"Binary Choice"-Modelle - Der Probit-Ansatz

Eine nicht direkt beobachtbare stochastische Variable y_i^* hängt von x_i ab:

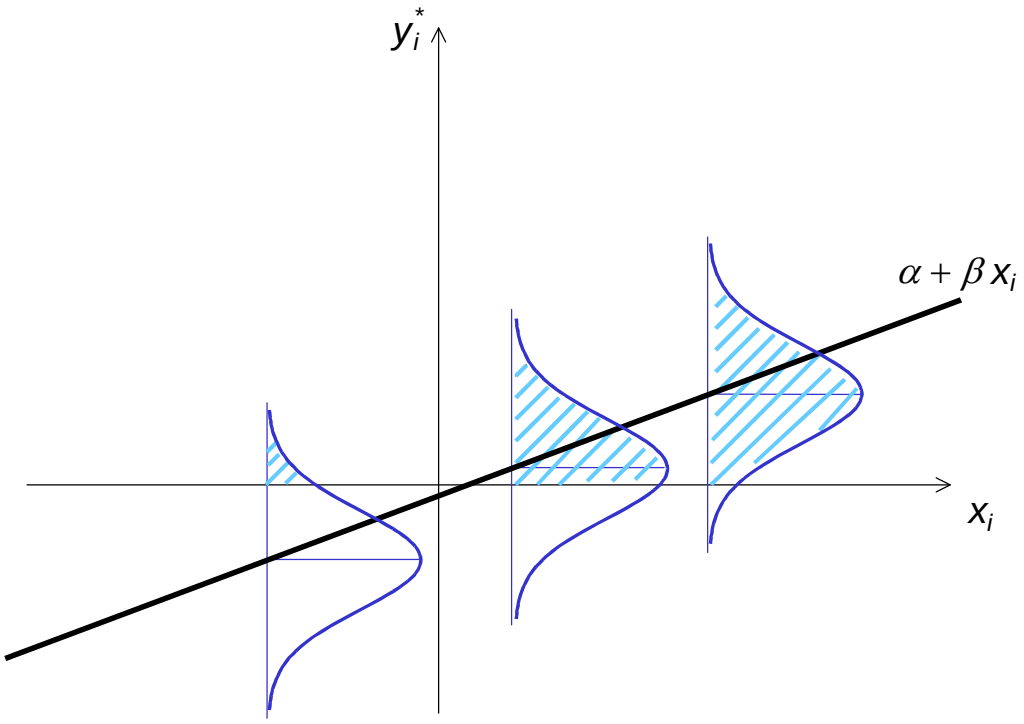
$$y_i^* = \alpha + \beta x_i + u_i \qquad u_i \sim N(0, \sigma^2) \tag{1}$$

Beobachtet wird eine binäre Variable y_i nach folgender Regel:

$$y_i = \begin{cases} 1 \ \textit{falls} \ y_i^* > 0 \\ 0 \ \textit{sonst} \end{cases} \tag{2}$$

Zum Beispiel: Pendler *i* wählt Auto (y_i = 1) oder Bus (y_i = 0) in Abhängigkeit des Zeitunterschieds $x_i = \textit{ZeitBus}_i - \textit{ZeitAuto}_i$ für seinen Arbeitsweg.)

In der grafisch dargestellten Situation steigt die Wahrscheinlichkeit $\text{Prob}(y_i^* > 0)$ mit wachsendem x_i (blau schraffierte Flächen):



Multiplikation von α, β und σ mit einer beliebigen Konstanten ist in diesem Modell beobachtungsäquivalent. Deshalb wird σ auf 1 normiert: $u_i \sim N(0, 1)$.

