Regression Models for Categorical Dependent Variables (Logit, Probit, and Related Techniques)

ZA Spring Seminar 2008

Andreas Diekmann Ben Jann

ETH Zurich
Switzerland

Cologne, February 25–29, 2008

Categorical Dependent Variables: Examples

- Participation in elections (no = 0, yes = 1)
- Voting in a multi-party system (polytomous variable)
- Labor force participation (no = 0, yes = 1; or: no = 0, part-time = 1, full-time = 2)
- Union membership
- Purchase of a durable consumer good in a certain time span
- Successfully selling a good on eBay
- Victim of a crime in last 12 month
- Divorce within five years after marriage
- Number of children
- Choice of traffic mode (by car = 0, public transportation = 1; or: by foot = 0, by bike = 1, public transportation = 2, by car = 3)
- Life satisfaction on a scale from 1 to 5

Categorical Dependent Variables: Models

Dependent Variable	**Method**
continuous, unbounded	linear regression (OLS)
binary (dichotomous)	logistic regression, probit, and related models
nominal (polytomous)	multinomial logit, conditional logit
ordered outcomes	ordered logit/probit, and related models
count data	poisson regression, negative binomial regression, and related models
limited/bounded	censored regression (e.g. Tobit)
(censored) duration data	survival models, event history analysis (e.g. Cox regression)
Independent Variables:	metric and/or dichotomous

Why multivariate methods with categorical or other types of variables?

Three Examples:

1. The smaller the proportion of calcium in the bones of an individual (X), the higher the number of unmarried aunts (Y).

(Krämer 1991: 127)

$r_{xy} < 0$

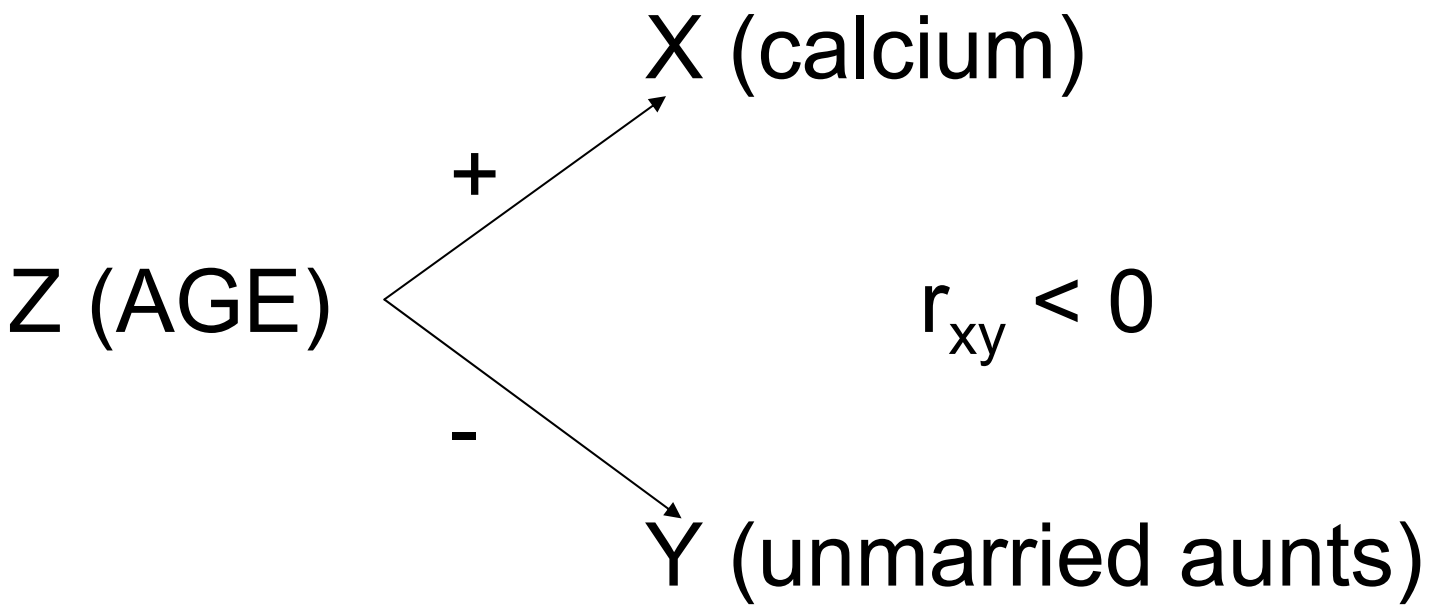
2. The larger the size of shoes (X), the higher the income (Y).

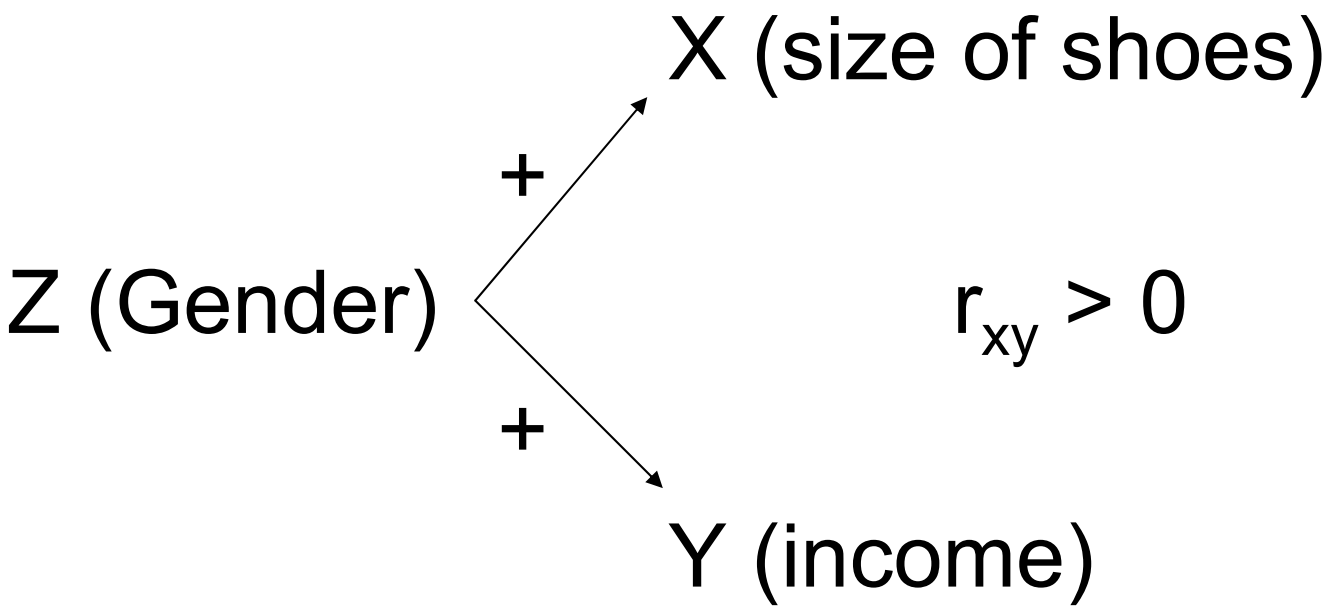
$r_{xy} > 0$

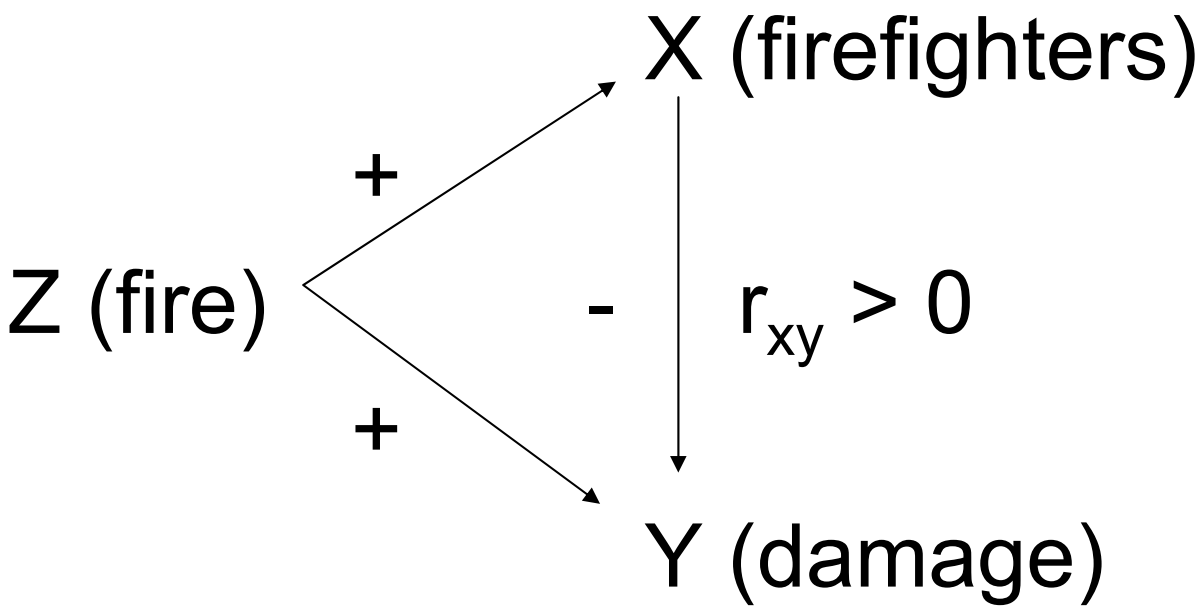
3. The larger a fire brigade, i. e. the more firefighters (X), the higher the damage by the fire (Y).

(Paul Lazarsfeld)

$r_{xy} > 0$


X (calcium)
+
Z (AGE)
$r_{xy} < 0$
-
Y (unmarried aunts)


X (size of shoes)
+
Z (Gender)
$r_{xy} > 0$
+
Y (income)


X (firefighters)
+
Z (fire)
-
$r_{xy} > 0$
+
Y (damage)

Experimental approach:

Control (small number of firefighters)
Experimental factor (large number of firefighters)

Essential: Randomization

x_1 o_1 R
x_2 o_2 R

Neutralization of other factors z1, z2, z3, …by randomization!

Serious problem in evaluation research
For example:

1. Labour Market: Occupational training and career success

2. Educational policies: Comparing state and private schools

3. Health policy: SWICA study: Fitness training and health conditions

4. Criminology: Imprisonment versus community work and recidivism rates

Causal effect or selection effect?

If experimental designs are not applicable:

Multivariate statistics

1. Which other factors besides X and Y may be relevant? (Theory)

2. Other factors have to be measured.

3. Include relevant other factors („controls“) in a statistical model

Linear Model

$y = b_0 + b_1x + b_2z + \varepsilon$

$r_{xy} > 0$, $b_1 = 0$, $b_2 > 0$ („spurious correlation“)

$b_{yx} = [s_y/s_x] \quad (r_{xy} - r_{xz} \bullet r_{yz}) / (1 - r^2_{xy})$

$r_{xy} > 0, b_1 = 0 \blacktriangleright r_{xy} = r_{xz} \bullet r_{yz}$

More general case:

$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_mx_m + \varepsilon$

Simple method: Split of bivariate tables by control variables (multivariate table analysis)

Problem: Easy, if one dichotomous control variable. Otherwise, a large data set is necessary.

Multivariate analysis:
Example with three categorical variables

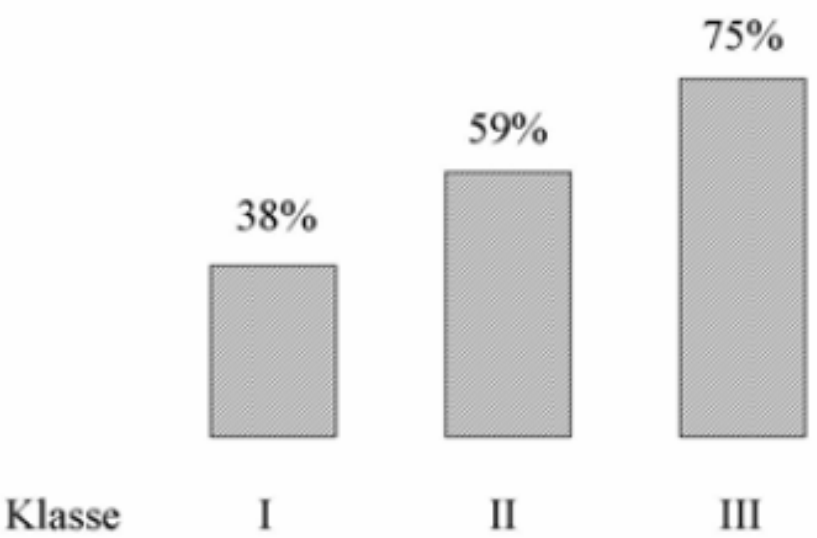
Data from an „unusual episode“

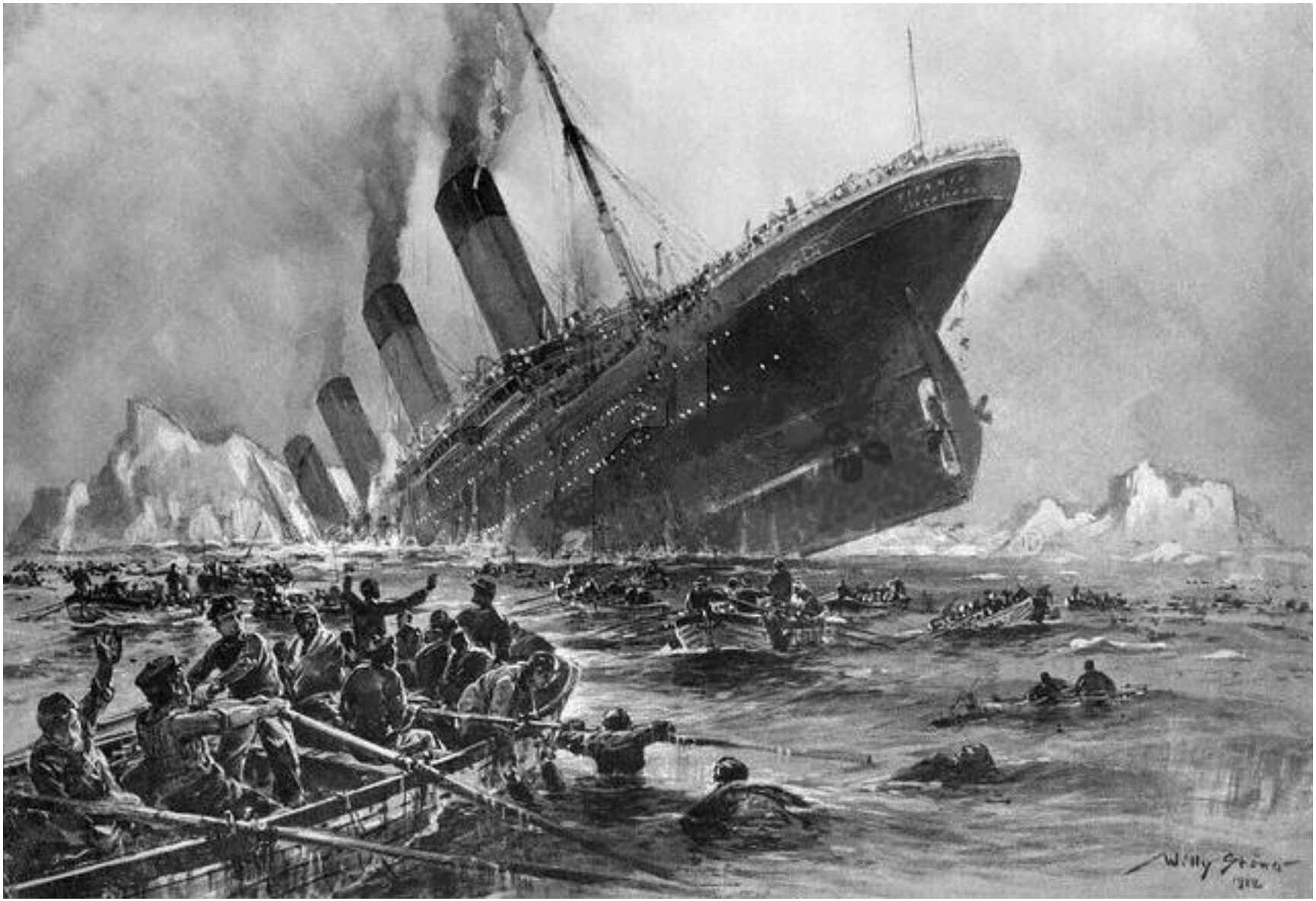
N = 2201 affected subjects for which data is available

Source: Robert M. Dawson, Journal of Statistics Education

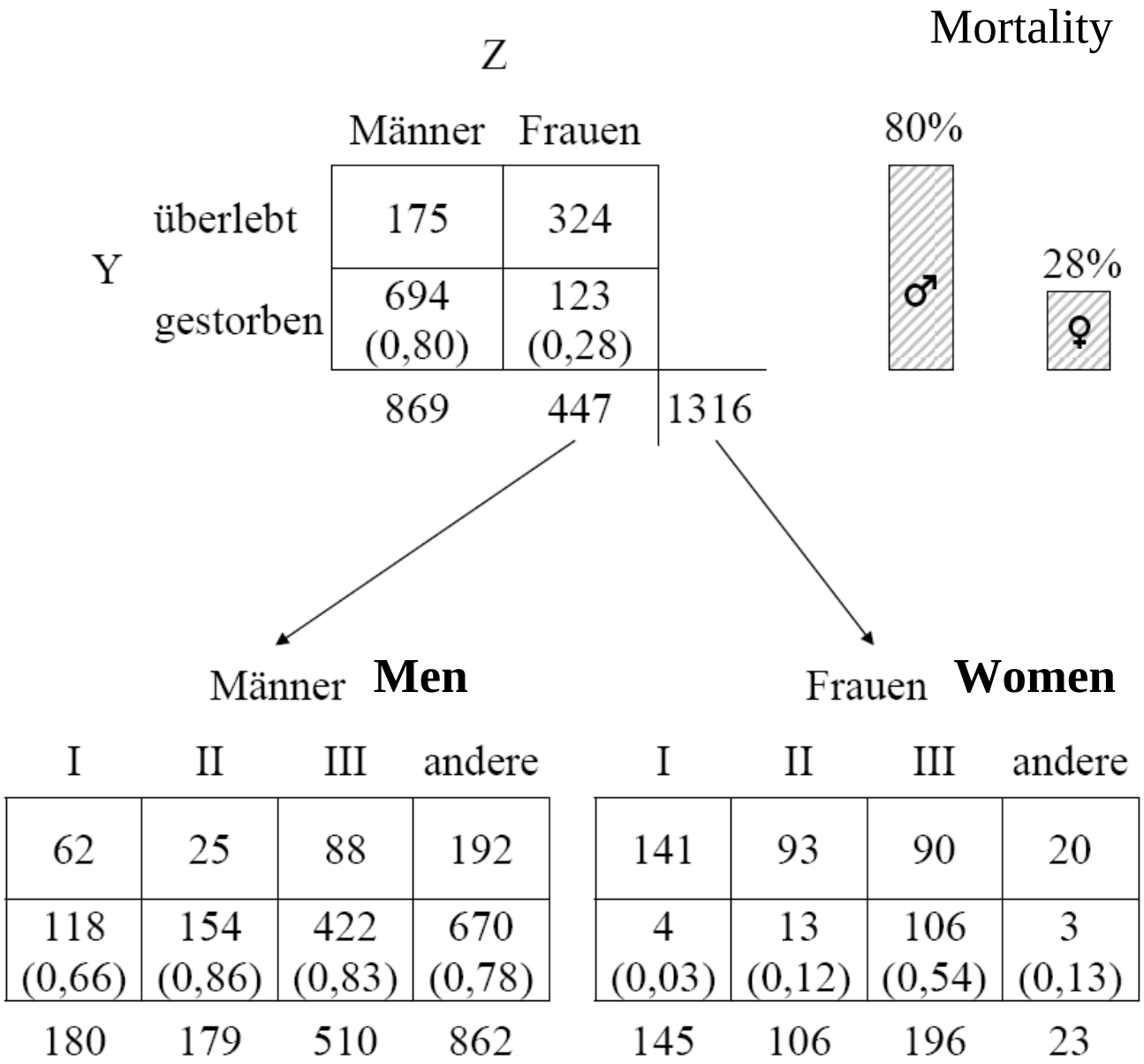
X social class

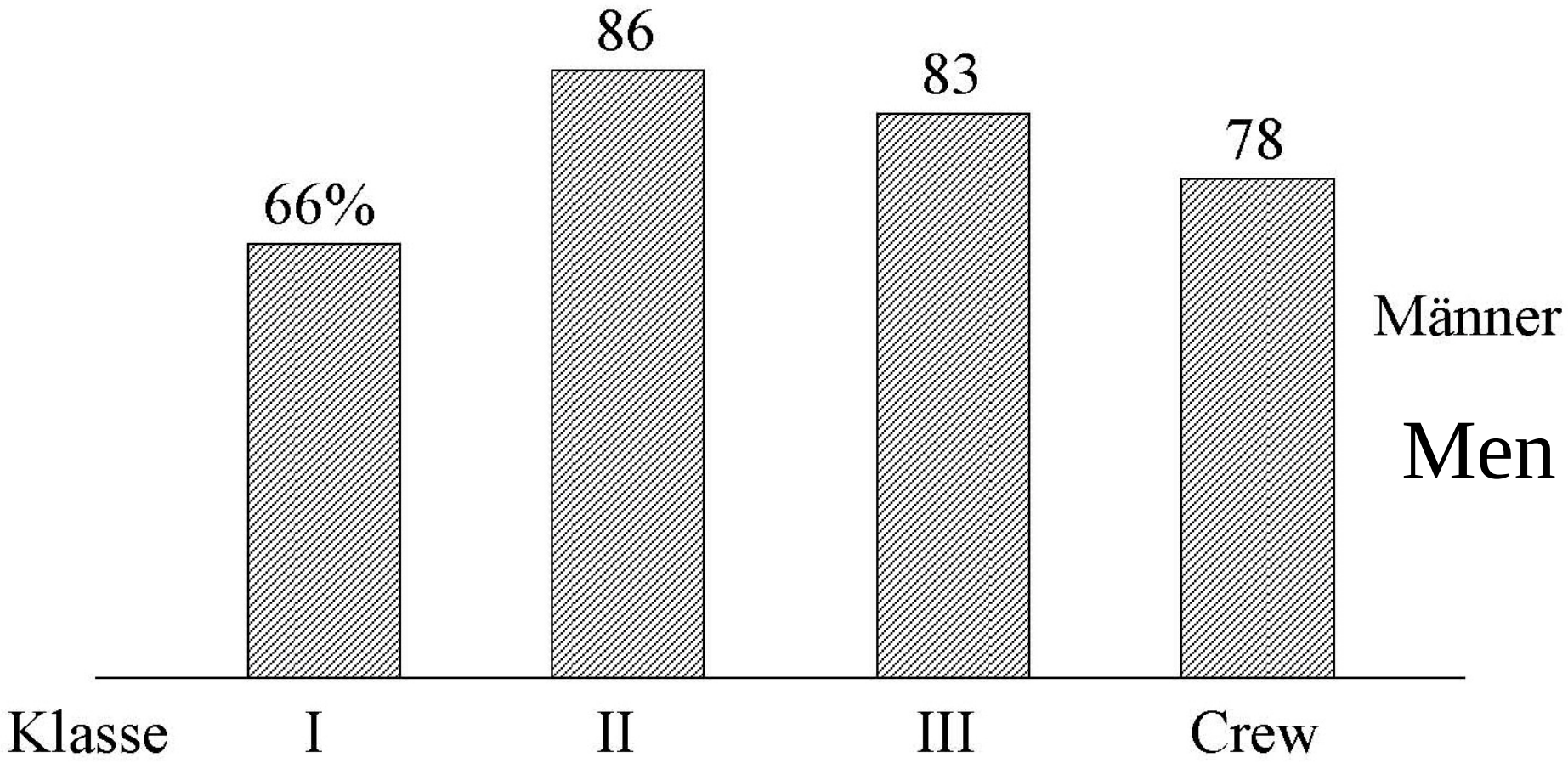
Y	I (high)	II	III (low)	
survived	203	118	178	
died	122 (0,38)	167 (0,59)	528 (0,75)	
	325	285	706	1316

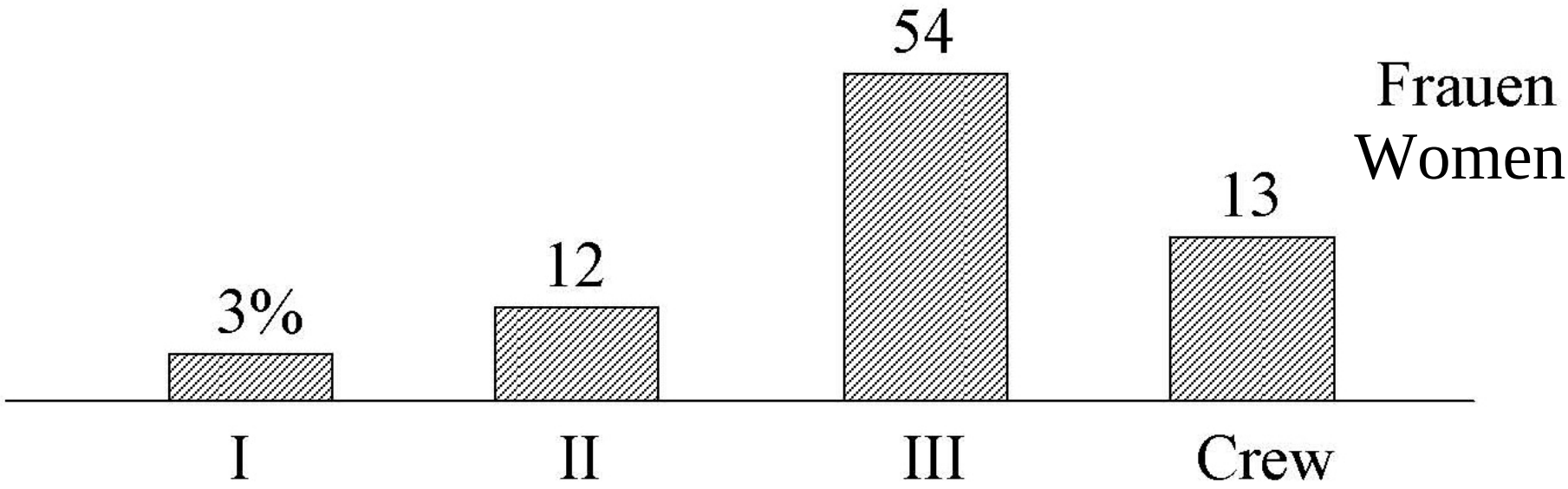




By Sex


Mortality
Z
Männer Frauen
80%
28%
Y
überlebt
175
324
gestorben
694
(0,80)
123
(0,28)
869
447
1316
♂
♀
Männer Men
Frauen Women
I II III andere
62 25 88 192
118 154 422 670
(0,66) (0,86) (0,83) (0,78)
180 179 510 862
I II III andere
141 93 90 20
4 13 106 3
(0,03) (0,12) (0,54) (0,13)
145 106 196 23


86
83
78
66%
Männer
Men
Klasse
I
II
III
Crew
Social class



Sex

Interaction of
Sex and class

Mortality

Multicausality and
interaction

Social
Class

Male, III. Class
died

Female, I. Class
survived

Course overview

Monday

- Morning: Lecture (Diekmann)
  - Introduction and overview
  - Linear regression (OLS)
  - Linear probability model (LPM)
  - Logistic regression model
- Afternoon: Lecture (Diekmann)
  - Logistic regression: Interpretation of coefficients
  - Maximum-likelihood estimation (MLE) of parameters

Tuesday

- Morning and afternoon: PC Pool (Jann)
  - Introduction to software (SPSS and Stata) and data files
  - Exercises for LPM and logistic regression
  - Post processing of results: Interpretation of coefficients
  - Post processing of results: Tabulation

Course overview (continued)

Wednesday

- Morning: Lecture (Jann)
  - Statistical inference: Wald-test and likelihood-ratio test (LR)
  - Evaluation of models: Goodness-of-fit measures (GOF)
  - Model specification
  - Model diagnostics
- Afternoon: PC Pool (Jann)
  - Exercises: Statistical inference, goodness-of-fit, specification, diagnostics

Course overview (continued)

Thursday

- Morning: Lecture (Jann)
  - Probit model
  - Logit/Probit as latent variable models and derivation from utility assumptions
  - Ordered Logit and Probit, multinomial and conditional Logit
- Afternoon: PC Pool (Jann)
  - Exercises for Probit, ordered Logit/Probit, multinomial and conditional Logit

Friday

- Morning: Lecture (Jann)
  - Outlook: Models for count data, panel data models, other topics
  - Final discussion

Textbooks and Overview Articles: English

- Aldrich, John H., and Forrest D. Nelson (1984). Linear Probability, Logit, and Probit Models. Newbury Park, CA: Sage.
- Amemiya, Takeshi (1981). Qualitative Response Models: A Survey. Journal of Economic Literature 19(4): 1483-1536.
- Borooah, Vani K. (2002). Logit and Probit. Ordered and Multinomial Models. Newbury Park, CA: Sage.
- Eliason, Scott R. (1993). Maximum Likelihood Estimation. Logic and Practice. Newbury Park, CA: Sage.
- Fox, John (1997). Applied Regression Analysis, Linear Models, and Related Methods. Thousand Oaks, CA: Sage. (Chapter 15)
- Greene, William H. (2003). Econometric Analysis. 5th ed. Upper Saddle River, NJ: Pearson Education. (Chapter 21)
- Gujarati, Damodar N. (2003). Basic Econometrics. 4th ed. New York: McGraw-Hill. (Chapter 15)
- Hosmer, David W., Jr., and Stanley Lemeshow (2000). Applied Logistic Regression. 2nd ed. New York: John Wiley & Sons.
- Jaccard, James (2001). Interaction Effects in Logistic Regression. Newbury Park, CA: Sage.

