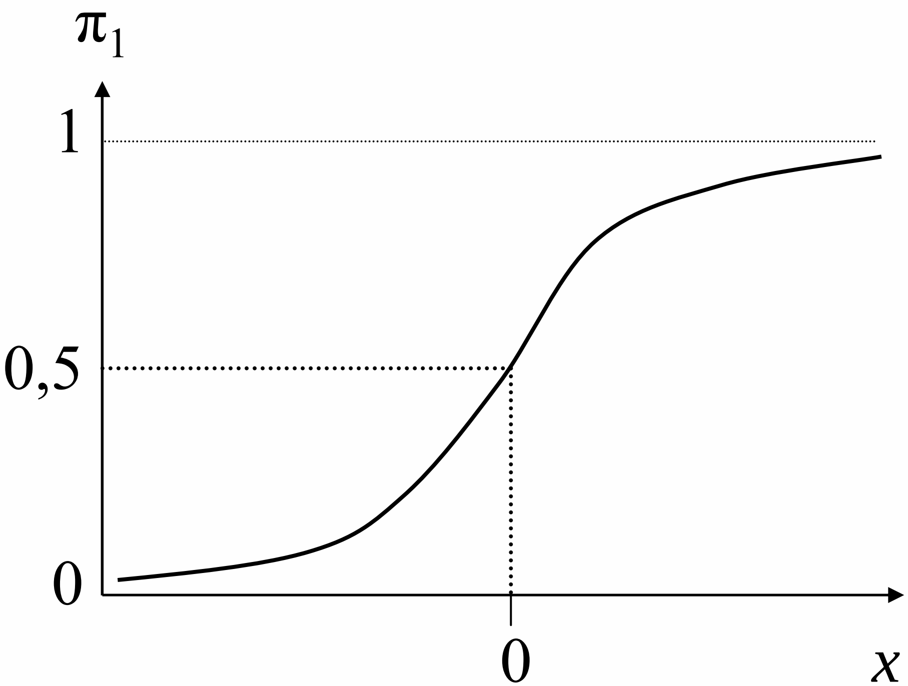
Lineare Regression

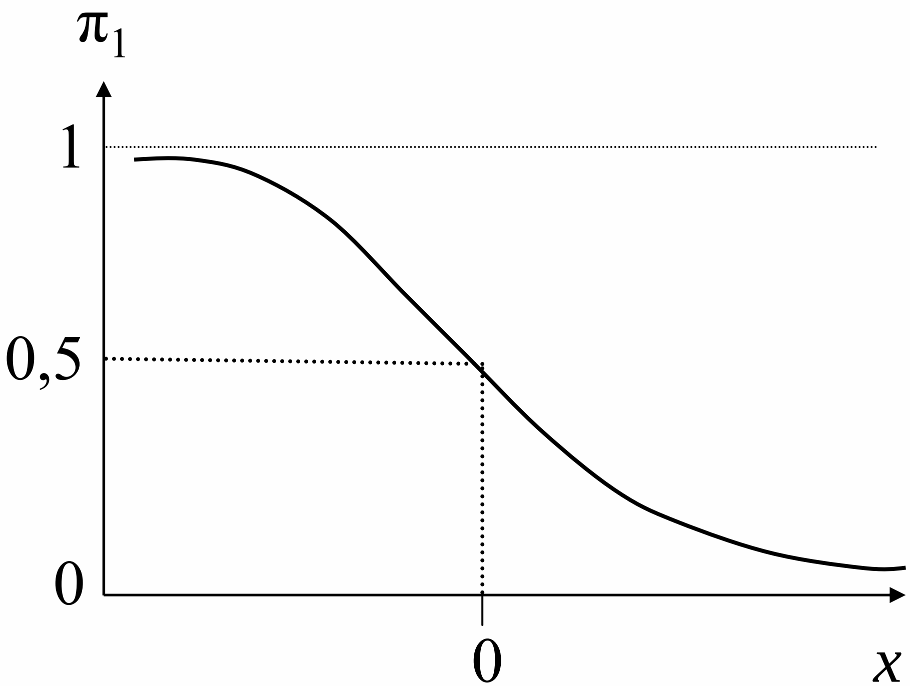
$$\pi_1 = \beta_0 + \beta_1 x$$

Logistische Regression (Logit-Modell)

$$\ln\left(\frac{\pi_1}{1-\pi_1}\right) = \beta_0 + \beta_1 x$$



$\beta_0=0;\ \beta_1=1$



$\beta_0=0;\ \beta_1=-1$

Logit (Log-Odd)

$$\ln\left(\frac{\pi_1}{1-\pi_1}\right) = \beta_0 + \beta_1 x$$

Odd

$$\frac{\pi_1}{1-\pi_1} = e^{\beta_0 + \beta_1 x} = e^{\beta_0} \times e^{\beta_1 x}$$

