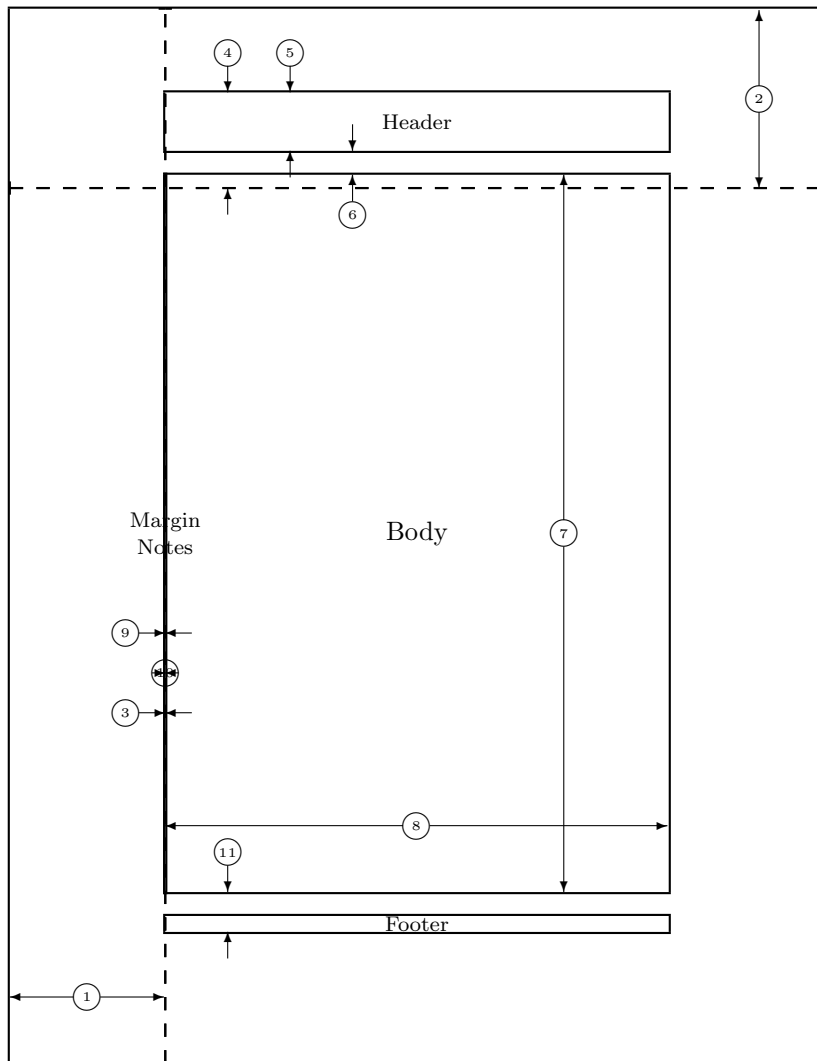


- |    |                      |    |                                  |
|----|----------------------|----|----------------------------------|
| 1  | one inch + \hoffset  | 2  | one inch + \voffset              |
| 3  | \oddsidemargin = 0pt | 4  | \topmargin = -72pt               |
| 5  | \headheight = 44pt   | 6  | \headsep = 18pt                  |
| 7  | \textheight = 540pt  | 8  | \textwidth = 379pt               |
| 9  | \marginparsep = 0pt  | 10 | \marginparwidth = 0pt            |
| 11 | \footskip = 30pt     |    | \marginparpush = 5pt (not shown) |
|    | \hoffset = 45pt      |    | \voffset = 63pt                  |
|    | \paperwidth = 614pt  |    | \paperheight = 794pt             |



- |    |                       |    |                                  |
|----|-----------------------|----|----------------------------------|
| 1  | one inch + \hoffset   | 2  | one inch + \voffset              |
| 3  | \evensidemargin = 0pt | 4  | \topmargin = -72pt               |
| 5  | \headheight = 44pt    | 6  | \headsep = 18pt                  |
| 7  | \textheight = 540pt   | 8  | \textwidth = 379pt               |
| 9  | \marginparsep = 0pt   | 10 | \marginparwidth = 0pt            |
| 11 | \footskip = 30pt      |    | \marginparpush = 5pt (not shown) |
|    | \hoffset = 45pt       |    | \voffset = 63pt                  |
|    | \paperwidth = 614pt   |    | \paperheight = 794pt             |

# Contents

<b>0</b>	<b>Regression Models For Categorical Dependent Variables</b>	<b>1</b>
0.1	The Linear Probability Model . . . . .	2
0.2	Basic Concepts . . . . .	6
0.2.1	Odds, Log-Odds und Odds-Ratios . . . . .	6
0.2.2	Excursus: The Maximum Likelihood Principle . . . . .	10
0.3	Logistic Regression with Stata . . . . .	13
0.3.1	The Coefficients Block . . . . .	15
	Sign Interpretation . . . . .	16
	Interpretation with Odds Ratios . . . . .	16
	Probability Interpretation . . . . .	17
0.3.2	The Iteration Block . . . . .	19
0.3.3	The Model Fit Block . . . . .	20
	Classification Tables . . . . .	21
	Pearson Chi Square . . . . .	24
0.4	Diagnostics of Logistic Regression . . . . .	25
0.4.1	Linearity . . . . .	25
0.4.2	Influential Cases . . . . .	30
0.5	Likelihood Ratio Test . . . . .	34
0.6	Refined Models . . . . .	36
0.7	Advanced techniques . . . . .	40
0.7.1	Probit Models . . . . .	40
0.7.2	Multinomial Logistic Regression . . . . .	43
0.7.3	Models for ordinal data . . . . .	47
0.8	Summary . . . . .	49

**Author index**

**53**

**Subject index**

**55**

# List of Tables

1	Probabilities, odds and logits . . . . .	8
---	--	---



# List of Figures

1	Sample of a dichotomous characteristic with the size of 3 . . . . .	11
---	---	----



# 0 Regression Models For Categorical Dependent Variables

The social sciences often have to deal with categorical dependent variables. These may be variables whose values may be dichotomous (e.g. rented flat yes or no), nominal (partisanship of the CDU, the SPD, or the Green Party) or ordinal (no concerns, some concerns, strong concerns). In this chapter we will present a number of procedures used to model variables such as these and will begin with a procedure for dichotomous dependent variables: “logistic regression”.

Logistic regression is, for the most part, similar to linear regression. Therefore, we will explain it as an analogy to the previous chapter. If you have no previous experience or knowledge of linear regression, then we would advise you to first read chapter ?? to page ??.

As is the case in linear regression, in logistic regression the dependent variables are predicted through a combination of independent variables. A combination of such variables is called a linear combination and looks like this:

$$b_0 + b_1x_{1i} + b_2x_{2i} + \dots + b_{K-1}x_{K-1,i}$$

Here  $x_{1i}$  is the value of the first independent variable for interviewee  $i$ ,  $x_{2i}$  is the respective value of the second independent variable and so on. The coefficients  $b_1, b_2, \dots, b_{K-1}$  represent the weights assigned to the variables.

In contrast to linear regression, in logistic regression one has to consider a particular transformation of the dependent variable. *Why* such a transformation is required and why linear regression cannot therefore be used is explained in section 0.1, while the transformation itself is explained in 0.2.1.

Section 0.2.2 contains an excursus for the method with which the logistic regression of the coefficients is determined. This section is slightly more difficult. As it is not required for an initial comprehension of logistic regression, you can skip this excursus for now.

Calculating a logistic regression with Stata is explained in section 0.3. This is followed by methods of verifying the basic assumptions of the model (section 0.4). The procedure for verifying the significance of the coefficients is discussed in section 0.5, while section 0.6 demonstrates a few possibilities for refining the modelling of correlations.

An overview of further procedures, in particular procedures for categorical variables with more than two values, can be found in section 0.7.

As it was true for our introduction in linear regression (chapter ??) additional reading is necessary to gain full a understanding of the techniques wie describe. Books that fit well to our approach are ??.

## 0.1 The Linear Probability Model

Why is linear regression not suitable for categorical dependent variables? Imagine you were employed by an international ship safety regulatory agency and were supposed to take a closer look at the sinking of the Titanic. You are supposed to find out whether the seafaring principle of “women and children first” was put into practice or whether there is any truth in the assumption made by the film ‘Titanic’, in which the first-class gentlemen took the places in the lifeboats at the expense of the third-class women and children.

For the purpose of this investigation, we have provided you with data on the sinking of the Titanic.<sup>1</sup> Open the file with<sup>2</sup>

```
. use titanic2, clear
```

and before you continue to read, make yourself familiar with the content of the dataset using the commands

```
. describe
. tab1 _all
```

You will discover that the file contains details on age (*age2*), gender(*sex*) and the passenger class (*class*) of the Titanic’s passengers. Furthermore, there is information for each passenger on whether or not they survived the catastrophe (*survived*).

In order to clarify the disadvantages of linear regression of categorical dependent variables, we will go through such a model. First, we will investigate whether children really were rescued more often than adults were. What would a scatterplot where the *Y* variable represents the variable for survival and *X* variable represents age look like? Try it out with a hand-drawn graph for this scatterplot.

You will notice that the points can only be entered on two horizontal lines: either at the value 0 (did not survive) or at 1 (survived). If children were actually rescued more often than adults, then the number of points on the 0-line should increase in relation to

---

<sup>1</sup>The data provided is real. The dataset and its exact description can be found at <http://amstat.org/publications/jse/archive.htm>. For didactic reasons, we have changed the original dataset in that we have divided adults and children into further fictional age groups, as the original set differentiates merely between adults and children. Our data package contains the Do-File, with which we conducted our “fabrication” (*crtitanic2.do*), as well as the original data set (*titanic.dta*).

<sup>2</sup>Please make sure that your working directory is *c:/data/kkstata*. Further information on this can be found at ??.

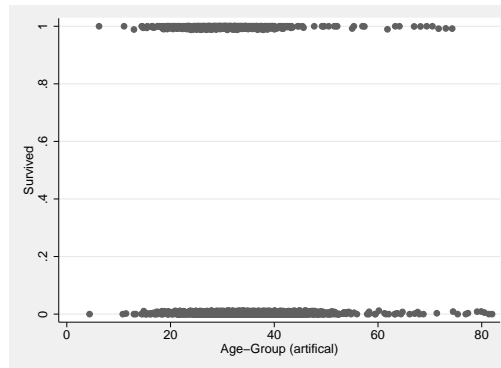
those on the 1-line the further right you go. To check whether your chart is correct use

```
. graph twoway scatter survived age2
```

In doing so you will establish that this diagram is not particularly informative, as the plot symbols are often directly marked on top of each other and that the number of data points therefore cannot to be seen.

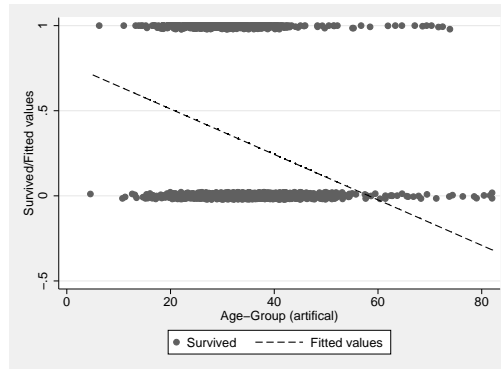
With the help of the `scatter` option `jitter()` it is possible to achieve a more informative diagram. With `jitter()`, a small random number is added to each data point, which reveals cases which previously covered one another up. Within the brackets is a number between 1 and 30 that controls the size of the random number; generally use small numbers if possible.

```
. graph twoway scatter survived age2, jitter(2)
```



On examining the chart, one receives the impression that there is in fact a negative correlation between ages and survival of the Titanic disaster. This impression is confirmed when you draw the regression line on the chart (also see section ??):

```
. regress survived age
. predict yhat
. graph twoway (scatter survived age2, jitter(2)) (line yhat age2, sort)
```



The chart reveals one central problem of linear regression for dichotomous dependent variables: The regression line in the illustration shows predicted values of under 0 from around the age of 60 onwards. What does this mean with regards to the content? Remind yourself of how the predicted values of dichotomous dependent variables are generally interpreted. Until now, we have understood the predicted values to be the estimated average extent of the dependent variables for the respective combination of independent variables. In this sense, one would say, for example, that the survival of a 5 year old averages at around 0.7. This is a less convincing interpretation if one considers that one can only survive or not survive; surviving a little bit does not exist.

However, the predicted value of the dichotomous dependent variable can also be interpreted in a different way. Here, one should make it clear to oneself what the arithmetic mean of a dichotomous variable with the values of 0 and 1 signifies. The variable *survived* has, for example, the arithmetic mean of 0.3230. This matches the share of passengers who are coded with 1 with the total number of passengers.<sup>3</sup> Therefore, the interpretation is as follows: the share of survivors in the dataset amounts to around 32 percent, or in other words, the probability that you will find a survivor in the dataset lies at 0.32. In general, the predicted values of the linear regression are estimates of the conditional mean of the dependent variable. This therefore allows you to use the probability interpretation for every value of the independent variable: the predicted value of around 0.7 for a 5 year-old means a probability of survival of 0.7. On the basis of this alternative interpretation, the linear regression model for dichotomous dependent variables is often called the “linear probability model”, or just LPM (?).

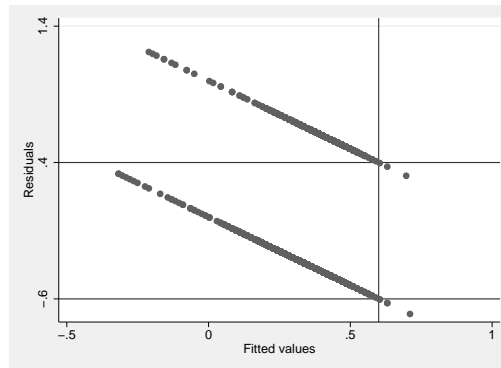
How can one interpret the negative predicted values for passengers over 60 with the help of the probability interpretation? In actual fact, not at all, as according to the mathematical definition of probabilities, they should basically be between 0 and 1. Given sufficient smaller or larger values of the  $X$  variable, a model that uses a *straight line* to represent probabilities will, however, inevitably produce values of over 1 or under 0. This is the *first problem* that affects OLS regression of dichotomous variables.<sup>4</sup>

<sup>3</sup>You can confirm this for yourself with `tab survived`.

<sup>4</sup>In practice, this is a problem of little importance when predicted values of over 1 or under 0 for realistic values of the independent variables do not appear. However, a model which would prevent

The *second problem* affects the homoscedastic assumption of linear regression that we introduced in section ???. According to this assumption, the variance of errors for all values of  $\hat{Y}$  should be constant. We suggested that the scatterplot of the residuals against the predicted values was an indication of a possible infringement of this assumption. You can achieve a graph such as this for our linear probability model through

```
. predict r, resid
. graph twoway scatter r yhat, yline(-.6 .4) ylab(-.6 .4 1.4) xline(.6)
```



In this graph, you can observe that only two possible residuals can appear for every predicted value. Less apparent is that both of these residuals result directly from the predicted values. If a survivor (*survived=1*) has a predicted value of 0.6 due to their age, they will have a residual of  $1 - 0.6 = 0.4$ . If you predict a value of 0.6 for an individual who did not survive (*survived=0*) you will receive a value of  $0 - 0.6 = -0.6$ .