Textbooks and Overview Articles: English

- Kohler, Ulrich, and Frauke Kreuter (2005). Data Analysis Using Stata. 2nd ed. College Station, TX: Stata Press. (Chapter 9)
- Liao, Tim Futing (1994). Interpreting Probability Models. Logit, Probit, and Other Generalized Linear Models. Newbury Park, CA: Sage.
- Long, J. Scott (1997). Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: Sage.
- Long, J. Scott, and Jeremy Freese (2005). Regression Models for Categorical Dependent Variables Using Stata. College Station, TX: Stata Press.
- Maddala, G. S. (1983). Limited Dependent and Qualitative Variables in Econometrics. Cambridge: Cambridge University Press.
- Menard, Scott (2002). Applied Logistic Regression Analysis. 2nd ed. Newbury Park, CA: Sage.
- Pampel, Fred C. (2000). Logistic Regression. A Primer. Newbury Park, CA: Sage.
- Verbeek, Marno (2004). A Guide to Modern Econometrics. West Sussex: John Wiley & Sons. (Chapter 7)
- Winship, Christopher, and Robert D. Mare (1984). Regression Models with Ordinal Variables. American Sociological Review 49(4): 512-525.
- Wooldridge, Jeffrey M. (2002). Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: The MIT Press. (Chapter 15)

Textbooks and Overview Articles: German

- Andreß, Hans-Jürgen, Jacques A. Hagenaars, und Steffen Kühnel (1997). Analyse von Tabellen und kategorialen Daten. Log-lineare Modelle, latente Klassenanalyse, logistische Regression und GSK-Ansatz. Berlin: Springer.
- Brüderl, Josef (2000). Regressionsverfahren in der Bevölkerungswissenschaft. S. 589-642 in: Ulrich Mueller, Bernhard Nauck, and Andreas Diekmann (Hg.). Handbuch der Demographie. Berlin: Springer. (Abschnitt 13.2)
- Backhaus, Klaus, Bernd Erichson, Wulff Plinke, und Rolf Weiber (2006). Multivariate Analysemethoden. Eine anwendungsorientierte Einführung. 11. Aufl. Berlin: Springer. (Kapitel 7)
- Kohler, Ulrich, und Frauke Kreuter (2006). Datenanalyse mit Stata. Allgemeine Konzepte der Datenanalyse und ihre praktische Anwendung. 2. Aufl. München: Oldenbourg. (Kapitel 9)
- Maier, Gunther, und Peter Weiss (1990). Modelle diskreter Entscheidungen: Theorie und Anwendung in den Sozial- und Wirtschaftswissenschaften. Wien: Springer.
- Tutz, Gerhard (2000). Die Analyse kategorialer Daten. Anwendungsorientierte Einführung in Logit-Modellierung und kategoriale Regression. München: Oldenbourg.

General Statistics Textbooks

English:

- Agresti, Alan, and Barbara Finlay (1997). Statistical Methods for the Social Sciences. Third Edition. Upper Saddle River, NJ: Prentice-Hall.
- Freedman, David, Robert Pisani, and Roger Purves (1998). Statistics. Third Edition. New York: Norton.

German:

- Fahrmeir, Ludwig, Rita Künstler, Iris Pigeot, and Gerhard Tutz (2004). Statistik. Der Weg zur Datenanalyse. 5. Auflage. Berlin: Springer.
- Jann, Ben (2005). Einführung in die Statistik. 2. Auflage. München: Oldenbourg.

Part I

General Statistical Concepts

- Probability and Random Variables
- Probability Distributions
- Parameter Estimation
- Expected Value and Variance

Probability and Random Variables

- Given is a random process with possible outcomes A, B, etc. The probability $\Pr(A)$ is the relative expected frequency of event A.
- Probability axioms:

$$\begin{aligned} \Pr(A) &\geq 0 \\ \Pr(\Omega) &= 1 \\ \Pr(A \cup B) &= \Pr(A) + \Pr(B) \quad \text{if } A \cap B = \emptyset \end{aligned}$$

  where Ω is the set of (disjunctive) elementary outcomes.
- Conditional probability $\Pr(A|B)$: Probability of A given the occurrence of B
- A **random variable** is a variable that takes on different values with certain probabilities.
  Imagine tossing a coin. The outcome of the toss is a random variable. There are two possible outcomes, heads and tails, and each outcome has probability 0.5.

Probability Distribution

- A random variable can be **discrete** (only selected values are possible) or **continuous** (any value in an interval is possible).
- The probability distribution of a random variable X can be expressed as probability mass or density function (PDF) or as cumulative probability function (CDF).

PDF

discrete: $f(x) = p(x) = \Pr(X = x)$

continuous: $\Pr(a \leq X \leq b) = \int_a^b f(x)\,dx$

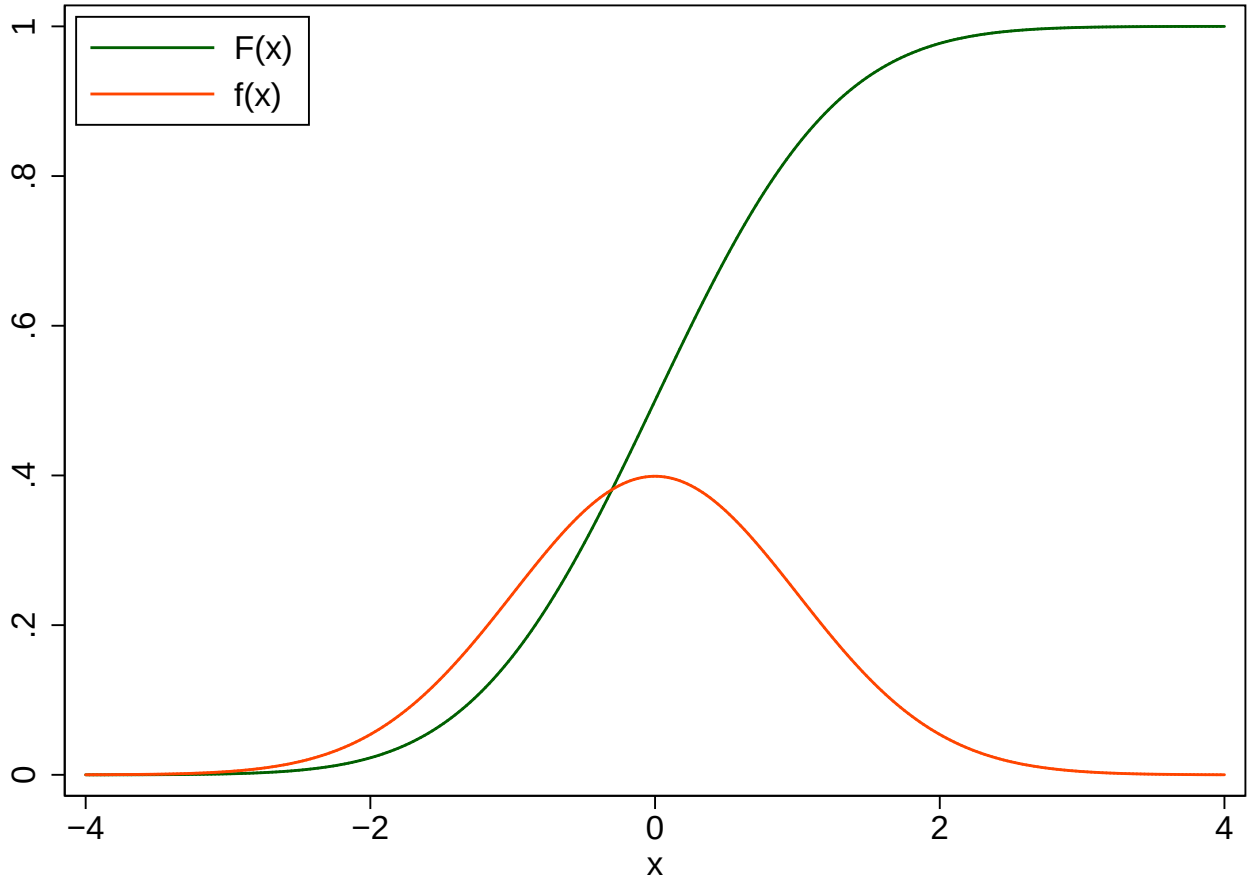
CDF

discrete: $F(x) = \Pr(X \leq x) = \sum_{x_i \leq x} p(x_i)$

continuous: $F(x) = \Pr(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$

Probability Distribution: Example

- Normal distribution and density (continuous)



$$f(x) = \phi(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Parameter Estimation

- Distributions can be described by parameters such as the expected value (the mean) or the variance. The goal in statistics is to estimate these parameters based on sample data.
- Desirable properties of an estimator $\hat{\theta}$ for parameter vector θ:
  1. Unbiasedness: $E(\hat{\theta}) = \theta$ – on average (over many samples) the estimator is equal to the true θ.
  2. Efficiency: The variance of the estimator (i.e. the variability of the estimates from sample to sample) should be as small as possible.
  3. Consistency: Unbiasedness is not always possible, but an estimator should at least be asymptotically unbiased (i.e. $E(\hat{\theta})$ should approach θ as $n \to \infty$).
- The sampling variance of an estimator determines the precision of the estimator and can be used to construct confidence intervals and significance tests. (The sampling variance itself is also unknown and has to be estimated.)

Expected Value and Variance

Expected Value $E(X) = \mu$

discrete: $E(X) = \sum_i x_i p(x_i)$ continuous: $E(X) = \int_{-\infty}^{+\infty} x f(x)\, dx$

Variance $V(X) = \sigma^2$

general: $V(X) = E[(X - \mu)^2] = E(X^2) - \mu^2$

discrete: $V(X) = \sum_i (x_i - \mu)^2 p(x_i)$ cont.: $V(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x)\, dx$

Sample estimators $\bar{X}$ and S_X

$$\hat{E}(X) = \bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i \qquad \hat{V}(X) = S_X = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2$$

Part II

Multiple Linear Regression

- Model
- Assumptions
- Interpretation
- Estimation
- Goodness-of-Fit

Linear Regression Model

- The conditional expectation of a continuous random variable Y is expressed as a linear function of the predictors $X_1, X_2, \ldots, X_m$.

$$E(Y_i|X_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_m X_{im}$$

where X_i stands for $(X_{i1}, \ldots, X_{im})$.

- Specific observations randomly deviate from the expected value so we add a random error term to the model:

Linear Regression Model (LRM)

$$Y_i = E(Y_i|X_i) + \epsilon_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_m X_{im} + \epsilon_i$$

$$E(\epsilon_i) = 0$$

Linear Regression Model

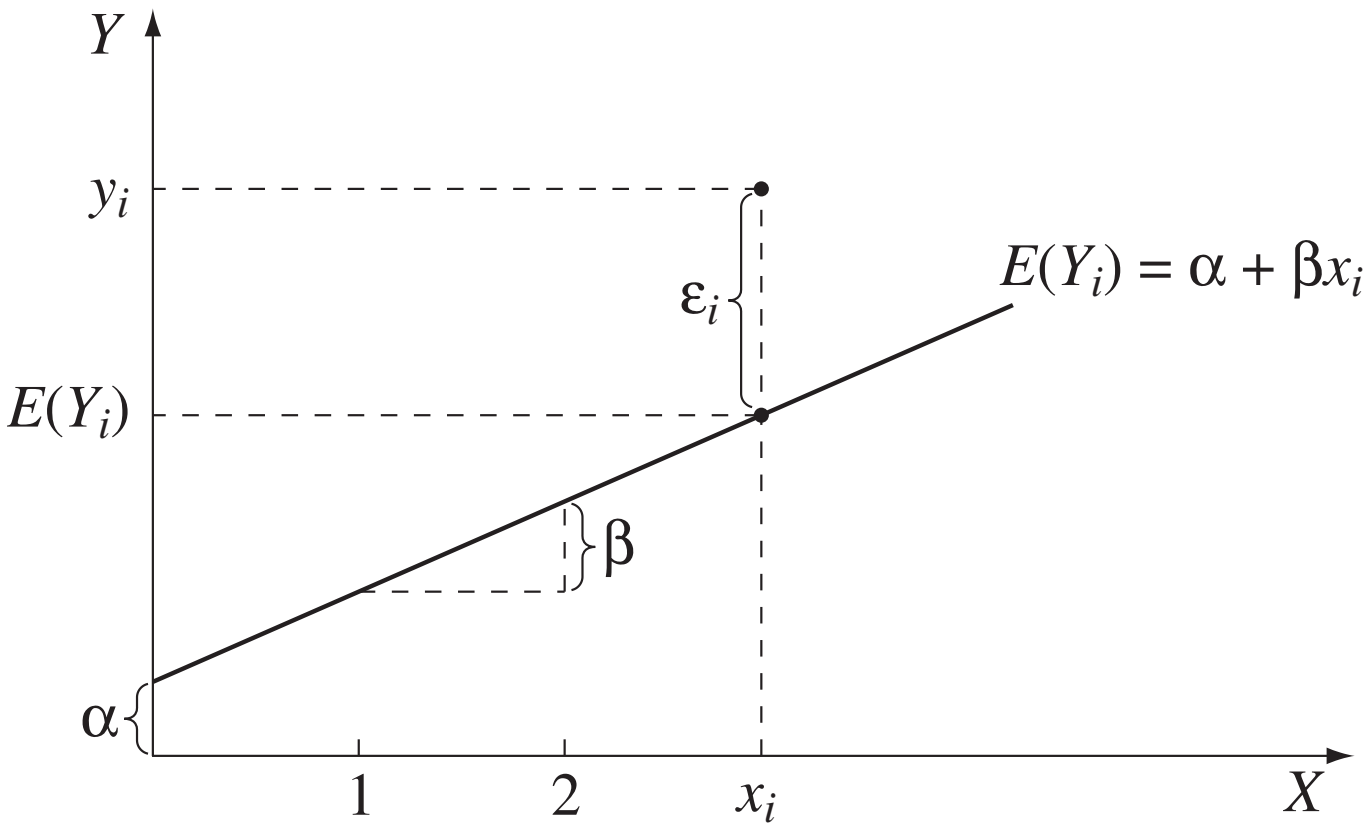
LRM in Matrix Notation

$$Y = X\beta + \epsilon$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1m} \\ 1 & X_{21} & X_{22} & \cdots & X_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nm} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Graphical Representation

Simple LRM



Assumptions

1. Conditional expectation of error term is zero

$$E(\epsilon_i) = 0$$

This implies

$$E(Y_i|X_i) = X_i'\beta$$

and also that the errors and predictors are uncorrelated:

$$E(\epsilon_i X_{ik}) = 0$$

(correct functional form, no relevant omitted variables, no reverse causality/endogeneity)

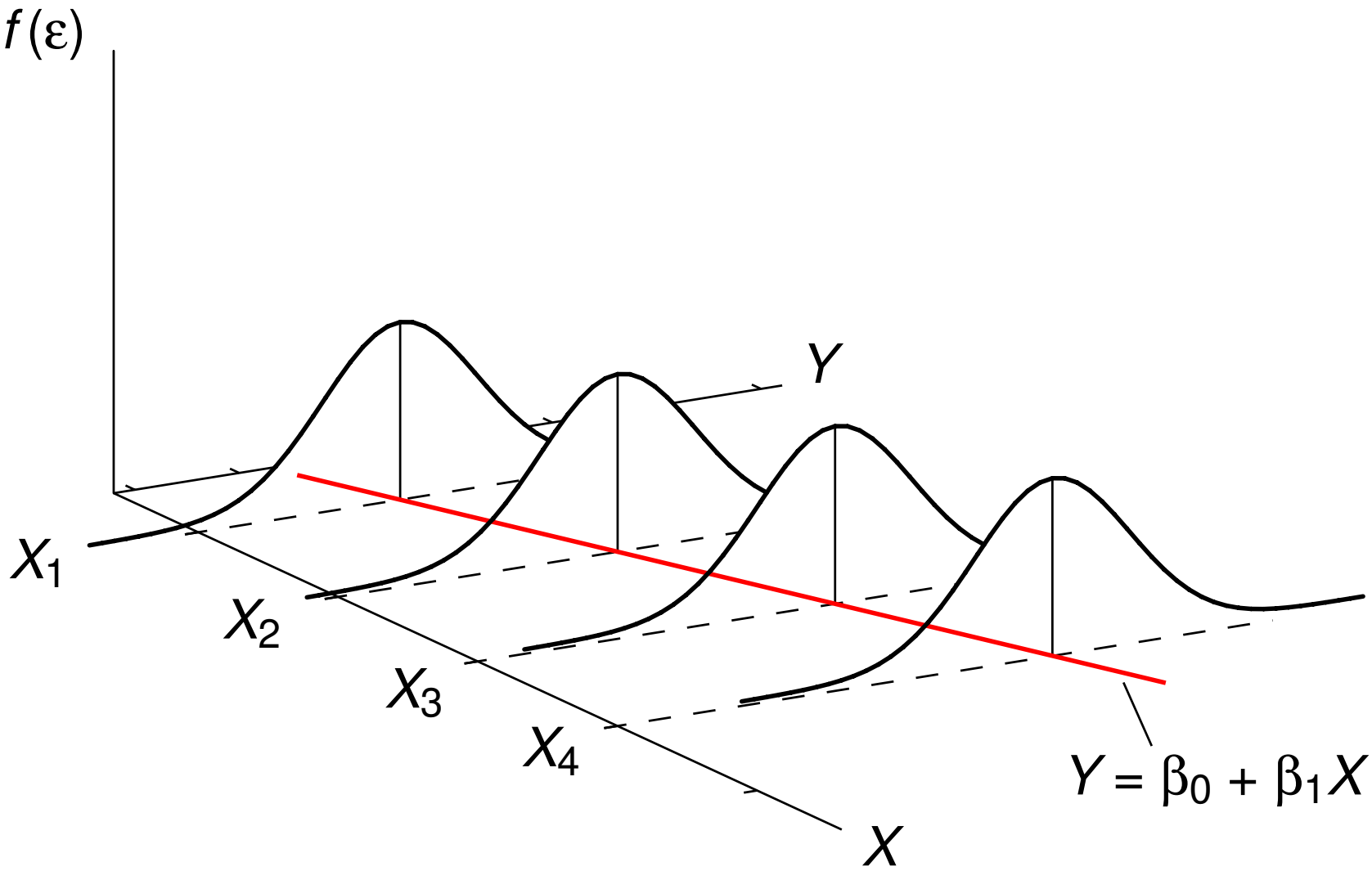
2. The errors have constant variance and are uncorrelated

$$E(\epsilon\epsilon') = \sigma^2 I$$

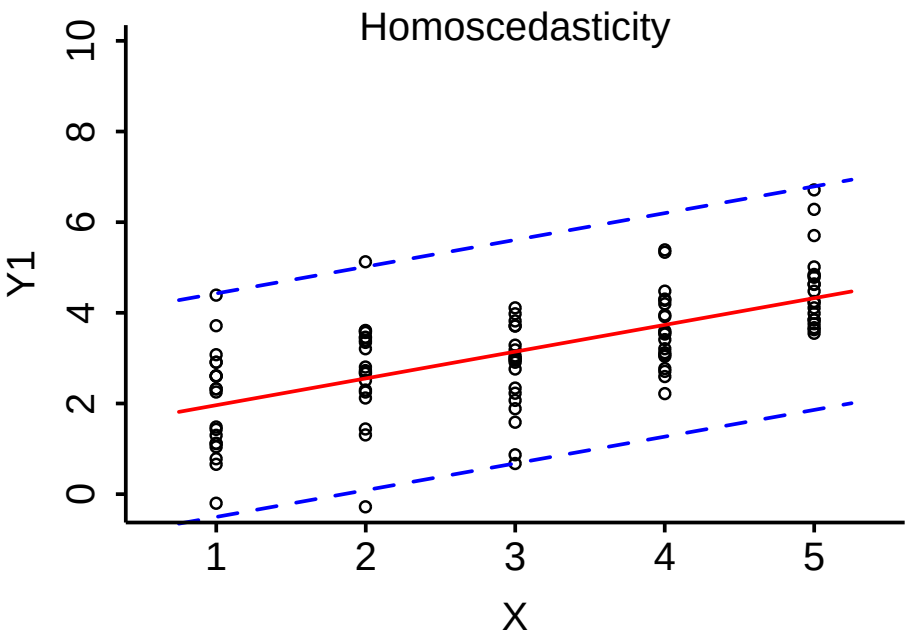
3. Errors are normally distributed:

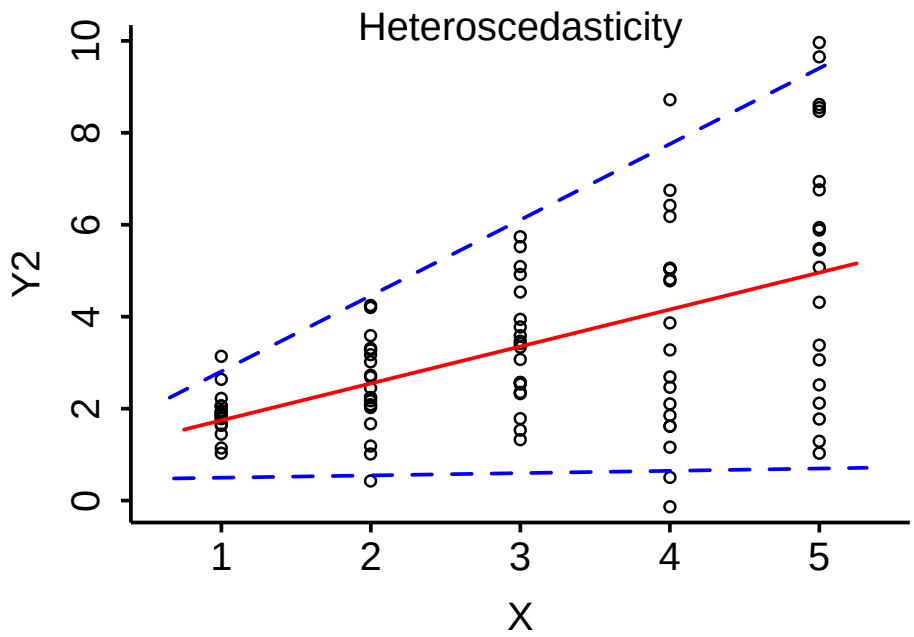
$$\epsilon_i \sim N(0, \sigma)$$

Assumptions



Heteroscedasticity





Interpretation

Model:

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m$$

Marginal Effect / Partial Effect

$$\frac{\partial E(Y|X)}{\partial X_k} = \beta_k$$

Discrete Change / Unit Effect (effect of $\Delta X_k = 1$)

$$\frac{\Delta E(Y|X)}{\Delta X_k} = E(Y|X, X_k + 1) - E(Y|X, X_k) = \beta_k$$

In linear regression: marginal effect = unit effect

$\Rightarrow$ effects are independent of value of X_k and Y (constant effects)

$\Rightarrow$ very convenient for interpretation

OLS Estimation

- The parameters of the linear regression model are usually estimated by the method of ordinary least squares (OLS).
- The objective of the method is to minimize the squared differences between the model predictions given the estimate and the observed data, i.e. to choose the $\hat{\beta}$ that minimizes

$$\sum_{i=1}^{N}(Y_i - X_i'\hat{\beta})^2$$

- In matrix notation, the OLS solution is as follows:

OLS Estimator

$$\hat{\beta} = (X'X)^{-1}X'Y$$

- The OLS estimator is unbiased and efficient (best linear unbiased estimator, BLUE), if assumptions (1) and (2) hold.

