

## 8 Logistische Regressionsanalyse

Die logistische Regressionsanalyse dient der Untersuchung des Einflusses einer quantitativen Variable auf eine qualitative (in unserem Fall dichotomen) Variable. Wir gehen also davon aus, dass die abhängige Variable nur zwei Kategorien (0 oder 1, ja oder nein) aufweist. Als Beispiel könnte man sich folgendes vorstellen: Es werden Personen, die drogenabhängig waren, gefragt, wie lange sie schon abstinent sind, und ob sie momentan noch Entzugserscheinungen haben. Als Grundlage für die logistische Regressionsanalyse dient eine Kontingenztafel. Es sei bemerkt, dass es zu jedem Wert der quantitativen Variable (im Beispiel die Abstinenzdauer) genügend viele Beobachtungen geben muss, ansonsten müssten Intervalle gebildet bzw. Werte zusammengefasst werden.

unabhängige Variable	Abhängig Variable	
	Anzahl "ja"	Anzahl "nein"
1	9	2
2	8	4
3	7	5
4	5	8
5	5	9
6	3	9
7	2	9
8	2	10
9	2	11

v1	v2
1	ja (insgesamt 9-mal)
.....	
1	nein (insgesamt 2-mal)
.....	
.....	
9	nein (insgesamt 11-mal)
.....	

Wir könnten nun die Daten, wie oben rechts zu sehen ist, eingeben, was aber zu aufwändig ist. Aus diesem Grund geben wir die Tabelle direkt ein. Dies können wir auf der Seite <http://statistikpaket.de/LogReg/LogReg2.html> tun. Dazu müssen wir

neben „Anzahl Zeilen“ eine 9 eingeben. Danach kann man  
 →**Häufigkeiten eingeben** wählen, die Tabelle eingeben und dann  
 →**Berechnung starten** wählen.

Bevor wir zu der weiteren Interpretation der Ergebnisse kommen, gehen wir zunächst auf die Theorie ein. Der logistischen Regression liegt das folgende Modell zu Grunde:

$$P(Y = \text{„ja“} \mid x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Wir nehmen also an, dass die Wahrscheinlichkeit dafür, dass eine Personen mit „ja“ antwortet, unter der Bedingung, dass die unabhängige Variable den Wert x aufweist (z.B. Abstinenzdauer = 1 Jahr) eine Funktion der unabhängigen Variable x ist. Die oben zu sehende Funktion ist eine so genannte logistische Funktion. Falls  $\beta_1 > 0$  ist, so steigt die Funktion mit größer werdendem x an, d.h. die Wahrscheinlichkeit wird größer. In unserem Beispiel gehen wir von einem negativen  $\beta_1$  aus, da anzunehmen ist, dass die Wahrscheinlichkeit dafür, dass Entzugserscheinungen auftreten, mit der Zeit abnimmt. Von Interesse ist also zunächst, ob der Parameter  $\beta_1$  ungleich Null ist. Außerdem muss untersucht werden, ob das Modell angemessen ist. Dies kann wieder mit einem Modellanpassungstest überprüft werden.

Um die Parameter zu schätzen, wird zunächst die Wahrscheinlichkeit auf der linken Seite der oberen Gleichung durch die relative Häufigkeit der jeweiligen Personen ersetzt, die mit ja geantwortet haben. Es wird also die relative Anzahl für jede Zeile der oberen Tabelle gebildet. Die Werte (bis auf den Faktor 100%) finden Sie in der vierten Spalte der oberen Tabelle. Es ergibt sich somit die folgende Gleichung:

$$\hat{p}_i \approx \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}; i = 1, \dots, k \quad (*)$$

$k$  ist dabei die Anzahl der verschiedenen Beobachtungen der unabhängigen Variablen  $x$ . In unserem Beispiel ist  $k = 9$ ;  $\hat{p}_i$  ist die relative Häufigkeit der Personen die die Frage mit „ja“ beantwortet haben und für die gilt  $x = x_i$ . Im Beispiel ist  $x_1 = 1, x_2 = 2, \dots, x_9 = 9$ . Wir haben anstelle des Gleichheitszeichens das Zeichen  $\approx$  verwendet, da es durch die Schätzung der Wahrscheinlichkeiten natürlich Abweichungen geben kann. Sonst müsste man einen Fehlerterm  $e_i$  (analog zu dem linearen Regressionsmodell) einführen. Im Beispiel lautet die obere Gleichung für  $x = x_1 = 1$ :

$$0,8182 \approx \frac{e^{\beta_0 + \beta_1 \cdot 1}}{1 + e^{\beta_0 + \beta_1 \cdot 1}}$$