Wahrschein-lichkeit

$$\pi_1 = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \qquad \pi_1 = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Zu den Transformationen

(1) $$\ln\left(\frac{\pi_1}{1-\pi_1}\right) = \beta_0 + \beta_1 x$$

da: $e^{\ln Q} = Q$

(2) $$e^{\ln(\pi_1/(1-\pi_1))} = \frac{\pi_1}{1-\pi_1} = e^{(\beta_0+\beta_1 x)} = e^{\beta_0} \times e^{\beta_1 x} \qquad \Big| \cdot (1-\pi_1)$$

(3) $$\pi_1 = (1-\pi_1)e^{(\beta_0+\beta_1 x)}$$

(4) $$\pi_1 = e^{(\beta_0+\beta_1 x)} - \pi_1 e^{(\beta_0+\beta_1 x)} \qquad \Big| + \pi_1 e^{(\beta_0+\beta_1 x)}$$

Zu den Transformationen (Fortsetzung)

(5) $$\pi_1 + \pi_1 e^{(\beta_0 + \beta_1 x)} = e^{(\beta_0 + \beta_1 x)}$$

(6) $$\pi_1 \left(1 + e^{(\beta_0 + \beta_1 x)}\right) = e^{(\beta_0 + \beta_1 x)} \qquad \Big| \div \left(1 + e^{(\beta_0 + \beta_1 x)}\right)$$

(7) $$\pi_1 = \frac{e^{(\beta_0 + \beta_1 x)}}{1 + e^{(\beta_0 + \beta_1 x)}}$$

EQUAL

Wenn sich die Einkommen in unserer Gesellschaft stärker angleichen würden: Versprechen Sie sich davon Vorteile? [**1**=ja; **0**=nein]

EK

Monatl. Haushaltsnettoeinkommen, dividiert durch die Zahl der Haushaltsmitglieder, die zu diesem Haushaltseinkommen beitragen

Codes der erfragten Einkommensklassen

1 = *unter* 1.000 DM
2, …, 15 = 1.000 bis 8.000 DM, in Stufen à 500 DM
16 = *über* 8.000 DM

EQUAL

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	,00	141	43,8	45,8	45,8
	1,00	167	51,9	54,2	100,0
	Gesamt	308	95,7	100,0	
Fehlend	System	14	4,3		
Gesamt		322	100,0		

Deskriptive Statistiken

	Mittelwert	Standardabweichung	N
EQUAL	,5432	,49903	278
EK	4,6640	2,55603	278

EK

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	,50	3	,9	1,0	1,0
	1,00	10	3,1	3,4	4,5
	1,25	1	,3	,3	4,8
	1,50	9	2,8	3,1	7,9
	2,00	24	7,5	8,3	16,2
	2,33	2	,6	,7	16,9
	2,50	13	4,0	4,5	21,4
	2,67	2	,6	,7	22,1
	3,00	27	8,4	9,3	31,4
	3,33	1	,3	,3	31,7
	3,50	23	7,1	7,9	39,7
	3,67	3	,9	1,0	40,7
	4,00	30	9,3	10,3	51,0
	4,50	19	5,9	6,6	57,6
	5,00	35	10,9	12,1	69,7
	5,33	3	,9	1,0	70,7
	5,50	8	2,5	2,8	73,4
	6,00	16	5,0	5,5	79,0
	6,50	5	1,6	1,7	80,7
	7,00	15	4,7	5,2	85,9
	7,50	3	,9	1,0	86,9
	8,00	21	6,5	7,2	94,1
	9,00	7	2,2	2,4	96,6
	10,00	1	,3	,3	96,9
	11,00	3	,9	1,0	97,9
	12,00	1	,3	,3	98,3
	13,00	1	,3	,3	98,6
	14,00	1	,3	,3	99,0
	16,00	3	,9	1,0	100,0
	Gesamt	290	90,1	100,0	
Fehlend	System	32	9,9		
Gesamt		322	100,0		