Thus, the residuals are either  $1 - \hat{y}_i$  or  $-\hat{y}_i$ . The variance of the residuals is  $\hat{y}_i \times (1 - \hat{y}_i)$  and is therefore all the more larger, the nearer the predicted values approach 0.5. The residuals of the linear probability model are therefore per definition heteroskedastic. This problem is especially of practical use if one is interested in the confidence intervals of the  $b$  coefficients.

In conclusion, it can be stated that although a linear regression with a dichotomous dependent variable is possible, it basically leads to two problems. *Firstly*, with regards to the content not all predicted values can be interpreted and *secondly* this model does not allow for correct confidence intervals to be determined. In order to avoid these problems we require a model which only produces probabilities between 0 and 1, and also relies on assumptions which are maintained by the model. Both are fulfilled by logistic regression, the basic principles of which we will introduce now.

---

impossible probabilities such as these from the start seems sensible.

## 0.2 Basic Concepts

### 0.2.1 Odds, Log-Odds und Odds-Ratios

In the previous section, we found that the linear OLS regression of dichotomous dependent variables can produce unwanted predicted values. This is clearly due to our attempt to represent values between 0 and 1 with a straight line. The values that are calculated with the help of a linear regression<sup>5</sup> are basically not subject to any restrictions. This means that theoretically, values between  $-\infty$  and  $+\infty$  may emerge. Therefore, regression models that are based on a linear combination should only use dependent variables whose range of values are equally infinite. As the range of values for probabilities lies between 0 and 1 they are unsuitable as dependent variables. An alternative is the logarithmic chance. With the help of the Titanic data from the previous section we will show what this means.

We previously received indications that children had a higher chance of survival than adults did. Now we want to investigate whether the second part of the principle “women and children first” is valid. Therefore, we must next ask whether women were more likely to survive the Titanic disaster. You can get the initial indication of the chance of survival for women and men through a two-way table between *sex* and *survived*:

```
. tabulate sex survived, row
```

Key			
	<i>frequency</i>		
	<i>row percentage</i>		
Gender	Survived		Total
	no	yes	
women	126 26.81	344 73.19	470 100.00
man	1,364 78.80	367 21.20	1,731 100.00
Total	1,490 67.70	711 32.30	2,201 100.00

In section ?? we interpreted tables such as this with the help of row or column percentages. In this way, by using the available row percentages, we were able to determine that the overall share of survivors was around 32 percent, whereas that of the women was about 50 percentage points higher than that of the men (73 % compared to 21 %). Alternatively, one can do a similar comparison by dividing the number of survivors to the number of dead. For the women this would be 344 : 126.

```
. display 344/126
2.7301587
```

<sup>5</sup>we showed you a linear combination such as this at the beginning on page 1.

You will reach the same figure<sup>6</sup> if you divide the proportional values (in this case the row percentages)

```
. display .7319/.2681
2.7299515
```

One can interpret these ratios as follows: “for women, the probability of surviving in almost 3 times as high as the probability of dying.” In other words: “the probability of dying is around one third ( $1 : 2.73 = 0.366$ ) of the probability of surviving.” In practice one generally says that the odds of surviving is around 2.73 to 1, while the odds of dying lies at around 1 to 2.73.

Overall, this relationship can be written as:

$$\text{odds}_{\text{surviving}} = \frac{\text{Probability}_{\text{surviving}}}{\text{Probability}_{\text{dying}}} \quad (1)$$

or slightly shorter by using symbols instead of text:

$$\text{odds} = \frac{P(Y = 1)}{1 - P(Y = 1)} \quad (2)$$

The probabilities of survival  $P(Y = 1)$  and dying  $P(Y = 0)$  can be respectively found in the numerator and the denominator. Both of these probabilities add up to 1. There are no further options besides surviving and dying. Therefore, you can use “1 – the probability of survival” for the probability of dying.

You can also calculate the chance of survival for men: their odds of survival is considerable lower than that of the women:  $\frac{367}{1364} = .269$ . This means that for men the chances that they will be among the survivors stands at  $0.269 : 1$ . Or in other words, men are 3.72 times more likely to be among the victims.

Of course, you can compare the odds of survival for men and women with the help of a measured value. For instance, you can calculate how large the chance of survival is for men in comparison to those for women. To do this you divide the odds of the men through the odds of the women:

```
. display .269/2.73
.0985348
```

This relationship is called the “odds ratio”.

In our case we would say that the chance of survival for a man is 0.099 times than that of a women, or around ten times smaller than that of a woman. Apparently,

---

<sup>6</sup>The deviations are due to rounding up.

the principle of “women and children first” was adhered to. Whether this *appearance* actually holds is something that we will investigate in more detail further on (page 37).

However, first we should consider the suitability of odds for our statistical model. Do you remember the discussion from the previous section? There we looked at the changes in the probability in surviving the Titanic catastrophe depending on age. Here we experienced the problem that predicting these probabilities with a linear combination could result in values outside the definition range of probabilities. What would happen if we were to draw upon odds instead of probabilities?

Table 1: Probabilities, odds and logits

$P(Y = 1)$	$odd = \frac{P(Y=1)}{1-P(Y=1)}$	$\ln(odd)$
0.01	1/99 = .01	-4.60
0.03	3/97 = .03	-3.48
0.05	5/95 = .05	-2.94
0.20	20/80 = .25	-1.39
0.30	30/70 = .43	-0.85
0.40	40/60 = .67	-0.41
0.50	50/50 = 1.00	0
0.60	60/40 = 1.50	0.41
0.70	70/30 = 2.33	0.85
0.80	80/20 = 4.00	1.39
0.95	95/5 = 19.00	2.94
0.97	97/3 = 32.33	3.48
0.99	99/1 = 99.00	4.60

In the first column of table 1 we listed a number of selected probability values. You will see that at first the values increase slowly, then rapidly and finally slowly again. The values are between 0 and 1. First, if we presume that the values represent the chance of survival for passengers on the Titanic of different ages, then the first row would contain the group of the oldest passengers with the lowest chance of survival and the bottom row would contain the group of the youngest passengers with the highest chance of survival. Through equation (2) you can calculate for each of these groups the odds that an individual within one of the groups survived the Titanic catastrophe. Furthermore, imagine that each of these groups contains one hundred people. As the first group has a probability of 0.01 you would have one person out of one hundred would have survived. In other terms one to ninety-nine (1 : 99) and if you calculate 1/99, you receive the value 0.010101. You can perform this for the rows in the table. An inspection of the values of the odds indicates that they lie between 0 and  $+\infty$ . odds of 0 occurs if there are no survivors within a specific group, while odds  $+\infty$  occur when practically everybody survives. If the number of survivors is equal to the number of victims, then we receive odds of 1.

Odds are therefore *slightly* better suited than probabilities for being the dependent

variables in a regression model, as no matter how high the absolute value is when predicting with a linear combination, it will not be outside the definition range of the odds. However, a linear combination also allows for negative values, but negative odds do not exist. The problem can be avoided by using the (natural) logarithm of the odds. These values, called *logits*, were calculated in the last column of the table.

Now look at the values of the logits more closely: while the odds have a minimum boundary, the logarithmic values have no lower or upper boundaries. The logarithm of 1 is 0. The logarithm of numbers under 1 results in lower figures which stretch to  $-\infty$ , the nearer one approaches 0. The logarithm of numbers over 1 stretches towards  $+\infty$ . Note also the symmetry of the values. At a probability of 0.5 the chance lies at 1 : 1 or fifty:fifty. The logarithmic value lies at 0. If you look at the probabilities above and below 0.5, then you will see that at equal intervals of probabilities of the odds' logarithm, only the algebraic sign changes.

The logit is not restricted and has a symmetric origin. It is therefore very well suited to being represented by linear combination and hence better suited to being a dependent variable in a regression model. Unfortunately, the logit is not necessarily easy to interpret. You are unlikely to be understood by your employer if you inform them that the logarithmic chance of survival of a male Titanic passenger is  $-1.31$ , while that of a female passenger is  $+1.00$ . However, through a simple transforming of (2) you can convert values of the logits back into probabilities:

$$P(Y = 1) = \frac{e^L}{1 + e^L} \quad (3)$$

where  $L$  is logit and  $e$  is the Euler's Number ( $e \approx 2.718$ ). A functional graph of this transformation can be drawn as follows:

```
. graph twoway function y=exp(x)/(1+exp(x)), range(-10 10)
```

Looking at the graph another interesting characteristic of the logits becomes apparent: while the range of values of the logits have no upper or lower boundaries, the values of the probabilities calculated from the logits remain between 0 and 1. For logits between around  $-2.5$  and  $2.5$  the probabilities increase relatively rapidly, however the nearer one approaches the boundary values of the probabilities, the less the probabilities change. In other words: the probabilities asymptotically approach the values 0 and 1, however they *never* go over the boundaries. From this it can be deduced that on the basis of a linear combination, predicted logits can always be converted into probabilities within the permitted boundaries of 0 and 1.

Overall, we can be certain that logarithmic odds is well suited as a dependent variable for our regression model. The equation for such a model would look like this:

$$\hat{L}_i = b_0 + b_1x_{1i} + b_2x_{2i} + \dots + b_{K-1}x_{K-1,i} \quad (4)$$

This is the so-called logistic regression model or logit model. The formal interpretation of the  $b$  coefficients of this model is identical to that of the linear regression (OLS): when a  $X$  variable increases by one unit, the predicted values (the logarithmic odds) rise by  $b$  units.

Before we can use the logistic regression, it is worth considering the procedure through which the  $b$  coefficients of the equation (4) are determined. In the case of the linear regression we used the OLS procedure for the process of estimation. For the logistic regression we instead use the process of “maximum likelihood”. The logic of this process is somewhat more demanding than that of OLS, even though the basic principle is similar: one looks for the  $b$  coefficients which in a certain respect are optimal. We would like to explain the process in greater detail in the following excursus. However, you do not need to work through the excursus in order to understand the section that it precedes!

## 0.2.2 Excursus: The Maximum Likelihood Principle

While discussing the linear regression, we explained the OLS process used to determine the  $b$  coefficients. In principle, it would be possible to calculate the logarithmic odds for each combination of the independent variables and to use these in an OLS regression model. Nevertheless, for reasons that we will not explain here, a procedure such as this is not as *efficient* as the process of estimation that is applied in logistic regression: the maximum likelihood principle.<sup>7</sup> With the help of this technique, the  $b$  coefficients are determined in a way that the proportionate values which you observed become maximally probable. What does this mean? In order to answer this question we will first make a little detour:

On page 4, we informed you that the share of survivors on the Titanic amounts to 32.3 percent. Suppose that you had determined this figure from a sample of the passengers. In this case, you could ask yourself how likely such a share may be, when the *true* share of the survivors among the passengers amounts to, say, 60 percent? To answer this question you should consider the following: you select a single passenger from the population. If the share of survivors from the population amounts to 60 percent, then the probability that this passenger will be a survivor is 0.6 and the probability that they will be a victim is 0.4. Now select a second person from the population. This person can once again either be a survivor or a victim, whereby the probabilities remain the same (sampling “with replacement”). Select one more individual from the population and then have a short break.

During this break you should look at the figure 1. In it, we have conducted all possible samples with three observations. In total we obtained  $2^n = 2^3 = 8$  samples with a size of  $n = 3$ . In the first, we only sampled survivors (S). The probability that a sample randomly selects 3 survivors is  $0.6 \times .6 \times .6 = 0.6^3 = 0.216$ . In the second, third and fifth samples we randomly selected two survivors and one victim (V). Each of these

---

<sup>7</sup>?, 40–45 introduction to the maximum likelihood principle served as a model for the following section.

samples has the probability  $0.6 \times .6 \times .4 = 0.6^2 \times .4^1 = 0.144$ . In total, the probability of such a sample is  $0.144 \times 3 = 0.432$ .

The probabilities of samples 4, 6 and 7 are respectively,  $0.6 \times .4 \times .4 = 0.6 \times .4^2 = 0.096$ . In total the probability of samples such as these is therefore  $0.096 \times 3 = 0.288$ . Finally, there is sample 8, where the probability lies at  $0.4 \times .4 \times .4 = 0.4^3 = 0.064$ . If, based on the samples given in the mapping, we ask the question of how likely it is that one out of three survives, the answer is: as likely as samples 4, 6 and 7 together, i.e. 0.288.

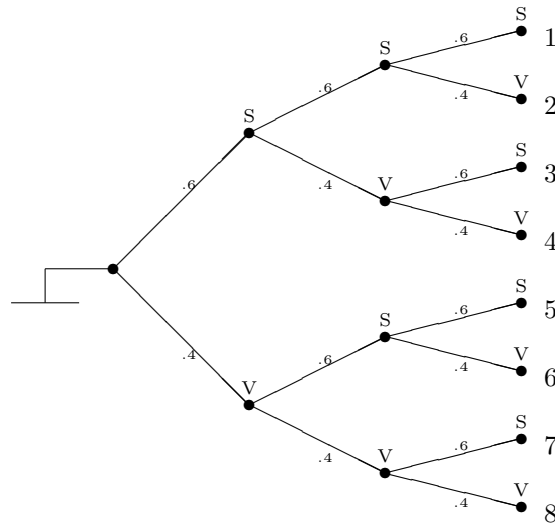


Figure 1: Sample of a dichotomous characteristic with the size of 3

Generally, the probability of a sample with a size of  $n$  taking place, where the dichotomous characteristic appears  $n$  times, lies at

$$P(h|\pi, n) = \binom{n}{h} \pi^h (1 - \pi)^{n-h} \quad (5)$$

where  $\pi$  defines the share of the dichotomous characteristic of the population. The term  $\binom{n}{h}$  stands for  $\frac{n!}{h!(n-h)!}$ . It enables us to calculate the number of potential samples in which the dichotomous characteristic appears  $n$  times. In Stata, the probability of the samples 4, 6 and 7 in our mapping can be calculated with this command:

```
. display comb(3,1) * .6^1 * .4^2
.288
```

In practice, for the most part we are not interested in this figure, instead our attention is on  $\pi$ , the characteristic's share of the population. Although  $\pi$  is unknown, given the example used, we can consider what value of  $\pi$  would make the given sample

most probable. For this, we can use various values for  $\pi$  in the equation (5) and then select the value which results in the highest probability. Formally, this means that we are searching for the value of  $\pi$  for which the likelihood

$$\mathcal{L}(\pi|h, n) = \binom{n}{h} \pi^h (1 - \pi)^{n-h} \quad (6)$$

is maximized and we can forego a calculation of  $\binom{n}{h}$  as this term remains constant for all values of  $\pi$ . Note that the likelihood is calculated with the same formula as in equation (5). However, the results of (5) add up all the possible values of  $h$  to the value 1, whereas this is not the case for the values of  $\mathcal{L}$  and all the possible values of  $\pi$ . Therefore, we linguistically differentiate between likelihood and probability.