Gemäss (1) gilt:

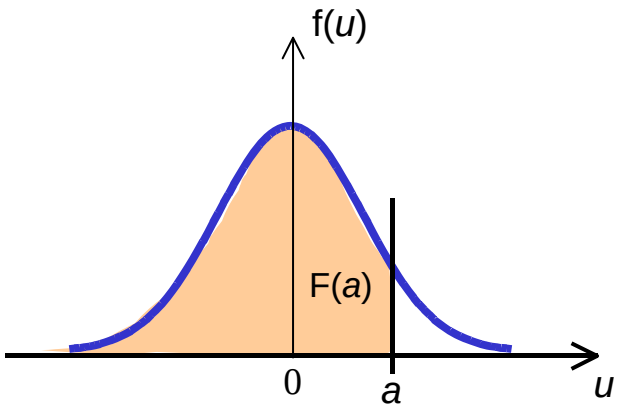
$$y_i^* = 0 \quad \Leftrightarrow \quad u_i = -(\alpha + \beta x_i) \quad \text{und} \quad y_i^* > 0 \quad \Leftrightarrow \quad u_i > -(\alpha + \beta x_i)$$

Folglich:

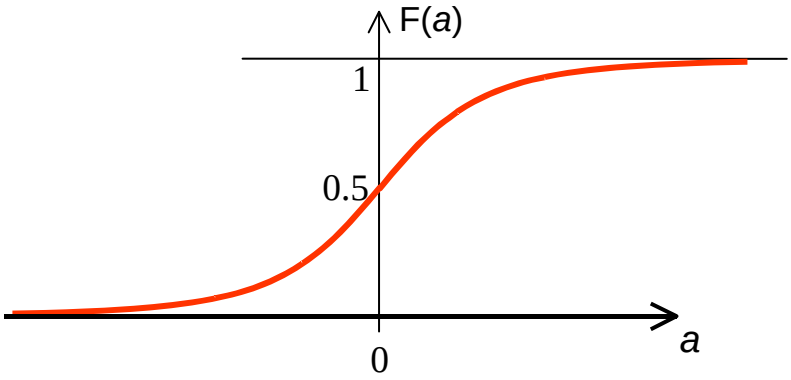
$$\text{Prob}(y_i = 1) \;=\; \text{Prob}(y_i^* > 0) \;=\; \text{Prob}(u_i > -(\alpha + \beta x_i))$$

Bezeichnet $f(u)$ die Dichtefunktion der Standardnormalverteilung, so gilt:

$$\text{Prob}(u < a) = \int_{-\infty}^{a} f(u)du \equiv F(a)$$



f(*u*): Dichtefunktion



F(*a*): Kumulative Verteilungsfunktion

Folglich kann man schreiben:

$$\text{Prob}(y_i = 1) \;=\; \text{Prob}(u_i > -(\alpha + \beta x_i)) \;=\; \text{Prob}(u_i < \alpha + \beta x_i) \;=\; F(\alpha + \beta x_i)$$

Schätzung der Modellparameter α und β nach dem "Maximum Likelihood"-Verfahren.

Exkurs: "Maximum Likelihood"-Schätzung

Allgemein formuliert werden in einer ML-Schätzung die Modellparameter so bestimmt, dass der vorliegenden Stichprobe maximale Wahrscheinlichkeit zukommt (Maximierung der Likelihood-Funktion). Unter der Annahme eines normalverteilten Störterms lässt sich praktisch jedes Modell mit diesem Verfahren schätzen. Häufig läuft eine ML-Schätzung auf eine einfache Kleinstquadrateschätzung hinaus (z.B. Regression). In bestimmten Fällen ist jedoch die Kleinstquadrateschätzung nicht anwendbar, so dass man explizit die Likelihood-Funktion des Modells maximieren muss (z.B. im Probit-Modell).

Zur Verdeutlichung des ML-Verfahrens zwei simple Beispiele:

1. Für eine Variable x liegt die folgende Stichprobe vor

x = 5, 6, 9, 3, 4, 2, 6.

Unter welcher der folgenden vier Verteilungsannahmen $N(\mu,\sigma^2)$ kommt der gemachten Beobachtungsreihe die grösste Wahrscheinlichkeit zu?

a) x ~ N(10,1) b) x ~ N(10,8) c) x ~ N(5,1) d) x ~ N(5,4)

Unter a) wäre es extrem unwahrscheinlich, eine Stichprobe wie die vorliegende zu ziehen; die Beobachtungswerte sind fast alle signifikant kleiner als 10. Unter b) erscheint die vorliegende Stichprobe etwas weniger unwahrscheinlich, weil die grössere Varianz mehr Spielraum für Abweichungen nach unten lässt. Unter c) liegt der Mittelwert der Normalverteilung beim Stichprobenmittel, die kleine Varianz macht jedoch die Beobachtungen 9 und 2 unwahrscheinlich. Somit entspricht d) am ehesten einer "Maximum Likelihood"-Schätzung. Die Parameter μ und σ^2 sind hier so festgelegt, dass das Ziehen einer Stichprobe wie der vorliegenden durchaus möglich erscheint.