Zur Schätzung kann man aus der Gleichung (\*), welche bezüglich der Parameter nicht linear ist, durch eine Transformation eine lineare Funktion erhalten. Dazu verwendet man die Umkehrfunktion der logistischen Funktion, die so genannte Logit-Funktion, die wie folgt definiert ist:  $\text{logit}(t) = \ln\left(\frac{t}{1-t}\right)$ .

Durch Anwendung der Logit-Funktion erhält man die folgende Gleichung:

$$\text{logit}(\hat{p}_i) \approx \beta_0 + \beta_1 x_i \quad (**)$$

Im Beispiel ergibt sich für  $x = 1$

$$1,5040774 \approx \beta_0 + \beta_1 \cdot 1$$

Die transformierte Gleichung kann nun als Matrix-Vektor Gleichung dargestellt werden:

$$\begin{pmatrix} \text{logit}(\hat{p}_1) \\ \text{logit}(\hat{p}_2) \\ \cdot \\ \cdot \\ \cdot \\ \text{logit}(\hat{p}_k) \end{pmatrix} \approx \begin{pmatrix} 1 & \mathbf{x}_1 \\ 1 & \mathbf{x}_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & \mathbf{x}_k \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

Oder kurz:  $\bar{y} = \text{logit}(\hat{\mathbf{p}}) \approx \mathbf{X}\bar{\beta}$

Die Parameterschätzung kann nun über die Methode der gewichteten kleinsten Quadrate durchgeführt werden, da die empirische Kovarianz-Matrix von  $\bar{Y}$  bekannt ist. Dabei werden die Werte in der zweiten Spalte der oberen Tabelle jeweils als Realisierungen von unabhängig binomialverteilten Zufallsvariablen angenommen, mit der geschätzten Varianz  $\hat{p}_i(1-\hat{p}_i)n_i$ . Hier ist  $n_i$  die Anzahl der Beobachtungen in der  $i$ -ten Gruppe bzw. Zeile der oberen Tabelle ( $n_1 = 9 + 2 = 11$ ,  $n_2 = 8 + 4 = 12, \dots$ ). Somit ergibt sich der Schätzer der Inversen der empirischen Kovarianz-Matrix:

$$\hat{\mathbf{V}}^{-1} = \begin{pmatrix} n_1 \hat{p}_1(1-\hat{p}_1) & 0 & \cdot & \cdot & 0 \\ 0 & n_2 \hat{p}_2(1-\hat{p}_2) & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & n_k \hat{p}_k(1-\hat{p}_k) \end{pmatrix}$$

Jetzt kann  $\bar{\beta}$  über die Methode der gewichteten kleinsten Quadrate geschätzt werden und es ergibt sich der Schätzer:  

$$\hat{\beta} = (X^t \hat{V}^{-1} X)^{-1} X^t \hat{V}^{-1} \bar{y}$$

Die Modellanpassung kann mit dem Wert sse der gewichteten Abweichungsquadrate überprüft werden. Dabei gilt:

$$sse = Q(\hat{\beta}) = (\bar{y} - X\hat{\beta})^t \hat{V}^{-1} (\bar{y} - X\hat{\beta})$$

Die Hypothesen zum zugehörigen Test lauten:

$H_0$ : Das Modell passt gegen $H_A$ : Das Modell passt nicht
---

Falls sse zu groß ist, so wird das Modell verworfen. Dabei ist sse die Realisierung einer unter  $H_0$  mit  $k-1$  Freiheitsgraden Chi-Quadrat verteilten Zufallsvariable.

Die Nullhypothese wird also auf dem Signifikanzniveau  $\alpha$  verworfen, falls gilt:  $p\text{-Wert} = 1 - F_{\chi_{k-2}}(sse) < \alpha$ .