You can manually act this out for sample 2 from figure 1 (2 survivors and 1 victim). In order to do so, create an artificial dataset with one hundred observations:

```
. clear
. set obs 100
obs was 0, now 100
```

Now generate the variable *pi* by rendering a series of possible values for  $\pi$ :

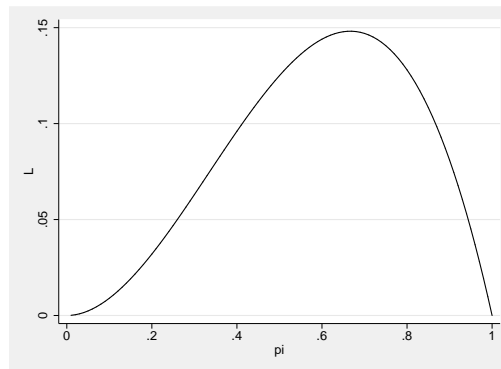
```
. generate pi = _n/100
```

As  $h$  and  $n$  are known from the sample, you can calculate the likelihood for the various values of  $\pi$ :

```
. generate L = pi^2 * (1 - pi)^(3-2)
```

With the help of a graph you can then analyze which  $\pi$  results in a maximal likelihood:

```
. graph twoway line L pi, sort
```



The maximum of the likelihood lies at around  $\pi = 0.66$ . This is the maximum likelihood estimate of the share of survivors from the population, given the sample contains two survivors and one victim.

How does one estimate the  $b$  coefficients of our regression model with the maximum likelihood principle from the equation (4)? The answer is simple. Instead of directly inserting the values for  $\pi$ , we calculate  $\pi$  with the help of our regression model. Here we insert (4) in (3)

$$\pi = \widehat{P}(Y = 1) = \frac{e^{b_0 + b_1 x_{1i} + \dots + b_{K-1} x_{K-1,i}}}{1 + e^{b_0 + b_1 x_{1i} + \dots + b_{K-1} x_{K-1,i}}} \quad (7)$$

and again in (6):

$$\begin{aligned} \mathcal{L}(b_k | f, n, m) &= \widehat{P}(Y = 1)^h \times \left(1 - \widehat{P}(Y = 1)\right)^{n-h} \\ &= \left(\frac{e^{b_0 + b_1 x_{1i} + \dots + b_{K-1} x_{K-1,i}}}{1 + e^{b_0 + b_1 x_{1i} + \dots + b_{K-1} x_{K-1,i}}}\right)^h \times \left(1 - \frac{e^{b_0 + b_1 x_{1i} + \dots + b_{K-1} x_{K-1,i}}}{1 + e^{b_0 + b_1 x_{1i} + \dots + b_{K-1} x_{K-1,i}}}\right)^{n-h} \end{aligned} \quad (8)$$

After this you can attempt to maximize this function by trying out different values of  $b_k$ . However, as is the case with OLS regression, it is better to reproduce the first derivative from  $b_k$  and to set the resulting standard equation as zero. The mathematical process is made easier when the log likelihood, i.e.  $\ln \mathcal{L}$  is used. Yet, you will not find an analytical solution in this procedure, as was the case in the linear OLS regression. For this reason iterative algorithms (which could be disrespectfully referred to as *astute trial and error*) are used to maximize the log likelihood.

We have introduced the maximum likelihood principle for logistic regression with dichotomous dependent variables. In principle, the process can be applied to various models. For this purpose, the equation (6) is adapted to the respective distributive assumption. Subsequently, the equation for the probability of the model is inserted through (6). The resulting likelihood function is then maximized through a preferably generic and rapid estimation algorithm. The logic behind this procedure is implemented in Stata with the command `ml`, which is described in detail in ?.

### 0.3 Logistic Regression with Stata

For a moment we would like to set aside our Titanic example in favor of an alternative. Consider that you assumed that when the age and household income of a surveyed individual increases, the probability of living in a freehold flat or house also increases. In addition, you expect that the share of individuals who own their own residence<sup>8</sup> to

<sup>8</sup>In the following we will refer to living in one's own freehold flat or house as *residence ownership*. In this respect children may also have proprietary housing. For household income we will use the word "income".

be higher in West Germany than it is in East Germany.

Now, please load our dataset *data1.dta*.

```
. use data1, clear
```

In order to check your assumption you can calculate a logistic regression model of proprietary housing against the independent variables of age, household income and an East-West variable. The Stata command for calculating logistic regression is `logit`. The syntax of the command is straightforward, as it matches all other model commands: after the command comes the name of the dependent variable, followed by a variable list with the names of the independent variables.

One peculiarity has to be noted with regards to the dependent variables: at least one category of the dependent variable should be 0, as `logit` models the logistic chance that the dependent variable is different from zero. Normally, one would use a dependent variable with the values 0 and 1, where the category assigned with 1 is generally labeled as success. Accordingly, the category labeled with 0 is seen as failure. For our example, the variable *owner* should be generated with the values of 1 for house owner and 0 for tenant. This is carried out effectively through:<sup>9</sup>

```
. generate owner = renttype == 1 if renttype < .
```

We generate the East-West variable analogously to the respective linear regression variable (page ??) through:

```
. generate east = state>=11 & state<=16 if state<.
```

It would also appear to be sensible to generate an age variable for our regression model from the year of birth variable available in our dataset.

```
. generate age = 1997-ybirth
```

Furthermore, it is recommended to center the two metrically independent variables *age* and *hhinc*, e.g. to deduct the mean of the variable from each value. The mean of centered variables is zero, which eases the interpretation of regression models at various points (?). The centering of *age* and *hhinc* can be conducted through the following commands:<sup>10</sup>

```
. summarize age if hhinc < . & owner < . & east < .
. generate cage = age - r(mean) if hhinc < . & owner < . & east < .
. summarize hhinc if age < . & owner < . & east < .
. generate chhinc = hhinc - r(mean) if age < . & owner < . & east < .
```

Afterwards you can calculate the following logistic regression:

---

<sup>9</sup>This command is dealt with in detail on page ?. The command `label list` determines the assignment of the values to labels (section ?).

<sup>10</sup>The Stata commands used here are explained in chapter ?.

```

. logit owner cage chhinc east
Iteration 0:  log likelihood = -2091.5129
Iteration 1:  log likelihood = -1930.1356
Iteration 2:  log likelihood = -1927.6015
Iteration 3:  log likelihood = -1927.5979
Logit estimates
Log likelihood = -1927.5979
Number of obs   =      3200
LR chi2(3)      =      327.83
Prob > chi2     =      0.0000
Pseudo R2      =      0.0784

```

owner	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
cage	.0189758	.0021862	8.68	0.000	.0146909	.0232608
chhinc	.0006504	.0000418	15.54	0.000	.0005684	.0007324
east	-.0583019	.0864511	-0.67	0.500	-.2277431	.1111392
_cons	-.6023514	.0462412	-13.03	0.000	-.6929826	-.5117202

The results table is very similar to the results table that you got to know from linear regression. At the bottom of the output you will find the coefficients block with the coefficients for the dependent variables and the constants. Above left you will find the iterations block with some results which are related to the maximum likelihood calculation, and above right you will find a model fit block. In the following sections we will individually discuss the blocks along the lines of our explanation of linear regression.

### 0.3.1 The Coefficients Block

The following should clarify the various possibilities for interpreting the  $b$  coefficients of the logistic regression. The  $b$  coefficients can be found in the first column of the coefficient block.<sup>11</sup> The  $b$  coefficients formally indicate how the predicted values change when the corresponding independent variable increase by one unit. This matches the interpretation of the  $b$  coefficients of the linear regression, although in this case the predicted values are the logarithmic odds of success, instead of being the mean of the dependent variable. For example, the interpretation of the regression coefficient *cage* is as follows: “The logarithmic odds of residence ownership rises on average by 0.0189758 if age increases by one year.” The interpretation of the regression coefficient of *chhinc* is analogous. Based on regression coefficient of *east* we can say: “With every one unit increase of the variable *east* the logarithmic chance of residence ownership falls on average by 0.0583019.” As *east* can only increase by one unit once, this can be reworded: “East Germans have, on average, a 0.06 smaller logarithmic chance of residence ownership than West Germans.” The regression constants provides the predicted value for those individuals surveyed, for whom all the other independent variables show the value

<sup>11</sup>In the second column you will find the standard errors of the regression coefficients, which will help you calculate significance test, as well as confidence interval limits. The interpretation of these figures corresponds with the respective figures in the linear regression (section ??). With regard to the significance of the coefficients it should, however, be noted that their evaluation within the framework of a logistic regression usually takes place with a likelihood ratio test (section 0.5).

0. Due to the centering this means that in this case: “the logarithmic chance of residence ownership for West German individuals with a mean age and mean income lies at  $-0.6023514$ .”

Overall, the information on the changes in the logarithmic chance of success appears to be slightly extrinsic. For this reason we require various workarounds to interpret the coefficients, which we will now introduce.

### Sign Interpretation

It is easier to limit oneself to the interpretation of signs and the relative size of the coefficients. A positive sign for the regression coefficient means that the probability or chance of residence ownership increases with the respective independent variable, whereas a negative sign means that the respective probability or chance decreases. The extent of the change is all the stronger, the higher the regression coefficient is. We can’t draw any conclusion from the exact extent of the change in probability. In our example, the probability of house ownership increases with age and with income. The probability of house ownership is lower in the East than it is in the West.

### Interpretation with Odds Ratios

With the help of the model equation we wanted to calculate the predicted the logarithmic chance of a West German with mean income and mean age. For the centered income variable (*chhinc*), the individuals surveyed whose income matched the mean have the value 0. The same applies for the centered age variable (*cage*), where the individuals with the mean age have the value 0. Lastly, West Germans surveyed also have the value 0 for the variable *east*. Therefore, all coefficients apart from the constants are omitted from the equation (4) when calculating the predicted logits. The predicted logarithmic chances (logits) for West Germans of mean age and with a mean income therefore match the regression constants.

By calculating the exponential function you can convert the uninformative logarithmic odds to odds:<sup>12</sup>

```
. display exp(_b[_cons])
.54752267
```

Accordingly, you can calculate the chance for those who are exactly one year older than the average:

```
. display exp(_b[_cons] + _b[cage]*1)
.55801156
```

The older person’s chance of residence ownership is therefore slightly larger than that of those with average age. We can use *odds ratio* (page 7) for a comparison of both

<sup>12</sup>We covered working with the saved coefficients in detail in section 0.1.

ages. In this case it amounts to:

```
. display exp(_b[_cons] + _b[cage])/exp(_b[_cons])
1.019157
```

This means that if the age increases by one year compared to the average then the chance of residence ownership increases by 1.02. An increase in age of 1 over the mean of 2 over the mean leads to a further multiplication of the *chance* by 1.02 and so on. As every one unit increase in age leads to an increase in chance of 1.02, one also talks of the multiplicative unit effect in connection with the odds ratio. This is in contrast to the additive unit effect of  $b$  coefficients.<sup>13</sup>

The calculative complexity of determining odds ratios can be noticeably reduced if one considers the following: in general, in order to determine the odds ratios we have first calculated the odds for a particular value of  $X$  and then for the value  $X + 1$ . After that we divided both results by each other, which can be presented as follows:

$$\text{odds-ratio} = \frac{e^{b_0+b_1(X+1)}}{e^{b_0+b_1X}} = \frac{e^{b_0+b_1X} e^{b_1}}{e^{b_0+b_1X}} = e^{b_1} \quad (9)$$

You can therefore also receive the odds ratio by computing the exponential function of the  $b$  coefficient.

Many logistic regression users prefer the interpretation of results in the form of the odds ratios. For this reason, in Stata the odds ratio can be directly displayed instead of the  $b$  coefficients. There are two ways to do this: *first* the odds ratio can be requested by the `logit` command option `or`.

```
. logit owner cage chhinc east, or
```

This option can also be used if the results of the last calculated model need to be displayed again: `logit, or`.

*Secondly* the command `logistic` can be used instead of `logit`. The command `logistic` works identically to `logit`, however it reports the odds-ratios instead of the  $b$  coefficient.

```
. logistic owner cage chhinc east
```

## Probability Interpretation

The third possibility for interpreting the coefficient is provided by the equation (3). There we showed you how to convert logits into probabilities. With Stata you can calculate these probabilities in the usual fashion. For example, the probability of residence ownership is available for West Germans with mean ages and incomes:

<sup>13</sup>With logarithmic chance,  $b$  units are added for every unit of the  $X$  variable, whereas the chance is multiplied by the odds ratio.

```
. display exp(_b[_cons])/(1 + exp(_b[_cons]))
.35380591
```

This means that the estimated share of house owners in the individuals surveyed with mean age and income amounts to around 35 percent.

The command `predict` enables you to generate a new variable which contains every observation for the predicted probabilities. Here, one would enter the command along with the name of the variable which should contain the predicted probabilities.

```
. predict Phat
```

We use the name *Phat*, in order to indicate that it deals with predicted probabilities. You can also calculate the predicted logits with the `xb` option of the `predict` command. In this case we would use *Lhat* as the variable name.

The problem with the interpretation of probabilities is that the probabilities do not increase at an equal rate with every increase of the independent variable. For example, please consider the following three probabilities where we have increased the age by 10 years at a time:

```
. display exp(_b[_cons] + _b[cage]*10)/(1 + exp(_b[_cons] + _b[cage]*10))
.39829047
. display exp(_b[_cons] + _b[cage]*20)/(1 + exp(_b[_cons] + _b[cage]*20))
.44452061
. display exp(_b[_cons] + _b[cage]*30)/(1 + exp(_b[_cons] + _b[cage]*30))
.49173152
```

Between West Germans with the mean age (see calculation on page 0.3.1) and those 10 years older the probability of residence ownership *increases* by around  $0.3983 - 0.3538 = 0.0445$ . Subsequently, the probability increases by  $0.4445 - 0.3983 = 0.0462$  and then by  $0.4917 - 0.4445 = 0.0472$ . An increase in age by 10 years at a time does not therefore lead to a consistent change in the predicted probability.

One way out is the graphic description of the probabilities in a conditional effects plot. As shown in section ??, it deals with a graph of the predicted values for various characteristics of the independent variables. Thus, one could, for example, generate a variable with income dependent predicted probabilities of West Germans with the mean age.<sup>14</sup>

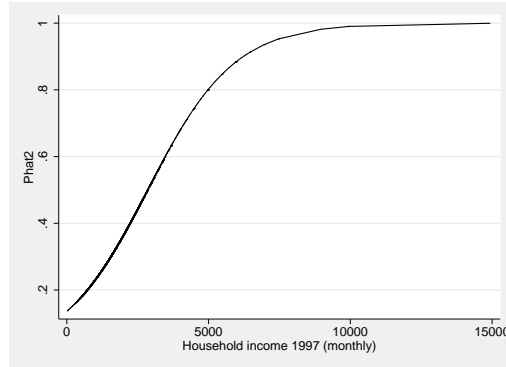
```
. generate Phat2 = exp(_b[_cons]+_b[chhinc]*chhinc)/ (1+exp(_b[_cons]+_b[chhinc]
> ]*chhinc))
```

and display this in a graph:

---

<sup>14</sup>At the mean age the value of the centred age variable is 0. Age is therefore omitted from the equation. The same applies for the variable *east*.

```
. graph twoway line Phat2 hhinc, sort
```



The graph clearly shows that the increase in probabilities is not identical for all income values. Depending on income, the probability of house ownership will rise either rapidly or slowly.

### 0.3.2 The Iteration Block

In the top-left section of the `logit` command output (see page 15) you will find a number of rows beginning with the word `iteration`. These figures are typical for models whose coefficients are determined by the maximum likelihood principle. As mentioned in our excursus on this procedure, when using the maximum likelihood principle there is no mathematical equation which can calculate the  $b$  coefficients. The coefficients are instead identified by repeated and targeted (iterative) testing. In other words, the testing is repeated until the values<sup>15</sup> in the individual iterations practically do not change at all. The interim results of these attempts are displayed in the iterations block.