Die Likelihood-Funktion entspricht dem Produkt der Dichtefunktionen für die 7 Beobachtungen, d.h. L = f(5)·f(6)·f(9)·f(3)·f(4)·f(2)·f(6) , und μ und σ^2 werden so festgelegt, dass L für die gegebene Stichprobe maximiert wird.

2. Eine Münze wird 4 mal geworfen. Sie fällt 3 mal auf Kopf und 1 mal auf Zahl. Wie gross ist die Wahrscheinlichkeit für dieses Ergebnis unter der Annahme, dass die Münze regulär ist (p = Prob(Kopf) = 0.5)?

$$\binom{4}{3} p^3(1-p)^1 = \binom{4}{3} 0.5^3 0.5^1 = 0.25 \qquad \text{(Binomial-Verteilung)}$$

Wie müssen wir p festlegen, damit der Stichprobe (3 mal Kopf, 1 mal Zahl) die grösstmögliche Wahrscheinlichkeit zukommt? Die Likelihood-Funktion lautet:

$$L = \binom{4}{3} p^3(1-p)^1$$

Maximierung von L (erste Ableitung = 0 setzen) führt zur *Lösung* p = 0.75. Der maximierte Wert der Likelihood-Funktion ist:

$$\binom{4}{3} 0.75^3 0.25^1 = 0.422$$

Die Likelihood-Funktion des Probit-Modells lautet:

$$L = \prod_{y_i=1} F(\alpha + \beta x_i) \prod_{y_i=0} \left(1 - F(\alpha + \beta x_i)\right) = L(\alpha, \beta) \qquad (3)$$

Die Likelihood-Funktion entspricht dem Produkt der Wahrscheinlichkeiten über alle Beobachtungen. Für die Beobachtungen mit $y_i = 1$ (Wahl Auto) bzw. $y_i = 0$ (Wahl Bus) erscheinen die entsprechenden Wahrscheinlichkeiten $\text{Prob}(y_i = 1) = F(\alpha + \beta x_i)$ bzw. $\text{Prob}(y_i = 0) = 1 - F(\alpha + \beta x_i)$.

Maximierung von *L* bezüglich α und β ergibt die Schätzwerte $\hat{\alpha}$ und $\hat{\beta}$.

Interpretation der Schätzergebnisse

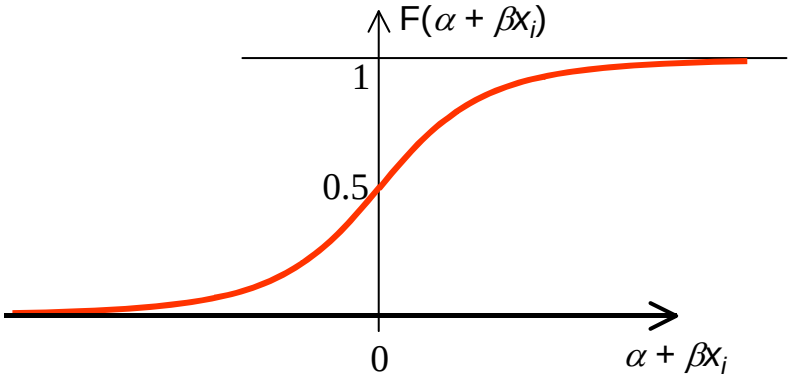
Anhand der Schätzung lässt sich der Einfluss einer marginalen Erhöhung von x_i auf die Wahrscheinlichkeit für Prob(y_i = 1) berechnen. Im Unterschied zu einer linearen Regression ist dieser Einfluss nicht einfach durch β gegeben, weil der Term βx_i innerhalb der kumulativen Normalverteilungsfunktion F(.) erscheint:

$$\text{Prob}(y_i = 1) = F(\alpha + \beta x_i)$$

Der marginale Einfluss von x_i auf $\text{Prob}(y_i = 1)$ ist durch

$$\frac{\partial F(\alpha + \beta x_i)}{\partial x_i} = f(\alpha + \beta x_i)\beta$$

gegeben (die Ableitung der kumulativen Verteilungsfunktion F entspricht der Dichtefunktion f). Der Effekt einer kleinen Änderung von x_i auf $\text{Prob}(y_i = 1)$ hängt somit vom Niveau von x_i ab. Er ist bei $\alpha + \beta x_i = 0$ am grössten, weil $f(.)$ an der Stelle 0 ein Maximum erreicht bzw. F(.) an der Stelle 0 maximale Steigung aufweist.