Lineare Regression

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,176[a]	,031	,027	,49218

a. Einflußvariablen : (Konstante), EK

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz
		B	Standardfehler	Beta		
1	(Konstante)	,703	,062		11,429	,000
	EK	-3,426E-02	,012	-,176	-2,962	,003

a. Abhängige Variable: EQUAL

n=278

Logistische Regression

Modellzusammenfassung

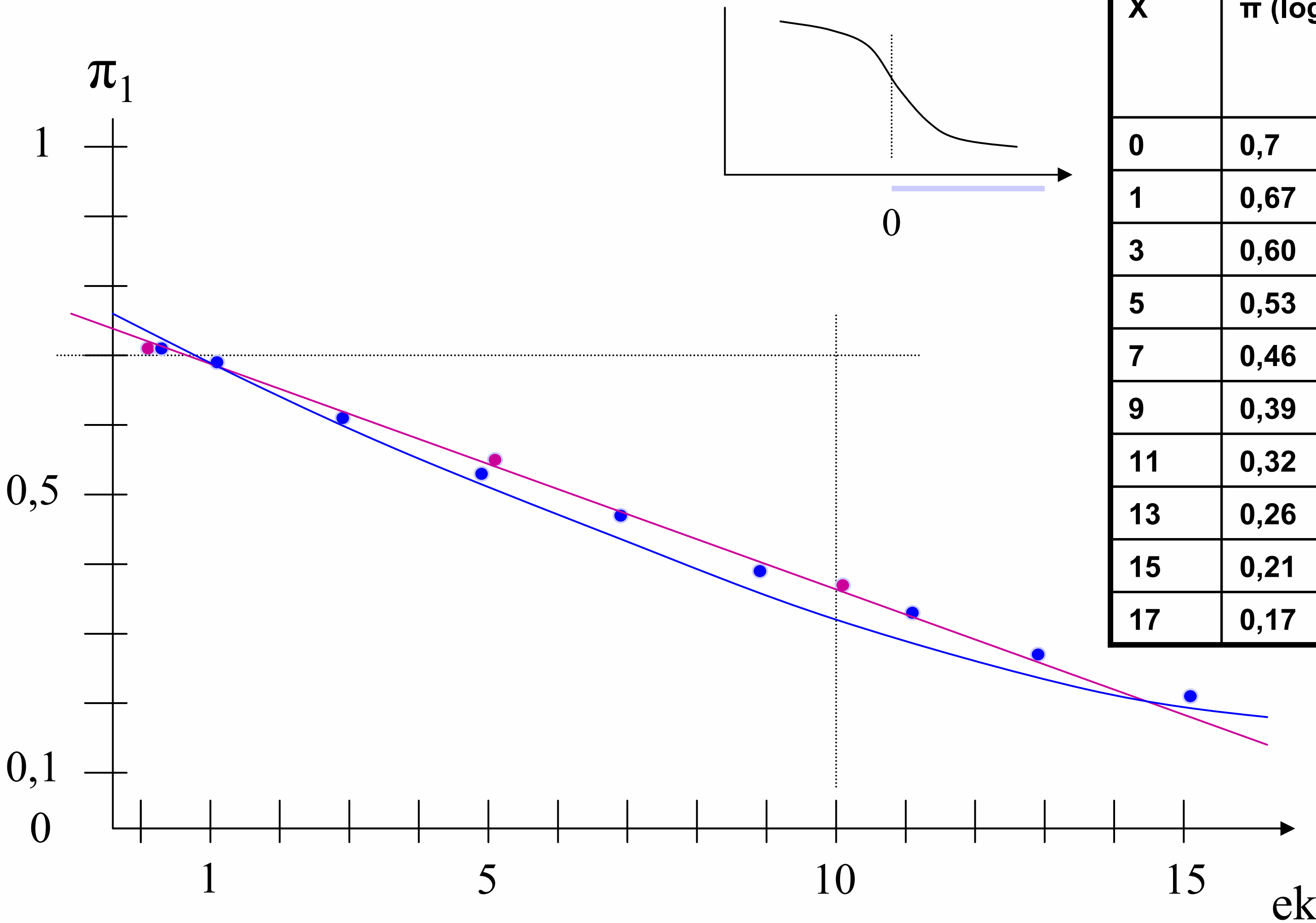
Schritt	-2 Log-Likelihood	Cox & Snell R-Quadrat	Nagelkerkes R-Quadrat
1	374,590	,031	,041

