nein (p2=1-p1)
Natürlich gilt: p2 = 1-p1
e =e-Zahl 2.718
c =Konstante

a(i) bezeichnet die Regressionskoeffizienten für die 2
Dummy-Variable des Wohnorts
b(j) bezeichnet die Regressionskoeffizienten für die 2
Dummy-Variable des Hausbesitz

es ist also:

a1=Regressionskoeffizient für "Stadt"
a2=Regressionskoeffizient für "Land"

E =Einkommen
β1=Regressionskoeffizient für Einkommen

R =Rueckrate
β2=Regressionskoeffizient für Rueckrate

L =Laufzeit
β3=Regressionskoeffizient für Laufzeit

Regressionskoeffizienten der nominalen Variablen

Der Regressionskoeffizienten a1=-0.43493 für "Stadt" und a2=0.43493 für "Land" haben folgende Bedeutung:

1. Das negative Vorzeichen von a1 drückt aus, daß Städter im Vergleich zur "Durchschnittsperson" das logarithmierte Wahrscheinlichkeitsverhältnis $\ln(p1/p2)$ aus Gleichung 1 verringern. Vereinfacht: Städter haben eine geringere Wahrscheinlichkeit ihren Kredit zurückzuzahlen. Umgekehrt drückt das positive Vorzeichen von a2 aus, daß Leute vom Land eine erhöhte Wahrscheinlichkeit haben ihren Kredit zurückzuzahlen.

2. Je (absolut) größer der Regressionskoeffizient ist, umso stärker ist diese Tendenz.

Regressionskoeffizienten der quantitativen Variablen

Der Regressionskoeffizient β1=0.68943 für "Einkommen" hat folgende Bedeutung: Wenn sich das Einkommen um 1 Einheit erhöht, dann erhöht sich das logarithmierte Wahrscheinlichkeitsverhältnis $\ln(p1/p2)$. Vereinfacht: Wenn sich das Einkommen um 1 Einheit erhöht, dann nimmt die Wahrscheinlichkeit zu, den Kredit zurückzuzahlen. Ein negatives Vorzeichen würde bedeuten, dass sich die Wahrscheinlichkeit verringert. Je (absolut) größer der Regressionskoeffizient ist, umso stärker ist diese Tendenz.

Der Risiko-Koeffizient exp (β)

Unser Beispiel ist relativ komplex. Wir haben 2 ursächliche nominale Variable und 3 ursächliche quantitative Variable.

Um unsere Erläuterung übersichtlich gestalten zu können, wollen wir ein anderes, einfacheres Beispiel betrachten, bei dem nur 1 ursächliche nominale und 1 ursächliche quantitative Variable vorhanden ist.

Die Variablen für unser vereinfachtes Beispiel sollen folgende sein:

Die Zielvariable ist Kredit-Rückzahlung: nein,
ja

Die unabhängige nominale Variable ist Beruf: Arbeiter,
Angestellter,
Selbständiger

Die unabhängige quantitative Variable ist: Einkommen
Sie wird in Einkommensklassen mit den Werten 1,2,3, ...,9 gemessen.

Almo liefert folgendes Ergebnis:

Ergebnisse für 2. Ausprägung "ja" der abhängigen Variablen "Rückzahlung"
(die Ausprägung "nein" wird als Referenzkategorie verwendet)

unabhängige Variable	Regress. Koeffiz.	"Risiko" exp (Regr.- koeffiz.)	relatives Risiko in %
c Konstante	1.88227	-	-
a1 Beruf:Arbeiter	1.37706	3.96324	296.32376
a2 Beruf:Angestellte	-0.92524	0.39644	-60.35623
a3 Beruf:Selbständige	-0.45182	0.63647	-36.35343
X Einkommen	-0.37586	0.68670	-31.33039

Die Logit-Modell-Gleichung ist folgende:

$$p1 = \frac{1}{1 + e^{-(c+a(i)+\beta \cdot x)}}$$

Man beachte:p1 ist die Wahrscheinlichkeit für die 2. Ausprägung "ja" der Zielvariablen "Rückzahlung". Mit p2 werden wir die Wahrscheinlichkeit für die Referenzkategorie "nein" bezeichnen

Diese Gleichung kann so umgewandelt werden, dass auf der rechten Seite ein linearer Ausdruck steht.

$$(1) \quad \ln(p1/p2) = c + a(i) + \beta X$$

p1=Wahrscheinlichkeit für Rückzahlung: ja
p2=Wahrscheinlichkeit für Rückzahlung: nein

Natürlich gilt: $p_2 = 1 - p_1$
 c =Konstante

$a(i)$ bezeichnet die Regressionskoeffizient für die 3
Dummy-Variable des Berufs (die den 3 Ausprägungen entsprechen)

es ist also:

a_1 =Regressionskoeffizient für "Arbeiter"
 a_2 =Regressionskoeffizient für "Angestellter"
 a_3 =Regressionskoeffizient für "Selbständiger"

X =Einkommen

β =Regressionskoeffizient für Einkommen

Für einen Arbeiter in der Einkommensklasse $X=4$ lautet also die Gleichung

$$(1a) \quad \ln(p_1/p_2) = c + a_1 + \beta X \\ = 1.88 + 1.38 - 0.38 \cdot 4$$

Gleichung 1 bzw. 1a kann so transformiert werden, dass der auf der linken Gleichungsseite stehende Logarithmus verschwindet.

$$(2) \quad p_1/p_2 = \exp(c) * \exp(a(i)) * \exp(\beta * X)$$

$\exp(\dots)$ = Exponentialfunktion von ...