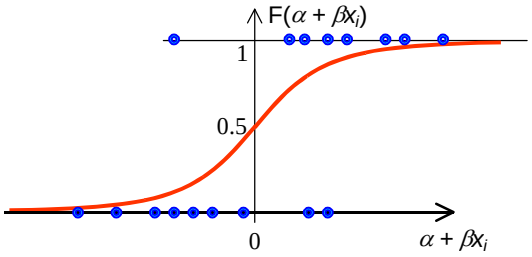
Bei mehreren erklärenden Variablen $x_{1i}, x_{2i}, x_{3i}, \ldots$ hängt der marginale Einfluss von x_{ji} auch vom Niveau aller anderen erklärenden Variablen ab.

Fit-Masse

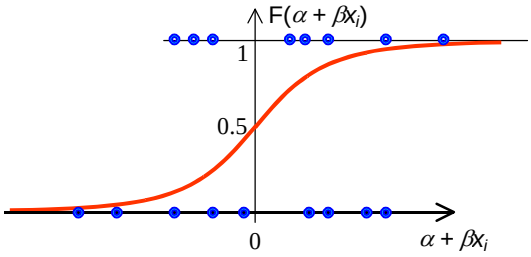
1. Prozentualer Anteil richtig klassifizierter Realisationen von y_i nach der Regel:

$$F(\alpha + \beta x_i) > 0.5 \Leftrightarrow y_i = 1 \qquad F(\alpha + \beta x_i) \leq 0.5 \Leftrightarrow y_i = 0$$

Relativ guter Fit:
3 Fehlklassifizierungen



Relativ schlechter Fit:
7 Fehlklassifizierungen



2. McFaddenR^2: Basiert auf einem Likelihood-Ratio Test. Die Schätzwerte für α und β maximieren die Likelihood-Funktion (unrestringiert: L_{UR}). Unter der Restriktion $\beta = 0$ (kein Einfluss von x) ist $F(\alpha)$ konstant, wobei α so festgelegt wird, dass $F(\alpha)$ dem Mittelwert der y_i entspricht (= Σy_i /n). In der Grafik entspricht dies einer horizontalen Linie, die nach dem ML-Kriterium umso weiter unten liegt, je grösser der Anteil der y_i auf der Null-Linie ist. Der unter dieser Restriktion erhaltene Wert der Likelihood-Funktion ist kleiner ($L_R < L_{UR}$). Er ist deutlich kleiner, wenn x_i viel erklärt. Ein intuitives Fit-Mass - analog zum üblichen R^2 - ist:

$$\text{McFadden}R^2 = 1 - \frac{\log(L_{UR})}{\log(L_R)}$$

Das McFaddenR^2 liegt zwischen 0 und 1. Falls x_i überhaupt nichts erklärt, ist $L_{UR} = L_R$ und somit McFadden$R^2 = 0$. Falls x_i die Realisationen von y_I perfekt erklärt, ist $L_{UR} = 1$, $\log(L_{UR}) = 0$ und folglich McFadden$R^2 = 1$. Warum ist L_{UR} ist in diesem Fall gleich 1? Weil in der Likelihood-Funktion (3) alle Terme gleich 1 sind. Numerisch bricht in diesem Grenzfall die Maximierung der Likelihood-Funktion allerdings zusammen.