Variablen in der Gleichung

		RegressionskoeffizientB	Standardfehler	Wald	df	Sig.	Exp(B)
Schritt 1[a]	EK	-,144	,051	8,080	1	,004	,866
	Konstante	,847	,266	10,160	1	,001	2,332

a. In Schritt 1 eingegebene Variablen: EK.

n=278



X	π (log)	π (lin)
0	0,7	0,7
1	0,67	0,67
3	0,60	0,60
5	0,53	0,53
7	0,46	0,47
9	0,39	0,40
11	0,32	0,33
13	0,26	0,26
15	0,21	0,19
17	0,17	0,13

Likelihood-Prinzip

Besagt, dass man aus der Menge der möglichen Parameter der Population denjenigen auswählen sollte, bei dem es am wahrscheinlichsten ist, dass man die vorliegende Stichprobe beobachtet.

Pr(f | π, n, M)

Wenn **Verteilungs**funktion M = Binomialverteilung:

$$\Pr\left(f \middle| \pi, n\right) = \binom{n}{f} \pi^{f} \left(1 - \pi\right)^{n-f}$$

Nach: Andreß/Hagenaars/Kühnel (1997) Analyse von Tabellen und Kategorialen Daten. Berlin: Springer, S. 33ff.

3 Gruppen (Populationen) á 10 Personen

G1: 3 von 10 haben gesuchte Eigenschaft
G2: 5 von 10 haben gesuchte Eigenschaft
G3: 8 von 10 haben gesuchte Eigenschaft

Es wird eine Stichprobe vom Umfang n=3 (mit Zurücklegen) gezogen; alle 3 Personen hatten die gesuchte Eigenschaft (*f*=3)

Frage: aus welcher der 3 Gruppen stammt die Stichprobe?

$$\Pr(f|\pi,n)=\binom{n}{f}\pi^{f}(1-\pi)^{n-f}$$

$$\Pr(3|\pi,3)=\binom{3}{3}\pi^{3}(1-\pi)^{3-3}$$

$$\Pr(3|0{,}30{,}3)=\binom{3}{3}0{,}30^{3}(1-0{,}30)^{3-3}=\underline{\underline{0{,}027}}$$

$$\Pr(3|0{,}50{,}3)=\binom{3}{3}0{,}50^{3}(1-0{,}50)^{3-3}=\underline{\underline{0{,}125}}$$

$$\Pr(3|0{,}80{,}3)=\binom{3}{3}0{,}80^{3}(1-0{,}80)^{3-3}=\underline{\underline{0{,}512}}$$

Pr(f | π, n, M) $$\Pr\left(f \mid \pi, n\right) = \binom{n}{f} \pi^{f} \left(1-\pi\right)^{n-f}$$

Likelihood-Funktion:

L(π|f, n, M) $$L\left(\pi \mid f, n\right) = \binom{n}{f} \pi^{f} \left(1-\pi\right)^{n-f}$$

$$L = K\pi^{f} \left(1-\pi\right)^{n-f}$$

Zu maximieren: $$\ln L = \ln\left(K\right) + f \ln \pi + \left(n-f\right)\ln\left(1-\pi\right)$$

Nach: Andreß/Hagenaars/Kühnel (1997) Analyse von Tabellen und Kategorialen Daten. Berlin: Springer, S. 33-45

Gütemaße

-2LL [Devianz] = -2 × LogLikelihood

(entspricht Fehlerquadratsumme der linearen Regression)

Chi²-verteilt mit df=N-K-1
N = Anzahl der Beobachtungen
K = Anzahl der unabh. Variablen

H_0: Modell besitzt perfekte Anpassung
H_1: Modell besitzt keine perfekte Anpassung

Perfekter Modellfit: Likelihood=1, Devianz=0

Likelihood-Ratio- bzw. Devianztest

$$LR = -2 \times \ln\left(L_r / L_u\right) = -2 \times \left[\ln\left(L_r\right) - \ln\left(L_u\right)\right] = -2 \times \ln\left(L_r\right) - 2 \times \ln\left(L_u\right)$$

L = Likelihood, r restringiertes u unrestringiertes Modell
ln L = LL = Loglikelihood

Likelihood-Ratio Chi-Quadrat $$L^2 = 2\sum_k f_k \ln\left(\frac{f_k}{\hat{F}_k}\right)$$

f beobachtete
F erwartete Häufigkeiten

(ML Schätzungen sind äquivalent mit der Minimierung von L^2);
L^2 = Spezialfall von LR, wenn das unrestringierte Modell saturiert ist*)

$$McFaddens - R^2 = 1 - \frac{LL_V}{LL_0}$$

LL_0: LogLikelihood des Nullmodells (nur mit Konstante)