Für unseren Arbeiter mit Einkommen $X=4$

$$(2a) \quad p_1/p_2 = \exp(c) * \exp(a_1) * \exp(\beta * X) \\ = \exp(1.88) * \exp(1.38) * \exp(-0.38 * 4) \\ = 6.62 * 3.96 * 0.22 \\ = 5.7886$$

Zuerst ist festzuhalten, dass sich die Interpretation auf die 2. Ausprägung der Zielvariablen also auf "Rückzahlung: Ja" bezieht.

p_1 ist also die Wahrscheinlichkeit für Rückzahlung: ja

p_2 ist also die Wahrscheinlichkeit für Rückzahlung: nein

Das Wahrscheinlichkeits-Verhältnis p_1/p_2 wird in der angelsächsischen Literatur "odds" genannt.

Wenn man p_1 als Gewinn-Wahrscheinlichkeit und p_2 als Verlust-Wahrscheinlichkeit interpretiert, dann könnte man p_1/p_2 als "Gewinn-zu-Verlust-Verhältnis" bezeichnen.

Ist die Zielvariable, wie in unserem Beispiel, dichotom, dann gilt

$$p_2 = 1 - p_1$$

Ist $p_1=0.5$ dann ist p_2 auch $=0.5$. Dann ist $p_1/p_2=1$. Das "Gewinn-zu-Verlust-Verhältnis" ist also ausgeglichen.

Ist $p_1=0.6666..$ dann ist $p_2=0.33333..$ Dann ist $p_1/p_2 = 2$. Die Gewinn-Chance ist 2 mal besser als die Verlust-Chance

In unserem Beispiel ist $p_1/p_2=5.7886$. Für unseren Arbeiter mit einem Einkommen von 4 gilt

also, dass seine Wahrscheinlichkeit den Kredit zurückzuzahlen 5.7886 mal größer ist als ihn nicht zurückzuzahlen.

Wie groß ist dann p_1 ?

Hier gilt die allgemeine Formel:

$$\begin{aligned} p_1 &= f / (1+f) \\ &= 5.7886 / (1+5.7886) \\ &= 0.853 \end{aligned}$$

wobei $f=p_1/p_2$

Die Wahrscheinlichkeit unseres Arbeiters mit Einkommen 4 den Kredit zurückzuzahlen ist also $p_1=0.853$.

Betrachten wir einige Werte von p_1

p_1	dann ist $p_2= 1-p_1$	"Gewinn-zu-Verlust-Verhältnis" p_1/p_2
0.1	0.9	0.111
0.2	0.8	0.250
0.3	0.7	0.429
0.4	0.6	0.667
0.5	0.5	1
0.6	0.4	1.500
0.7	0.3	2.333
0.8	0.2	4
0.9	0.1	9

Betrachten wir nun wieder Gleichung 2 bzw. 2a. Alle Arbeiter haben - im Vergleich zum Durchschnitt aller Untersuchungspersonen - eine um den Faktor $\exp(a_1) = 3.96$ erhöhtes Wahrscheinlichkeits-Verhältnis p_1/p_2 , d.h. ihre Wahrscheinlichkeit den Kredit zurückzuzahlen ist erhöht.

Dieser Faktor wird in der Literatur gelegentlich "Risiko" genannt. Auch der Begriff "Effekt-Koeffizient" wird gelegentlich gebraucht (so bei D. Urban: Logit-Analyse, Gustav Fischer, Stuttgart, 1993).

Wäre $\exp(a_1)=1$, dann würden sich die Arbeiter so verhalten wie der Durchschnitt.

Wir definieren nun als

$$\text{relatives Risiko} = (\exp(a(i)) - 1) * 100$$

Für die Arbeiter finden wir dann

$$\begin{aligned} \text{relatives Risiko} &= (\exp(a_1) - 1) * 100 \\ &= (3.96 - 1) * 100 \\ &= 296 \end{aligned}$$

Wir können jetzt formulieren: Arbeiter haben ein um 296 % höheres Risiko einen Kredit zurückzuzahlen als die durchschnittliche Untersuchungsperson.