Residuals and R-Squared

- Predictions based on estimated parameters:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_m X_{im}$$

- Residuals: Deviation between predictions and observed data

$$r_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \cdots - \hat{\beta}_m X_{im}$$

- The mechanics of OLS are such that the sum of residuals is zero and the sum of squared residuals is minimized.
- As a measure of **goodness-of-fit** between the data and the model, the R-squared is used: Proportion of variance of Y which is "explained" by the model.

R-squared

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^{N} r_i^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

Part III

Linear Probability Model (LPM)

- Model
- Advantages
- Example
- Problems
- WLS estimation

Linear Probability Model (LPM)

Given is a dichotomous (binary) dependent variable Y:

$$Y_i = \begin{cases} 1 & \text{Event} \\ 0 & \text{No event} \end{cases}$$

- **Expected value:** The probability distribution of Y is given as $\Pr(Y_i = 1) = \pi_i$ for value 1 und $\Pr(Y_i = 0) = (1 - \pi_i)$ for value 0, $\pi_i \in [0, 1]$, therefore

$$E(Y_i) = 1 \cdot \pi_i + 0 \cdot (1 - \pi_i) = \Pr(Y_i = 1)$$

$\Rightarrow$ the expected value of Y is equal to the probability of Y being 1

Linear Probability Model (LPM)

- Because $E(Y_i) = \Pr(Y_i = 1)$ it seems reasonable to model $\Pr(Y_i = 1)$ using standard linear regression techniques:

$$\Pr(Y_i = 1|X_i) = E(Y_i|X_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_m X_{im}$$

- Adding the error term yields the Linear Probability Model, which models $\Pr(Y_i = 1)$ as a linear function of the independent variables:

Linear Probability Model (LPM)

$$\begin{aligned} Y_i &= \Pr(Y_i = 1|X_i) + \epsilon_i = E(Y_i|X_i) + \epsilon_i \\ &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_m X_{im} + \epsilon_i \end{aligned}$$

where $E(\epsilon_i) = 0$ is assumed.

- The model parameters can be estimated using OLS.

Are the OLS parameter estimates unbiased?

The model is

$$Y_i = \Pr(Y_i = 1|X_i) + \epsilon_i = Z_i + \epsilon_i$$

where

$$Z_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_m X_{im}$$

Therefore:

$$\epsilon_i = \begin{cases} -Z_i & \text{if } Y_i = 0 \\ 1 - Z_i & \text{if } Y_i = 1 \end{cases}$$

It follows:

$$E(\epsilon_i) = \Pr(Y_i = 0|X_i)(-Z_i) + \Pr(Y_i = 1|X_i)(1 - Z_i) = 0$$

because $\Pr(Y_i = 1|X_i) = Z_i$ by definition. Since $E(\epsilon_i) = 0$, the OLS estimates are unbiased. (Depending on the model being correct!)

Advantages of the LPM

- Easy to interpret:
  - β_k is the expected change in $\Pr(Y = 1)$ for a unit increase in X_k

$$\Pr(Y = 1|X_k + 1)) - \Pr(Y = 1|X_k) = \beta_k$$

  - β_k is the partial (marginal) effect of X_k on $\Pr(Y = 1)$

$$\frac{\partial \Pr(Y = 1|X)}{\partial X_k} = \beta_k$$

- Simple to estimate via OLS.
- Good small sample behavior.
- Often good as a first approximation, especially if the mean of $\Pr(Y = 1)$ is not close to 0 or 1 and effects are not too strong.
- Applicable in cases where the Logit (or similar) fails (e.g. if Y constant for one value of a categorical predictor, see Caudill 1988; in badly behaved Randomized Response data)

Polytomous dependent variable

LPM can also be used to model a polytomous variable that has $J > 2$ categories (e.g. religious denomination).

One equation is estimated for each category, i.e.

$$
\begin{aligned}
Y_{i1} &= \beta_{01} + \beta_{11} + X_i + \cdots + \epsilon_{i1} \\
Y_{i2} &= \beta_{02} + \beta_{12} + X_i + \cdots + \epsilon_{i2} \\
&\vdots \\
Y_{iJ} &= \beta_{0J} + \beta_{1J} + X_i + \cdots + \epsilon_{iJ}
\end{aligned}
\qquad \text{where } Y_{iJ} = \begin{cases} 1 & \text{if } Y_i = j \\ 0 & \text{else} \end{cases}
$$

As long as constants are included in the separate models, OLS ensures that

$$\sum_{j=1}^{J} \hat{Y}_{ij} = \sum_{j=1}^{J} (\hat{\beta}_{0j} + \hat{\beta}_{1j} + X_i + \cdots +) = 1$$

Example of the LPM: Labor Force Participation (Long 1997:36-38)

TABLE 3.1 Descriptive Statistics for the Labor Force Participation Example

Name	*Mean*	*Standard Deviation*	*Minimum*	*Maximum*	*Description*
LFP	0.57	0.50	0.00	1.00	1 if wife is in the paid labor force; else 0
K5	0.24	0.52	0.00	3.00	Number of children ages 5 and younger
K618	1.35	1.32	0.00	8.00	Number of children ages 6 to 18
AGE	42.54	8.07	30.00	60.00	Wife's age in years
WC	0.28	0.45	0.00	1.00	1 if wife attended college; else 0
HC	0.39	0.49	0.00	1.00	1 if husband attended college; else 0
LWG	1.10	0.59	−2.05	3.22	Log of wife's estimated wage rate
INC	20.13	11.63	−0.03	96.00	Family income excluding wife's wages

NOTE: $N = 753$.

Example of the LPM: Labor Force Participation (Long 1997:36-38)

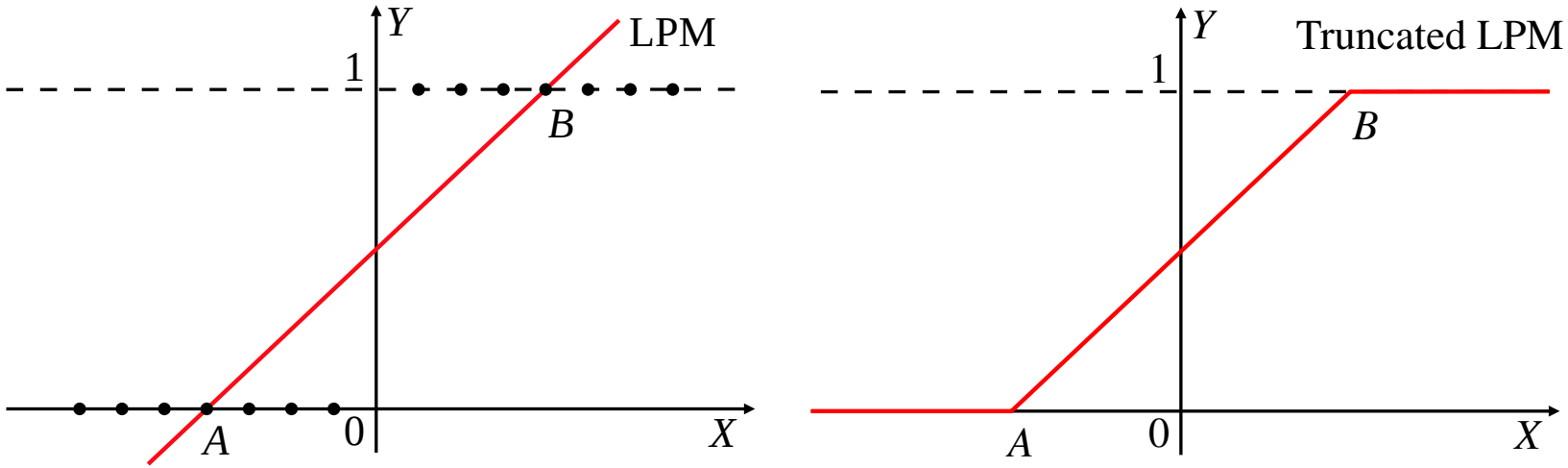
TABLE 3.2 Linear Probability Model of Labor Force Participation

Variable	β	β^{S_x}	t
Constant	1.144	—	9.00
K5	−0.295	−0.154	−8.21
K618	−0.011	−0.015	−0.80
AGE	−0.013	−0.103	−5.02
WC	0.164	—	3.57
HC	0.019	—	0.45
LWG	0.123	0.072	4.07
INC	−0.007	−0.079	−4.30

NOTE: $N = 753$. β is an unstandardized coefficient; β^{S_x} is an x-standardized coefficient; t is a t-test of β.

Problems of LPM

- **Nonsensical predictions.** Predictions from the LPM can be below 0 or above 1, which contradicts Kolmogorov's probability axioms.



- **Unreasonable functional form.** The LPM assumes that effects are the same over the whole probability range, but this is often not realistic.
- Usefulness of R^2 is questionable. Even a very clear relationships do not result in a high R^2-values.

Problems of LPM

- **Nonnormality.** The errors are not normally distributed in the LPM, since ϵ_i can only have two values: $-Z_i$ or $1 - Z_i$.
  Consequences:
  - normal-approximation inference is invalid in small samples
  - more efficient non-linear estimators exist (LPM is not BUE)
- **Heteroskedasticity.** The errors are do not have constant variance

$$\begin{aligned} V(\epsilon_i) = E(\epsilon_i^2) &= \Pr(Y_i = 0)[-Z_i]^2 + \Pr(Y_i = 1)[1 - Z_i]^2 \\ &= [1 - \Pr(Y_i = 1)][\Pr(Y_i = 1)]^2 + \Pr(Y_i = 1)[1 - \Pr(Y_i = 1)]^2 \\ &= \Pr(Y_i = 1)[1 - \Pr(Y_i = 1)] = Z_i(1 - Z_i) \end{aligned}$$

since $E(\epsilon_i) = 0$ and $\Pr(Y_i = 1) = Z_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_m X_{im}$.

$\Rightarrow$ OLS estimator is inefficient (and standard errors are biased)

WLS estimation of LPM

- Goldberger (1964) proposed to use a two-step, weighted least squared approach (WLS) to correct for heteroskedasticity.
  - Step 1: Estimate standard OLS-LPM to obtain an estimate of the error variance

$$\hat{V}(\epsilon_i) = \hat{Y}_i(1 - \hat{Y}_i), \quad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_m X_{im}$$

  - Step 2: Use the estimated variance in a weighted least squares approach to estimate the heteroskedasticity-corrected LPM parameters. (Intuitively: smaller variance, more reliable observation.) This is equivalent to dividing all variables (including the constant) by the square-root of the variance and then fit an OLS model to these modified data.

$$w_i Y_i = \beta_0 w_i + \beta_1 w_i X_{i1} + \cdots + \epsilon_i w_i, \quad w_i = \sqrt{\hat{Y}_i(1 - \hat{Y}_i)^{-1}}$$

The procedure increases the efficiency of the LPM (but standard errors are still somewhat biased since the true variance is unknown). Problematic are observations for which $\hat{Y}_i$ is outside $[0, 1]$ (negative variance estimate!).

Part IV

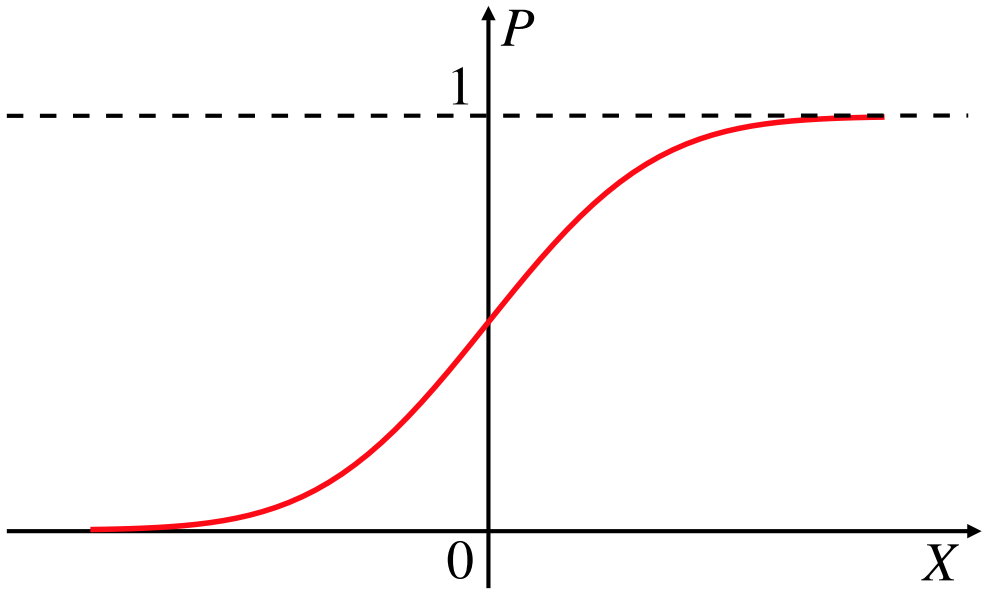
Logistic Regression: Model

- Nonlinear Relationship
- The Logit Model
- Log Odds Form
- Simple Example

Nonlinear Effect on $P(Y = 1)$

- The linear effects assumed in the LPM do often not make much sense and a more reasonable probability model should
  1. ensure that $\Pr(Y = 1)$ always lies within $[0, 1]$ and
  2. use a nonlinear functional form so that effects diminish if $\Pr(Y_i = 1)$ gets close to 0 or 1

  Usually it is also sensible to assume symmetry.
- Example for such a nonlinear function:



The Logit Model

A suitable function to model the relationship between $\Pr(Y_i = 1)$ and the independent variables is the logistic function.

The logistic function

$$Y = \frac{e^Z}{1 + e^Z} = \frac{1}{1 + e^{-Z}}$$

Parameterizing Z as a linear function of the predictors yields the Logit model.

The Logit Model

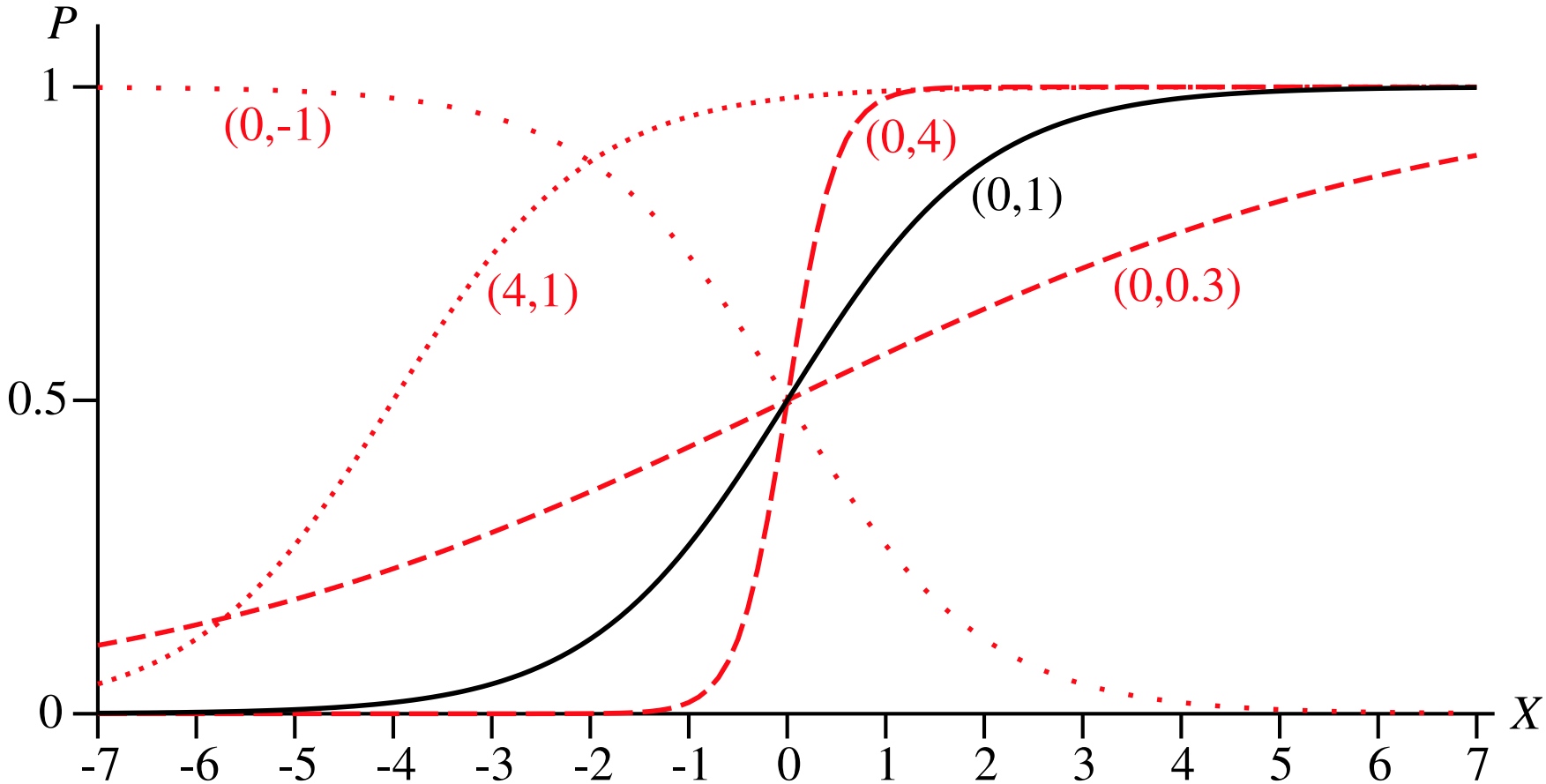
$$\Pr(Y_i = 1|X_i) = E(Y_i|X_i) = \frac{1}{1 + e^{-Z_i}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_m X_{im})}}$$

The Logit Model

Illustration of

$$\Pr(Y_i = 1|X_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_i)}}$$

for different choices of β_0 and β_1:



The Logit Model

The logistic model is intrinsically linear and can be re-expressed as:

The Logit Model: Log Odds Form

$$\ln\left(\frac{\Pr(Y_i = 1|X_i)}{1 - \Pr(Y_i = 1|X_i)}\right) = Z_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_m X_{im}$$

That is, the log of the odds of $Y = 1$ is expressed in the Logit model as a linear function of the predictors. ($\Pr(x)/[1 - \Pr(x)]$ is called the odds of event x.)

Derivation:

$$P = \frac{e^Z}{1 + e^Z} = 1 + \frac{e^Z}{1 + e^Z} - \frac{1 + e^Z}{1 + e^Z} = 1 - \frac{1}{1 + e^Z} \quad \Rightarrow \quad 1 - P = \frac{1}{1 + e^Z}$$

$$\Rightarrow \quad e^Z = \frac{1}{1 - P} - 1 = \frac{P}{1 - P} \quad \Rightarrow \quad Z = \ln\left(\frac{P}{1 - P}\right)$$

The function $f(x) = \ln(x/(1 - x))$ is sometimes called the *logit* function; $L = \ln(x/(1 - x))$ is called the *logit* of x (Berkson 1944, 1951).

A Simple Example

Data:

id	educ	income	sex	car	rural
1	9	3500	0	1	1
2	8	2400	0	0	1
3	18	5200	1	1	1
4	9	3200	0	0	0
5	9	2300	0	0	0
6	10	4500	1	1	1
7	18	12000	0	1	0
8	10	6500	1	1	1
9	9	99999	0	1	1
10	9	99999	0	0	0
11	9	2300	0	0	1
12	10	99999	1	0	0
13	13	4600	0	1	1
14	10	1600	1	1	0
15	9	2900	1	0	0

Income = 99999 is missing value

car by sex (income $\neq$ 99999, $n = 12$):

		sex = 0	sex = 1
car	0	4 4/7 = 0.5714	1 1/5 = 0.20
	1	3 3/7 = 0.4286	4 4/5 = 0.80

Logit: $\Pr(\texttt{car} = 1) = \frac{1}{1+e^{-[\beta_0+\beta_1 \texttt{sex}]}}$

LPM: $E(\texttt{car}) = \beta_0 + \beta_1 \texttt{sex}$

$\Rightarrow \quad \hat{\beta}_0, \hat{\beta}_1 ?$

A Simple Example - Estimation

- LPM
  - $\text{sex} = 0$:
$$E(\text{car}) = \beta_0 \quad \Rightarrow \quad \hat{\beta}_0 = 0.4286$$
  - $\text{sex} = 1$:
$$E(\text{car}) = \beta_0 + \beta_1 \quad \Rightarrow \quad \hat{\beta}_1 = 0.80 - 0.4286 = 0.3714$$

  $$\Rightarrow \quad \widehat{E}(\text{car}) = 0.4286 + 0.3714 \cdot \text{sex}$$
- Logit
  - $\text{sex} = 0$:
$$0.4286 = \frac{1}{1 + e^{-\hat{\beta}_0}} \quad \Rightarrow \quad \hat{\beta}_0 = \ln\left(\frac{0.4286}{1 - 0.4286}\right) = -0.2876$$
  - $\text{sex} = 1$:
$$\hat{\beta}_0 + \hat{\beta}_1 = \ln\left(\frac{0.80}{1 - 0.80}\right) \quad \Rightarrow \quad \hat{\beta}_1 = \ln 4 - (-0.2876) = 1.6739$$

  $$\Rightarrow \quad \widehat{\Pr}(\text{car} = 1) = \frac{1}{1+e^{-[-0.2876+1.6739 \cdot \text{sex}]}}$$

A Simple Example - Interpretation

- LPM
  - Effect of shift from $\texttt{sex} = 0$ to $\texttt{sex} = 1$ ($\Delta X = 1$):

$$\Delta\widehat{E}(\texttt{car}) = +0.3714$$

  - Partial effect:

$$\frac{\partial\widehat{E}(\texttt{car})}{\partial\texttt{sex}} = 0.3714$$

- Logit
  - Effect of shift from $\texttt{sex} = 0$ to $\texttt{sex} = 1$ ($\Delta\texttt{sex} = 1$):

$$\Delta\widehat{\Pr}(\texttt{car}) = 0.80 - 0.4286 = 0.3714$$

  - Partial effect (let $P = \widehat{\Pr}(\texttt{car} = 1)$):

$$\frac{\partial P}{\partial\texttt{sex}} = \beta_1 \cdot P(1-P), \quad \text{for } P = .5: \quad \frac{\partial P}{\partial\texttt{sex}} = 1.6739 \cdot .5(1-.5) = 0.42$$

Effect of sex: Probability of driving a car increases by 0.37 (= percentage difference d%).

Part V

Logistic Regression: Interpretation of Parameters

- Qualitative Interpretation
- Effect on the Log of the Odds
- Odds Ratio
- Marginal Effects and Elasticities
- Predictions and Discrete Change Effect
- Relative Risks

Non-Linearity

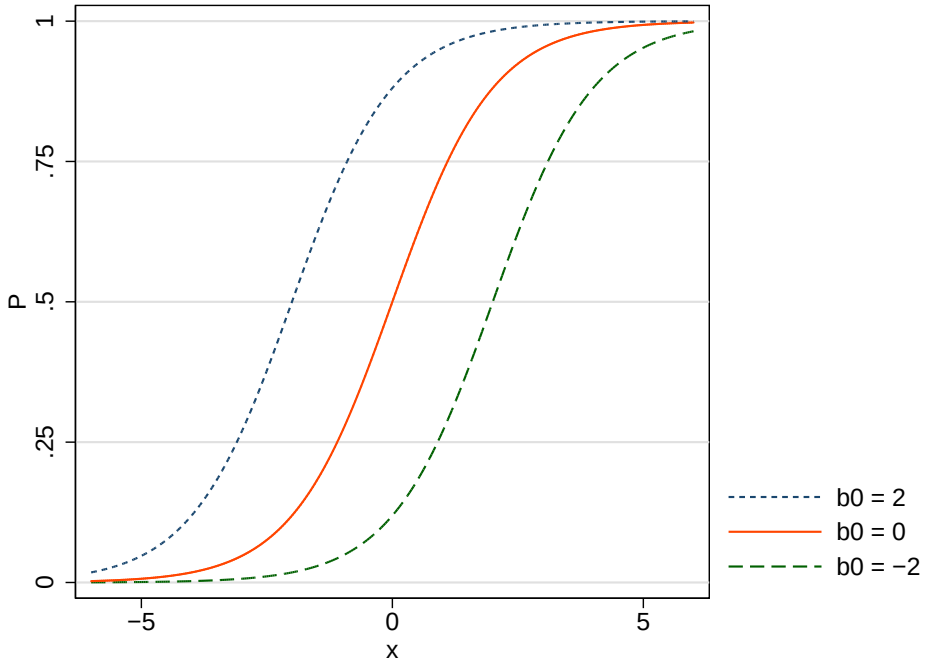
- The relationship between $\Pr(Y = 1)$ and the predictors in a Logit model is non-linear (S-shaped).
- Therefore: The effect of a predictor on $\Pr(Y = 1)$ depends on the level $\Pr(Y = 1)$ of, i.e. the effect is not constant.
- This makes interpretation more difficult than for linear regression.