The first and last figures of the iteration block are in some respects similar to the figures given in the Anova block of the linear regression (section ??). The linear regression Anova block contained figures for TSS, RSS and MSS. TSS was the sum of the squared residuals if one predicted all the values of the dependents variables through arithmetic means. RSS was the sum of the squared residuals from the regression model and MSS was the difference between TSS and RSS. MSS is therefore the amount of fewer errors that we make when using the regression model instead of the mean for predicting the dependent variable.

In the logistic regression model the residuals used to determine the regression coefficient have no formal importance. Instead, the likelihood, a sort of probability, is established for possible regression coefficients, this is then followed by the selection of the respective coefficients which overall are most *probable*. The logarithm of this probability is displayed in the iteration block. Two of these probabilities are of particular

<sup>15</sup>Readers of this excursus should note that the values are those of  $\ln \mathcal{L}$ .

interest, namely the first and the last. The first likelihood shows how probable it is that all  $b$  coefficients of the logistic regression apart from  $b_0$  equal 0 ( $\mathcal{L}_0$ ). The last likelihood shows how likely the set at the end of the selected  $b$  coefficients is ( $\mathcal{L}_K$ ). It is clear that the probability of the  $b$  coefficients selected at the end should at least be slightly larger than the probability that all  $b$  coefficients are 0. The larger the difference between the first and last probabilities, the stronger the advantage of the model with independent variables compared to the “null model”.

All in all, a sort of analogy appears for TSS to  $\mathcal{L}_0$ , for RSS to  $\mathcal{L}_K$  and for MSS to  $\mathcal{L}_0 - \mathcal{L}_K$ .

Apart from the first and last log likelihoods, the rest of the figures in the iteration block are of little interest. There is one exception. Under certain circumstances, the maximum likelihood process delivers a solution for the  $b$  coefficients, which is not optimal. This may occur if the domain within which the coefficients are being searched for is difficult. This all sounds very abstract and we do not wish to delve any further. However, it is important for us to note here that the number of iterations may provide us with indications of a difficult search domain. An exact number after which iterations appear to be too large cannot be given. Nevertheless, you should generally bargain for more iterations if the number of independent variables increases.

In such cases you might change your dataset slightly. Delete a few randomly selected cases and recalculate your model; or marginally change the values of a few variables and recalculate your model. If the results remain constant you can set your mind at rest and interpret the results of your first model. If the results noticeably change then things may be problematic. You should then consult further reading (??).

### 0.3.3 The Model Fit Block

$r^2$  was a significant measured value for the linear regression model fit. The reason for the sizable practical significance of  $r^2$  is presumably because  $r^2$  indicates, on the one hand, the clear boundaries of 0 and 1, and on the other, a clear interpretation of the *share of declared variance*. There is no comparable generally accepted measured value for logistic regression. Instead, a series of measured values and concepts have been suggested, some of which we would like to introduce to you.

We would like to start with the measured value which is already contained in the model fit block of `logit` version: the pseudo- $r^2$  ( $p^2$ ). Nevertheless, it is already a mistake to speak of *the* pseudo- $r^2$ . There are a number of various pseudo- $r^2$  values (??). Therefore, one should always indicate which pseudo- $r^2$  is being referred to. The one issued by Stata is the one suggested by ? which is why we want to refer to it as  $p_{MF}^2$ .

McFadden’s  $p_{MF}^2$  is calculated in a direct analogy to the linear regression  $r^2$ . Subject to the equation (??)  $r^2 = \frac{MSS}{TSS} = 1 - \frac{RSS}{TSS}$ . Correspondingly:

$$p_{MF}^2 = \frac{\ln \mathcal{L}_0 - \ln \mathcal{L}_K}{\ln \mathcal{L}_0} = 1 - \frac{\ln \mathcal{L}_K}{\ln \mathcal{L}_0} \quad (10)$$

where  $\mathcal{L}_0$  is the likelihood that all coefficients apart from the constants are 0 and  $\mathcal{L}_K$  is the likelihood of the calculated models. As is the case in  $r^2$ ,  $p_{MF}^2$  lies within the boundaries of 0 and 1, however the interpretation of the content is disproportionately more problematic. “The higher, the better” is pretty much the only thing that one can say of  $p_{MF}^2$ . In our example (page 15), the value of  $p_{MF}^2$  at around 0.08 is rather small.

The likelihood ratio  $\chi^2$  value ( $\chi_{\mathcal{L}}^2$ ) is an additional indicator of the quality of the overall model besides McFadden’s Pseudo  $r^2$ . It, too, touches on the difference between the first and last figures of the iteration block. However, unlike  $p_{MF}^2$  this difference is not standardized for the values between 0 and 1, instead it is merely multiplied by  $-2$ :

$$\chi_{\mathcal{L}}^2 = -2(\ln \mathcal{L}_0 - \ln \mathcal{L}_K) \quad (11)$$

The reason behind this is that  $\chi_{\mathcal{L}}^2$  undergoes a  $\chi^2$  distribution through this multiplication. In a similar way to the  $F$  value in linear regression, this enables  $\chi_{\mathcal{L}}^2$  to be used for investigating the following hypothesis: “The inclusion of the independent variable in the regression model does not increase the probability of the model” or in other words, “in the population, all coefficients apart from the constants are equal to zero.” The probability of this hypothesis being applicable is also accounted for in the model fit block (“Prob > chi2”). In the case under consideration it is practically 0. Therefore we can assume that at least one of the two  $b$  coefficients in the population is not 0. As is the case for the linear regression  $F$  test, the rejection of the null hypothesis is in *no* way sufficient for us to be satisfied with the results.

Similarly to linear regression, judgment on the suitability of a model should not be based purely on the basis of the measured values within the model fit block. This information should be taken more seriously when working within the framework of logistic regression, as there is no generally accepted measured value such as  $r^2$ . For this reason we would like to explain some further measured values that are not part of the model fit block.

## Classification Tables

The fit of the linear regression model was primarily assessed on the basis of the residuals  $(y - \hat{y})^2$ . In logistic regression one can interpret the residuals as the difference between the real values and the so-called classified values. With classification, every observation is assigned one of the two values of dependent variable. The value 1 is normally assigned when the model predicts a probability of over 0.5, whereas an observation is assigned the value 0 if a probability of under 0.5 is predicted. For example, if carried out manually it looks like this:<sup>16</sup>

<sup>16</sup>The variable *Phat* contains the regression model’s predicted probabilities. We have generated it with `predict Phat` on page 18.

```
. generate ownerhat = Phat >= .5 if Phat < .
```

The classified values generated in this way are typically presented in a classification table. This is a simple cross-classified table with the classified values and the original values:

```
. tabulate ownerhat owner, cell column
```

Key			
	<i>frequency</i>		
	<i>column percentage</i>		
	<i>cell percentage</i>		
ownerhat	owner		Total
	0	1	
0	1,858	829	2,687
	90.77	71.90	83.97
	58.06	25.91	83.97
1	189	324	513
	9.23	28.10	16.03
	5.91	10.12	16.03
Total	2,047	1,153	3,200
	100.00	100.00	100.00
	63.97	36.03	100.00

A number of values can now be read off this table. The sensitivity and the specificity of the model are of particular importance for people in the medical profession. It deals with the column percentage rate of the main diagonal in the above-generated table. Sensitivity is the share of observations classified as residence owners within the observations who are actually residences owners. Specificity is the share of observations classified as tenants among those who are actual tenants.

The so-called count  $r^2$  is commonly used in the social sciences. It deals with the share of overall correctly predicted observations. You can determine these by adding the overall shares in the main diagonal in the above-generated table. However, it is easier to use the command `lstat`, with which you can receive the table in a slightly different order, as well as derive the sensitivity, specificity, count  $r^2$  and a some further figures:

```
. lstat
```

```
Logistic model for owner
```

Classified	True		Total
	D	-D	
+	324	189	513
-	829	1858	2687
Total	1153	2047	3200

```
Classified + if predicted Pr(D) >= .5
```

True D defined as owner != 0		
Sensitivity	Pr( +  D)	28.10%
Specificity	Pr( -  ~D)	90.77%
Positive predictive value	Pr( D  +)	63.16%
Negative predictive value	Pr(~D  -)	69.15%
False + rate for true ~D	Pr( +  ~D)	9.23%
False - rate for true D	Pr( -  D)	71.90%
False + rate for classified +	Pr(~D  +)	36.84%
False - rate for classified -	Pr( D  -)	30.85%
Correctly classified		68.19%

The classification table shows that we have a total of 513 classified as 1. For 324 observations this corresponds to the true value, for 189 it does not. We have assigned the value 0 to 2,687 observations, which turned out to be correct for 1,858 of the observations. In total we correctly classified  $r_{count}^2 = \frac{324+1858}{3200} = 68.19$  percent of the observations. This figure can be found at the end of the `lstat` readout.

As a result, our model does not look too bad. However, please note that you are also able to correctly classify some cases without knowledge of the independent variable. If you only know the distribution of the dependent variable, then you will make fewer errors if you assign all observations to the most frequent category. If we were to predict all the observations in the case under consideration as tenants, then we would already be correct in  $\frac{2047}{3200} = 63.97$  percent of the cases. By comparing the correct classification by means of the marginal distribution with correct classification with the knowledge of the independent variable one can calculate the so-called Adjusted Count  $r^2$  (?, 108):

$$r_{AdjCount}^2 = \frac{\sum_j n_{jj} - \max_c(n_{+c})}{n - \max_c(n_{+c})} \quad (12)$$

where  $n_{+c}$  is the sum of column  $c$  and  $\max_c(n_{+c})$  is the column with the higher value of  $n_{+c}$ .  $\sum_j n_{jj}$  is the sum of cases in the main diagonal of the classification table, i.e. the amount of correctly classified cases. In our example, we receive a  $r_{AdjCount}^2$  of:

```
. display ((324 + 1858) - 2047)/(3200 - 2047)
.11708586
```

This means that with knowledge of the independent variables, errors in prediction drop by 12 percent in comparison to prediction based solely on the basis of marginal distribution. You can receive the Adjusted Count  $r^2$ , as well as other model fit measured values (including AIC and BIC) through the Scott Long and Jeremy Freese's Ado package `fitstat`. The Ado package can be obtained through the SSC-archive (see section ??).

## Pearson Chi Square

A second group of measured values is based on the so-called Pearson residuals. In order to understand these it is necessary to clarify the term covariate pattern. A covariate pattern is defined as every possible combination of a model's independent variables. In our example, this is every possible combination of the values of household income, age and region. Every covariate pattern occurs  $m_j$ -times whereby  $j$  serially numbers every covariate pattern that occurs. Through

```
. predict cpatt, number
. list cpatt
```

you can view all the covariate patterns which occur in our example along with the up-to-date number (variable *cpatt*).

Thus, the Pearson residuals are based on a comparison of the number of successes<sup>17</sup>  $y_j$  within the pattern  $j$  with the predicted number of successes  $m_j\hat{P}_j$  within the same pattern. The Pearson residual looks like this:

$$r_{P(j)} = \frac{(y_j - m_j\hat{P}_j)}{\sqrt{m_j\hat{P}_j(1 - \hat{P}_j)}} \quad (13)$$

whereby  $\hat{P}_j$  stands for the predicted probability of a success for the pattern  $j$ . The multiplication of  $\hat{P}_j$  with the number of cases per pattern  $m_j$  results in the predicted number of successes in pattern  $j$ . As the number of successes in a covariate pattern can only be underestimated or overestimated by the model, there is exactly one residual for every covariate pattern.

```
. predict pres, resid
```

allows you to generate a variable with the Pearson-residual. The sum of the square of this variable over all covariate patterns produces the Pearson Chi Square statistic. A formal test of this statistic can be received by:

```
. lfit
Logistic model for owner, goodness-of-fit test
      number of observations =      3200
  number of covariate patterns =      3183
      Pearson chi2(3179) =      3218.80
      Prob > chi2 =              0.3066
```

The test which forms the basis of this hypothesis is the conformity of predicted and observed frequencies. A high probability for such a  $\chi^2$  value signals small differences between the observed and the estimated frequencies. Inversely, a low probability that

<sup>17</sup>We mean the observations on the dependent variable with the value 1.

the difference between observed and estimated values cannot be explained by random process. Be careful when interpreting this probability as “significance”: a probability of  $\chi^2$  under 0.05 may indicate that the model apparently does not represent reality, however values over 5 percent do not inevitably mean that the model is acceptable. A probability of, shall we say, 6 percent is still fairly small, even though the difference between observed and estimated values is completely random.

The  $\chi^2$  test is less suitable when the number of covariate patterns (here: 3,183) is close to the number of observation contained in the model (here: 3,200). ?, 140–145 have therefore suggested a change to the test which would entail sorting the data by the predicted probabilities and dividing them in  $g$  approximately equally sized groups. This is followed by a comparison for each group of the frequency of the actually observed successes in each group with the frequency estimated by the model. A high probability resulting from the test statistic signals a likewise small difference between the observed and the estimated frequencies.

You may obtain the Hosmer-Lemeshow test by using the command `lfit` together with the option `group()`. The number of groups in which the data should be divided into is entered between the brackets.  $g = 10$  is often used.

```
. lfit, group(10)
```

## 0.4 Diagnostics of Logistic Regression

Logistic regression assumes a linear correlation between logarithmic chance of a success and every independent variable. The coefficients of such a model only represent reality in a meaningful way if this linear assumption is actually applicable. It is therefore necessary to test the validity of this assumption before a linear regression model can be interpreted. This is where the first part of the following diagnostic procedure serves its purpose.

The second part deals with the problem of influential observations. By this, we mean observations that strongly influence the results of a statistical procedure. In influential observations this is usually down to a few small outliers. This is problematic when the conclusions of the content are heavily dependent on these few observations. Occasionally these outliers turn out to be incorrectly entered data, however more often than not these outliers indicate that variables are missing from the model.

### 0.4.1 Linearity

We used the first clue for the discovery of non-linear correlations in the linear regression scatterplots with the dependent variable and all the independent variables, whereby the form of the relationship was clarified with the help of a scatterplot smoother. You can also use the appropriate scatterplots for logistic regression, however you should consider two particularities. *Firstly*, the median trace used within the framework of the

linear regression as a scatterplot smoother is of little use for dichotomous variables, as in practice the median can only take the values of 0 and 1.<sup>18</sup> *Secondly*, the functional form of the scatterplot does not have to be linear, as linearity is only assumed with regards to the logits. The functional form between the probabilities and the independent variable has the shape of an S (see the graph on page 9).

You may use a local mean-regression as the scatterplot smoother instead of the median trace. Here, the  $X$  variable is divided into bands in the same way as for the *median trace* and arithmetic mean of each dependent variable is calculated for each of these bands. These means are then entered against the respective independent variable in a graph.

In order for the functional form of a regression model to become applicable, the graph should show the local mean regression to have the  $S$ -shaped curve of the illustration on page 0.2.1. However, it should be taken into consideration that the illustrations often only depict a small section of the  $S$ -shape. If the band means restrict themselves to between 0.2 and 0.8, then the mean regression should run almost linearly. However,  $U$ -shaped, reverse  $U$ -shaped and other non-continuous curves are certainly problematic.