Beispiel Verkehrsmittelwahl

ZeitAuto	ZeitBus	x ZeitBus-ZeitAuto	y 1=Auto, 0=Bus
52.9	4.4	-48.5	0
4.1	28.5	24.4	0
4.1	86.9	82.8	1
56.2	31.6	-24.6	0
51.8	20.2	-31.6	0
0.2	91.2	91.0	1
27.6	79.7	52.1	1
89.9	2.2	-87.7	0
41.5	24.5	-17.0	0
95.0	43.5	-51.5	0
99.1	8.4	-90.7	0
18.5	84.0	65.5	1
82.0	38.0	-44.0	1
8.6	1.6	-7.0	0
22.5	74.1	51.6	1
51.4	83.8	32.4	1
81.0	19.2	-61.8	0
51.0	85.0	34.0	1
62.2	90.1	27.9	1
95.1	22.2	-72.9	0
41.6	91.5	49.9	1

Dependent Variable: Y
Method: ML - Binary Probit
Sample: 1 21
Included observations: 21
Convergence achieved after 5 iterations
Covariance matrix computed using second derivatives

Variable		Coefficient	Std. Error	z-Statistic	Prob.
C	alfa =	-0.064434	0.399244	-0.161390	0.8718
X	beta =	0.029999	0.010287	2.916279	0.0035

Mean dependent var	0.476190	S.D. dependent var	0.511766
S.E. of regression	0.310890	Akaike info criterion	0.777634
Sum squared resid	1.836405	Schwarz criterion	0.877112
Log likelihood	-6.165158	Hannan-Quinn criter.	0.799223
Restr. log likelihood	-14.53227	Avg. log likelihood	-0.293579
LR statistic (1 df)	16.73423	McFadden R-squared	0.575761
Probability(LR stat)	4.30E-05		

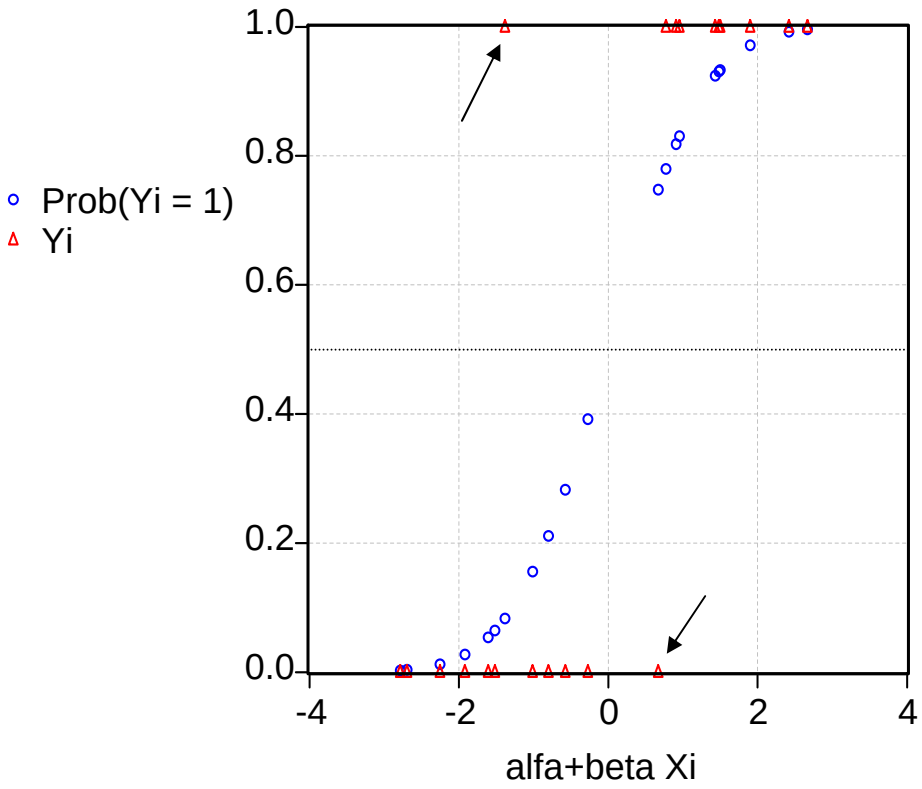
Obs with Dep=0	11	Total obs	21
Obs with Dep=1	10		

Dependent Variable: Y
Method: ML - Binary Probit
Sample: 1 21
Included observations: 21
Prediction Evaluation (success cutoff C = 0.5)