The Constant

- Consider the following simple Logit model:

$$\text{logit}(\Pr[Y = 1|X]) = \beta_0 + \beta_1 X$$

- A change in β_0 simply sifts the entire probability curve (an increase in β_0 shifts the curve left(!), and vice versa):



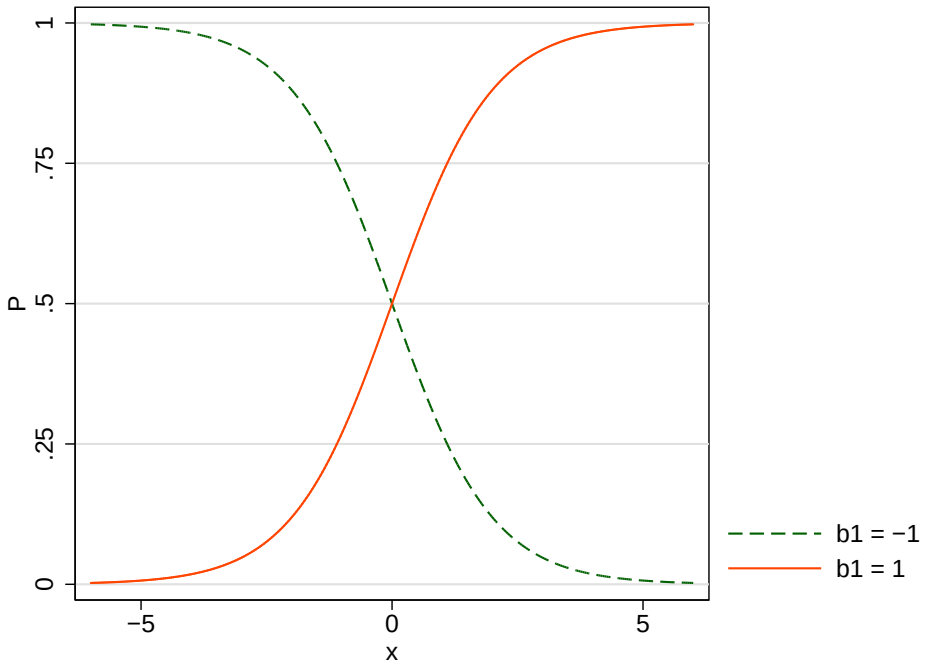
When $\beta_0 = 0$, the curve passes through point $(0, .5)$.

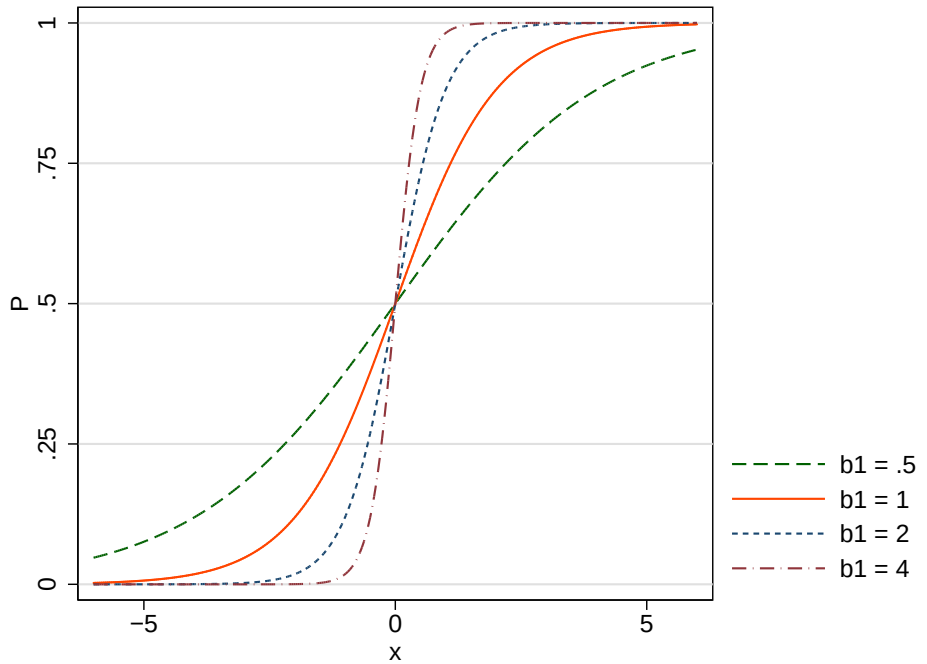
($-\beta_0/\beta_1$ is equal to the value of X for which $P = 0.5$)

- All else equal, a higher β_0 means that the general level of $\Pr(Y = 1)$ is higher.

Slope Parameters: Sign and Size

- The sign of β_1 determines the direction of the probability curve. If β_1 is positive, $\Pr(Y = 1|X)$ increases as a function of X (and vice versa).
- The size of β_1 determines the steepness of the curve, i.e. how quickly $\Pr(Y = 1|X)$ changes as a function of X.





- Sizes of effects can be compared for variables that are on the same scale (e.g. compare the effect of the same variable in two samples).

Effect on Log of the Odds

- The Logit model is

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$$

where P stands for $\Pr(Y = 1|X)$. Therefore,

$$\partial \ln\left(\frac{P}{1-P}\right) / \partial X_k = \beta_k$$

$\Rightarrow \beta_k$ is the marginal effect of X_k on the log of the odds (the *logit*).

- Increase in X_k by t units changes the log of the odds by $t \cdot \beta_k$
- But what does this mean? Log-odds are not very illustrative

Example

Travel to work:

$$Y = \begin{cases} 1 & \text{public transportation} \\ 0 & \text{by car} \end{cases}$$

$$\Pr(Y = 1) = \frac{1}{1 + e^{-Z}}, \quad Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$$

$$Z = 0.01 + 0.079 \cdot \text{educ} + 0.02 \cdot \text{age}$$

- Increase in educ by one unit (one year) increases the logit by 0.079.

Odds Ratio (Effect Coefficient, Factor Change)

- The Logit model can be rewritten in terms of the odds:

$$\frac{P}{1-P} = e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots)} = e^{\beta_0} \cdot e^{\beta_1 X_1} \cdot e^{\beta_2 X_2} \cdot \ldots.$$

- Effect of increase in X_1 by one unit:

$$\frac{P(X_1+1)}{1-P(X_1+1)} = e^{\beta_0} \cdot e^{\beta_1 (X_1+1)} \cdot e^{\beta_2 X_2} \cdot \ldots = \frac{P}{1-P} \cdot e^{\beta_1}$$

$\Rightarrow$ change in X_1 has a multiplier effect on the odds: The odds are multiplied by factor e^{β_1}

Odds Ratio / Effect Coefficient / Factor Change Coefficient

$$\alpha_k = \frac{\frac{P(X_k+1)}{1-P(X_k+1)}}{\frac{P}{1-P}} = e^{\beta_k}$$

- Example: $e^{0.079} = 1.08$

Odds Ratio (Effect Coefficient, Factor Change)

- Percent change coefficient $(e^{\beta_k} - 1) \cdot 100$
- Example: $(e^{0.079} - 1) \cdot 100 = 8$
  $\Rightarrow$ increase of education by one year increases odds by 8%
- Assume the proportion of public transportation is 25% (odds = 1 : 3)

$$\frac{P}{1-P} = \frac{0.25}{1-0.25} = \frac{1}{3} = 0.333$$

$$\frac{1}{3} \cdot 1.08 = 0.36$$

$\Rightarrow \Delta\text{educ} = 1$: Odds increase from 0.33 to 0.36 in favor of public transportation

Approximation

$$(e^{\beta_k} - 1) \approx \beta_k \quad \text{if} \quad |\beta_k| < 0.1$$

Standardized Factor Change

- To make effects comparable, it is sometimes sensible to weight them by the standard deviations of the X.

Standardized Factor Change

$$\alpha_k^{s_k} = e^{\beta_k \cdot s_k}$$

where

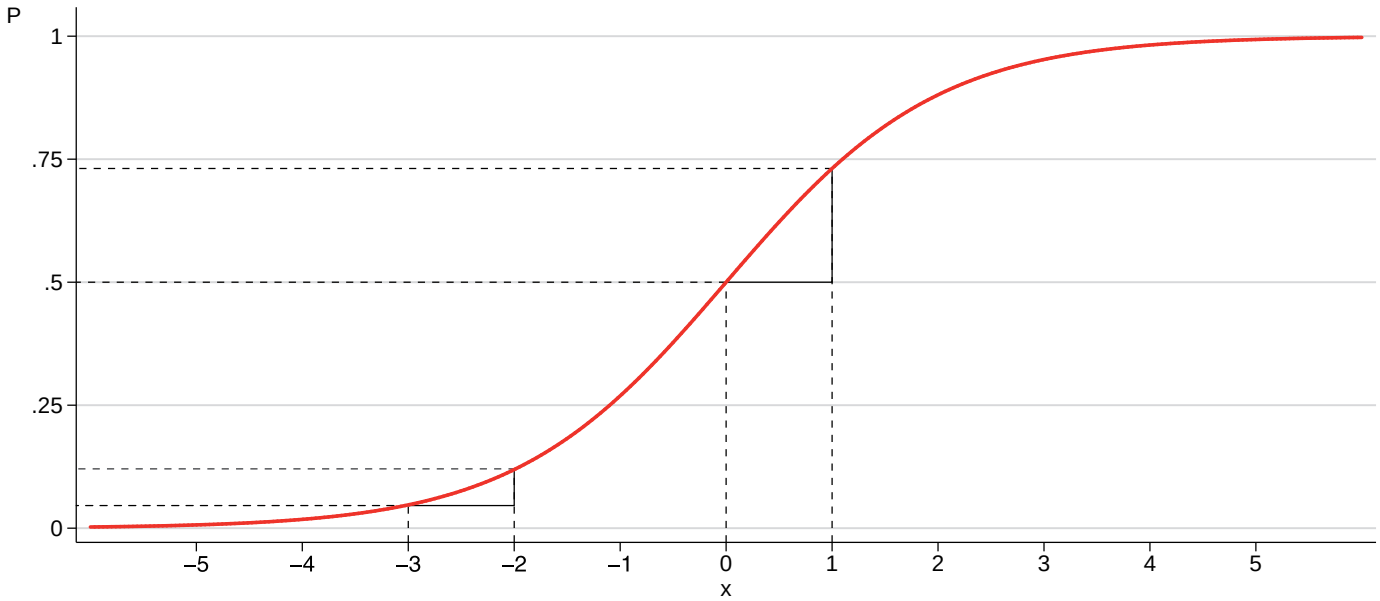
$$s_k = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

is the standard deviation of X_k

- Interpretation: Effect of a standard deviation increase in X on the odds $P/(1-P)$
- The procedure makes not much sense for binary predictors, since in this case the standard deviation has not much meaning.

Marginal/Partial Effect

- Odds may be more intuitive then Log-Odds, but what we really are interested in are the effects on the probability $P(Y = 1)$.
- Unfortunately $P(Y = 1)$ is a non-linear function of X so that the effect on $P(Y = 1)$ not only depends on the amount of change in X, but also on the level of X at which the change occurs.



- A first step in the direction of interpreting effects on the probability scale is to compute the first derivative of the function at different positions.

Marginal/Partial Effect

- In linear regression: $\frac{\partial Y}{\partial X_k} = \beta_k$
- In logistic regression:

$$P = \frac{1}{1 + e^{-[\beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m]}} = \frac{1}{1 + e^{-Z}} = \left(1 + e^{-Z}\right)^{-1}$$

$$\frac{\partial P}{\partial X_k} = -\frac{1}{(1 + e^{-Z})^2} \cdot (-\beta_k) e^{-Z}$$

$$= \beta_k \cdot \underbrace{\frac{1}{1 + e^{-Z}}}_{P} \cdot \underbrace{\frac{e^{-Z}}{1 + e^{-Z}}}_{1-P} = \beta_k \cdot P(1 - P).$$

Marginal Effect in Logistic Regression

$$\frac{\partial P}{\partial X_k} = \beta_k \cdot P(1 - P)$$

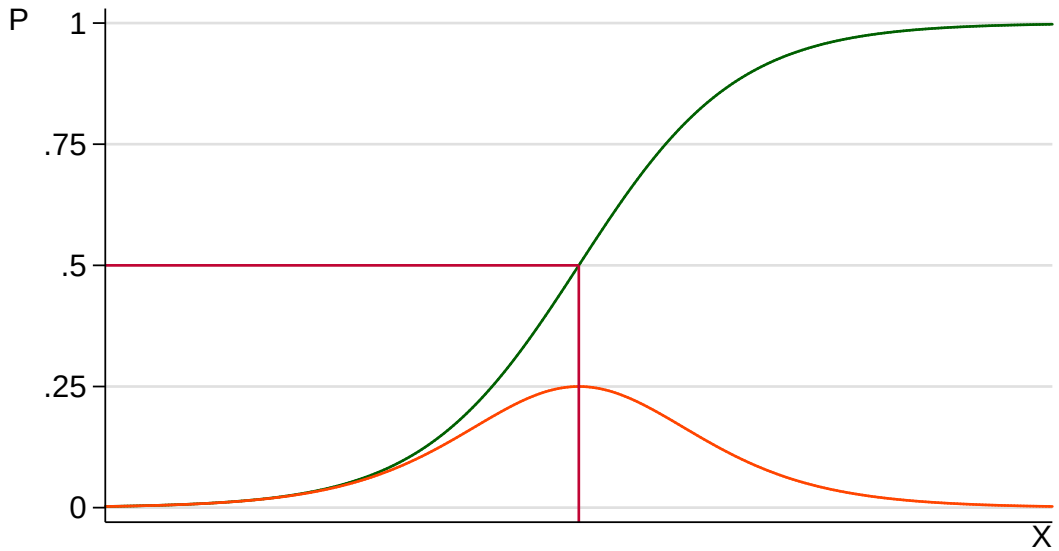
Marginal/Partial Effect

- Maximum marginal effect at $P = 0.5$ (the inflection point of the probability curve)

$$\beta_k \cdot \frac{1}{2}\left(1 - \frac{1}{2}\right) = \beta_k \cdot \frac{1}{4}$$

Divide-by-four rule

The maximum marginal effect of X_k is equal to $\beta_k/4$.



Marginal/Partial Effect

- Relative magnitudes of marginal effect for two variables

$$\frac{\partial P/\partial X_k}{\partial P/\partial X_j} = \frac{\beta_k \cdot P(1-P)}{\beta_j \cdot P(1-P)} = \frac{\beta_k}{\beta_j}$$

The ratio of coefficients reflects the relative magnitudes of the marginal effect on $\Pr(Y = 1)$.

Marginal/Partial Effect

- Average marginal effect: The value of the marginal effect differs depending on the values of the predictors. There are two main methods to compute the "average" marginal effect based on the observed sample.
  1. Marginal effect at the mean of the predictors (marginal effect for the "average" individual)

$$\Pr(Y = 1|\bar{X}) = \frac{1}{1 + e^{-[\beta_0 + \beta_1 \bar{X}_1 + \cdot + \beta_m \bar{X}_m]}}$$

$$\frac{\partial \Pr(Y = 1|\bar{X})}{\partial X_k} = \beta_k \cdot \Pr(Y = 1|\bar{X})(1 - \Pr(Y = 1|\bar{X}))$$

  2. Average of the marginal effect over all observations (sample average of individual marginal effects).

$$\frac{1}{N} \sum_{i=1}^{N} \frac{\partial \Pr(Y = 1|X_i)}{\partial X_k} = \frac{1}{N} \sum_{i=1}^{N} \beta_k \cdot \Pr(Y = 1|X_i)(1 - \Pr(Y = 1|X_i))$$

Marginal/Partial Effect

- Marginal effect at the mean: Example

		sex = 0	sex = 1
car	0	4	1
	1	3	4

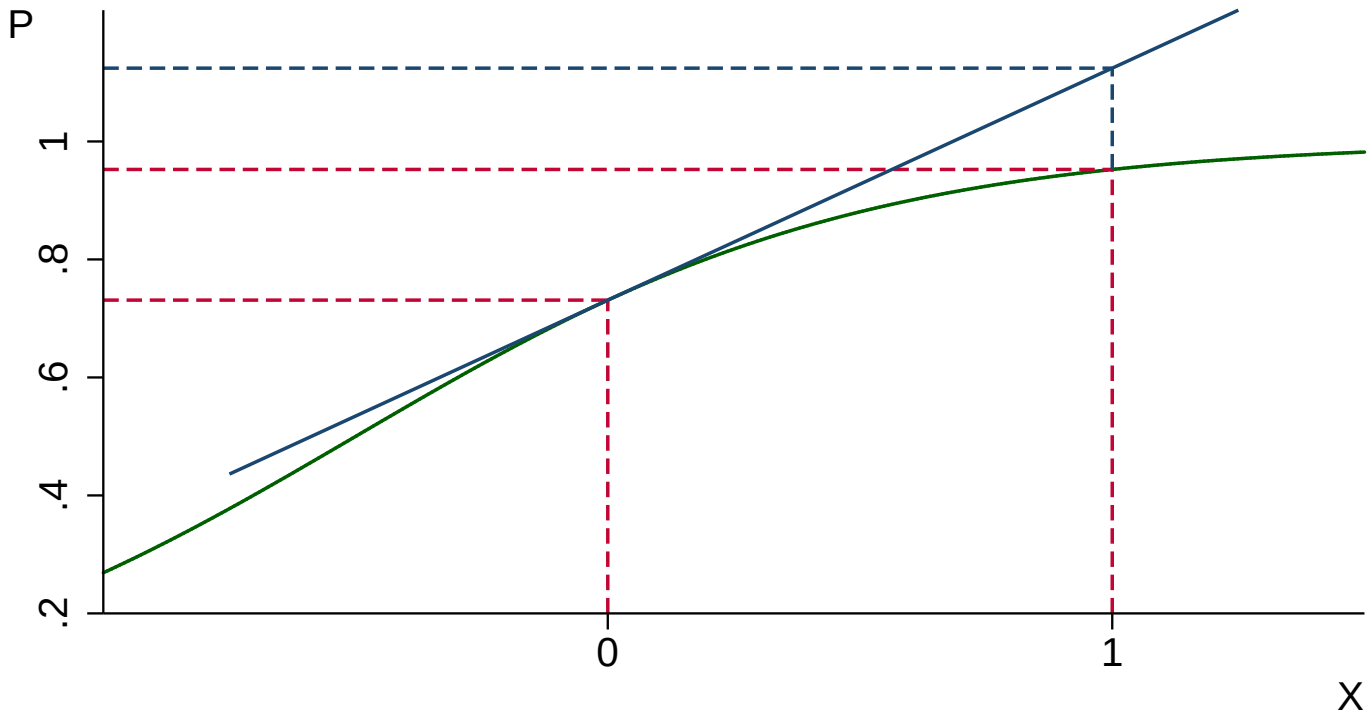
$$\overline{\texttt{sex}} = 5/12 = 0.416$$

$$\Pr(\texttt{car} = 1|\overline{\texttt{sex}}) = \frac{1}{1 + e^{-[-0.2876+1.6739 \cdot 0.416]}} = 0.601$$

$$\frac{\partial \Pr(\texttt{car} = 1|\overline{\texttt{sex}})}{\partial \texttt{sex}} = 1.6739 \cdot 0.601(1 - 0.601) = 0.40$$

Marginal/Partial Effect: Problems

- Marginal effects at the mean of the predictors often do not make much sense. For binary variables, as in the example above, the mean does not correspond to an observable value. In general, $\bar{X}$ may not be a good description of the "typical" or "average" observation.
- Marginal effects are often only crude approximations of the "real" effects on the probability (especially for binary predictors).



Elasticity

- The elasticity, which is closely related to the marginal effect, can be computed if the predictor has **ratio scale** level (metric variable with a natural zero-point).

Elasticity in logistic regression

$$\frac{\partial P/P}{\partial X_k/X_k} = \frac{\partial P}{\partial X_k}\frac{X_k}{P} = \beta_k X_k(1-P)$$

- Example: In traffic mode choice, let X_k be the price or travel duration. The elasticity can then (approximately) be interpreted as the percent-percent effect $\Rightarrow$ percent change of P if X_k is increased by one percent.

Semi-Elasticity

- Since P is on an absolute scale in the $[0, 1]$ interval, the semi-elasticity is usually better interpretable.

Semi-elasticity in logistic regression

$$\frac{\partial P}{\partial X_k / X_k} = \frac{\partial P}{\partial X_k} X_k = \beta_k X_k P(1 - P)$$

- Interpretation: How much does P change (approximately) if X_k is increased by one percent.
- Example: If X_k is the price of a bus ticket and the semi-elasticity is -0.02, then the probability to use the bus decreases by two percentage points if the price is increased by one percent.
- Note: Similar to marginal effects, elasticities and semi-elasticities depend on the values of the predictors in logistic regression.

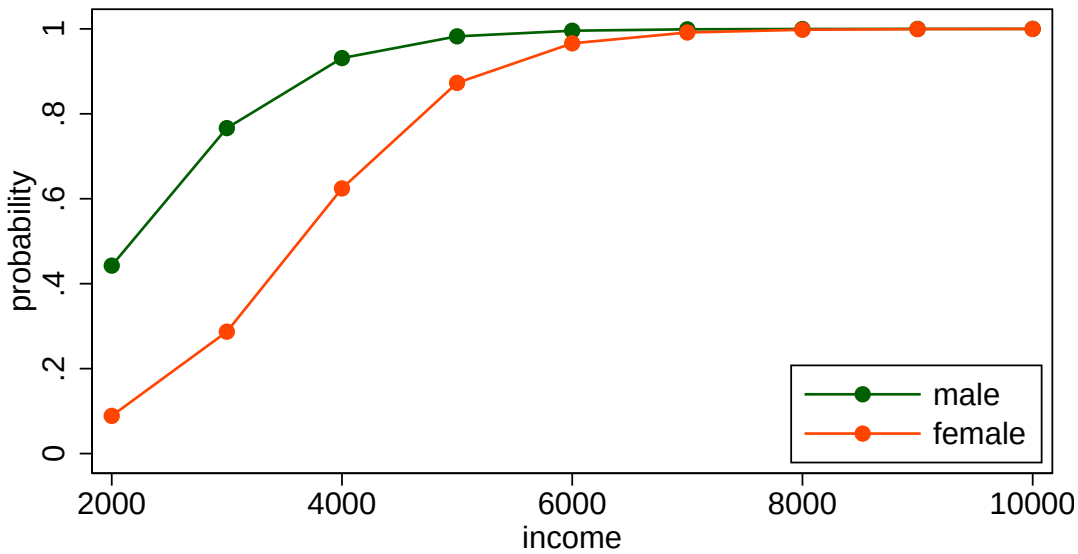
Predictions

- The most direct approach to interpret the effects of the covariates on the probability is to compute the predicted probabilities for different sets of covariate values.

Prediction given vector X_i

$$\hat{P}_i = \widehat{\Pr}(Y = 1|X_i) = \frac{1}{1 + e^{-[\beta_0 + \beta_1 X_{i1} + \cdot + \beta_m X_{im}]}}$$

- The predictions can then be reported in tables or in plots.
- Example: Predicted probability of driving a car depending on sex and age.



Discrete Change Effect

- Based on predicted probabilities, the effect of a discrete change in an independent variable, holding all other covariates constant, can be computed.

Discrete Change Effect

$$\frac{\Delta \Pr(Y = 1|X)}{\Delta X_k} = \Pr(Y = 1|X, X_k + \delta) - \Pr(Y = 1|X, X_k)$$

- Unit change effect: Effect of a one unit change in X_k (Petersen 1985)

Unit Change Effect

$$\frac{\Delta \Pr(Y = 1|X)}{\Delta X_k} = \Pr(Y = 1|X, X_k + 1) - \Pr(Y = 1|X, X_k)$$

- As outlined above, the partial change (marginal) effect does not equal the unit change effect in the logit model.

Discrete Change Effect

- Variants of discrete change effects

0-1 change (for binary covariates)

$$\Pr(Y = 1|X, X_k = 1) - \Pr(Y = 1|X, X_k = 0)$$

centered unit change

$$\Pr(Y = 1|X, \bar{X}_k + 0.5) - \Pr(Y = 1|X, \bar{X}_k - 0.5)$$

standard deviation change

$$\Pr(Y = 1|X, \bar{X}_k + 0.5s_k) - \Pr(Y = 1|X, \bar{X}_k - 0.5s_k)$$

minimum-maximum change

$$\Pr(Y = 1|X, X_k = X_k^{\max}) - \Pr(Y = 1|X, X_k = X_k^{\min})$$

Discrete Change Effect

- Simple example

$$\Pr(\texttt{car} = 1|\texttt{sex}) = \frac{1}{1 + e^{-[-0.2876+1.6739\cdot\texttt{sex}]}}$$

$$\texttt{sex} = 0: \quad P_0 = \frac{1}{1 + e^{-[-0.2876]}} = 0.4286$$

$$\texttt{sex} = 1: \quad P_1 = \frac{1}{1 + e^{-[-0.2876+1.6739]}} = 0.80$$

$$\Rightarrow \quad \Delta P = 0.80 - 0.4286 = 0.37$$

Discrete Change Effect

- Values of the other covariates? $\Rightarrow$ usually set to their mean

$$\Pr(Y = 1|\bar{X}, X_k + \delta) - \Pr(Y = 1|\bar{X}, X_k)$$

- In some cases, it might be more reasonable to use the median or the mode for selected variables.
- If the model contains categorical variables (e.g. sex), it can also be illustrative to compute separate sets of discrete change effects for the different groups.
- Furthermore, it can also be sensible to compute the "average" discrete change effect over the sample:

$$\frac{1}{N}\sum_{i=1}^{N}(\Pr(Y = 1|X_i, X_{ik} + \delta) - \Pr(Y = 1|X_i, X_{ik}))$$

Relative Risks

- In cases where $\Pr(Y = 1)$ is very low (e.g. accident statistics), it can be reasonable to express effects in terms of relative risks.
- Example: Parachute Type A versus Type B

$$P_A = 4 \cdot 10^{-6}, \quad P_B = 2 \cdot 10^{-6}$$

$\Rightarrow$ discrete change effect: $P_A - P_B = 0.000002$

$\Rightarrow$ relative risks: $R = P_A / P_B = 2$

(For fun see Smith and Pell, BMJ 2003, who complain about the lack of randomized controlled trials on parachute effectiveness.)

Part VI

Logistic Regression: Estimation

- Maximum Likelihood Estimation
- MLE for Linear Regression
- MLE for Logit
- Large Sample Properties of MLE
- Summary

Maximum Likelihood Estimation (MLE)

Although it would be possible to use the (weighted nonlinear) least squares method, a Logit model is typically estimated by the maximum likelihood method.