Zu beachten ist, dass die Bezugskategorie der Durchschnitt aller Untersuchungs-personen ist. Dies ist in Almo der Fall, wenn die 0,1,-1 - Kodierung der Dummies der unabhängigen nominalen Variablen verwendet wird. Dies ist die Voreinstellung in Almo.

Wird die 0,1 - Kodierung verwendet, dann wird (standardmäßig) die letzte Dummy, in unserem Beispiel die Selbständigen, auf 0 gesetzt. Sie erscheint dann auch gar nicht in der Ergebnis-Ausgabe.

Almo liefert folgendes Ergebnis (verkürzt):

Ergebnisse für 2. Auspräg. "ja" der abhängigen Variablen "Rückzahlung"

unabhängige Variable	Regress. Koeffiz.	"Risiko" exp(Regr.-koeffiz.)	relatives Risiko
c Konstante	1.43044	-	-
a1 Beruf:Arbeiter	1.82889	6.22695	522.69462
a2 Beruf:Angestellte	-0.47341	0.62287	-37.71264
X Einkommen	-0.37586	0.68670	-31.33039

Die Selbständigen sind jetzt die Bezugskategorie. Die Arbeiter haben im Vergleich zu den Selbständigen eine um 522 % erhöhte Wahrscheinlichkeit den Kredit zurückzuzahlen und die Angestellten eine um 37.7 % reduzierte Wahrscheinlichkeit.

In Almo ist es bei der 0,1 - Kodierung möglich, entweder die erste oder die letzte Dummy zu eliminieren.

Allgemein gilt:

- a. Bei der 0,1 - Kodierung ist die Bezugskategorie die eliminierte Dummy.
- b. Bei der 0,1,-1 - Kodierung ist die Bezugskategorie der Durchschnitt aller Untersuchungspersonen.

Risiko bei quantitativen Variablen

Betrachten wir nochmals obige Gleichung (2)

$$(2) \quad p_1/p_2 = \exp(c) * \exp(a(i)) * \exp(\beta * X)$$

Das Einkommen unseres Arbeiters ist $X=4$.

Der Ausdruck $\exp(\beta * X)$ ist also $\exp(-0.37586 * 4) = 0.22236$

Wenn sich das Einkommen dieser Person um 1 Einheit erhöht, dann ist der Ausdruck $\exp(\beta * X) = \exp(-0.37586 * 5) = 0.15270$

Wenn wir für $X=5$ obige Gleichung (2) für unsere Person ausrechnen, dann erhalten wir

$$p_1/p_2 = 3.9750$$

Für $X=4$ haben wir oben errechnet

$$p_1/p_2 = 5.7886$$

So hat sich also $p1/p2$ um den multiplikativen Faktor

$$3.9750 / 5.7886 = 0.68670$$

verringert. Und das ist genau das in obiger Tabelle angegebene Risiko $\exp(\beta)$.

Risiko-Werte unter 1 führen zu einer Verringerung von $p1/p2$. D.h. $p1$ wird kleiner und $p2$ wird größer.

Risiko-Werte über 1 führen zu einer Erhöhung von $p1/p2$. D.h. $p1$ wird größer und $p2$ wird kleiner.

Wir können nun den Begriff "Risiko" ($=\exp(\beta)$) bei ursächlichen quantitativen Variablen allgemein definieren.

Nimmt die ursächliche quantitative Variable X um 1 Einheit zu, dann nimmt das Wahrscheinlichkeits-Verhältnis $p1/p2$ um den multiplikativen Faktor $\exp(\beta)$ zu.

Wir können diese Zunahme bzw. Abnahme auch in Prozentwerten ausdrücken. Sie beträgt dann $100(\exp(\beta)-1)$. Das ist das relative Risiko.

Betrachten wir für Arbeiter die Werte, die sich gemäß Gleichung 2 für Einkommenswerte X von 0 bis 6 ergeben.

<u>X</u>	<u>p1/p2</u>	<u>Multiplikator</u>
0	26.0326	
1	17.8765	0.6867
2	12.2758	0.6867
3	8.4298	0.6867
4	5.7886	0.6867
5	3.9750	0.6867
6	2.7297	0.6867

Das Wahrscheinlichkeits-Verhältnis $p1/p2$ einer nachfolgenden Einkommensstufe entsteht durch Multiplikation mit $\exp(\beta)=0.6867$ des Wahrscheinlichkeits-Verhältnis $p1/p2$ der vorhergehenden Einkommensstufe.

Literatur:

Almo-Dokument Nr.

9. Logitanalyse

24. Statistische Datenanalyse Teil I, Data Mining I

25. Statistische Datenanalyse Teil II, Data Mining II

Arminger, Küsters: Statistische Verfahren zur Analyse qualitativer Variablen, Bergisch Gladbach, 1986

G.S. Maddala: Limited-dependent and qualitative variables in econometrics, Cambridge, 1990

Dieter Urban: Logit-Analyse, Gustav Fischer, Stuttgart, 1993