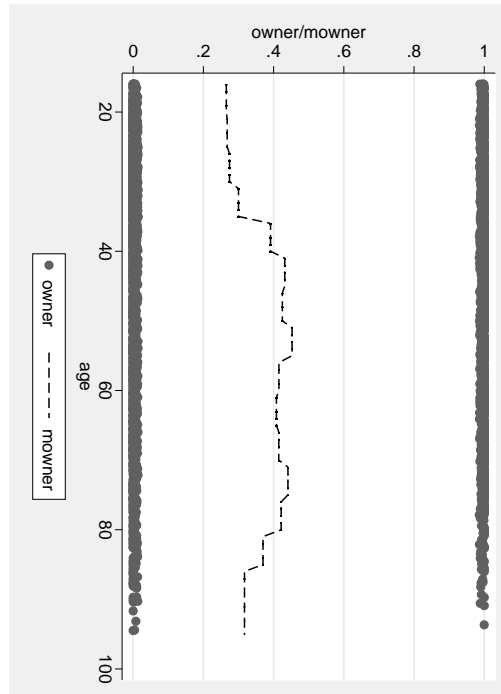
The simple local mean-regression process does not exist as a stand-alone command in Stata, but it can be quickly carried out manually with a simplified version:<sup>19</sup>

```
. generate groupage = autocode(age,15,16,90)
. egen mowner = mean(owner), by(groupage)
. graph twoway (scatter owner age, jitter(2)) (line mowner age, sort)
```

---

<sup>18</sup>The value 0.5 may occur if there is an equal number of 0 and 1 values

<sup>19</sup>For the function `autocode()` see page???. For the command `egen` see section ?? on page ??.

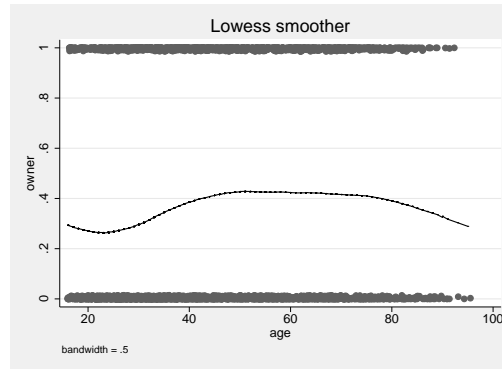


In this graph the mean of residence ownership increases with age, then remains constant and drops with the oldest individuals surveyed. This is referred to as a reverse U-shaped correlation, and certainly does not match the correlation assumed by a logistic regression.

Cleveland's (1979) Locally Weighted Scatterplot Smoother (LOWESS)<sup>20</sup> is better suited to investigating the functional form of correlation. You can receive this smoother through the `twoway`-plottype `lowess` or by the statistical graph command `lowess`. We will forego displaying the calculation of this smoothers, and just refer to the excellent representation of the logic behind LOWESS in `?`. One should only note that the level of smoothing is adjusted through the `bwidth()` option in the form of a number between 0 and 1. High numbers lead to increased smoothing and vice versa. Furthermore, one should note that LOWESS is an intensive calculative process and it may take some time before the following graph is displayed on your screen:

```
. lowess owner age, jitter(2) bwidth(.5)
```

<sup>20</sup>The process has recently also become to be known as *loess*. We use the older term as it corresponds to the name of the Stata plottype.



This graph also displays a reverse U-shaped correlation between residence ownership and age. The middle age groups have a higher probability of residence ownership than the upper and lower age groups. The youngest individuals surveyed, who presumably still live with their parents, also live relatively frequently in their own houses or flats.

Both graphs show a correlation which contradicts the S-shaped correlation required by logistic regression. As is the case with linear regression, U-shaped correlations can be modelled through the generation of polynomials. Nevertheless, before you do this, you should check if the U-shaped correlation is still visible *under control* of the household income. This is carried out by using the logic of the above-demonstrated local mean-regression in the regression model.<sup>21</sup> Here you replace the age variable of your regression model with a set of dummy-variables (cf. page 37 and section ??.) for the grouped version of the age variable on page 26:

```
. tabulate groupage, gen(aged)
. logit owner aged2-aged15 chhinc east
```

By calculating this regression model you will receive a total of 14  $b$  coefficients for the age variable. Every one of these  $b$  coefficients indicates how much higher the logarithmic chance of residence ownership is for the respective age group compared to the youngest surveyed individual. When the correlation between age and (the logarithmic chance of) residence ownership is linear, then the age  $b$  coefficients should increase continuously and steadily. This does not appear to be the case for the coefficients in front of us. This can easily be evaluated by graphically depicting the rise of the  $b$  coefficients. This is done with the following commands:

```
. matrix b = e(b)'
. svmat b, names(b)
```

(The symbol ' in the `matrix` command is a simple quotation mark. It can be

<sup>21</sup>For the following process cf. ?, 90. Alternatively, scatterplots with scatterplot smoothers of the dependent variables can be used against an independent variable for various combinations of the other independent variables (?, 253). A process related to the component plus residual plot (page ??) is demonstrated by ?.

found next to the “a” button on english keyboards. In this case it stands for a specific calculative operation, the so-called transposition. It is explained further down.)

*Explanation:* In Stata, the coefficients of statistical models are stored in a matrix, in the row vector  $\mathbf{e}(\mathbf{b})$  to be precise. This is nothing more than the stored results to which we have repeatedly referred you. Matrixes and vectors are of particular interest as they contain numerous saved results. The vector  $\mathbf{e}(\mathbf{b})$  is, for example, a list of all the coefficients within the regression model. With

```
. matrix list e(b)
```

you can take a closer look at  $\mathbf{e}(\mathbf{b})$ .

Similarly to the saved results, you can also calculate with the matrixes and vectors, which is what the `matrix` commands are for.<sup>22</sup> In the commands above we used a `matrix` command to transpose the row vector  $\mathbf{e}(\mathbf{b})$  into the column vector  $\mathbf{b}$ , i.e. we turned rows into columns and columns into rows. This is important, as the coefficients are next to each other in row vector  $\mathbf{e}(\mathbf{b})$ . In contrast, in the newly generated column vector the coefficients are under one other.

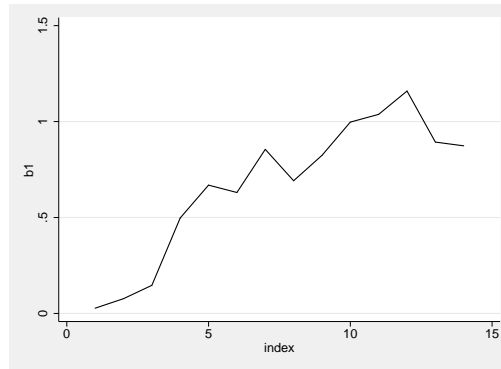
The column vector is therefore nothing more than a list of numbers. This list of numbers is then written as a variable within our dataset with the command `svmat`, whereby the option `names` specifies what the variable will be called. Please note that Stata automatically adds the figure 1 onto the name chosen by you. This occurs because the command can also save matrixes with a number of columns as variables, whereby every column in the matrix becomes a variable.

Once the `svmat` command has been carried out your dataset will contain the new variable `b1`. This variable contains the 14 age  $b$  coefficients, the  $b$  coefficient for household income and the  $b$  coefficient for the constants. The first 14 figures of the variable `b1` are the age coefficients. As the coefficients are sorted by age, a graphical depiction of the coefficients against the case number is sufficient:

```
. generate index = _n
. graph twoway line b1 index in 1/14, sort
```

---

<sup>22</sup>You can receive an overview of the `matrix` commands with the help of `help matrix`.



The graph shows a falling logarithmic chance of residence ownership for the last two age groups. In this respect the slight reverse *U*-shaped correlation remains. The inclusion of a quadratic term for age within a regression model results in a slight (albeit significant) improvement in the model fit. We will discuss this further in section 0.5 on page 34 and on page `pagereftxt:age2`.

## 0.4.2 Influential Cases

Influential data points are observations which heavily influence the  $b$  coefficients of a regression model. As explained on page ??, influential observations are observations that exhibit an unusual combination of values for the  $X$  variable (leverage), as well as an unusual characteristic (given the  $X$  values) of the  $Y$  variable (discrepancy). Correspondingly, the measured value of *Cook's D* is calculated by multiplying leverage and discrepancy.

The use of this concept is somewhat more problematic in logistic regression than it is in linear regression as the measurement of leverage and discrepancy is only approximately possible (?, 459). In Stata, approximation of the leverage values is available through.

```
. logit owner cage chhinc east
. predict leverage, hat
```

Note that you must recalculate the original model with *cage* and *chhinc* as independent variables, since `predict` always refers to the last calculated regression model. The last entered model is the one with dummy-variables for age. After that you can receive the standardized residuals as an approximation to discrepancy through

```
. predict spres, rstandard
```

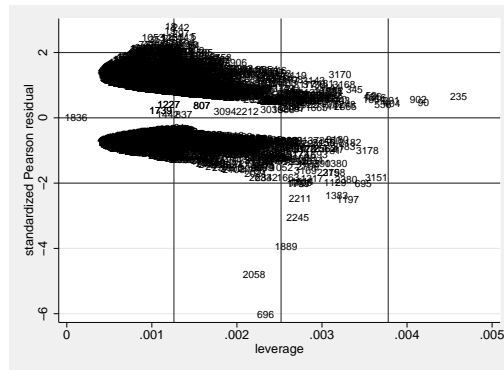
In logistic regression all the standardized residuals for observations are identical within the same combination of independent variables. The same also applies for the leverage values. In order to isolate those covariate patterns which exhibit high leverage

and discrepancy values, one can produce a graph with the standardized residuals compared to the leverage values. `? , 461` uses a diagram with vertical lines on the double and triple values of the mean of the leverage value. To prepare for this graph we first calculate the mean of the variable leverage. We save the mean, as well as its doubled and tripled values in the local macros 'a', 'b' and 'c' (see Chapter ??).

```
. summarize leverage
. local a = r(mean)
. local b = 2 * r(mean)
. local c = 3 * r(mean)
```

Following on from this we generate the graph with the standardized residuals against the leverage-values. In order to generate vertical lines (`xline()`) we turn to the recently defined local macros 'a', 'b' and 'c'. We use the number of covariate patterns as the plot symbol. These are found in the variable `cpatt`, which we generated on page 24:

```
. scatter spres leverage, xline('a' 'b' 'c') yline(-2 0 2) mlabel(cpatt) mlabpo
> s(0) ms(i)
```



Eight covariate patterns in the graph are particularly conspicuous: both patterns with the lowest standardized residuals, and the six patterns with standardized residuals under  $-2$  and leverage values over twice the average. The following command shows that invariably the latter consists of observations from West Germany with comparatively high income that do not exhibit residence ownership

```
. list cpatt owner age hhinc east if leverage > 'b' & spres < -2
```

	cpatt	owner	age	hhinc	east
279.	1889	0	48	6965	0
421.	2245	0	56	5970	0
1127.	1197	0	35	5970	0
2382.	2211	0	55	5326	0
2526.	1795	0	46	4977	0
3005.	1383	0	38	5721	0

The model therefore appears to be unsuitable for explaining cases such as these.

In linear regression the influence of individual observations on the regression result was determined by Cook's D (section ??). This dealt with the multiplication of leverage and discrepancy. An analogue measured value for logistic regression is

$$\Delta\beta \underbrace{\frac{r_{P(j)}^2}{(1-h_j)^2}}_{\text{Discrepance}} \times \underbrace{h_j}_{\text{Leverage}} \quad (14)$$

where  $h_j$  is the value for the leverage. In Stata, you can save this measured value by

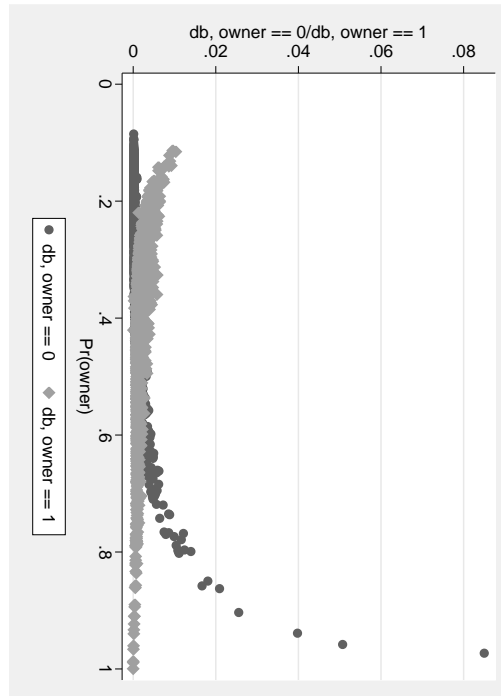
```
. predict db, dbeta
```

as a variable under the name *db*. A scatterplot of  $\Delta\beta$  against the predicted probabilities is generally used as a graphical illustration of  $\Delta\beta$ , whereby observations of success are given different colors or symbols to those of failure. The `separate` command is particularly useful for the latter<sup>23</sup>:

```
. separate db, by(owner)
. graph twoway scatter db0 db1 Phat
```

---

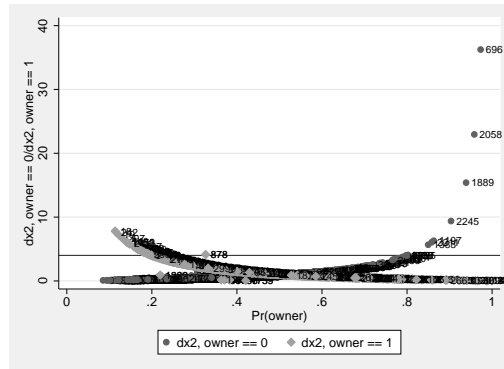
<sup>23</sup>To explain `separate` we would like to point you in the direction of `help separate`. The variable *Phat* was generated on page 18 with `predict Phat`.



In this plot the points from two curves. The curve from the bottom-left to the top-right consists of all tenants, while the curve from the top-left to the bottom-right consists of all residence owners. Several noticeable covariate patterns can be observed for tenants in this illustration, which are predicted by the model with having a high probability of residence ownership. If one enters the number of covariate patterns into the graph instead of the plot symbols then it becomes apparent that these are patterns that were already noticeable in the previous analysis.

The Pearson residuals allow for a further test statistic for influential observations. As was illustrated in section 0.3.3, the sum of the squared Pearson residuals is a measured value for the deviation of the predicted values from the observed values. The contribution of every covariate pattern to this measured value matches the square of the Pearson residual. If one divides this contribution by  $1 - h_j$ , then one gets  $\Delta\chi_{P(j)}^2$ , a measured value which indicates the change in the Pearson-Chi square statistic when the covariate pattern  $j$  is removed from the dataset. The scatterplot of  $\Delta\chi_{P(j)}^2$  against the predicted probabilities is well suited to the discover of answer patterns which would be hard to predict through the model. Here it would be useful to enter Hosmer/Lemeshow's raw threshold value of  $\Delta\chi_{P(j)}^2$  of four into the graph (? , 163):

```
. predict dx2, dx2
. separate dx2, by(owner)
. graph twoway scatter dx20 dx21 Phat, yline(4) mlabel(cpatt cpatt)
```



Once again a number of covariate patterns stand out, and again they are *the usual suspects*: patterns for which residence ownership was incorrectly predicted. If one can eliminate data errors then one should investigate if a variable important to the model was left out. This could be a subgroup that does not at all apply to the correlation between age, household income, region and residence ownership.