General estimation principle:

- Least Squares (LS): minimize the squared deviations between observed data and predictions
- Maximum Likelihood (ML): maximize the likelihood (i.e. the probability) of the observed data given the estimate

MLE – Example: Proportion of men

- If π is the proportion of men in the population, then the probability of having k men in a sample of size N is

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

(k follows a binomial distribution). Assumption: independent sampling with identical probability (i.i.d.)

- Assume $k = 2$ and $N = 6$. We can now ask: What value for π makes this outcome most likely? The answer is found by maximizing the **likelihood function**

$$L(\pi|k, N) = \binom{N}{k} \pi^k (1-\pi)^{N-k}$$

with respect to π. $\Rightarrow$ Choose π so that the first derivative (gradient) is zero: $\partial L(\pi|k, N)/\partial\pi = 0$

MLE – Example: Proportion of men

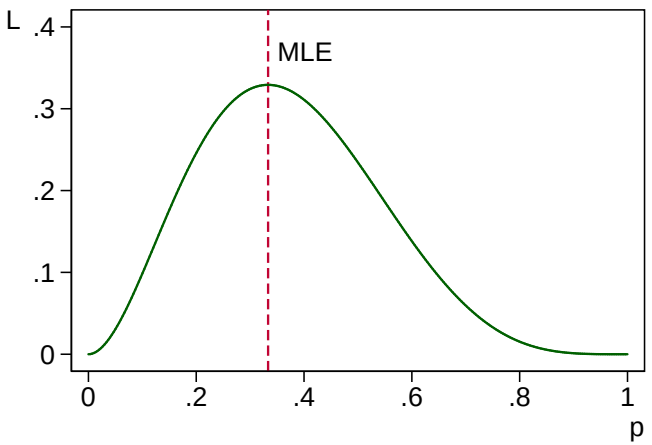
- Instead of maximizing $L(\pi|k,N)$ directly, we can also maximize the logarithm of $L(\pi|k,N)$, which is generally easier:

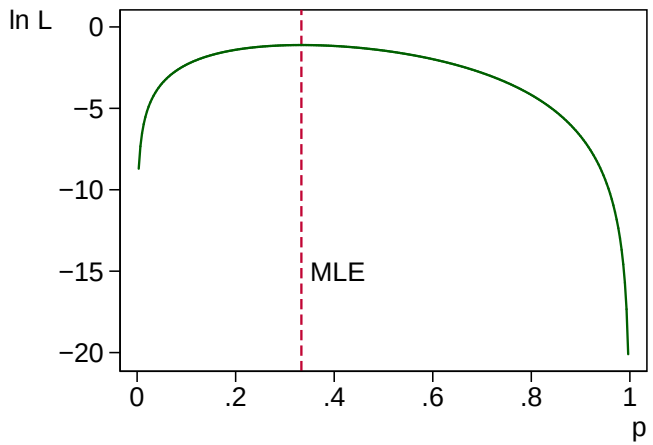
$$\ln L(\pi|k,N) = \ln\binom{N}{k} + k\ln(\pi) + (N-k)\ln(1-\pi)$$

$$\frac{\partial \ln L(\pi|k,N)}{\partial \pi} = 0 + \frac{\partial k\ln(\pi)}{\partial \pi} + \frac{\partial (N-k)\ln(1-\pi)}{\partial \pi}$$

$$= \frac{k}{\pi} + \frac{\partial (N-k)\ln(1-\pi)}{\partial (1-\pi)}\frac{\partial (1-\pi)}{\partial \pi} = \frac{k}{\pi} - \frac{N-k}{1-\pi}$$

$$\text{set } \frac{\partial \ln L}{\partial \pi} = 0 \quad \Rightarrow \quad \hat{\pi} = \frac{k}{N} = \frac{2}{6} = \frac{1}{3}$$





MLE for Linear Regression

- Model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_m X_{im} + \epsilon_i = X_i'\beta + \epsilon_i$$
$$\epsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma) \quad \Rightarrow \quad Y_i \overset{\text{i.i.d}}{\sim} N(X_i'\beta, \sigma)$$

- The probability density function for Y_i is

$$f(Y_i|X_i'\beta, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_i - X_i'\beta)^2} = \frac{1}{\sigma}\phi\left(\frac{Y_i - X_i'\beta}{\sigma}\right)$$

  where $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ is the standard normal density, so that

$$L(\beta, \sigma|Y, X) = \prod_{i=1}^{N} \frac{1}{\sigma}\phi\left(\frac{Y_i - X_i'\beta}{\sigma}\right)$$

- The $\hat{\beta}$ that maximizes L also minimizes $\sum(Y_i - X_i'\beta)^2 \Rightarrow$ MLE = OLS in this case (but only if ϵ is assumed to be normally distributed)

MLE for Logit

- Let $P_i = \Pr(Y_i = 1) = \left(1 + e^{-(X_i'\beta)}\right)^{-1}$. The likelihood of a specific observation can then be written as

$$\Pr(Y_i|X_i,\beta) = P_i^{Y_i}(1 - P_i)^{1-Y_i} = \begin{cases} P_i & \text{if } Y_i = 1 \\ 1 - P_i & \text{if } Y_i = 0 \end{cases}$$

and the likelihood and log likelihood functions are

$$L(\beta|Y,X) = \prod_{i=1}^{N} P_i^{Y_i}(1 - P_i)^{1-Y_i}$$

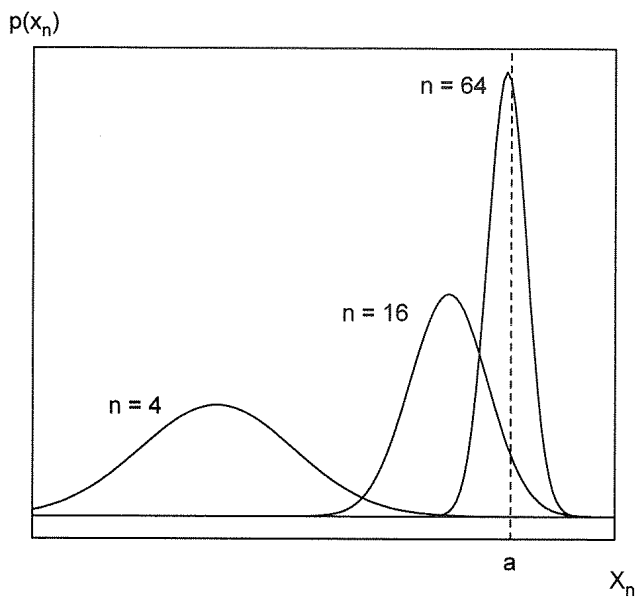
$$\begin{aligned} \ln L(\beta|Y,X) &= \sum_{i=1}^{N} [Y_i \ln P_i + (1 - Y_i) \ln(1 - P_i)] \\ &= \sum_{i=1}^{N} Y_i X_i'\beta - \sum_{i=1}^{N} \ln\left(1 + e^{X_i'\beta}\right) \end{aligned}$$

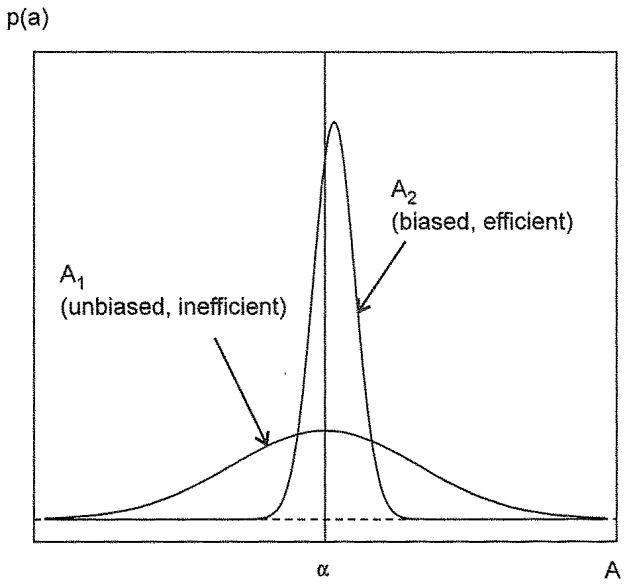
MLE for Logit

- Usually no closed form solution exists for the maximum of $\ln L$ so that numerical optimization methods are used (e.g. the Newton-Raphson method). The general procedure is to start with an initial guess for the parameters and then iteratively improve on that guess until approximation is "good enough".
- Care has to be taken if the (log) likelihood function (surface) is not globally concave, i.e. if local maxima exist.
- Luckily, the Logit model has a globally concave log likelihood function, so that a single global maximum exists and the solution is independent of the starting values.

Large Sample Properties of MLE

- **Consistency:** asymptotically unbiased (expected value of MLE approaches true value with increasing N)
- **Efficiency:** asymptotically minimal sampling variance
- **Normality:** estimates are asymptotically normally distributed (MLE are BAN, Best Asymtotic Normal estimator)
  $\Rightarrow$ statistical inference can be based on normal theory (e.g. Wald test for single coefficients, likelihood ratio test for nested models)





(Fox 1997)

MLE: summary

Benefits of MLE

- very flexible and can handle many kinds of models
- desirable large sample properties

Drawbacks of MLE

- may be biased and inefficient in small samples (MLE-Logit with $N < 100$ not recommended; depends on model complexity; see e.g. Peduzzi et al. 1996, Bagley et al. 2001)

  Example: MLE estimate of variance of x is $1/N \sum(x_i - \bar{x})^2$, unbiased estimate is $1/(N-1) \sum(x_i - \bar{x})^2$
- requires distributional assumptions
- generally no closed form solutions (but computers do the job)
- numerical algorithms may not converge in some cases (e.g. in Logit if there is perfect classification so that $|\beta| \to \infty$)

Part VII

Logistic Regression: Inference

- MLE and Statistical Inference
- Significance Tests and Confidence Intervals
- Likelihood Ratio Tests
- Wald Test

MLE and Statistical Inference

MLE theory shows:

- The sampling distribution of ML parameter estimates is asymptotically normal.

  $\Rightarrow$ Therefore, statistical tests and confidence intervals can be based on an estimate for the variance of the sampling distribution.

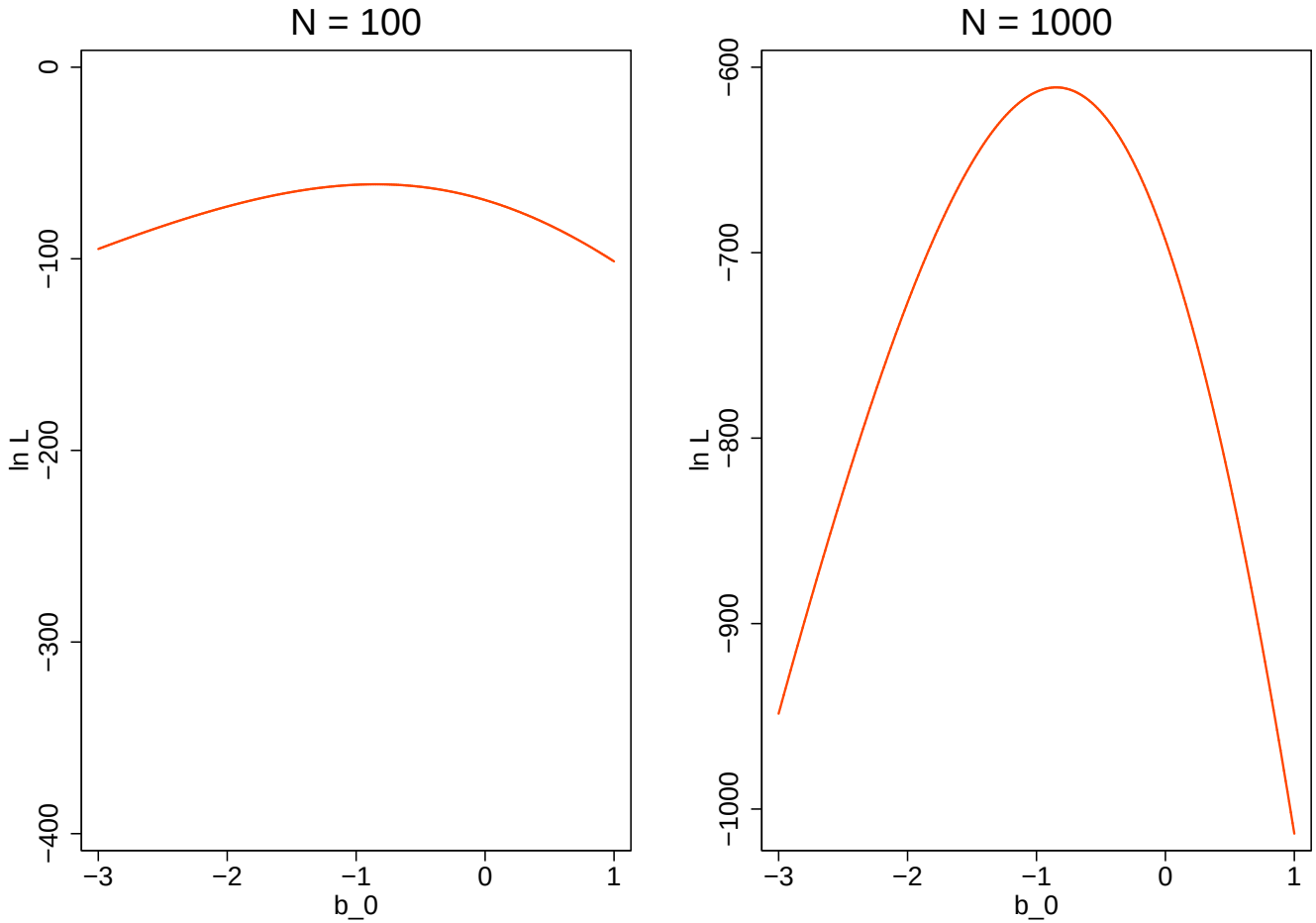
- The variance-covariance matrix of an ML estimator for a parameter vector θ is given as the negative of the inverse of the expected value of the matrix of second derivatives of the log likelihood function.
  (The matrix of second derivatives of $\ln L(\theta)$ is called the *Hessian*. The negative of the expectation of the Hessian is called the *information matrix*.)

  Intuitive explanation: The second derivative indicates the curvature of log likelihood function. If the function is flat, then there is much uncertainty in the estimate. Variance reflects uncertainty.

- Cautionary note: Results are only asymptotic, large N required ($N > 100$)

MLE and Statistical Inference

- Curvature of the log likelihood function and sample size:



Model: $\Pr(Y = 1) = \frac{1}{1+e^{-\beta_0}}$

MLE and Statistical Inference

Example: Logit model with only a constant β_0, i.e. $\pi = \Pr(Y = 1)$ is a constant.

- MLE of π: $\hat{\pi} = \frac{k}{N}$, where k is the observed number of events
- *What are the variance and standard deviation (= standard error) of the sampling distribution of $\hat{\pi}$?*

$$L(\pi|k, N) = \binom{N}{k} \pi^k (1-\pi)^{N-k}$$

$$\ln L(\pi|k, N) = \ln \binom{N}{k} + k \ln(\pi) + (N-k) \ln(1-\pi)$$

$$\ln L' = \frac{\partial \ln L}{\partial \pi} = \frac{k}{\pi} - \frac{N-k}{1-\pi} = k\pi^{-1} - (N-k)(1-\pi)^{-1} \quad \Rightarrow \hat{\pi} = \frac{k}{N}$$

$$\ln L'' = \frac{\partial^2 \ln L}{\partial \pi^2} = -k\pi^{-2} - (N-k)(1-\pi)^{-2} = -\left[\frac{k}{\pi^2} + \frac{N-k}{(1-\pi)^2}\right]$$

Note: $\ln L'' < 0$ for all π (concave), i.e. $\ln L(\hat{\pi})$ is maximum.

MLE and Statistical Inference

$$\ln L'' = -\left[\frac{k}{\pi^2} + \frac{N-k}{(1-\pi)^2}\right] = -\frac{(1-\pi)^2 k + p^2(N-k)}{\pi^2(1-\pi)^2} \cdot \frac{N}{N}$$
$$= -\frac{\pi^2 - 2\pi\frac{k}{N} + \frac{k}{N}}{\frac{1}{N}\pi^2(1-\pi)^2}$$

$$E[k/N] = \pi$$

$$E[\ln L''] = -\frac{\pi^2 - 2\pi^2 + \pi}{\frac{1}{N}\pi^2(1-\pi)^2} = -\frac{\pi - \pi^2}{\frac{1}{N}\pi^2(1-\pi)^2} = -\frac{\pi(1-\pi)}{\frac{1}{N}\pi^2(1-\pi)^2}$$
$$= -\frac{1}{\frac{1}{N}\pi(1-\pi)}$$

MLE and Statistical Inference

Variance: take negative of inverse of $E[\ln L'']$

$$V(\hat{\pi}) = -\frac{1}{E[\ln L'']} = \frac{\pi(1-\pi)}{N}$$

Variance estimate: plug in estimate for π

$$\hat{V}(\hat{\pi}) = \frac{\hat{\pi}(1-\hat{\pi})}{N} = \frac{\frac{k}{N}\left(1-\frac{k}{N}\right)}{N} \qquad \widehat{SE}(\hat{\pi}) = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}}$$

$\Rightarrow (1-\alpha)$ confidence interval for $\hat{\pi}$

$$\hat{\pi} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal (e.g. $z_{0.975} = 1.96$ for the 95% CI)

MLE and Statistical Inference

In general for $\beta = [\beta_1, \beta_2, \ldots, \beta_m]^T$

Hessian

$$H = \frac{\partial^2 \ln L}{\partial \beta^2} = \begin{bmatrix} \frac{\partial^2 \ln L}{\partial \beta_1 \partial \beta_1} & \frac{\partial^2 \ln L}{\partial \beta_1 \partial \beta_2} & \cdots & \frac{\partial^2 \ln L}{\partial \beta_1 \partial \beta_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \ln L}{\partial \beta_m \partial \beta_1} & \frac{\partial^2 \ln L}{\partial \beta_m \partial \beta_2} & \cdots & \frac{\partial^2 \ln L}{\partial \beta_m \partial \beta_m} \end{bmatrix}$$

MLE Variance-Covariance Matrix

$$V(\hat{\beta}) = \begin{bmatrix} V(\hat{\beta}_1) & V(\hat{\beta}_1, \hat{\beta}_2) & \cdots & V(\hat{\beta}_1, \hat{\beta}_m) \\ \vdots & \vdots & \ddots & \vdots \\ V(\hat{\beta}_m, \hat{\beta}_1) & V(\hat{\beta}_m, \hat{\beta}_2) & \cdots & V(\hat{\beta}_m) \end{bmatrix} = -\frac{1}{E[H]}$$

Significance Test for a Single Regressor

- According to ML theory, parameter estimate β_k is asymptotically normal, i.e.

$$\hat{\beta}_k \overset{a}{\sim} N(\beta_k, V(\hat{\beta}_k))$$

- Using the variance estimate $\hat{V}(\hat{\beta}_k)$ we can therefore construct a significance test for β_k in the usual way.
- Let the null hypothesis be $H_0 : \beta_k = \beta_k^0$. The test statistic then is

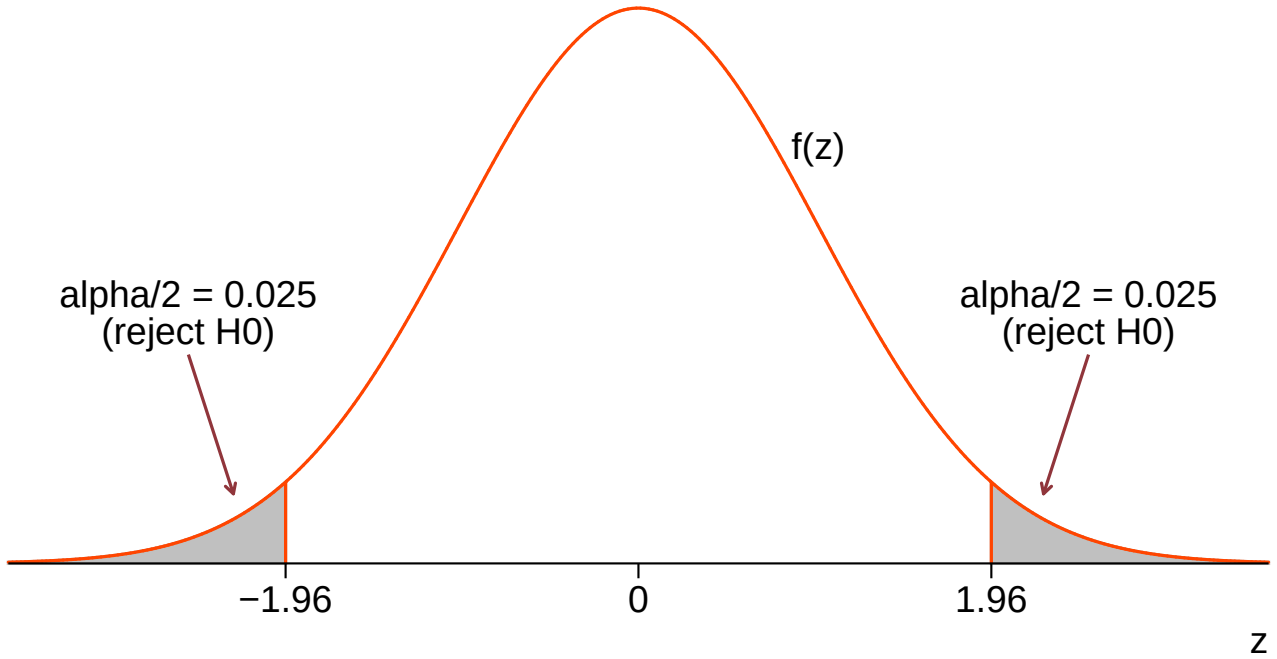
$$Z = \frac{\hat{\beta}_k - \beta_k^0}{SE(\hat{\beta}_k)} \quad \text{with} \quad SE(\hat{\beta}_k) = \sqrt{\hat{V}(\hat{\beta}_k)}$$

  (SE = standard error).
- The null hypothesis is rejected on significance level α if $|Z| > z_{1-\alpha/2}$ where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$-quantile of the standard normal distribution.

Significance Test for a Single Regressor

Usually: Test against zero on a 5% level

- $H_0 : \beta_k = 0$ (i.e. $\beta_k^0 = 0$)
- 5% significance level ($\alpha = 0.05$)
- Test statistic: $Z = \dfrac{\hat{\beta}_k}{SE(\hat{\beta}_k)} \overset{a}{\sim} N(0,1)$
- reject H_0 if $|Z| > 1.96$



Confidence Interval for a Single Regressor

$(1-\alpha)$-Confidence Interval

$$\hat{\beta}_k \pm z_{1-\alpha/2} \cdot SE(\hat{\beta}_k)$$

Usually: 95%-Confidence Interval

$$\left[\hat{\beta}_k - 1.96 \cdot SE(\hat{\beta}_k)\ ,\ \hat{\beta}_k + 1.96 \cdot SE(\hat{\beta}_k)\right]$$

Likelihood Ratio Tests

- The ratios of the likelihoods of "nested" models can be used to perform significance tests for general hypotheses.
- Model A is "nested" in model B if it can be formed by imposing constraints to model B. Example:

$$M_1\text{: logit}[P(Y=1)] = \beta_0 + \beta_1 X_1$$
$$M_2\text{: logit}[P(Y=1)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

  Model M_1 is equal to model M_2 with constraint $\beta_2 = 0$. Therefore, M_1 is nested in M_2.
- The log likelihood of the constrained model M_1 cannot be larger than the log likelihood of the unconstrained model M_2:

$$\ln L(M_1) \leq \ln L(M_2) \quad \text{or} \quad \ln L(M_2) - \ln L(M_1) \geq 0$$

- The approach can be used to test hypotheses involving multiple parameters, e.g. a hypothesis that $\beta_1 = \beta_2 = \beta_3 = 0$ or a hypothesis that $\beta_1 \geq \beta_2$.

Likelihood Ratio Tests

Likelihood Ratio Test

General result: The likelihood ratio statistic

$$LR = G^2 = -2 \ln \left(\frac{L(M_C)}{L(M_U)} \right) = 2 \ln L(M_U) - 2 \ln L(M_C)$$

where M_C is the constrained (restricted, nested, null) model and M_U is the unconstrained (full) model, is asymptotically chi-square distributed with degrees of freedom equal to the number of independent constraints.

- Example applications:
  - overall LR test of the full model
  - LR test of a single coefficient
  - LR test of a subset of coefficients
  - LR test of equality of two coefficients

Overall LR Test of Full Model

Is the model significant at all? Does the model "explain" anything?

- Null hypothesis:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_m = 0$$

- Models:

$$M_C: \text{logit}[P(Y=1)] = \beta_0$$
$$M_U: \text{logit}[P(Y=1)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m$$

- Test statistic:

$$LR = 2 \ln L(M_U) - 2 \ln L(M_C) \overset{a}{\sim} \chi^2(m)$$

- Reject H_0 at the α level if $LR > \chi^2_{1-\alpha}(m)$

[Note: The log likelihood of the null model in this case is $\ln L(M_C) = \ln \binom{N}{k} + k \ln(k/N) + (N-k) \ln(1-k/N)$.]

LR Test of a Single Coefficient

Is coefficient β_k significant?

- Null hypothesis:

$$H_0: \beta_k = 0$$

- Models:

$$M_C: Z = \beta_0 + \cdots + \beta_{k-1} X_{k-1} + \qquad \beta_{k+1} X_{k+1} \cdots + \beta_m X_m$$
$$M_U: Z = \beta_0 + \cdots + \beta_{k-1} X_{k-1} + \beta_k X_k + \beta_{k+1} X_{k+1} \cdots + \beta_m X_m$$

- Test statistic:

$$LR = 2 \ln L(M_U) - 2 \ln L(M_C) \quad \overset{a}{\sim} \quad \chi^2(1)$$

- Reject H_0 at the α level if $LR > \chi^2_{1-\alpha}(1)$

The LR test of a single coefficient is asymptotically equivalent to the test based on the Hessian discussed above.

LR Test of a Subset Of Coefficients

Is any coefficient in a set of coefficients different from zero? Does the set of coefficients jointly "explain" anything?

Note: This is different than testing each parameter separately!