	Estimated Equation			Constant Probability		
	Dep=0	Dep=1	Total	Dep=0	Dep=1	Total
P(Dep=1)<=C	10	1	11	11	10	21
P(Dep=1)>C	1	9	10	0	0	0
Total	11	10	21	11	10	21
Correct	10	9	19	11	0	11
% Correct	90.91	90.00	90.48	100.00	0.00	52.38
% Incorrect	9.09	10.00	9.52	0.00	100.00	47.62
Total Gain*	-9.09	90.00	38.10			
Percent Gain**	NA	90.00	80.00			

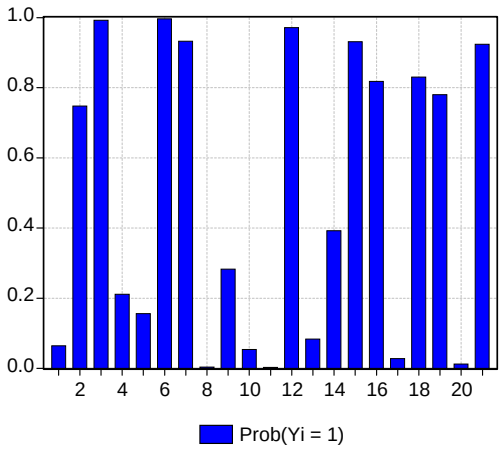
Von den 21 Beobachtungen sind 2 fehlklassifiziert. Für eine Beobachtung ist Prob(y_i = 1) > 0.5, das Individuum nimmt aber trotzdem den Bus. Für eine Beobachtung ist Prob(y_i = 1) < 0.5, das Individuum fährt aber trotzdem mit dem Auto zur Arbeit.

Die zwei Fehlklassifikationen können auch grafisch gezeigt werden:

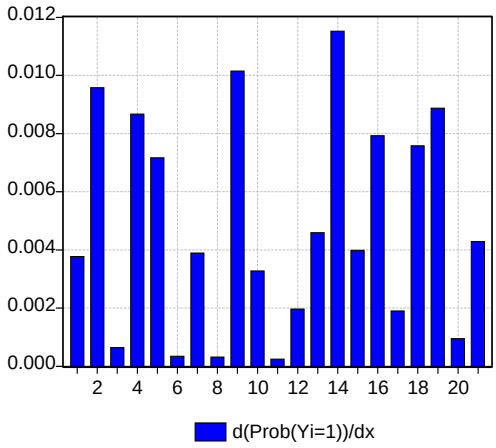


Wahrscheinlichkeiten der 21 Pendler, mit dem Auto zu fahren

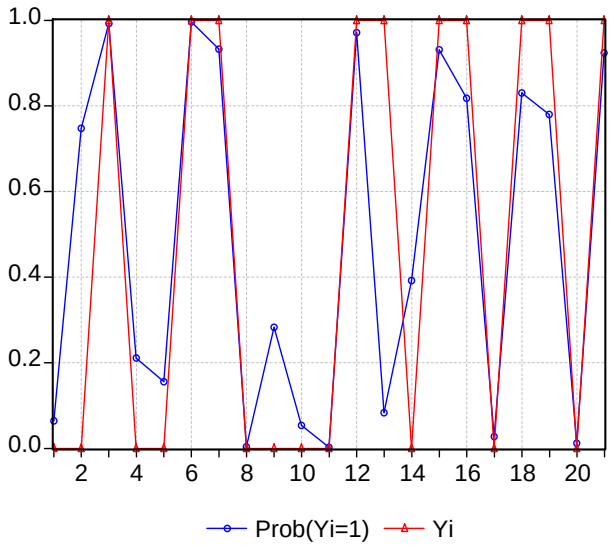
Geschätzte Wahrscheinlichkeiten



Effekt einer Erhöhung von x_i um 1
(1 Min. zusätzlicher Zeitnachteil Bus)



Die Pendler Nr. 2 und 13 sind fehlklassifiziert:



Weitere mögliche Fragestellungen:

Warum ist der Effekt eines zusätzlichen 1-minütigen Zeitnachteils des Busses bei Pendler Nr. 14 am grössten und bei den Pendlern Nr. 6 und 8 klein?

Schätzen Sie ab, wie viele Pendler auf den Bus umsteigen würden, falls sich die Bus-Fahrzeiten halbieren!