## 0.5 Likelihood Ratio Test

Most social sciences datasets are samples of a large population. If one calculates the  $b$  coefficients of a logistic regression for the samples then it is likely that they will differ from the *true* values in the population. One therefore often asks in which band the  $b$  coefficients lie if one takes into account the random sample fluctuations. Frequently this is because one wants to make sure that a coefficient in the population is not 0, i.e. one wants to be sure that an independent variable also has influence on the dependent variable in the population. If this is the case, then it is considered to be a “significant” effect.

The Wald test offers the opportunity to determine the significance of a coefficient. It is displayed in the coefficient block of the logistic regression results (page 15). For the Wald test the coefficient is first divided by an estimate of the standard error (“Std. Err”)<sup>24</sup>. The result of this calculation assumes it follows a normal distribution. Based on the normal distribution probability function, one can evaluate how likely the observed  $b$  coefficient is, if the true value in the population is 0. The advantage of this procedure is that the result is immediately available in the logistic regression results table; you will find it in the coefficient block column labeled “ $P > |z|$ ” (see the results table on page 15). However, the Wald test is said to occasionally incorrectly displays some regression coefficients as insignificant(?, 17)<sup>25</sup>.

The investigation into the significance of a  $b$  coefficient in logistic regression is therefore often followed by the likelihood ratio test. This investigates whether the quality of

<sup>24</sup>cf. ?, 89–93 for calculating the standard error.

<sup>25</sup>Cf. ?, 97 for a somewhat different position.

a regression model is actually improved by an expansion, regardless of which type, or whether this merely complicates the model.

In section 0.3.3 we showed you the calculation of  $\chi_{\mathcal{L}}^2$ . This was a measured value which compares the likelihood of the calculated model with that of a model in which all of the coefficients apart from the constant were set at 0. The larger the difference between the likelihood of our calculated model and that of the null model, the higher the estimate of the significance of our model.

Due to the  $\chi_{\mathcal{L}}^2$  value displayed by the Stata model fit block, the question remains about whether a combination of independent variables achieves an increase in knowledge when compared to a null model. The same logic allows us to rephrase the question and ask: “Does the fit of a model on residence ownership against household income increase if we introduce the additional variable of age?” In order to look at this question one can carry out a calculation analogously to the test of overall model, by using the  $-2$  multiplied difference between the logarithmic likelihood of the model without age ( $\ln \mathcal{L}_{\text{without}}$ ) and the same model with age ( $\ln \mathcal{L}_{\text{with}}$ ):

$$\chi_{\mathcal{L}(\text{Diff})}^2 = -2(\ln \mathcal{L}_{\text{without}} - \ln \mathcal{L}_{\text{with}}) \quad (15)$$

Like  $\chi_{\mathcal{L}}^2$ , this test statistic also follows a  $\chi^2$  distribution, whereby the number of degrees of freedom is the difference in the number of parameters between the two models.

Calculating  $\chi_{\mathcal{L}(\text{Diff})}^2$  with Stata is carried out with the command `lrtest`. In our example we want to investigate the significance of the age effects. In order to do so, the model is first calculated with the variable that is to be investigated:

```
. logit owner cage chhinc east
```

This model is then internally stored with the command `estimates store`. To do so we need to choose a name for the stored model, and we propose the name `full` here:

```
. estimates store full
```

Subsequently, the reduced model is calculated.

```
. logit owner chhinc east
```

Afterward we can use `lrtest` to test the difference between this model and the formerly stored model. Therefore we simply list the name of the stored model (`full`) and the name of the model against which it should be compared. If we do not list a second name the most recent model is used:

```
. lrtest full
likelihood-ratio test                LR chi2(1) =    76.89
(Assumption: . nested in full)      Prob > chi2 =    0.0000
```

The probability of receiving a  $\chi_{\mathcal{L}(\text{Diff})}^2$  value of 76.89 in our sample is very small

when the age coefficient in the population is 0. We can therefore be fairly certain that the age coefficient is not 0. However, this does not reveal anything about the level of influence of age on residence ownership.

When using the likelihood ratio test, one should note that only models that are nested can be compared with one another. This means, in particular, that the full model must contain all the variables of the reduced model. Furthermore, both models must be calculated for the same observations. The latter may be problematic if, for example, some observations in your saturated model have to be excluded due to missings, while they may however be included in the reduced model by leaving out a variable. Remember that in such cases, Stata displays a warning (“observations differ”).

If you are interested in a varied combination of independent variables or in a comparison of models with varying samples, the likelihood ratio test cannot be used. The two information criteria BIC (Bayesian Information Criterion) and AIC (Akaike’s Information Criterion) are often used instead. AIC and BIC are obtained through the `fitstat` command mentioned earlier (page 23). An excellent introduction into the statistical foundations of these indices is provided by ?.

## 0.6 Refined Models

Similarly to the linear regression model, the logistic regression model can also be expanded in various ways to allow for the investigation of complicated causal hypotheses. We would like to discuss three such expansions in the following: the specification of non-linear relationships, the comparison of sub-groups (categorical variables) and the investigation of varying correlations between sub-groups (interaction effects). It should be noted here that the procedures for expanding the model do not differ from those for linear regression. In this respect, we would merely like to show a few examples in depth that have already appeared within the framework of the concepts introduced in linear regression.

### Non-linear relationships

During the diagnosis of our regression model, we came across signs of a U-shaped correlation between age and the logarithmic chance of residence ownership (section 0.4.1). In this respect, U-shaped correlations are only *one* form of non-linear relationships. Logarithmic or hyperbolic relationships also occur relatively frequently. One should note that the model assumption of logistic regression is only violated if these relationships appear between the logits and the independent variables. With respect to probabilities, logarithmic or hyperbolic relationships are to a certain extent already taken into account by the S-shape distribution of the logit transformation.

There are various possibilities of taking into account non-linear relationships. If we have an assumption as to *why* residence ownership is rarer for old people than it is for middle-aged people, then it is best to incorporate the respective variable into the

regression model. If, for example, one suspects that the observed decline is a consequence of the *migration due to old-age care reasons*, then it would perhaps be sensible to introduce dummy-variables for living in one's own residence or that of one's children, in nursing homes, old people's homes, etc. If the suspicion were accurate then the non-linear curve in multiple models would disappear.

If the variables that could explain the non-linear curve are not available, then the non-linear course cannot be directly taken into account. One possibility would be to group the respective independent variable and to introduce it into the model as a set of dummy-variables. We talked about a strategy such as this on page 28. A more economical possibility would be to use transformations or the polynomials of the independent variables. The rules for linear regression are applicable here too: in the case of hyperbolic relationships the  $X$  variable is squared and in the case of logarithmic relationships the  $X$  variable is logarithmised. For U-shaped relationships we use the squared  $X$  variable in addition to the original variable.

In order to model the U-shaped relationship between residence ownership and age one would proceed as follows:

```
. generate cage2 = cage^2
(140 missing values generated)
. logit owner cage cage2 chhinc east, nolog
```

Logit estimates		Number of obs	=	3200
		LR chi2(4)	=	332.12
		Prob > chi2	=	0.0000
		Pseudo R2	=	0.0794
Log likelihood = -1925.4516				

owner	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
cage	.0209879	.0024164	8.69	0.000	.016252	.0257239
cage2	-.0002469	.0001196	-2.06	0.039	-.0004814	-.0000124
chhinc	.0006378	.0000422	15.10	0.000	.000555	.0007206
east	-.0675665	.0866436	-0.78	0.435	-.2373849	.1022519
_cons	-.5198484	.0608729	-8.54	0.000	-.6391571	-.4005397

It would be best to display the results of this regression model with a conditional effects plot (section ??).

## Categorical Independent Variables

Categorical variables are assimilated into logistic regression in the same way as in linear regression (section ??). This means that a set of dummy-variables is generated from a categorical variable and is then introduced into the model with the omission of a *reference category*.

Here we would like to continue our investigation into the Titanic catastrophe (section 0.1). The question we asked was whether the seafaring principle of "women and children first" was put into practice or whether there is any truth in the assumption made by

the film ‘Titanic’, where the first-class gentlemen took the places in the lifeboats at the expense of the third-class women and children.

We previously established that women and children evidently really did have better chances of survival than men did (and adults respectively). We want to look into more closely and for this, we will use the original dataset:

```
. use titanic, clear
```

This dataset contains dichotomous variables for age (adults vs. children), sex and survival, as well as a categorical variable for first class passengers (1), second class passengers (2), third class passengers (3) and crew (4).

The assumption made by the film ‘Titanic’ is that besides gender and age, (passenger) class was also a criterion for a place in the lifeboats. Verification of this speculation can be carried out with a logistic regression model of survival against age, sex and class. In order to assimilate the independent variable *class* into the regression model, it first has to be transferred into a set of dummy-variables. We do this with:<sup>26</sup>

```
. tabulate class, gen(class)
```

This command generates four dummy-variables with the names *class1* to *class4*. You can now use them in your regression model. As is the case in linear regression, you have to however use one of the variables as the reference category. This variable is then not included in the model. In this case, we want to use the first class passengers as the reference category.

```
. logit survived age sex class2-class4, nolog
Logit estimates                               Number of obs   =    2201
                                                LR chi2(5)      =    559.40
                                                Prob > chi2     =    0.0000
Log likelihood = -1105.0306                    Pseudo R2      =    0.2020
```

survived	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	-1.061542	.2440257	-4.35	0.000	-1.539824	-.5832608
sex	-2.42006	.1404101	-17.24	0.000	-2.695259	-2.144862
class2	-1.018095	.1959976	-5.19	0.000	-1.402243	-.6339468
class3	-1.777762	.1715666	-10.36	0.000	-2.114027	-1.441498
class4	-.8576762	.1573389	-5.45	0.000	-1.166055	-.5492976
_cons	3.10538	.2981829	10.41	0.000	2.520952	3.689808

According to the sign, it appears that the survival chance for adults was smaller than those of the children and that the survival chance for the men was smaller than those of the women. So far for the principle of “women and children first”. However, at the same time it becomes apparent that the first class passengers have the largest chance of survival, as in comparison the rest all had less chance of survival. The third class passengers had the smallest chance of survival, in fact their chance of survival were even

<sup>26</sup>An alternative would be the command `xi`, see section ??.

smaller than those of the crew. In conclusion, one can state that women and children were indeed favored for rescue, but that gender and age were not the only criteria for a place in the lifeboats.

Let us imagine for one second that these data were a sample of the Titanic's passengers. If this was the case, then we would certainly ask ourselves, if it was possible that we would derive the coefficients from the class dummies, if *in reality* there was no correlation whatsoever between survival and class. To answer this, we use likelihood ratio test mentioned above. To do this we first save the last calculated model as a saturated model,

```
. estimates store full
```

and then calculate the respective model without the class-dummies:

```
. logit survived age sex
```

A comparison of both models with the likelihood ratio test shows that it is highly unlikely that the class variable has no influence whatsoever on the population.

```
. lrtest full
likelihood-ratio test                LR chi2(3) =    119.03
(Assumption: . nested in full)      Prob > chi2 =    0.0000
```

## Interaction Effects

The logistic regression model calculated in the previous section shows one more weakness. It assumes that the sex plays the same role for adults and children. However, when putting the principle of “women and children first” into practice, children should be preferentially treated, regardless of their sex. Sex should primarily be used for adults as a criterion for a place in the lifeboats.

If this is transferred to the logic of regression model then it means that the coefficient for sex should be smaller for children than it is for adults. Sex should only be a criterion for a place in the lifeboats for adults or in other words: the effect of sex on survival varies with age. Effects of independent variables that vary between sub-groups are called interaction effects.

The modelling of interaction effects in logistic regressions matches those in linear regression models. Multiplying the variables involved in the interaction effect generates interaction terms:

```
. use titanic, clear
. tabulate class, gen(class)
. generate menage = sex * age
```

After this has been done, the interaction terms can be assimilated into the model:

```
. logit survived sex age menage class2-class4, nolog
Logit estimates                               Number of obs =      2201
                                                LR chi2(6)      =      577.41
                                                Prob > chi2     =      0.0000
Log likelihood = -1096.0213                    Pseudo R2      =      0.2085
```

survived	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
sex	-.7150863	.406223	-1.76	0.078	-1.511269 .0810961
age	.1099979	.335319	0.33	0.743	-.5472153 .7672111
menage	-1.902104	.4330925	-4.39	0.000	-2.750949 -1.053258
class2	-1.033786	.1998153	-5.17	0.000	-1.425417 -.6421551
class3	-1.810499	.1759416	-10.29	0.000	-2.155338 -1.46566
class4	-.8033246	.1598088	-5.03	0.000	-1.116544 -.4901051
_cons	2.071621	.3528719	5.87	0.000	1.380005 2.763237

The only coefficients of interest here are those of *sex*, *age* and *menage*. The effects of *sex* and *age* reproduce the effect for passengers for whom the interaction variable is 0. The effect of *sex* therefore indicates how much lower the logarithmic chance of survival is for male children compared to female children. It shows that male children had a lower chance of survival than female children did. The same interpretation is applied to the effect of *age*. Age increases the chance of surviving for women; therefore, adult women had a greater chance of surviving than girls.

The interaction effect indicates how much the influence of *sex* changes when one considers adults instead of children. If male children already had a  $-.72$  smaller logarithmic chance of survival than female children, then this would yield a  $-.72 + (-1.90) = -2.62$  smaller log-chance of survival for male adult compared to a female adult. Therefore the survival chance of men was only around fourteenth ( $e^{-2.62} = 0.07 = 1/14$ ) of that of the women.

## 0.7 Advanced techniques

Stata allows numerous related models to be calculated in addition to logistic regression. Unfortunately, there is not enough space in this book in order to show them in detail. However, in this section we would like to describe the fundamental ideas behind the most important procedures. For further information we will specifically refer you to the entry in the alphabetically sorted Stata Reference Manual [R] allocated to the respective command. There you will also find further bibliographies on the individual processes.

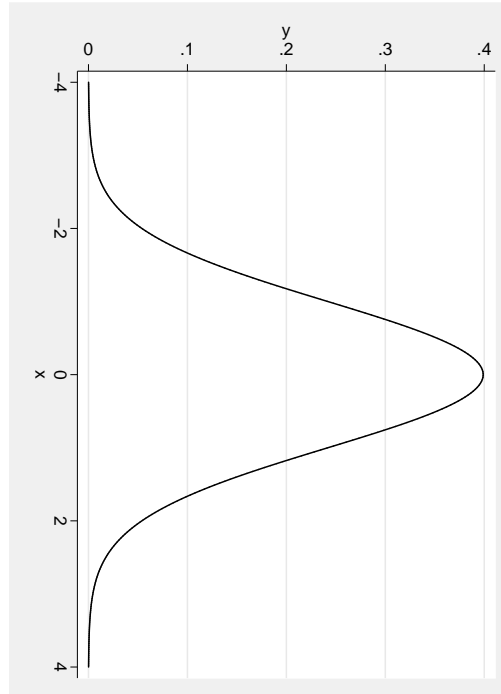
### 0.7.1 Probit Models

In the logistic regression model, we attempted to predict the probability of a success through a linear combination of one or more independent variables. In order to be sure that the predicted probabilities remained between the limits of 0 and 1, the probability of the success underwent a logit transformation. However, the logit transformation is

not the only possibility for achieving this. An alternative is the probit transformation used in probit models.