- Null hypothesis:

$$H_0: \beta_k = \cdots = \beta_l = 0$$

- Models:

$$M_C: Z = \beta_0 + \cdots + \qquad\qquad \cdots + \beta_m X_m$$
$$M_U: Z = \beta_0 + \cdots + \beta_k X_k + \cdots + \beta_l X_l + \cdots + \beta_m X_m$$

- Test statistic:

$$LR = 2 \ln L(M_U) - 2 \ln L(M_C) \quad \overset{a}{\sim} \quad \chi^2(l - k + 1)$$

- Reject H_0 at the α level if $LR > \chi^2_{1-\alpha}(l - k + 1)$

LR Test of Equality of Two Coefficients

Is the difference between two coefficients significant?

- Null hypothesis:

$$H_0: \beta_k = \beta_l$$

- Models:

$$M_C: Z = \beta_0 + \cdots + \gamma(X_k + X_l) + \cdots + \beta_m X_m$$
$$M_U: Z = \beta_0 + \cdots + \beta_k X_k + \beta_l X_l + \cdots + \beta_m X_m$$

- Test statistic:

$$LR = 2 \ln L(M_U) - 2 \ln L(M_C) \quad \overset{a}{\sim} \quad \chi^2(1)$$

- Reject H_0 at the α level if $LR > \chi^2_{1-\alpha}(1)$

The trick is to use the sum of the two variables in the restricted model.

See Jann (2005) and Gelman and Stern (2006) for some general comments on interpreting differences between coefficients.

Wald Test

- An alternative, asymptotically equivalent approach to testing the difference between nested models is the Wald test, which is based on the variance-covariance matrix of the estimates.
- Example: For the test of the null hypothesis that β_1 is equal to β_2, the Wald statistic is

$$W = \frac{(\hat{\beta}_1 - \hat{\beta}_2)^2}{\hat{V}(\hat{\beta}_1) + \hat{V}(\hat{\beta}_2) + \widehat{Cov}(\hat{\beta}_1, \hat{\beta}_2)} \quad \overset{a}{\sim} \quad \chi^2(1)$$

- See, e.g., Long (1997: 89pp.) for the general form of the test and further details. The Wald test is analogous to the generalized *F*-test in linear regression (Greene 2003:95pp.).
- The advantage of the Wald test over the LR test is that only the full model has to be estimated.
- Some people prefer the *LR* test over the Wald test (especially in moderate samples) because there is (weak) evidence that it is more efficient and behaves less erratic. But note that both procedures can be quite off in small samples.

Relation between LR test and Wald test

- LR test: Difference between $\ln L(\hat{\beta})$ and $\ln L(\beta^0)$
- Wald test: Difference between $\hat{\beta}$ and β^0 weighted by curvature of log likelihood, $\partial^2 \ln L / \partial \beta^2$

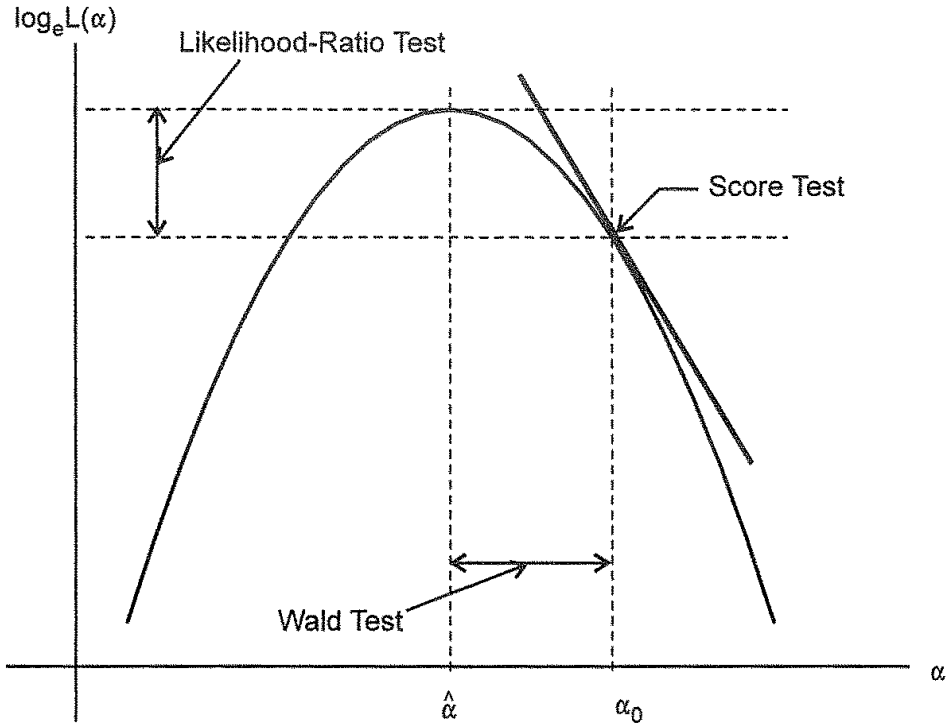


Figure D.16. Tests of the hypothesis H_0: $\alpha = \alpha_0$: The likelihood-ratio test compares $\log_e L(\hat{\alpha})$ with $\log_e L(\alpha_0)$; the Wald test compares $\hat{\alpha}$ with α_0; and the score test examines the slope of $\log_e L(\alpha)$ at $\alpha = \alpha_0$.

(Fox 1997)

Significance Test of a Single Regressor in SPSS

- SPSS reports a Wald statistic with one degree of freedom for the individual tests of the single regressors.
- The Wald statistic is equivalent to the square of Z of the usual significance test discussed above in this case.

$$W = \left(\frac{\hat{\beta}_k}{SE(\hat{\beta}_k)} \right)^2 \quad \overset{a}{\sim} \quad \chi^2(1)$$

- Reject H_0: $\beta_k = 0$ if W is larger than the $(1 - \alpha)$ quantile of the chi-squared distribution with 1 degree of freedom.
- Note: The square of a standard normally distributed variable is chi-square distributed with 1 degree of freedom.

Part VIII

Logistic Regression: Diagnostics and Goodness-of-Fit

- Pearson Chi-Square and Deviance
- Residuals and Influence
- Classification Table
- Goodness-of-Fit Measures

Pearson Chi-Square and Deviance

- Two concepts to measure discrepancy between model and data
  - Pearson statistic: difference between observed data and predicted probabilities
  - Deviance: difference between the likelihood of a "saturated" model (= perfect model) and the likelihood of the fitted model
- To formalize the concepts it is useful to think in "covariate patterns" (distinct patterns of values in X) rather then observations.
- Notation
  - J: number of covariate patterns
  - if some observations have the same X values, then $J < N$
  - if each observation has its own unique covariate pattern, then $J = N$
  - X_j: jth covariate pattern, $j = 1, \ldots, J$
  - n_j: number of observations with covariate values equal to X_j
  - $\bar{Y}_j$: the mean of Y (proportion of ones) among the observations with with covariate pattern X_j

Pearson Chi-Square and Deviance

Pearson X^2

- Difference between observed proportions $\bar{Y}_j$ and predicted probabilities $\hat{\pi}_j = (1 + e^{-X_j'\beta})^{-1}$

$$E_j = \bar{Y}_j - \hat{\pi}_j, \quad \bar{Y}_j \in [0,1],\ \hat{\pi}_j \in [0,1] \Rightarrow E_j \in [-1,1]$$

Pearson statistic

$$X^2 = \sum_{j=1}^{J} r_j^2, \quad r_j = \sqrt{n_j}\frac{\bar{Y}_j - \hat{\pi}_j}{\sqrt{\hat{\pi}_j(1-\hat{\pi}_j)}}$$

- Interpretation: Sum of variance weighted residuals
- r_j is called the "Pearson residual"
- If $J = N$: $X^2 = \sum_{i=1}^{N} r_i^2, \quad r_i = (Y_i - \hat{\pi}_i) \Big/ \sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}$

Pearson Chi-Square and Deviance

Deviance

- Discrepancy between the log likelihood $L(\hat{\beta})$ of the estimated model and the log likelihood L_S of a "saturated" model (i.e. a model with one parameter per covariate pattern)

Deviance

$$D = -2 \ln\left(\frac{L(\hat{\beta})}{L_S}\right) = 2[\ln L_S - \ln L(\hat{\beta})]$$

- Since
$$\ln L_S = \sum_j n_j[\bar{Y}_j \ln(\bar{Y}_j) + (1-\bar{Y}_j)\ln(1-\bar{Y}_j)]$$
$$\ln L(\hat{\beta}) = \sum_j n_j[\bar{Y}_j \ln(\hat{\pi}_j) + (1-\bar{Y}_j)\ln(1-\hat{\pi}_j)]$$

$$D = \sum_{j=1}^{J} d_j^2, \quad d_j = \pm\sqrt{2n_j\left[\bar{Y}_j \ln\left(\frac{\bar{Y}_j}{\hat{\pi}_j}\right) + (1-\bar{Y}_j)\ln\left(\frac{1-\bar{Y}_j}{1-\hat{\pi}_j}\right)\right]}$$

where sign of the deviance residual d_j agrees with the sign of $\bar{Y}_j - \hat{\pi}_j$

Pearson Chi-Square and Deviance

Deviance

- If $J = N$:

$$D = -2 \ln L(\hat{\beta}) = \sum_{i=1}^{N} d_i^2$$

with

$$d_i = \pm \sqrt{-2 \left[Y_i \ln(\hat{\pi}_i) + (1 - Y_i) \ln(1 - \hat{\pi}_i) \right]}$$

since

$$\ln L_s = \sum_{i=1}^{N} \left[Y_i \ln(Y_i) + (1 - Y_i) \ln(1 - Y_i) \right] = 0$$

.

Pearson Chi-Square and Deviance

Overall goodness-of-fit test:

- Large values for X^2 or D provide evidence against the null hypothesis that the model fits the observed data.
- Think of a $J \times 2$ table with distinct covariate patterns in the rows and the values of Y (0 and 1) in the columns. X^2 and D are the Pearson statistic and the likelihood ratio statistic for the test of the difference between the observed cell frequencies and the expected cell frequencies using the fitted model.
- The null hypothesis is rejected if $X^2 > \chi^2_{1-\alpha}(J - m - 1)$ or $D > \chi^2_{1-\alpha}(J - m - 1)$, respectively, where m is the number of regressors in the model.
- However, the test is only valid in a design with many observations per covariate pattern. For example, if the model contains continuous regressors, J will increase with N and the cell counts remain small. As a consequence, X^2 and D are *not* asymptotically chi-square distributed.

Pearson Chi-Square and Deviance

Hosmer-Lemeshow test

- To avoid the problem of growing J with increasing N, Hosmer and Lemeshow proposed a test based on grouped data (1980; Lemeshow and Hosmer 1982)
  - The data are divided into g (approx.) equally sized groups based on percentiles of the predicted probabilities ($\Rightarrow g \times 2$ table)
  - In each group the expected and observed frequencies are computed for $Y = 0$ and $Y = 1$.
  - A Pearson X^2 statistic is computed from these cell frequencies in the usual manner.
  - The null hypothesis is rejected if the statistic exceeds the $(1 - \alpha)$-quantile of the χ^2 distribution with $g - 2$ degrees of freedom.
- For alternative approaches see the discussion in Hosmer and Lemeshow (2000: 147pp.).

Residuals and Influence

- A goal of regression diagnostics is to evaluate whether there are single observations that badly fit the model and exert large influence on the estimate (see, e.g., Fox 1991, Jann 2006). We have less confidence in estimates if they strongly depend on only a few influential cases.
  - Outliers: Observations for which the difference between the model prediction and the observed value is large.
  - Influence: The effect of an individual observation (or a group of observations) on the estimate.
- A popular measure is Cook's Distance as an overall measure for the influence of an observation on the estimated parameter vector.
- Outliers or influential observations can be due to data errors. But they can also be a sign of model misspecification.
- The regression diagnostics tools, which were developed for linear regression, can be translated to logistic regression (Pregibon 1981; good discussion in Hosmer and Lemeshow 2000).

Residuals and Influence

[For simplicity, assume $J = N$ for the following discussion.]

- As noted before, if Y is a binary variable, the difference between Y_i and $\pi_i = \Pr(Y_i = 1)$ is heteroscedastic:

$$V(Y_i|X_i) = V(Y_i - \pi_i) = \pi_i(1 - \pi_i)$$

- This suggests using the *Pearson residual*

$$r_i = \frac{Y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}}$$

  as a standardized measure for the discrepancy between an observation and the model prediction.

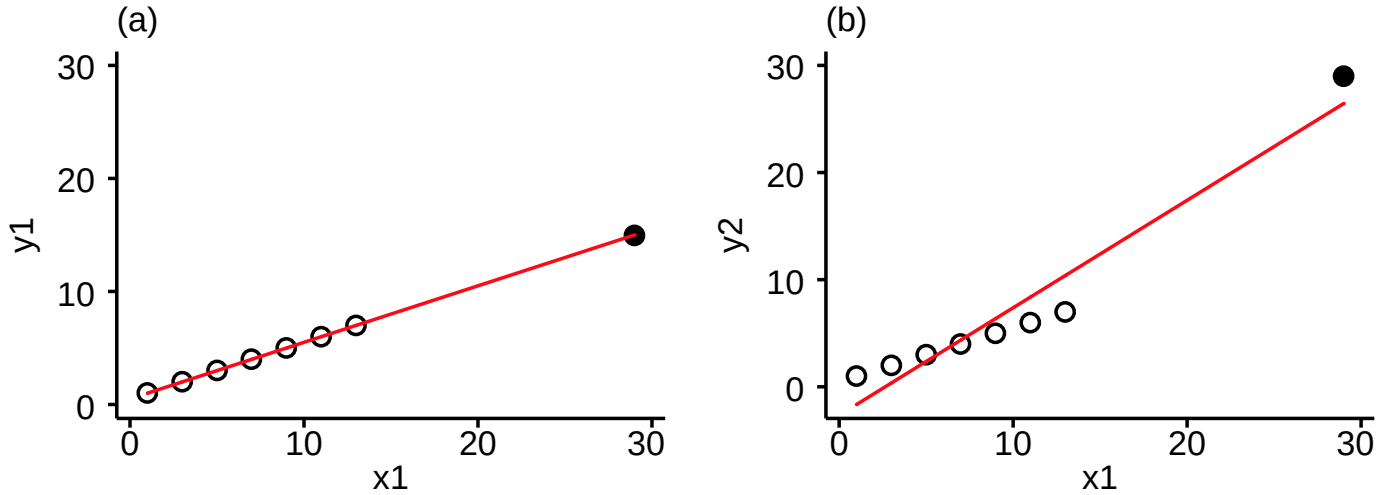
- However, because $\hat{\pi}_i$ is an estimate and the observed Y_i has influence on that estimate, $V(Y_i - \hat{\pi}_i) \neq \pi_i(1 - \pi_i)$ so that the variance of r_i is not 1 and the residuals from different observations cannot be compared.

Residuals and Influence

- An improved measure is the *standardized Pearson residual*

$$r_i^S = \frac{r_i}{\sqrt{1 - h_{ii}}}$$

  where $h_{ii} = \hat{\pi}_i(1 - \hat{\pi}_i)X_i' \hat{V}(\hat{\beta}) X_i$
- r_i^S has variance 1 and can be compared across observations. A large value of r_i^S indicates bad fit for observation i.
- h_{ii} is called the leverage or hat value and is a measure for the potential influence of an observation on the estimate. The leverage mainly depends on how "atypical" an observation's X values are.



Residuals and Influence

- Observations with a large residual and a large leverage exert strong influence on the model estimate.
- A measure to summarize this influence is Cook's Distance

$$\Delta\beta_i = \frac{r_i^2 h_{ii}}{(1-h_{ii})^2} = \frac{(r_i^S)^2 h_{ii}}{1-h_{ii}}$$

- Other popular measures are the influence on the Pearson chi-square statistic or the Deviance:

$$\Delta X_i^2 = \frac{r_i^2}{1-h_{ii}} = (r_i^S)^2 \qquad \Delta D_i \approx \frac{d_i^2}{1-h_{ii}}$$

- Large values indicate strong influence.
- Problematic observations can be easily spotted using index plots of, say, $\Delta\beta_i$. Another popular graph is to plot the statistic against the predicted probabilities and use different colors/symbols for $Y_i = 0$ and $Y_i = 1$.

Residuals and Influence: Example

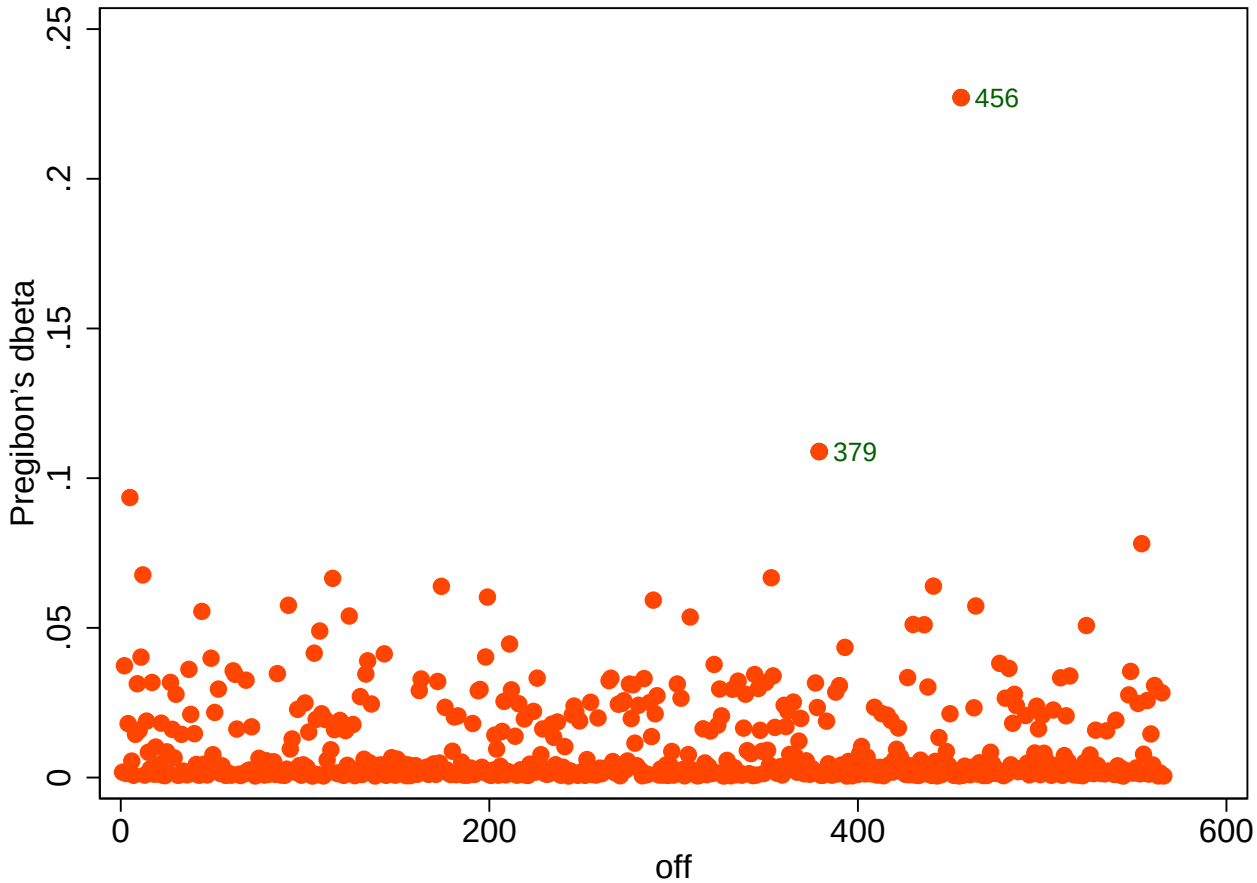
- "Did you ever cheat on your partner?" by age, sex, etc.
- Model estimates:

```
. logit cheat male age highschool extraversion cheat_ok, nolog
Logistic regression                               Number of obs   =        566
                                                  LR chi2(5)      =      44.33
                                                  Prob > chi2     =     0.0000
Log likelihood =  -294.4563                       Pseudo R2       =     0.0700
```

cheat	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
male	-.0289207	.2194078	-0.13	0.895	-.4589521	.4011108
age	.0303383	.0075396	4.02	0.000	.015561	.0451155
highschool	-.1937281	.2138283	-0.91	0.365	-.6128239	.2253677
extraversion	.2514374	.0835575	3.01	0.003	.0876676	.4152072
cheat_ok	.2733903	.0739183	3.70	0.000	.128513	.4182675
_cons	-3.793885	.5430144	-6.99	0.000	-4.858174	-2.729597

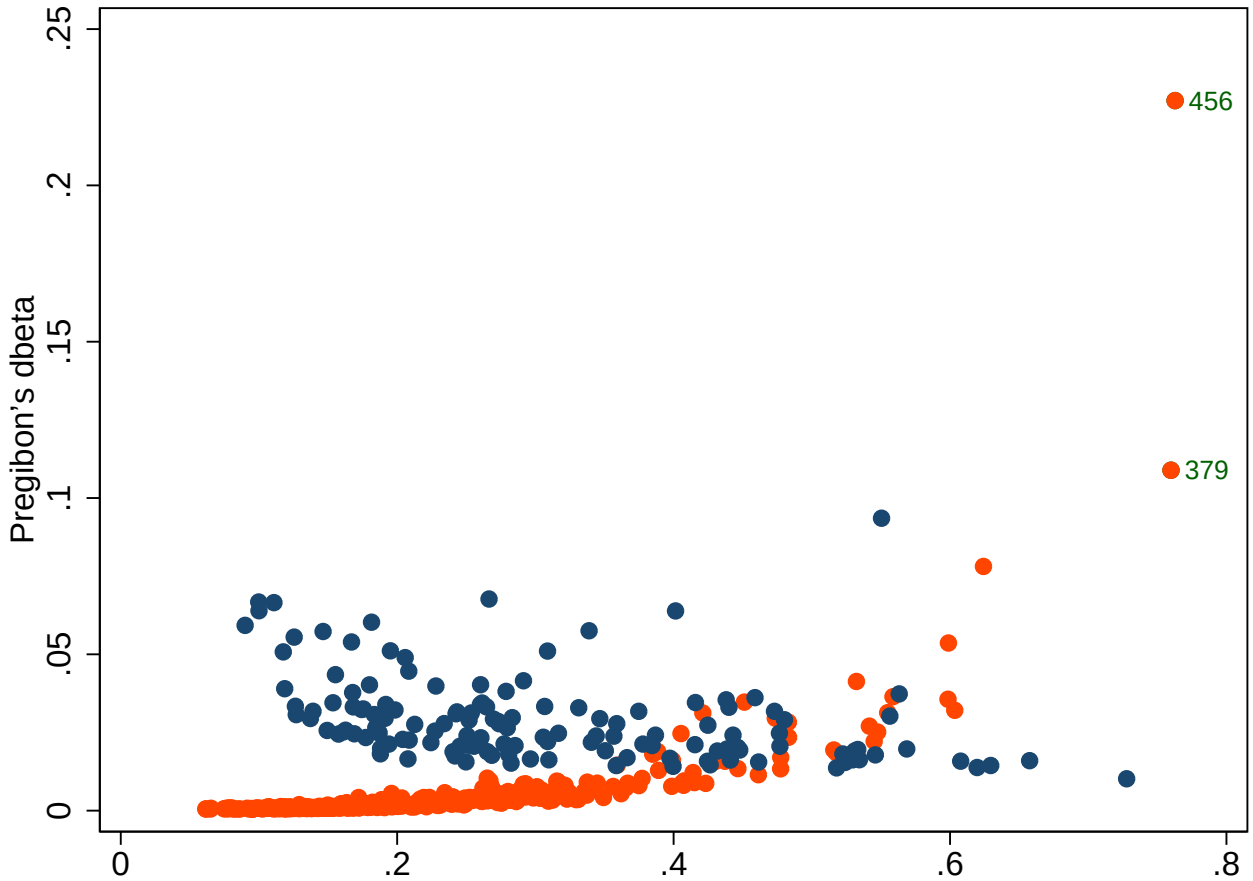
Residuals and Influence: Example

- Index plot of $\Delta\beta$



Residuals and Influence: Example

- Plot of $\Delta\beta$ by prediction



Residuals and Influence: Example

- List problematic observations:

```
. list dbeta cheat male age highschool extraversion cheat_ok if dbeta>.1
```

	dbeta	cheat	male	age	highsc~l	extrav~n	cheat_ok
379.	.1088898	0	0	53.70551	0	6.666667	6
456.	.2271219	0	0	107	1	4.333333	3

- Observation 379 seems to be okay although somewhat atypical. However, there appears to be a data error in observation 456 (age $>$ 100). The data are from an online survey and respondents had to pick their birth year from a dropdown list. Value 0 was stored if the respondent did not make a selection. These invalid observations were accidently included when computing the age variable.

What to do about outliers/influential data?

- Take a close look at the data.
- Correct, or possibly exclude, erroneous data (only in clear-cut cases).
- Compare models in which the outliers are included with models in which they are excluded. If the main results do not change you're on the save side.
- Think about the outliers and improve your model/theory because ...

> *An apparently wild (or otherwise anomalous) observation is a signal that says: "Here is something from which we may learn a lesson, perhaps of a kind not anticipated beforehand, and perhaps more important than the main object of the study." (Kruskal 1960)*

Goodness-of-Fit: Classification Table

- Overall goodness-of-fit: How well does the estimated model describe the observed data?
- One way to asses the goodness-of-fit is to compare the observed data with the "maximum probability rule" predictions

$$\hat{Y}_i = \begin{cases} 0 & \text{if } \hat{\pi}_i \leq 0.5 \\ 1 & \text{if } \hat{\pi}_i > 0.5 \end{cases}$$

- Classification table: Table summarizing the number of "true" and "false" predictions

		predicted: 0	predicted: 1
observed	0	true	false
	1	false	true

- Problem: Prediction table not very useful for extreme distributions (i.e. if $\bar{Y}$ is close to 0 or 1).

Classification Table: Example

- Example ($H = 169$):

		predicted 0	predicted 1
observed	0	25	47
	1	15	82

- Percentage of true predictions using the model:

$$\frac{25 + 82}{169} = 0.63 = 63\%$$

- However: Unconditional prediction (i.e. predict modal category for all observations, here: $Y = 1$) yields

$$\frac{15 + 82}{169} = 0.57 = 57\%$$

- $\Rightarrow$ model improves proportion of true predictions by 6 percentage points.