Zusätzlich kann überprüft werden, ob die Parameter signifikant von Null verschieden sind. Dabei ist, wie bereits beschrieben, der Parameter  $\beta_1$  von besonderem Interesse. Die Hypothesen zu diesem Test lauten:

$H_0$ : $\beta_1 = 0$ gegen $H_A$ : $\beta_1 \neq 0$
--

Dabei wird die Nullhypothese auf dem Signifikanzniveau  $\alpha$  verworfen, falls gilt:

$$p\text{-Wert} = 1 - F_{\chi_1} \left( \left( \frac{\hat{\beta}_1}{\sqrt{\hat{\text{Var}}(\hat{\beta}_1)}} \right)^2 \right) < \alpha$$

Die geschätzte Varianz von  $\hat{\beta}_1$  unter der Wurzel im Nenner des oberen Ausdrucks erhält man, indem man die empirische Kovarianz-Matrix des Vektors  $\hat{\beta}^t = (\hat{\beta}_0 \hat{\beta}_1)$  bildet. Diese ergibt sich durch  $(X^t \hat{V}^{-1} X)^{-1}$ . Das Element (2,2) dieser Matrix enthält einen Schätzer für die Varianz von  $\hat{\beta}_1$ . Analog kann ein Test für  $\beta_0$  durchgeführt werden. Hier muss das Element (1,1) dieser Matrix verwendet werden.

Betrachten wir nun die Ausgabe in unserem Beispiel.

## Logistische Regression

Kreuztabelle mit absoluten Häufigkeiten:

	0	1	Summe
1	9	2	11
2	8	4	12
3	7	5	12
4	5	8	13
5	5	9	14
6	3	9	12
7	2	9	11

8	2	10	12
9	2	11	13
Summe	43	67	110

Die relativen Häufigkeiten für die erste Kategorie:

	0
1	0.8182
2	0.6667
3	0.5833
4	0.3846
5	0.3571
6	0.25
7	0.1818
8	0.1667
9	0.1538

Parameter	Schätzer	geschätzte Standardabweichung	Prüfgröße für Test (H0: Parameter = 0; gegen H1: Parameter <math>< 0</math>)	p-Wert
$b_0$	1.4628716646698	0.4921956208004	8.8335844442838	0.003
$b_1$	-0.398583004027	0.093671974008416	18.105820356834	0

**Modellanpassung:** sse: 1.1326604161, p-Wert (H0: Modell passt; gegen H1: Modell passt nicht): 0.9924 (df = 7)

Die über das Modell geschätzten Wahrscheinlichkeiten für die erste Kategorie:

	0
1	0.7435
2	0.6605
3	0.5664
4	0.4672
5	0.3705
6	0.2832
7	0.2096
8	0.1511
9	0.1068

Wie Sie sehen, ist der Schätzer für  $\hat{\beta}_0 \approx 1,46287$  zu finden unter intercept und für  $\hat{\beta}_1 \approx -0,39858$  zu finden unter slope. Da der Schätzer für  $\beta_1$  negativ ist, fällt die Wahrscheinlichkeit mit der Zeit der Abstinenzdauer. Unter Std. Estimator finden Sie die empirische Standardabweichung der Schätzer. Unter Prob sehen Sie die p-Werte. Wie zu sehen ist, ist  $\beta_1$  signifikant von Null verschieden (und das auf jedem gängigen Signifikanzniveau). Wählen wir  $\alpha = 5\%$ , so gilt: p-Wert =  $0 < 0,05$  (der p-Wert ist nicht exakt Null, nur der auf 4 Nachkommastellen gerundete p-Wert), womit die Nullhypothese, dass der Parameter Null ist, verworfen werden kann. Das Modell selbst kann nicht verworfen werden, denn hier ist ein p-Wert von  $0,9924$  zu sehen. Wählen wir ein für Anpassungstests übliches hohes Signifikanzniveau  $\alpha = 20\%$ , so kann die Nullhypothese (das Modell passt) nicht verworfen werden, da  $0,9924 > 0,20$ .