In order to get some idea of this transformation, one should first visualise the density function of a standard distribution:

```
. graph twoway function y = 1/sqrt(2*_pi) * exp(-.5 * x^2), range(-4 4)
```



You can interpret this graph in the same way as a histogram or a kernel density estimator (section ??), i.e. in this variable the values around 0 occur most often and the larger or smaller they become, the rarer they are.

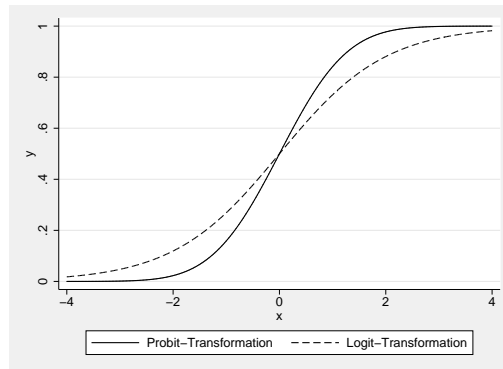
Suppose you randomly selected an observation from the variable  $X$ . How large would the probability be of selecting an observation that had a value of less than  $-2$ ? As values under  $-2$  do not occur very often in the  $X$  variable, the intuitive answer is: not very probably. If you want to know the exact answer then you can determine the probability through distribution function tables for standard normal distribution or through the Stata command

```
. display normprob(-2)
.02275013
```

The probability of selecting an observation with a value of less than or equal to  $-2$  from a standard normal variate would accordingly be 0.023. You can repeat the same

calculation for any value of  $X$ . Do this and enter the calculated probabilities against the values of  $X$  in a scatterplot. This results in the distribution function for the standard normal distribution  $\Phi$  depicted in the following graph:

```
. tw (function y = normprob(x), range(-4 4)) (function y = exp(x)/(1+exp(x)), r
> ange(-4 4)), legend(lab(1 "Probit-Transformation") lab(2 "Logit-Transformatio
> n"))
```



The function shows a S-shaped curve, similar to the probabilities assigned to the logits, which we have also included in the graph.

With the distribution function for the standard normal distribution, one can transform values between  $-\infty$  and  $+\infty$  into values between 0 and 1. Correspondingly, the inverse of the distribution function for the standard normal distribution ( $\Phi^{-1}$ ) converts probabilities between 0 and 1 for a success ( $P(Y = 1)$ ) into values between  $-\infty$  and  $+\infty$ , which is similar to the logit transformation. The values of this probit transformation are also suitable as dependent variables of a linear model.<sup>27</sup> This is the probit model:

$$\Phi^{-1}[\hat{P}(Y = 1)] = b_0 + b_1x_{1i} + b_2x_{2i} + \dots + b_{K-1}x_{K-1,i} \quad (16)$$

Estimating the  $b$  coefficients of this model is done through the maximum likelihood principle. Interpretation of the coefficients is along the lines of the known regression models. However, in each the value of the inverse distribution function of the normal distribution increases by  $b$  units. Through the distribution function for the standard normal distribution, one can calculate the changes in the probabilities of success. Usually, the predicted probabilities of probit models are, to large extent, identical to those of logit models, and the coefficients generally have a 0.58-fold value of that of the logit models coefficients (?, 49). In this respect, the probit model is rather an alternative to logistic regression than a continuous.

<sup>27</sup>You can think of them as z-scores.

The Stata command used to calculate probit models is `probit`. The syntax of the command is the same as all the model commands in Stata. You may calculate the last calculated logit model (page 40) as a probit model as an example:

```
. probit survived sex age menage class2-class4, nolog
Probit estimates                    Number of obs   =      2201
                                   LR chi2(6)        =      573.48
                                   Prob > chi2       =      0.0000
Log likelihood = -1097.988          Pseudo R2      =      0.2071
```

survived	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
sex	-.4554407	.2533275	-1.80	0.072	-.9519534	.041072
age	.084439	.2076534	0.41	0.684	-.3225543	.4914322
menage	-1.102542	.2673721	-4.12	0.000	-1.626582	-.5785024
class2	-.6327645	.1193993	-5.30	0.000	-.8667829	-.3987462
class3	-1.02884	.0999088	-10.30	0.000	-1.224657	-.8330219
class4	-.5055305	.0962533	-5.25	0.000	-.6941836	-.3168775
_cons	1.223367	.2160143	5.66	0.000	.7999871	1.646748

## 0.7.2 Multinomial Logistic Regression

Multinomial logistic regression is used when the dependent variable exhibits more than two categories that cannot be ranked. An example for this would be party preference with values for the German parties CDU, SPD and for any other parties.

The main problem with using multinomial logistic regression is in the interpretation of the coefficients and this will be the focus point of this section. Nevertheless, in order gain understanding of this problem it is essential to acquire at least an intuitive idea of the statistical fundamentals of the process. These fundamentals will be discussed shortly (? , cf.).

In multinomial logistic regression, one predicts the probability for every value of the dependent variable. One could initially calculate a binary<sup>28</sup> logistic regression for every value of the dependent variable. In our example, we could calculate three separate logistic regressions: a logistic regression with the dependent variable CDU against non-CDU, logistic regression with the dependent variable SPD vs. non-SPD and finally a logistic regression with the dependent variable for the other parties against the non-other parties:

<sup>28</sup>In order to differentiate it from multinomial logistic regression, we call the logistic regression of a dichotomous dependent variable as a binary logistic regression.

$$\begin{aligned}
\ln \frac{P(Y = \text{CDU})}{P(Y = \text{non-CDU})} &= b_0^{(1)} + b_1^{(1)}x_{1i} + b_2^{(1)}x_{2i} + \dots + b_{K-1}^{(1)}x_{K-1,i} \\
\ln \frac{P(Y = \text{SPD})}{P(Y = \text{non-SPD})} &= b_0^{(2)} + b_1^{(2)}x_{1i} + b_2^{(2)}x_{2i} + \dots + b_{K-1}^{(2)}x_{K-1,i} \\
\ln \frac{P(Y = \text{other})}{P(Y = \text{non-other})} &= b_0^{(3)} + b_1^{(3)}x_{1i} + b_2^{(3)}x_{2i} + \dots + b_{K-1}^{(3)}x_{K-1,i}
\end{aligned} \tag{17}$$

The superscript in brackets means that the  $b$  coefficients differ between the individual regression equations:  $b_k^{(1)} \neq b_k^{(2)} \neq b_k^{(3)}$ . To simplify the notation we refer to  $b_1^{(1)} \dots b_{K-1}^{(1)}$  as  $\mathbf{b}^{(1)}$  and refer to the sets of  $b$  coefficients from the other two equations as  $\mathbf{b}^{(2)}$  and  $\mathbf{b}^{(3)}$  respectively.

Every one of the unconnected regressions allows for a calculation of a predicted probability of every value of the dependent variable. However, these predicted probabilities do *not* all add up to 1. However, this should be the case, as one of the three possibilities, *SPD*, *CDU* or *Other*, *must*<sup>29</sup> be present.

For this reason would it appear to be sensible to jointly define  $\mathbf{b}^{(1)}$ ,  $\mathbf{b}^{(2)}$  and  $\mathbf{b}^{(3)}$  and to adhere to the rule which states that the predicted probabilities must add up to 1. Unfortunately, a model which would joint define  $\mathbf{b}^{(1)}$ ,  $\mathbf{b}^{(2)}$  and  $\mathbf{b}^{(3)}$  is not definitely solvable. In order to do so, the number of coefficients one is required to calculate must be reduced. This is done by setting some of the coefficients at an arbitrary value. Usually in multinomial logistic regression, the coefficients in one of the equations is set at 0, i.e. one could set  $\mathbf{b}^{(1)} = 0$ . The model is now identified and the remaining coefficients can be calculated using the maximum likelihood principle. However, one has to take into account the zero setting of  $\mathbf{b}^{(1)}$  when interpreting  $\mathbf{b}^{(2)}$  and  $\mathbf{b}^{(3)}$ . This is the reason for the above-mentioned difficulty in interpreting the coefficients.

Let us show you an example of interpreting coefficients. For this please load *data1.dta*

```
. use data1, clear
(SOEP'97 (Kohler/Kreuter))
```

and generate a new variable for partisan choice with values for the CDU, the SPD and the other parties from the original variable for party preferences (*np9402*). One way of doing this is:

```
. generate party = np9402
(1965 missing values generated)
. recode party 2 3 =1 1=2 4/8 = 3
(part: 1375 changes made)
. label define party 1 "CDU" 2 "SPD" 3 "Other"
. label value party party
```

---

<sup>29</sup>In this case, we are disregard the possibility of no party preference. If one does not then would have to calculate a further regression model for this alternative. The predicted probabilities for the four regression should then add up to 1.

This creates the variable *party* with the value 1 for the CDU/CSU, the value 2 for the SPD and the value 3 for the other parties. Interviewees without a party preference have a missing value.

The Stata command for multinomial logistic regression is `mlogit`. The syntax for the command is the same as the syntax of all model commands, i.e. the dependent variable follows the command and is in turn followed by the list of independent variables. With the `base()` option, you can select the equation for whom the *b* coefficients are set at 0.

Let us calculate a multinomial logistic regression for party preference against education (in years of education) and the year of birth. In this case, the *b* coefficients of the equation for the CDU are set at 0:

```
. mlogit party yedu ybirth, base(1) nolog
Multinomial logistic regression           Number of obs =      1360
                                           LR chi2(4)      =      92.69
                                           Prob > chi2     =      0.0000
                                           Pseudo R2      =      0.0325

Log likelihood = -1379.4301
```

	party	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
SPD	yedu	-.0039571	.0255271	-0.16	0.877	-.0539892	.0460751
	ybirth	.0126934	.0034126	3.72	0.000	.0060047	.019382
	_cons	-24.54483	6.627974	-3.70	0.000	-37.53542	-11.55424
Other	yedu	.1305466	.0295313	4.42	0.000	.0726663	.188427
	ybirth	.0352889	.0046591	7.57	0.000	.0261573	.0444206
	_cons	-71.04625	9.092251	-7.81	0.000	-88.86673	-53.22576

(Outcome party==CDU is the comparison group)

The readout of the command is similar to that of the binary logistic regression. In contrast to binary logistic regression, the coefficient block is split into two parts. The upper part contains the coefficients of the equation for the SPD, while the lower part contains the coefficients of the equation for the other parties. The coefficients of the equation of the CDU were set at 0 and are therefore not displayed.

As a result of setting  $b^{(CDU)} = 0$  one can interpret the coefficients of the other two equations in relation to the CDU supporters. By this we mean that coefficients in the equation for the SPD indicate how much the logarithmic chance of preferring the SPD *and not the CDU* changes when the independent variables increase by one unit. The equation for the other parties indicates changes in the logarithmic chance of preferring the other parties *and not the CDU*.

The interpretation of the coefficients for a multinomial logistic regression is problematic in as much that the sign interpretation cannot be used. The negative sign for length of education in the SPD equation does not necessarily mean that the probability

of a preference for the SPD declines with education. In our regression model this can be demonstrated with the example of the coefficient for the variable *ybirth* from the equation for the SPD. Let us shorten the probability of preferring the SPD to  $P_{SPD}$  and the probability of preferring the CDU to  $P_{CDU}$ . The above-mentioned  $b$  coefficient can be written as

$$b_{ybirth}^{(SPD)} = \ln \left( \frac{\hat{P}_{SPD|ybirth+1}}{\hat{P}_{CDU|ybirth+1}} \right) - \ln \left( \frac{\hat{P}_{SPD|ybirth}}{\hat{P}_{CDU|ybirth}} \right) \\ \ln \left( \frac{\hat{P}_{SPD|ybirth+1}}{\hat{P}_{SPD|ybirth}} \times \frac{\hat{P}_{CDU|ybirth}}{\hat{P}_{CDU|ybirth+1}} \right). \quad (18)$$

Through this, the  $b$  coefficient for year of birth in the equation for the SPD is, on the one hand, dependent on the change in probability of SPD preference with the year of birth. On the other hand, it is also dependent on the respective change in probability for choosing the CDU. In contrast to the binary logit model, in the multinomial logit model the change in the probability of CDU preference is not completely dependent on the change in the probability of SPD preference. In this respect, the  $b$  coefficient is solely, mainly or partly dependent on the probability relationship in the base category.

In order to avoid misinterpretations of the multinomial logistic regression, we recommend you use the conditional effects plot for the predicted probabilities.<sup>30</sup> Most of the time, these can be quickly generated, providing that there aren't any high demands required for the optical design of the graph.

For this, you should first generate the predicted probabilities of the model with `predict`. As there is a predicted probability for every value of the dependent variable, you will have to provide three variable names for the predicted probabilities.

```
. predict PCDU PSPD POther
```

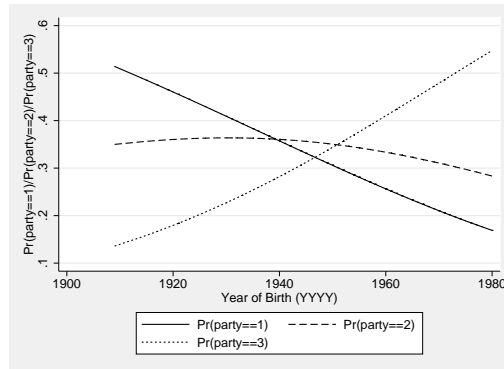
In order to illustrate the effect of the year of birth, these variables are entered against the year of birth, whereby length of education is fixed at a particular value.

If you fix length of education at the highest value (18 years), you can establish that the probability of preferring the SPD declines with the year of birth, despite the regression model indicating a (“significant”) positive effect for the year of birth.<sup>31</sup>

```
. graph twoway line PCDU PSPD POther ybirth if yedu==18, sort
```

<sup>30</sup>One alternative is the *method of recycled predictions* which is described in [R] `mlogit`. A further alternative is the calculation of marginal effects with the command `mfxf compute`.

<sup>31</sup>In the legend, `Pr(partei==1)` stands for the CDU, `Pr(partei==2)` stands for the SPD and `Pr(partei==3)` stands for the other parties.



Above we said that conditional effects plot can easily produced “most of the time”. Problems occurs when too many variables are included in the model and as a result the fixing of the values of the independent variables means there are too few observations left over for a sensible plot. The method of recycled predictions described in [R] `mlogit` appears to be good way of solving this problem. An even more powerful possibility for producing conditional effects plots is given by means of the command `prgen`, which is down loadable from the SSC-archive (see section ??) and more fully described by ?.

### 0.7.3 Models for ordinal data

Models for ordinal data are used when the dependent variable indicates an ordinal scale level. This means that the dependent variable has more than two values that can be ranked, whereby the size of the difference between the two is of no importance. An example would be the question regarding concerns about the increase of crime, which one could answer with “no concerns”, “moderate concerns” or “strong concerns”. In a dataset these could be allocated the values of 0, 1 and 2. One could equally use the values of 0, 10 and 12. A statistical model should therefore produce the same results for both codings.

In principle, there are two strategies available for modelling ordinal dependent variables. The first uses multinomial logistic regression, whereby certain constraints are imposed upon the coefficients (stereotype model). The second strategy is generalization of binary logistic regression for variables with more than two values (proportional odds model). ? discusses the implementation prerequisites for both models.