Goodness-of-Fit Measures

- It may be desirable to summarize the overall goodness-of-fit of a model using a single number.
- In linear regression this is done by the R-squared.
- A number of fit measures imitating the R-squared have been developed for logistic regression (and other models).
- Critique: Scalar measures of fit should always be interpreted in context. How high the value of such a measure should be for the model to be a "good" model strongly depends on the research topic and the nature of the data.

Goodness-of-Fit Measures: R^2 in Linear Regression

- Various interpretations are possible for the R-squared, but in comparison to linear regression these interpretations lead to different measures in logistic regression. Two of the interpretations are:
  - Proportion of explained variation in Y

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

  - Transformation of the likelihood ratio (assumption: $\epsilon \sim N(0, \sigma)$; L_0 is the likelihood of a model with just the constant)

$$R^2 = 1 - (L_0/L)^{2/N}$$

- $R^2 \in [0, 1]$, with higher values indicating better fit.

Goodness-of-Fit Measures: Pseudo-R^2

- R^2 measures can be constructed for logistic regression (or other models) by analogy to one of the interpretations of R^2 in linear regression.
- These measures are called **pseudo-R-squared**.
- Some desirable properties:
  - normalized to $[0, 1]$
  - clear interpretation of values other than 0 and 1

Pseudo-R^2: Explained Variation

- Efron's pseudo-R^2: explained variation based on predicted probabilities

$$R^2_{\text{Efron}} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{\pi}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \quad \in [0, 1]$$

- In the case where J, the number of distinct covariate patterns, is smaller then N, the statistic should be computed as (see Hosmer and Lemeshow 2001:165)

$$R^2_{SSC} = 1 - \frac{\sum_{j=1}^{J}[n_j(\bar{Y}_j - \hat{\pi}_j)]^2}{\sum_{j=1}^{J}[n_j(\bar{Y}_j - \bar{Y})]^2}$$

where n_j is the number of observations with covariate pattern X_j and $\bar{Y}_j$ is the mean of Y among these observations.

Pseudo-R^2: Explained Variation

- McFadden's pseudo-R^2 (sometimes called the "likelihood ratio index"): use log likelihood in analogy to the sum of squares
  - $\ln L_0$: log likelihood of model with just a constant as the total sum of squares
  - $\ln L_1$: log likelihood of the fitted model as the residual sum of squares

$$R^2_{\text{McF}} = \frac{\ln L_0 - \ln L_1}{\ln L_0} = 1 - \frac{\ln L_1}{\ln L_0} \quad \in [0, 1]$$

- The maximum value of 1 can only be reached if J, the number of distinct covariate patterns is equal to N (see Hosmer and Lemeshow 2001:165 for a correction). In general, high values for R^2_{McF} are hard to reach and already values within 0.2 and 0.4 usually indicate a very good fit.
- Interpretation of values other than 0 and 1 not clear.

Pseudo-R^2: Explained Variation

- McFadden's pseudo-R^2 increases if variables are added to the model. A proposed correction is

$$\tilde{R}^2_{\text{McF}} = 1 - \frac{\ln L_1 - m}{\ln L_0}$$

  where m is the number regressors. If adding variables to a model, $\tilde{R}^2_{\text{McF}}$ will only increase if the log likelihood increases by more than 1 for each added variable.
- $\tilde{R}^2_{\text{McF}}$ can be used to compare models (if they are based on the same set of observations)

Pseudo-R^2: Transformation of Likelihood Ratio

- Maximum likelihood pseudo-R^2 / Cox and Snell's pseudo-R^2

$$R^2_{\text{ML}} = 1 - \left(\frac{L_0}{L_1}\right)^{2/N} = 1 - e^{-LR/N}$$

  where LR is the likelihood ratio chi-square statistic.

- Cragg and Uhler pseudo-R^2 / Nagelkerke pseudo-R^2: The maximum of R^2_{ML} is $1 - (L_0)^{2/N}$. This suggests the following correction

$$R^2_{\text{N}} = \frac{R^2_{\text{ML}}}{1 - (L_0)^{2/N}}$$

  so that the measure can take on values between 0 and 1.

Predictive Pseudo-R^2 based on Classification Table

- The information can in the classification table (see above) can be used to construct a R^2 that reflects the prediction errors according to the "maximum probability rule"
- Let $\hat{Y}_i = 0$ if $\hat{\pi}_i \leq 0.5$ and $\hat{Y}_i = 1$ if $\hat{\pi}_i > 0.5$, then

$$R^2_{\text{Count}} = \frac{\#(Y_i = \hat{Y}_i)}{N}$$

- Even without explanatory variables (i.e. if we use the mode of outcome categories as prediction for all observations) R^2_{Count} is at least 50%.

Predictive Pseudo-R^2 based on Classification Table

- Let $M = \max[\#(Y = 0), \#(Y = 1)]$, then

$$\tilde{R}^2_{\text{Count}} = \frac{\#(Y_i = \hat{Y}_i) - M}{N - M}$$

- $\tilde{R}^2_{\text{Count}}$ has a PRE interpretation (Proportional Reduction in Error). It indicates the proportional reduction in prediction errors compared to a model with only a constant.
- Problem: $\tilde{R}^2_{\text{Count}}$ is not very sensitive and is often 0 (especially if $\Pr(Y = 1)$ is generally close to 0 or 1).

Information Measures

- Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)

$$\text{AIC} = \frac{-2 \ln L(\hat{\beta}) + p}{N}$$

$$\text{BIC} = -2 \ln L(\hat{\beta}) + p \ln(N)$$

  where p is the number of parameters in the model
- General interpretation: the smaller AIC or BIC, the better the model fit
- AIC and BIC can be used to compare models (also non-nested models). The model with the smaller AIC or BIC is preferred.
- Interpretation of BIC differences (strength of evidence favoring the model with the smaller BIC): 0-2 weak, 2-6 positive, 6-10 strong, > 10 very strong

Part IX

Logistic Regression: Specification

- Specification Error
- Excursus: Categorical Predictors
- Nonlinearity
- Non-Additivity and Interaction Effects
- Numerical Problems: Zero Cells, Complete Separation, Collinearity

Specification Error

- Assuming the Logit model is essentially correct, i.e. that the model has the general form $\text{logit}[\Pr(Y = 1)] = X\beta$, we can still misspecify the right hand side (RHS) of the equation.
- Some specification errors are:
  - **Omitted variables:** Hard to detect with statistical approaches since this is more of a theoretical issue. If an important variable Z that affects both Y and X is missing in the model, then the estimate of the effect of X on Y will be biased (since it also contains the "indirect" effect of Z on Y through X).
  - **Nonlinearity:** The effect of X on $\text{logit}[\Pr(Y = 1)]$ might be nonlinear. This is also a theoretical issues to some degree, but departures from linearity can be detected statistically. General procedure: Model the effect nonlinearly and compare the results.
  - **Non-additivity:** Assume the model contains X_1 and X_2. The effect of X_1 is assumed to be independent of the value of X_2. This might not be true. Non-additive model can be constructed using interaction terms.

Excursus: Categorical Predictors

- Binary categorical variables (e.g. sex) can be included in a Logit model without problem. The effect measures the difference in the log odds between the two groups defined by the variable.
- What to do if an independent variable X is categorical and has more than two categories (e.g. religious denomination)? We cannot just include the variable in the model because the parameter estimate would arbitrary (e.g. a change in the distances between the codes for the categories, which would not change the meaning of the variable, would change yield a different parameter estimate).
- The solution is to divide the variable into separate indicator variables, one for each category.

Excursus: Categorical Predictors

- To be precise, if the variable has K categories, we only need $K - 1$ indicator variables. One of the categories is chosen as the reference category.
- For example, define

$$X_j = \begin{cases} 1 & \text{if } X = j \\ 0 & \text{else} \end{cases}, \qquad j = 1, \ldots, K-1$$

- The effect β_j then measures the difference between group j and the reference group K. (You can also use any other category as the reference category.)

Excursus: Categorical Predictors

- The coding above is called the dummy or indicator coding. Other coding schemes can be used. Two examples are given bellow.
- Effect coding: deviation from “grand mean” (the effect for the reference category is the negative of the sum of the effects for the other categories)

$$X_j = \begin{cases} 1 & \text{if } X = j \\ -1 & \text{if } X = k \\ 0 & \text{else} \end{cases}, \qquad j = 1, \ldots, K-1$$

- Split coding for ordinal variables: the parameters represent the differences form one category to the next

$$X_j = \begin{cases} 1 & \text{if } X \geq j \\ 0 & \text{else} \end{cases}, \qquad j = 2, \ldots, K$$

Nonlinear Effects

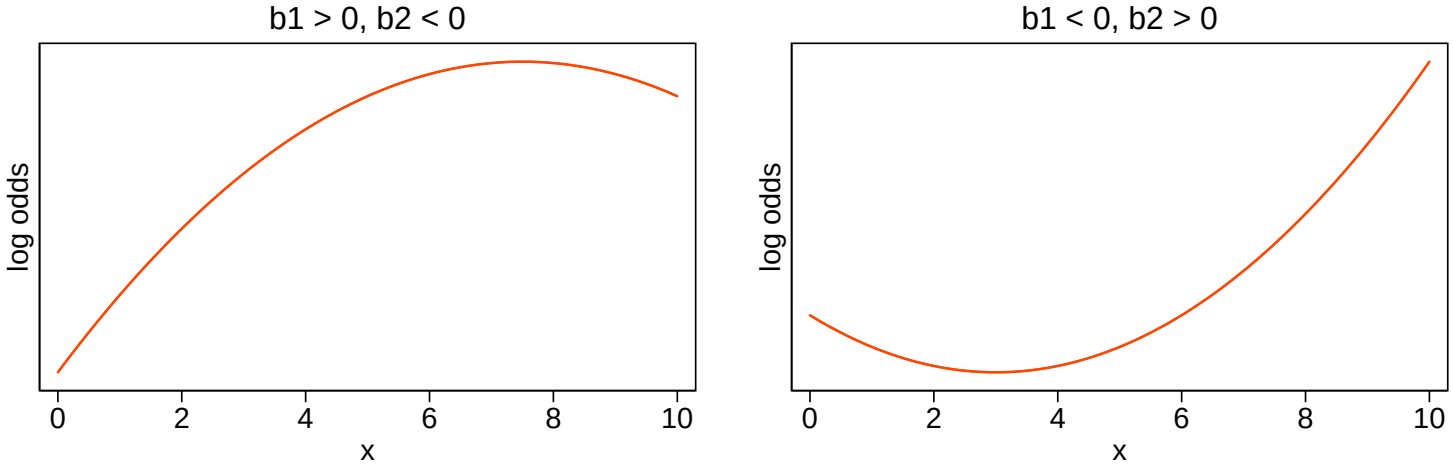
- The effect of a variable X on the log odds may be nonlinear.
- Some strategies to model nonlinear effects:
  - categorize the variable and use dummies (or splines); very flexible, however, choice of categories is arbitrary
  - if a variable is discrete and has only few values, the most flexible way to model the effect is to use separate dummies for the values
  - polynomials (quite flexible, but bad for extrapolation)
  - nonlinear transformations of variables
  - combinations of the above
- For data exploration, i.e. as a first check for nonlinearity, "non-parametric" approaches may be useful. For example, it is often helpful to visualize relationships using scatterplot smoothers such as the the Lowess (see Fox 2000 for an overview).

Nonlinear Effects: Polynomials

- A nonlinear relationship can be modeled by adding powers of X to the equation

$$\text{logit}[\Pr(Y = 1)] = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \ldots$$

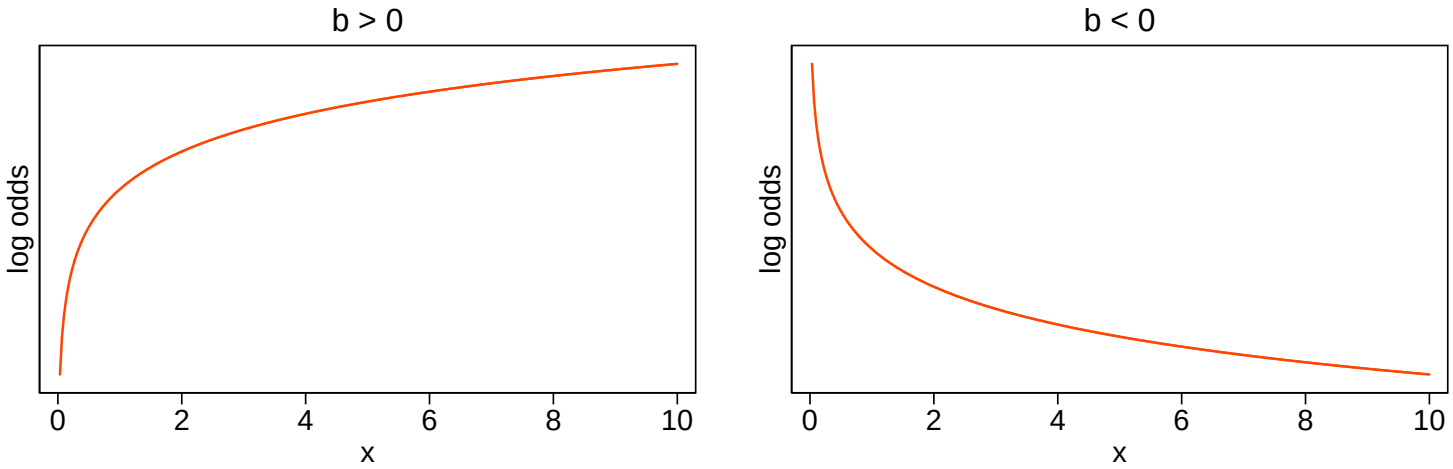
- Most common is the quadratic model model $\Rightarrow$ parabolic effect (higher order models are usually hard to interpret)



Nonlinear Effects: Transformations

- Example: take the logarithm of X

$$\text{logit}[\Pr(Y = 1)] = \beta_0 + \beta_1 \ln(X)$$



- Interpretation: Effect of proportional change in X
  - If X is increased by one percent, then the log odds change by approx. $\beta/100$ (to be precise: $\beta \cdot \ln(1.01)$)
  - Odds ratio: a one percent increase in X changes the odds by approx. β_1 percent (the odds ratio is 1.01^{β}) $\Rightarrow$ elasticity

Non-Additivity and Interaction Effects

- The effect of a variable X_1 may depend on the value of another variable X_2. Think of the effect of getting married on the probability to be in the labor force. Probably the effect is different for men and women.
- Non-additivity: This means that if X_1 and X_2 change simultaneously, then the effect of this simultaneous change is not simply the sum of the two separate effects.
- Non-additive models can be fitted by including products of variables ("interaction terms"):

$$\text{logit}[P(Y = 1)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

- For example, the effect of X_1 in this model is $\beta_1 + \beta_3 X_3$
- When computing results such as marginal effects or discrete change effects, these dependencies have to be taken into account.

Zero Cells

- For categorical covariates there is sometimes no variation in Y for one of the groups. The Logit model does not converge in this case because $|\beta| \to \infty$
- Some software packages automatically detect the problem and remove the corresponding observations and variables from the model, but not all packages do.
- Detection: many iterations, huge parameter estimates, huge or missing standard errors
- Solutions: combine categories if it makes sense, exact logistic regression, Bayesian methods

Complete Separation

- A similar problem occurs if there is complete separation for a continuous variable, e.g. if $Y = 0$ for $X < x^*$ and $Y = 1$ for $X > x^*$.
- This means that X is a "perfect" predictor.
- Congratulations!
- However, usually such situations arise due to errors by the researcher.

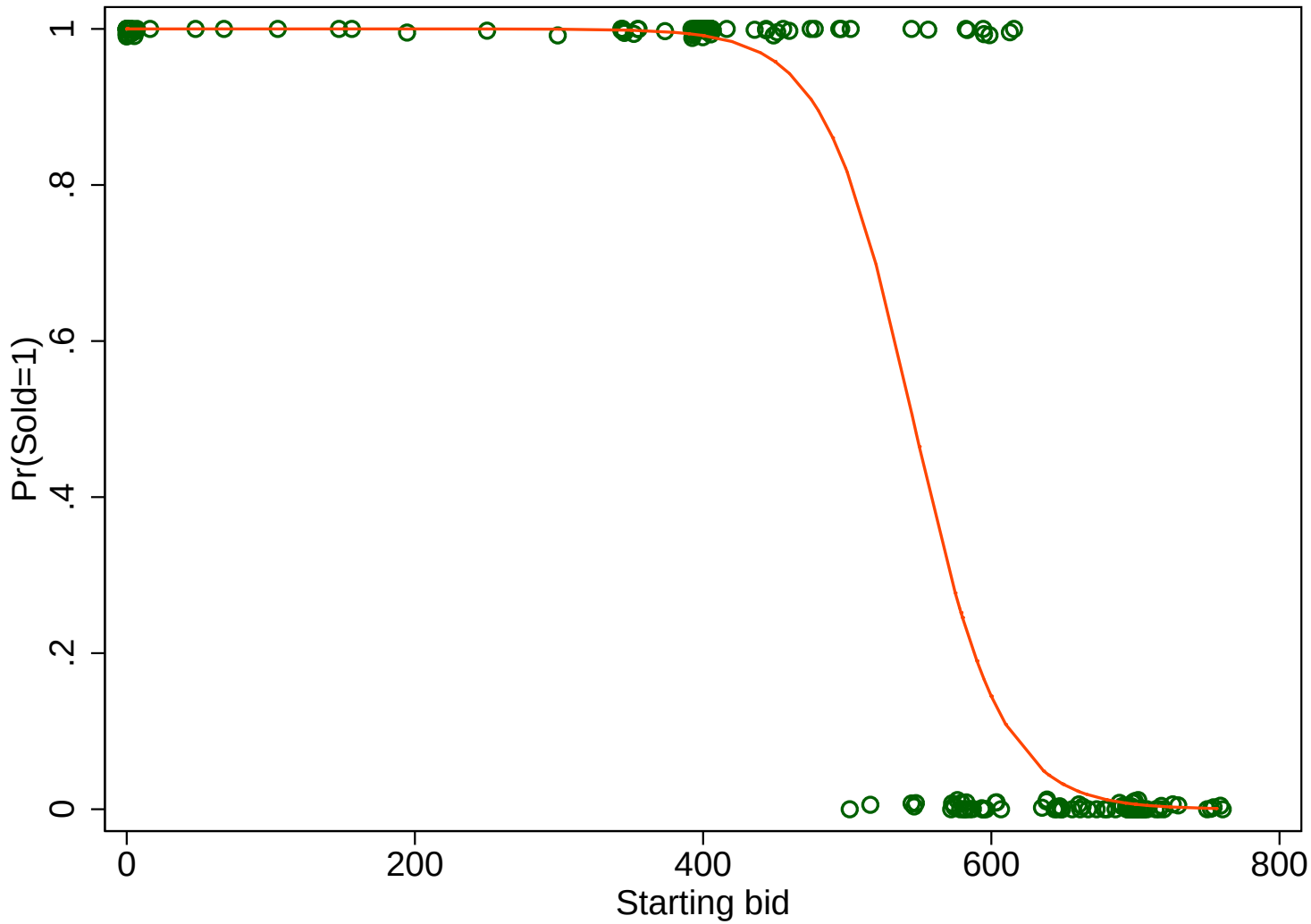
Complete Separation

- Example for almost complete separation: Effect of minimum price in an online auction on whether the product was sold

```
. logit sold ratings stbid date, nolog
Logistic regression                               Number of obs   =        167
                                                  LR chi2(3)      =     192.90
                                                  Prob > chi2     =     0.0000
Log likelihood = -19.301401                       Pseudo R2       =     0.8333

------------------------------------------------------------------------------
        sold |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ratings |   .0319446   .0144293     2.21   0.027     .0036638    .0602255
       stbid |   -.050882   .0121951    -4.17   0.000     -.074784   -.0269799
        date |  -.0507112   .0223791    -2.27   0.023    -.0945734   -.0068489
       _cons |   26.89204   6.552966     4.10   0.000     14.04846    39.73562
------------------------------------------------------------------------------
Note: 0 failures and 21 successes completely determined.

. logit sold ratings date, nolog
Logistic regression                               Number of obs   =        167
                                                  LR chi2(2)      =       5.23
                                                  Prob > chi2     =     0.0731
Log likelihood = -113.13655                       Pseudo R2       =     0.0226

------------------------------------------------------------------------------
        sold |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ratings |   .0099287   .0084343     1.18   0.239    -.0066023    .0264597
        date |   .0112851   .0052426     2.15   0.031     .0010097    .0215604
       _cons |  -.0953283   .1809954    -0.53   0.598    -.4500728    .2594163
------------------------------------------------------------------------------
```

Complete Separation or Perfect Determination

- Example: Effect of minimum price in an online auction on whether the product was sold



Collinearity

- Complete collinearity: If a predictor is a linear combination of one or more other predictors in the equation then the single effects cannot be separated. Example: K dummies for a variable with K categories.

$$d_K = 1 - \sum_{k=1}^{K-1} d_k$$

- Almost complete collinearity: If two (or more) covariates are closely related, then there is only little information to separate the effects of the covariates ($\Rightarrow$ large standard errors)
- The solution to collinearity problems depends on context.
  - If the variables are different measures for the same: use only one or use an index
  - If variables overlap by definition: create non-overlapping variables

Collinearity

- Example: Effect of net minimum price and gross minimum price in an online auction (the gross price includes shipping)

```
. logit sold netminprice grossminprice, nolog
Logistic regression                               Number of obs   =        167
                                                  LR chi2(2)      =     184.47
                                                  Prob > chi2     =     0.0000
Log likelihood = -23.515457                       Pseudo R2       =     0.7968

------------------------------------------------------------------------------
        sold |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 netminprice |   .1549288   .0818699     1.89   0.058    -.0055332    .3153908
grossminpr~e |  -.1935158   .0864308    -2.24   0.025    -.3629171   -.0241146
       _cons |   24.09443   5.751801     4.19   0.000     12.82111    35.36775
------------------------------------------------------------------------------
Note: 0 failures and 19 successes completely determined.

. gen shipping = grossminprice - netminprice

. logit sold netminprice shipping, nolog
Logistic regression                               Number of obs   =        167
                                                  LR chi2(2)      =     184.47
                                                  Prob > chi2     =     0.0000
Log likelihood = -23.515457                       Pseudo R2       =     0.7968

------------------------------------------------------------------------------
        sold |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 netminprice |   -.038587   .0087499    -4.41   0.000    -.0557365   -.0214376
    shipping |  -.1935158   .0864308    -2.24   0.025    -.3629171   -.0241146
       _cons |   24.09443   5.751801     4.19   0.000     12.82111    35.36775
------------------------------------------------------------------------------
Note: 0 failures and 19 successes completely determined.
```

Part X

Probit and Latent Variable Model

- Alternatives to Logit
- The Probit Model
- Latent Variable Model
- Example: Logit versus Probit

Alternatives to Logit

- The Logit model is

$$\Pr(Y = 1|X) = F(X\beta) = \frac{\exp(X\beta)}{1 + \exp(X\beta)}$$

- We can use other (typically s-shaped) functions in place of $F(X\beta)$ instead of the logistic function

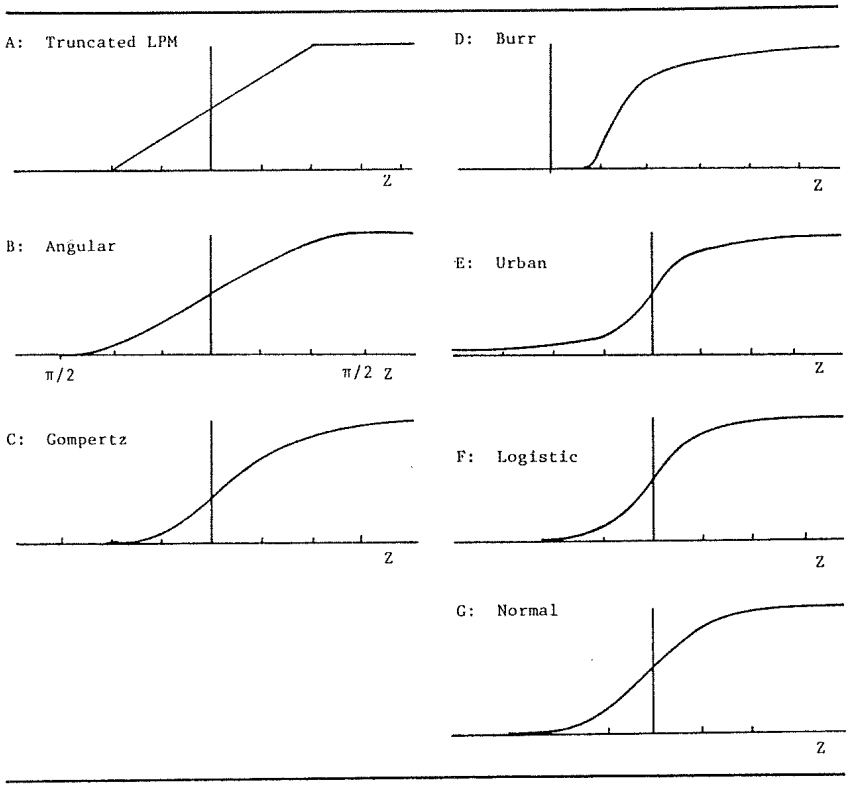


Figure 3. Graphs of Alternative Specifications

(Aldrich/Nelson 1984)