LL_V: LogLikelihood des vollständigen Modells

$$Cox\,\&\,Snell - R^2 = 1 - \left[\frac{L_0}{L_V}\right]^{\frac{2}{N}}$$

L_0: Likelihood des Nullmodells (nur mit Konstante)

L_V: Likelihood des vollständigen Modells

N: Stichprobenumfang

*Vgl. Andreß/Hagenaars/Kühnel op.cit., S. 39 und 45;
Backhaus et al. (2003) Multivariate Analysemethoden. Berlin: Springer, S. 437 - 442

$$Nagel\ker ke - R^2 = \frac{R^2}{R^2_{\max}}$$

$$R^2_{\max} = 1 - \left(L_0\right)^{\frac{2}{N}}$$

L_0: Likelihood des Nullmodells (nur mit Konstante)

N: Stichprobenumfang

Wald-Statistik

$$W = \left(\frac{b_k}{s_{b_k}}\right)^2$$

- asymptotisch Chi²-verteilt,
- bei df=1 Freiheitsgrad

Erweiterung um erklärende Variablen

$$\ln\left(\frac{\pi_1}{1-\pi_1}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

$$\frac{\pi_1}{1-\pi_1} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k} = e^{\beta_0} \times e^{\beta_1 x_1} \times e^{\beta_2 x_2} \times \ldots \times e^{\beta_k x_k}$$

$$\pi_1 = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k}}$$

Polytome abhängige Variable

mit ***I*** Ausprägungen

Die $\frac{I \times (I-1)}{2}$ unterscheidbaren Odds bzw. Logits

sind durch ***I*** – 1 Modellgleichungen darzustellen.

z.B.:

$$\ln\left(\frac{\pi_1}{\pi_3}\right) = \beta_{0,1/3} + \beta_{1,1/3}x_1 \qquad \ln\left(\frac{\pi_2}{\pi_3}\right) = \beta_{0,2/3} + \beta_{1,2/3}x_1$$

Kategoriale unabhängige Variablen

Logit Modell

$$\ln\left(\frac{\pi_1}{1-\pi_1}\right) = \ln\left(\frac{m_{q1}}{m_{q2}}\right) = \beta_0 + \beta_1 x_{1q}$$

Loglineares Modell:

$$\ln\left(m_{ql}\right) = \lambda + \lambda_q^X + \lambda_l^Y + \lambda_{ql}^{XY}$$

		Y_l	
		$l=1$	$l=2$
X_q	$q=1$	m_{q1}	m_{q2}
	$q=2$	m_{q1}	m_{q2}

Zur Äquivalenz von loglinearen und Logit Modellen:

$$\ln\left(\frac{m_{q1}}{m_{q2}}\right) = \beta_0 + \beta_1 x_q = 2\lambda_1^Y + 2\lambda_{q1}^{XY} \qquad \text{wenn:} \quad \underline{\lambda_1^Y = -\lambda_2^Y} \quad \text{und} \quad \underline{\lambda_{q1}^{XY} = -\lambda_{q2}^{XY}}$$

$$\ln\left(\frac{m_{q1}}{m_{q2}}\right) = \ln(m_{q1}) - \ln(m_{q2})$$

$$\ln(m_{q1}) = \lambda + \lambda_q^X + \lambda_1^Y + \lambda_{q1}^{XY} \qquad \ln(m_{q2}) = \lambda + \lambda_q^X + \lambda_2^Y + \lambda_{q2}^{XY}$$

$$\ln(m_{q1}) - \ln(m_{q2}) = \lambda - \lambda + \lambda_q^X - \lambda_q^X + \lambda_1^Y - \underline{\lambda_2^Y} + \lambda_{q1}^{XY} - \underline{\lambda_{q2}^{XY}}$$

$$= \lambda_1^Y - \left(\underline{-\lambda_1^Y}\right) + \lambda_{q1}^{XY} - \left(\underline{-\lambda_{q1}^{XY}}\right) = 2\lambda_1^Y + 2\lambda_{q1}^{XY}$$

	d=1	d=2	~~d=3~~	d=1	d=2
q=1	1	0	~~0~~	1	0
q=2	0	1	~~0~~	0	1
q=3	**0**	**0**	~~1~~	**-1**	**-1**
	Dummykodierung			Effektkodierung	