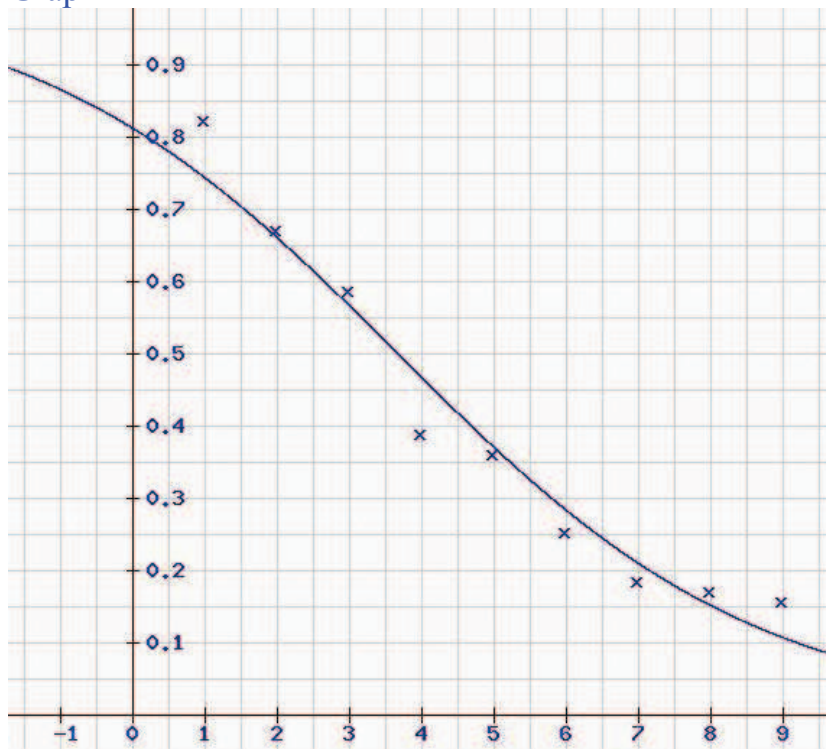
Wollen Sie die Wahrscheinlichkeit für das Auftreten von Entzugserscheinungen nach 15 Jahren über das Modell schätzen (d.h.  $P(Y = \text{„ja“} \mid 15)$ ), so ergibt sich

$$\frac{e^{1,46287-0,39858 \cdot 15}}{1 + e^{1,46287-0,39858 \cdot 15}} \approx 0,01081579 \approx 1,08\%$$

Also treten Entzugserscheinungen nach 15 Jahren noch in ca. 1,08% aller Fälle auf.

Unten sehen Sie die logistische Kurve und die geschätzten Wahrscheinlichkeiten für „ja“ zu den verschiedenen Abstinenzdauern. Wie Sie sehen, wird die Wahrscheinlichkeit für das Auftreten von Entzugserscheinungen nach 9 Jahren noch auf mehr als 10% geschätzt.

### Graph



### Umsetzung mit SAS:

```
data dat1;
input anzahl x y;
datalines;
9 1 0
2 1 1
8 2 0
4 2 1
7 3 0
5 3 1
5 4 0
8 4 1
5 5 0
9 5 1
3 6 0
9 6 1
2 7 0
9 7 1
2 8 0
10 8 1
2 9 0
11 9 1
run;

proc catmod data=dat1 order = data;
  direct x;
  model y = x / WLS;
  weight anzahl;
run;
```

**SAS-Output zur Prozedur CATMOD:**

Das SAS System

Die Prozedur CATMOD

<b>Datenübersicht</b>			
<b>Response</b>	y	<b>Response-Ausprägungen</b>	2
<b>Gewichtungsvariable</b>	anzahl	<b>Grundgesamtheiten</b>	9
<b>Datei</b>	DAT1	<b>Gesamthäufigkeit</b>	110
<b>Anzahl der fehlenden Werte</b>	0	<b>Beobachtungen</b>	18

<b>Grundgesamtheitsprofile</b>		
<b>Stichprobe</b>	<b>x</b>	<b>Stichprobengröße</b>
1	1	11
2	2	12
3	3	12
4	4	13
5	5	14
6	6	12
7	7	11
8	8	12
9	9	13

<b>Responseprofile</b>	
<b>Abhängige</b>	<b>y</b>
1	0
2	1

Varianzanalyse			
Quelle	Freiheits- grade	Chi-Quadrat	Pr > ChiSq
<b>Konstante</b>	1	8.83	0.0030
<b>x</b>	1	18.11	<.0001
<b>Residuum</b>	7	1.13	0.9924

Analyse der gewichteten Kleinste-Quadrate-Schätzer				
Parameter	Schätzwert	Standard- fehler	Chi- Quadrat	Pr > ChiSq
<b>Konstante</b>	1.4629	0.4922	8.83	0.0030
<b>x</b>	-0.3986	0.0937	18.11	<.0001