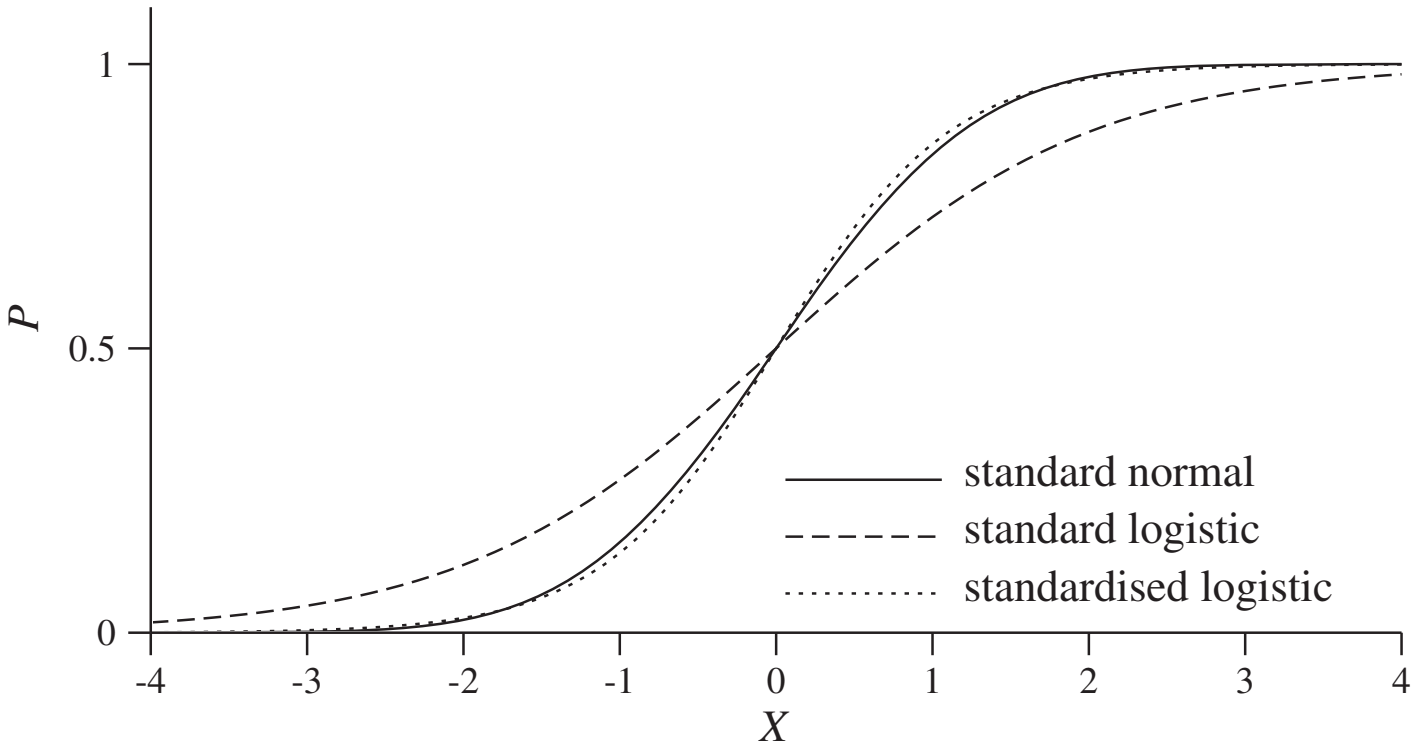
The Probit Model

- One such alternative is the Probit model which uses the cumulative normal distribution

$$\Pr(Y = 1|X) = \Phi(Z) = \int_{-\infty}^{Z} \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)\, du$$

or: $\Phi^{-1}(\Pr(Y = 1|X)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m$

- The Logit and Probit are very similar:



The Probit Model

- Interpretation is of coefficients is similar to the Logit model
- Marginal effect:

$$\frac{\partial \Pr(Y = 1|X)}{\partial X_k} = \left[\frac{1}{\sqrt{2\pi}} \exp(-Z^2/2) \right] \beta_j = \phi(Z)\beta_j$$

- Discrete change effect:

$$\frac{\Delta \Pr(Y = 1|X)}{\Delta X_j} = \Phi[\cdots + \beta_k(X_k + 1) + \cdots] - \Phi[\cdots + \beta_k X_k + \cdots]$$

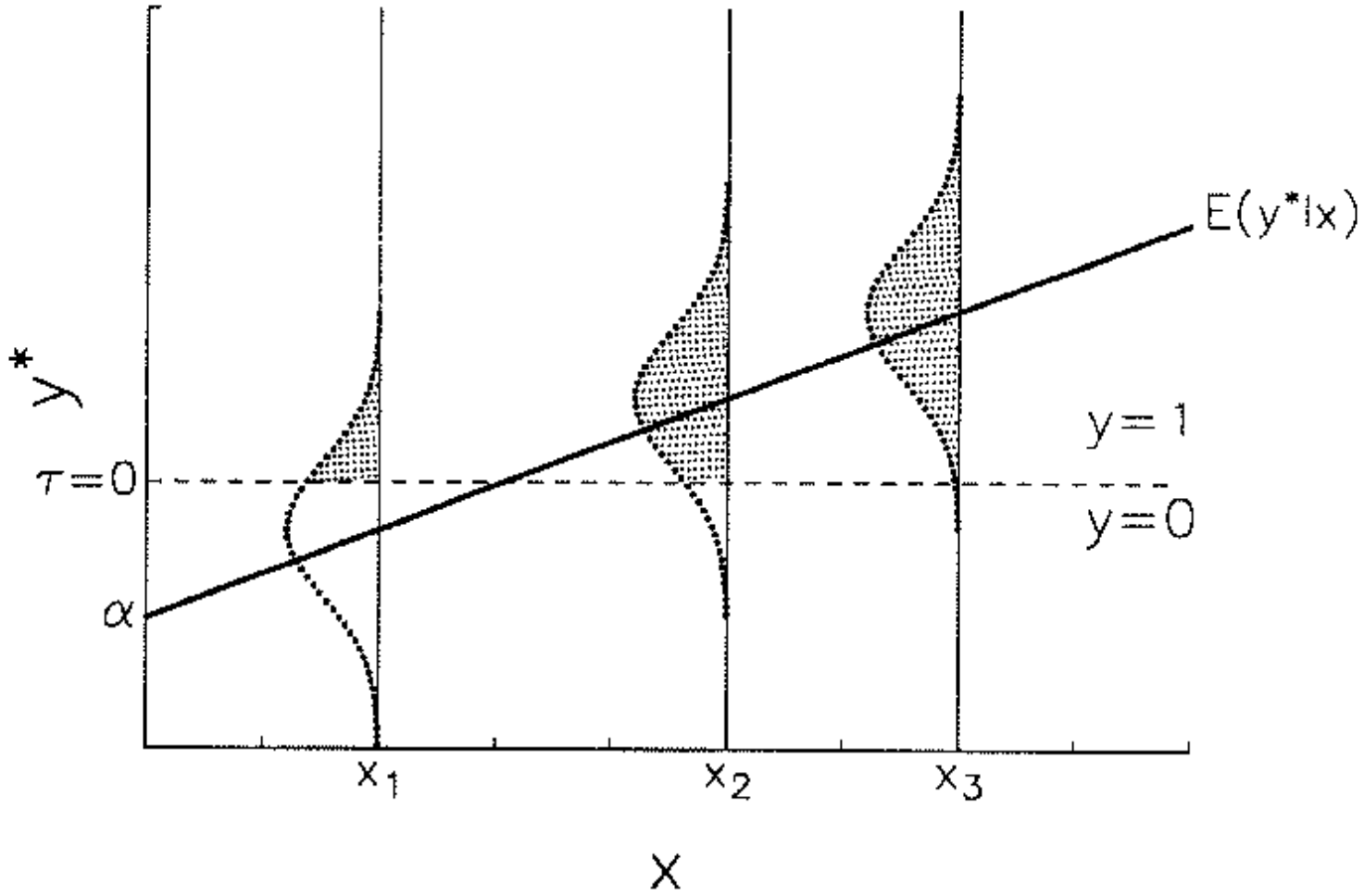
Latent Variable Model

- The Probit model and the Logit model can be expressed as latent variable models.
- The conceptual idea is that there is an unobserved variable Y^* that reflects the propensity of Y to take on value 1 and is related to the covariates in a linear way.
- The model is:

$$Y^* = X\beta + \epsilon \qquad \text{with} \qquad Y = \begin{cases} 1 & \text{if } Y^* > 0 \\ 0 & \text{if } Y^* \leq 0 \end{cases}$$

- Discrete choice models: Y^* can be interpreted as the difference between the utilities of the two alternatives (plus error)

Latent Variable Model



(Long 1997)

Latent Variable Model

- Expressed this way, the only difference between the Logit model and the Probit model then is that they make a different assumption about the distribution of ϵ
- Probit: ϵ has a standard normal distribution, $N(0,1)$
- Logit: ϵ has a standard logistic distribution, $\exp(\epsilon)/[1+\exp(\epsilon)]$
- The standard deviation of ϵ in the Logit model is $\pi/\sqrt{3}$. Therefore the coefficients from the Logit model are usually about $\pi/\sqrt{3} \approx 1.8$ times larger than the coefficients from the Probit model.

Logit versus Probit

$$\beta^{\text{Logit}} \approx 1.8 \cdot \beta^{\text{Probit}}$$

Example: Logit versus Probit

```
. estout, cells("b Xmfx_dydx(label(MargEfct))" "t(par)")
```

	LOGIT		PROBIT	
	b/t	MargEfct	b/t	MargEfct
educ	.1493077	.0285554	.0791481	.0258561
	(3.673453)		(3.582843)	
age	-.0380774	-.0072824	-.020756	-.0067806
	(-5.212636)		(-5.211616)	
k5	-1.171423	-.2240373	-.6776337	-.2213693
	(-8.855418)		(-9.153815)	
k618	-.2328272	-.0445287	-.1340417	-.0437887
	(-2.983747)		(-2.888898)	
protestant	.1803333	.0344891	.1072967	.0350517
	(1.201285)		(1.217839)	
inc1000	-.066502	-.0127187	-.0405756	-.0132552
	(-3.474943)		(-3.585695)	
_cons	1.490499		.9182765	
	(2.62552)		(2.928594)	

Part XI

Generalizations

- Ordered Logit/Probit
- Multinomial Logit
- Conditional Logit

More Than Two Categories

- Up to now we assumed that Y, the dependent variable, has only two values, 0 and 1.
- The discussed approaches can be generalized to polytomous variables with more than two values.

Ordered Logit/Probit

- The values of Y have a natural order.
- Examples:
  - "How satisfied are you with your life?" 1 = "not at all", 2 = "a bit', 3 = "very much"
  - Social class: lower, middle, upper
- Such variables are often analyzed using linear regression.
- However, linear regression makes the assumption that the "distances" between the categories are the same.
- Since this assumption may be wrong, linear regression might be biased, and it is worthwhile to analyze such data using techniques that do not assume interval data.
- In general, using linear regression is less problematic if Y has many values.

Ordered Logit/Probit

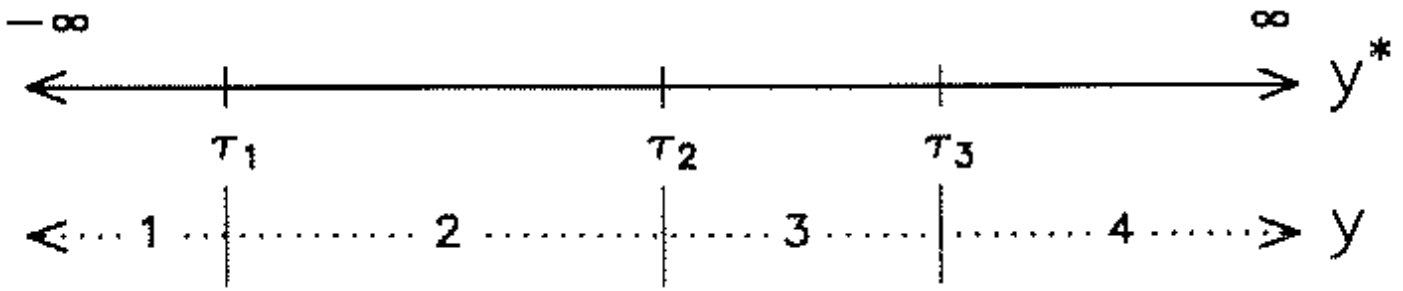
- Ordered Logit/Probit can be represented using the latent variable model.

$$Y^* = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m + \epsilon$$

with

$$Y = \begin{cases} 1 & \text{if } Y^* < \tau_1 \\ 2 & \text{if } \tau_1 \leq Y^* < \tau_2 \\ \cdots & \\ J & \text{if } \tau_{J-1} \leq Y^* \end{cases}$$

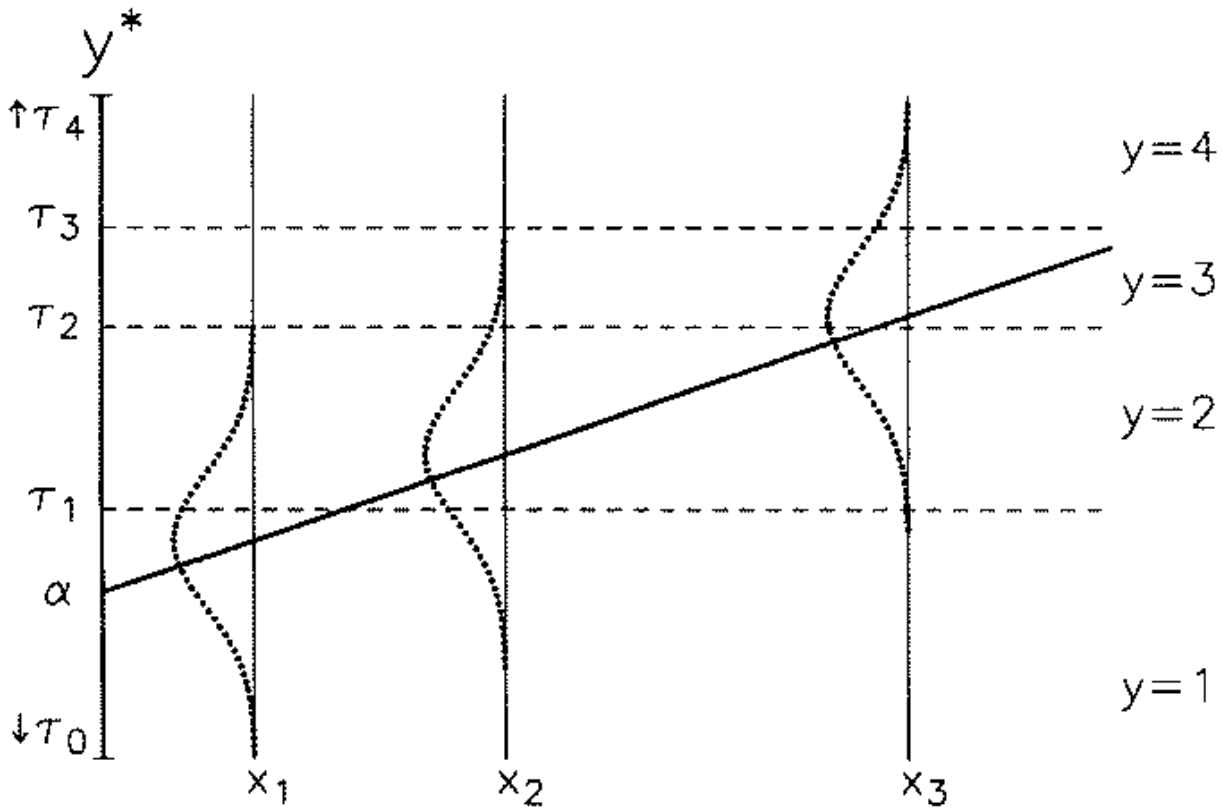
- $\tau_1, \ldots, \tau_{J-1}$ are the thresholds that map the continuous Y^* to the discrete Y



(Long 1997)

Ordered Logit/Probit

- The difference between ordered Logit and ordered Probit lies again in the assumed distribution for the error term.
  - Probit: ϵ has a standard normal distribution
  - Logit: ϵ has a standard logistic distribution



(Long 1997)

Ordered Logit/Probit: Interpretation

- The raw coefficients of ordered Logit/Probit are expressed in terms of the latent variable Y^*.
- The signs of the coefficients have a clear interpretation: If the effect is positive, the probability Y to take on a higher value increases (and vice versa).
- In the case of the ordered Logit an odds ratio interpretation is possible: If X_k is increased by one unit, then the odds of Y being equal to j or less are changed by factor $\exp(-\beta_k)$.
- For detailed interpretation, marginal effects and discrete change effects on the probabilities of the categories can be computed (see Long 1997 for details). However, note that these effects are category specific (separate effect for each category).

Ordered Logit/Probit: Interpretation

- The probability of $Y = j$ given variables X is

$$\Pr(Y = j|X) = F(\tau_j - X\beta) - F(\tau_{j-1} - X\beta)$$

  where $\tau_0 = -\infty$ and $\tau_j = \infty$ and $F()$ denotes the cumulative normal distribution or logistic.

- Marginal effect

$$\frac{\partial \Pr(Y = j|X)}{\partial X_k} = \beta_k \left[f(\tau_{j-1} - X\beta) - f(\tau_j - X\beta) \right]$$

- Discrete change

$$\frac{\Delta \Pr(Y = j|X)}{\Delta X_k} = F(\tau_j - \cdots - \beta_k(X_k + 1) - \dots) - F(\tau_{j-1} - \cdots - \beta_k X_k - \dots)$$

- Note that the sign of a marginal effect or discrete change effect may be different than the sign of β_k. The sign may even change over the range of X_k.

Ordered Logit/Probit: Example

```
. fre weffort, nomissing
weffort — work effort beyond what is required
```

	Freq.	Percent	Cum.
1 none	124	6.01	6.01
2 only a little	175	8.49	14.50
3 some	927	44.96	59.46
4 a lot	836	40.54	100.00
Total	2062	100.00	

```
. esttab, mtitles eqlabels(none) wide compress
```

	(1) LRM		(2) OPROBIT		(3) OLOGIT	
female	0.106*	(2.50)	0.147*	(2.50)	0.229*	(2.33)
parttime	-0.303***	(-6.57)	-0.418***	(-6.62)	-0.704***	(-6.56)
selfemp	0.224***	(4.19)	0.374***	(4.91)	0.688***	(5.28)
educ	0.0158*	(2.02)	0.0194	(1.81)	0.0300	(1.65)
_cons	3.027***	(31.20)				
cut1			-1.365***	(-9.96)	-2.469***	(-10.33)
cut2			-0.862***	(-6.40)	-1.484***	(-6.47)
cut3			0.464***	(3.47)	0.733**	(3.23)
N	2062		2062		2062	

```
t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

Multinomial Logit

- Nominal variable: The categories have no natural order.
- The multinomial Logit is used for nominal variables or for ordered variables where the ordering is questionable (e.g. occupational attainment)
- The model has many parameters! There is one set of parameters for each category of Y.

Multinomial Logit

- The probabilities of the different outcomes of Y are expressed as

$$\Pr(Y = j|X) = \frac{\exp(X\beta_j)}{\sum_{\ell=1}^{J} \exp(X\beta_\ell)}$$

  with β set to zero for one of the outcomes.
- The outcome for which the β vector is set to zero is called the "base outcome" or the "reference category"
- The parameter estimates of the multinomial Logit therefore express differences compared to the base outcome.
- Note that the binary Logit model is a special case of the multinomial Logit.

Multinomial Logit: Odds

- The odds of outcome j versus outcome m are

$$\frac{\Pr(Y=j|X)}{\Pr(Y=m|X)} = \frac{P_j}{P_m} = \frac{\frac{\exp(X\beta_j)}{\sum_{\ell=1}^{J} exp(X\beta_\ell)}}{\frac{\exp(X\beta_m)}{\sum_{\ell=1}^{J} exp(X\beta_\ell)}} = \frac{\exp(X\beta_j)}{\exp(X\beta_m)}$$

- Taking the logarithm yields

$$\ln(P_j/P_m) = X(\beta_j - \beta_m)$$

and the partial derivative is:

$$\frac{\partial \ln(P_j/P_m)}{\partial X_k} = \beta_{kj} - \beta_{km}$$

Multinomial Logit: Odds

- The parameters in the multinomial Logit can therefore be interpreted as follows.
- Log Odds: If X_k is increased by one unit the log of the odds of outcome j against outcome m changes by $\beta_{kj} - \beta_{km}$.
- Odds: If X_k is increased by one unit the odds of outcome j against outcome m (the "relative risk ratio") changes by factor $\exp(\beta_{kj} - \beta_{km})$.
- If m is the base outcome: $\beta_{km} = 0$
  - Log Odds: If X_k is increased by one unit the log of the odds of outcome j against the base outcome changes by β_{kj}.
  - Odds: If X_k is increased by one unit the odds of outcome j against the base outcome changes by $\exp(\beta_{kj})$.

Multinomial Logit: Partial Effect and Discrete Change

- Also for the multinomial Logit we can compute marginal effects and discrete change effects. Similar to the ordered Logit/Probit, the direction of these effects may not correspond to the signs of the coefficients and the direction may also depend on the values of the covariates for which the effects are computed.
- Recall that $\Pr(Y = j|X) = \exp(X\beta_j) \Big/ \sum_{\ell=1}^{J} \exp(X\beta_\ell)$
- Partial effect:

$$\frac{\partial \Pr(Y = j|X)}{\partial X_k} = \Pr(Y = j|X) \left[\beta_{kj} - \sum_{\ell=1}^{J} \beta_{k\ell} \Pr(Y = \ell|X) \right]$$

- Discrete change:

$$\frac{\Delta \Pr(Y = j|X)}{\Delta X_k} = \Pr(Y = j|X, X_k + 1) - \Pr(Y = j|X)$$

Multinomial Logit: Example

- Labor force status by sex and education

```
. fre lfstatus, nomissing
lfstatus
```

	Freq.	Percent	Cum.
1 full time	1267	44.60	44.60
2 part time	538	18.94	63.53
3 selfemployed	285	10.03	73.57
4 not in labor force	751	26.43	100.00
Total	2841	100.00	

Multinomial Logit: Example

```
. quietly mlogit lfstatus female educ
. quietly prchange
. quietly estadd prchange, adapt
. estout, cell("b se t(fmt(2)) p(fmt(3)) dc(label(DisChg))")
```

	b	se	t	p	DisChg
full time					
female	0	.	.	.	-.4035242
educ	0	.	.	.	.0224239
_cons	0	.	.	.	
part time					
female	2.76204	.1427824	19.34	0.000	.2668968
educ	.0272368	.0250682	1.09	0.277	.017066
_cons	-2.856912	.3279246	-8.71	0.000	
selfemployed					
female	.4184821	.1375259	3.04	0.002	-.052557
educ	.0128819	.0274198	0.47	0.638	.0086798
_cons	-1.78688	.340294	-5.25	0.000	
not in lab~e					
female	1.725244	.1025924	16.82	0.000	.1891844
educ	-.0981816	.0243676	-4.03	0.000	-.0481697
_cons	-.28136	.2915872	-0.96	0.335	

Conditional Logit

- The conditional Logit is a variant on the multinomial Logit that can be applied if the data contain alternative specific information.
  - Multinomial Logit: X variables are characteristics of the individuals; they vary over individuals
  - Conditional Logit: Additional variables Z that reflect characteristics of the categories of Y (the “alternatives”)
- Examples are the traffic mode choice where travel costs and time are known for each mode or the choice of consumer goods or services where various characteristics of the good or service are known.
- The shape of the data set is different for conditional Logit than for multinomial Logit:
  - Multinomial Logit: one row per individual
  - Conditional Logit: one row per alternative (including a variable identifying the individual and a variable indicating the chosen alternative)

Conditional Logit

- The probability of choosing alternative J given variables X that vary by individual and variables Z that vary by alternative (and individual) is expressed as

$$\Pr(Y = j|X, Z) = \frac{\exp(X\beta_j + Z_j\gamma)}{\sum_{\ell=1}^{J} \exp(X\beta_\ell + Z_\ell\gamma)}$$

- Odds interpretation of γ: If the difference in Z_k between alternative j and alternative m is increased by 1 unit then the odds of alternative j over alternative m change by factor $\exp(\gamma_k)$

Further Topics

- Models for count data
- Models for limited or truncated data
- Complex samples, weights
- Panel data models, multilevel models
- GLM (Generalized Linear Models)
- Log-linear models
- Nonparametric methods
- Causal inference, endogeneity, sample selection correction

Additional References

References cited in the lectures but not listed in textbooks section:

- Bagley, Steven C., Halbert White, and Beatrice A. Colomb (2001). Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. Journal of Clinical Epidemiology 54: 979-985.
- Berkson, Joseph (1944). Application of the Logistic Function to Bio-Assay. Journal of the American Statistical Association 39(227): 357-365.
- Berkson, Joseph (1951). Why I Prefer Logits to Probits. Biometrics 7(4): 327-339.
- Caudill, Steven B. (1988). An Advantage of the Linear Probability Model over Probit or Logit. Oxford Bulletin of Economics & Statistics 50(4): 425-427.
- Fox, John (1991). Regression Diagnostics. Newbury Park, CA: Sage.
- Fox, John (2000). Nonparametric Simple Regression. Smoothing Scatterplots. Thousand Oaks, CA: Sage.
- Gelman, Andrew, and Hal Stern (2006). The Difference Between “Significant” and “Not Significant” is not Itself Statistically Significant. The American Statistician 60(4): 328-331.

- Goldberger, Arthur Stanley (1964). Econometric Theory. New York: John Wiley & Sons.
- Golding, Jean, Rosemary Greenwood, Karen Birmingham, and Martin Mott (1992). Childhood cancer, intramuscular vitamin K, and pethidine given during labour. BMJ 305: 341-346.
- Hosmer, David W., and Stanley Lemeshow (1980). Goodness of fit tests for the multiple logistic regression model. Communications in Statistics - Theory and Methods 9(10): 1043 - 1069.
- Jann, Ben (2005). Comment: Earnings Returns to Education in Urban China: A Note on Testing Differences among Groups. American Sociological Review 70(5): 860-864.
- Jann, Ben (2006). Diagnostik von Regressionsschätzungen bei kleinen Stichproben. Kölner Zeitschrift für Soziologie und Sozialpsychologie Sonderheft 44/2004: 421-452.
- Kruskal, William H. (1960). Some Remarks on Wild Observations. Technometrics 2(1): 1-3.

- Lemeshow, Stanley, and David W. Hosmer, Jr. (1982). A Review of Goodness of Fit Statistics for Use in the Development of Logistic Regression Models. American Journal of Epidemiology 115(1): 92-106.
- Peduzzi, Peter, John Concato, Elizabeth Kemper, Theodore R. Hiolford, and Alvan R. Feinstein (1996). A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis. Journal of Clinical Epidemiology 49(12): 1373-1379.
- Petersen, Trond (1985). A Comment in Presenting Results from Logit and Probit Models. American Sociological Review 50(1): 130-131.
- Pregibon, Daryl (1981). Logistic Regression Diagnostics. The Annals of Statistics 9(4): 705-724.
- Smith, Gordon C. S., and Jill P. Pell (2003). Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. BMJ 327: 1459-1461.