The logic behind the stereotype model is easily explained. In multinomial logistic regression, every value of the dependent variable acquires its own coefficients. This is why the length of education in the regression model on page 45 had a negative effect on the chance of preferring the SPD (and not the CDU), and at the same time have a positive effect on the chance of preferring another party (and not the CDU). If the dependent variable indicates the presence of ranking then one would normally not expect a directional change in the effects. For example, consider the variable for concerns about the increase of crime (`np9506`). This variable contains the values 1 for

no concerns, 2 for moderate concerns and 3 for strong concerns. Please calculate a multinomial logistic regression for this variable against the length of education. Before you do this, you should, however, mirror the variable *np9506* so that high values stand for strong concerns and vice versa:

```
. generate worries = 4 - np9506
. mlogit worries yedu, base(1)
```

You will get a coefficient of around  $-.05$  in the equation for moderate concerns and  $-.11$  in the equation for strong concerns. The direction of the effects does not change here. One would say that this comes as little surprise, as if education reduces the chance of having moderate concerns (and not no concerns), then education should also reduce the chance of having strong concerns (and not no concerns). However, if one calculates a multinomial logistic regression, then this assumption is ignored. Nevertheless, one can include such assumptions in the model by imposing so-called constraints on the  $b$  coefficients.

Through constraints one can establish certain structures for the  $b$  coefficients before calculating a model. One could, for example, establish that education reduces the chance of having moderate concerns (and not no concerns) to the same extent as it does for having strong concerns (and not moderate concerns). In this case, the coefficient of education for strong concerns would have to be exactly twice as large as the coefficient of education for moderate concerns. With the `constraint` command you can set this structure for the `mlogit` command. With

```
. constraint define 1 [3]yedu = 2*[2]yedu
```

you define the constraint nr. 1, which states that coefficient of the variable *yedu* in the third equation should be exactly twice as large as the coefficient of the variable *yedu* in the second equation. You can use constraint nr. 1 through the `constraints` option of the `mlogit` command. Here you would enter the number of the constraint you wish to use in the parentheses.

```
. mlogit worries yedu, base(1) constraints(1)
```

If you calculate this model, you will discover that it is almost identical to the previous model. However, it is far more economical, as in principle only one education coefficient has to be calculated. The other coefficient is derived from the ordinal structure of the dependent variable and our assumption that education proportionately increases concerns.

Establishing specific constraints which take into account the ordinal structure of the dependent variable is one way of modeling the ordinal dependent variable. Nevertheless, the constraint is just one example of numerous alternatives.

A different approach is followed by the proportional odds model. In the proportional odds model the value of the ordinal variable is understood as being the result of the categorisation of an underlying metric variable. With regards to our example, one

could assume that answers in the *worries* variable only provide a rough indication of the attitudes towards the increase in crime. The attitude of people probably varies between infinitely numerous concerns and no concerns whatsoever, and can take any value in between. Through the default answers no, moderate and strong concerns, the variable is already categorised during the interview, whereby we don't know exactly with which attitude values ( $\kappa_1$ ,  $\kappa_2$ ) we should undertake the categorisation. All that we can say is that our observed ordinal variable ( $Y$ ) receives the value  $k = 1$ , when the unknown attitude  $E$  plus a measuring error  $u$  lies below the unknown boundary  $\kappa_1$ . The observed ordinal variable acquires the value  $k = 2$  when it lies between the unknown boundaries  $\kappa_1$  and  $\kappa_2$ , and the value  $k = 3$  when the attitude lies over  $\kappa_2$ .

You should now remind yourself of the predicted values ( $\hat{L}$ ) of the binary logistic regression. These values lay between  $-\infty$  and  $+\infty$ . In this respect, one could understand these predicted values as being the unknown metric attitude  $E$ . If one were to know the value of  $\kappa_k$ , by assuming a specific distribution for measuring error  $u$ , one could determine the probability for each observation, with which  $Y$  one would obtain the value  $k$ . In a proportional odds model one estimates the values of  $\kappa_k$ , as well as a linear combination of independent variables for  $Y$ .

An example may clarify this. The command for the *proportional odds* model in Stata is `ologit`. The syntax of the command is the same as all other model commands: The dependent variable follows the command and is in turn followed by the list of independent variables. For this, we will calculate the same model as above:

```
. ologit worries yedu
```

The predicted value of this model for interviewees with ten years of education is  $S_{10} = -0.068 \times 10 = -0.68$ . The value for  $\kappa_1$  and  $\kappa_2$  are provided underneath the coefficient block. The probability that interviewees with a predicted value of  $-0.68$  are classified as individuals with moderate concerns matches the probability of  $-0.68 + u_j \leq -1.196$ . Or in other words, the probability of  $u_j \leq -1.128$ . If one assumes that the error is the result of logistic distribution, then the probability is  $1/(1 + e^{-1.128})0.76$ .

## 0.8 Summary

`logit y x1 x2` Calculates a logistic regression of the dependent variable  $y$  on the independent variables  $x1$  and  $x2$ .

`logit y x1 x2, or` Calculates a logistic regression of the dependent variable  $y$  on the independent variables  $x1$  and  $x2$ . The *odds ratio* is listed in the results table.

`logistic y x1 x2` Identical to `logit y x1 x2, or`.

`predict Phat` Saves the predicted probabilities of the last regression model in a new variable called *Phat*. The name of the new variable is user-defined.

`predict statvar, statistic` Saves the values of a selected statistic in the new variable *statvar*. The name of the new variable is user-defined.

**lfit** Calculates the Pearson  $\chi^2$  test.

**lstat** Calculates the classification table.

The following statistics are available in connection with logistic regression as an option of **predict**:

**Xb** predicted logits

**deviance** deviance residuals

**resid** Pearson residuals

**rstandard** standardized Pearson residuals

**dx2** Hosmer-Lemeshow goodness-of-fit statistic

**dbeta** Pregibons Delta-Beta (goodness-of-fit statistic)

**number** Continuous covariate pattern number

## References



## **Author index**



# Subject index

- $\Delta\beta$ , 283–284
- $\Delta\chi^2$ , 284
- $\beta$ , see regression (stand. coefficient)
- `*`, see do-files (comments)
- `+`, see operators
- `//`, see do-files (comments)
- `#delimit`, 37–40
- `&`, see operators
- `_N`, 89–90, 102
- `_all`, 51
- `_b[ ]`, 193
- `_merge` (variable), 322, 324
- `_n`, 88–89, 102
- `^`, see operators
- `~`, see operators
- `|`, see operators
- `||`, 110
  
- Academic Technology Service, see ATS
- added-variable plot, 214–216
- additive index, 85
- adjusted count  $r^2$ , 275
- adjusted  $r^2$  = adjusted  $r^2$ , 202–203
- ado-directories, 379
- ado-files
  - basics, 351–353
  - programming, 353–370
- aggregate, see collapse
- AIC, 275
- Akaike’s information criterion, see AIC
- Aldrich-Nelson’s  $p^2$ , 340
- alphanumeric variables, see strings
- Analysis of Variance, see regression
- `angle()` (axis-labels sub-option), 132
- ANOVA, see regression
- Anscombe quartet, 206–208, 233–234
- anylist, 68–69
  
- `append`, 330–333
- arithmetic mean, see average
- arithmetical expressions, see expressions
- `ascategory` (dot option), 157–158
- ASCII files, 303–312
- ATS, 374
- augmented component-plus-residual plot,
  - 211
- `autocode()` (function), 161
- autocorrelation, see regression (autocorrelation)
- average, 16, 162, 165
- `avplots`, 214–215
- `aweight` (weighting type), 73–74
- axis
  - labels, 114, 130–132
  - scales, 124–126
  - titles, 114, 133–135
  - transformations, 125–126
  
- balanced panel data, 324, 326
- `bands()` (mbands option), 210
- `bar` (graph type), 111, 115, 155–156, 168
- bar charts, 155–156, 168
- batch jobs, see do-files
- Bayesian information criterion, see BIC
- Bernoulli’s distribution, see binomial distribution
- beta, see regression (stand. coefficient)
- bias, 207–208
- BIC, 275
- `bin()` (histogram-option), 174
- binary variables, see variables (dummy)
- binomial distribution, 262–263
- BLUE, see Gauss-Markov assumptions
- Bookstore, 373

- bootstrap**, 237
- box** (graph type), 111, 171–172
- box plots, 22–23, 171–172
- Box-Cox transformation, 234
- bcskew0**, 234
- branch, 366–367
- break, *see* commands (break)
- browse**, 302
- by prefix, 18–19, 64–66, 88–92, 102, 165–166
- by()** (graph option), 138–139
- by()** (tabstat-option), 166
- bysort**, 65
- byte (Storage-Type), 108
  
- calculator, *see* pocket calculator
- caption()** (graph option), 136–137
- capture**, 42
- categories, 146–147
- cd**, 10
- center, *see* variables (center)
- chi-square
  - likelihood-ratio, 151
  - Pearson, 151
- classification tables, 273–275
- clock position, 128–129
- cluster samples, 238–239
- cmdlog**, 34–36
- CMYK, 118
- CNEF, 330
- coefficient of determination, 197–198
- collapse**, 88
- comma separated values, *see* spreadsheet format
- command + ., *see* commands (break)
- command line, *see* windows (command)
- commands
  - abbreviations, 14, 50
  - access previous, 14
  - break, 14
  - e-class, 77–78
  - end of commands, 37–40
  - external, 50, 353, 374
  - internal, 50, 353, 374
  - long, *see* do-files (line breaks)
  - r-class, 77–78
  - search, 24
- comments, *see* do-files (comments)
- component-plus-residual plot, 211–212
- compound quotes, 364–365
- compress**, 333
- compute, *see* generate
- cond()** (function), 364
- conditional-effects plot, 229–231
- conditions, *see* if qualifier
- confidence interval, 235–236
- connect()** (scatter option), 119–122
- connected** (plotype), 119–122
- contingency table, *see* frequency table (two-way)
- contract**, 73
- Cook's-D, 216–220
- cooksd** (predict-option), 217
- correlation
  - coefficient, 186–187
  - negative, 186
  - positive, 186
  - weak, 186
- count  $r^2$ , 274–275
- covariate, *see* variables (independent)
- covariate pattern, 276
- cplot**, 187
- cprplot**, 211–212
- Cramer's V, 151
- cross-tabs, *see* tables
- Ctrl + Break, *see* commands (break)
- Ctrl + C, *see* commands (break)
- cumulated probability function, *see* probit model
- CVS, *see* spreadsheet format
  
- data editor, 312–314
- data matrix, 302–303
- data package, 3
- data region, 113
- data types, *see* storage type
- datasets
  - ASCII files, 305–312
  - combine, 317–318, 320–333
  - describe, 11–12

- export, 334
- hierarchical, 91, 327–330
- import, 304–305
- load, 11
- non-machine-readable, 312–317
- oversized, 336–337
- panel data, 241
- preserve, 71
- rectangular, 302
- reshape, 241–245
- restore, 71
- save, 29, 333–334
- sort, 15
- titanic, 254
- dates
  - combining datasets, 316
  - combining datasets, 317
  - elapsed dates, 100–101
  - from strings, 101
- degrees of freedom, *see* `df`
- delete, *see* erase
- density, 173, 175–177
- describe, 9, 11–12
- destring, 96–97
- `df`, 197
- $DF\beta$ , 213–214
- `dfbeta`, 213–214
- dictionary, 309–312
- `dir`, 10, 28
- directory
  - change, 10
  - contents, 10–11
  - working directory, 4, 10–11, 63
- discard, 353
- discrepancy, 218, 281
- `discrete` (histogram option), 152–153
- `display`, 59, 343–344
- distributions
  - describe, 145–183
  - grouped, 159–162
- `do`, 27
- do-files
  - analyzing, 43–44
  - basics, 26–28
  - comments, 37
  - create, 43–44
  - editors, 26–27
  - error messages, 28
  - execute, 27–28
  - exit, 42–43
  - from interactive work, 31–36
  - line breaks, 37–40
  - master, 44–47
  - organization, 43–47
  - set more off, 40–41
  - version control, 40
- `doedit`, 26, 32
- `dot` (graph type), 111, 157–158, 168–170
- dot charts, 157–158, 168–170
- double (Storage-Type), 108
- `drop`, 12, 51
- dummy-variables, *see* variables (dummy)
- `e()` (saved results), 77–79
- `e(b)` (saved result), 280
- e-class, *see* commands (e-class)
- `edit`, 312–314
- `egen`, 93–95
- EMF, 142–143
- Encapsulated PostScript, *see* EPS
- `encode`, 96–97
- endogenous variable, *see* variables (dependent)
- enhanced metafile, *see* EMF
- Epanechnikov kernel, 177
- EPS, 142–143
- erase, 47, 322
- `ereturn list`, 78–79
- error components model, 249–251
- error messages
  - ignore, 42
  - invalid syntax, 17
- Excel files, 303–304
- `exit` (in Do-Files), 42–43
- exit Stata, 28–29
- exogenous variable, *see* variables (independent)
- `expand`, 73
- export, *see* datasets (export)

- expressions, 59–62
- extensions, 63
- F-Test, 198–199
- FAQ, 24, 373
- fence, 171–172
- filenames, 62–63
- findit**, 379–380
- Fisher’s exact test, 151
- five-number-summary, 165, 171
- fixed effects model, 245–249
- fixed format, 309–312
- float()** (function), 108
- float** (Storage-Type), 108
- foreach**, 66–69, 325–326
- forvalues**, 69–70
- free format, 307–309
- frequencies
  - absolute, 147–148
  - conditional, 150–151
  - relative, 147–148
- frequency tables, 20–21
  - one-way, 147–148
  - two-way, 149–152
- frequency-weights, *see* weights
- function** (plottype), 111
- functions, 61–62
- fweight** (weighting type), 71–73
- fxsize()** (graph option), 140–141
- fysize()** (graph option), 140–141
- gamma coefficient, 151
- Gauss curve, *see* normal distribution
- Gauss distribution, *see* normal distribution
- Gauss-Markov assumptions, 206
- GEE, 251
- gen()** (tabulate-Option), 226–227
- generalized estimation equations, *see* GEE
- generate**, 25, 81–92
- generate()** (tab option), 155–156
- gladder** (statistical graph), 113
- global, 343
- graph, 109–143
- graph region, 113, 123
- graphs
  - 3D, 113
  - combining, 139–141
  - connecting points, 120–121
  - elements, 113–115
  - export, 142–143
  - multiple, 137–141
  - overlay, 137–138
  - print, 141–142
  - titles, 136–137
  - types, 111–113
  - weights, 216
- grid lines, 126, 132
- grouping
  - by quantiles, 160
  - intervals with arbitrary width, 161–162
  - intervals with same width, 160–161
- GSOEP, 4, 12, 103, 105, 238, 240–241, 318–320, 330
- help, 23–25
- help-files, 370
- histogram**, 152–154, 173–175
- histogram** (plottype), 111
- homoscedasticity, *see* regression (homoscedasticity)
- Hosmer-Lemeshow test, 277
- Huber/White/sandwich estimator, 223–224
- if** (branch), 366–367
- if qualifier, 17–18, 56–59
- imargin()** (graph combine option), 140
- importieren, *see* data (import)
- in qualifier, 15, 55–56
- index()** (function), 97–98
- infile**, 308–312
- influential cases, 212–221, 281–284
- input**, 314–315
- inputting data, 312–317
- insheet**, 306–307
- inspect**, 146
- interaction terms, 227–229, 289–290
- invnorm()** (function), 84

- `iscale()` (graph combine option), 140
- iteration block, 271–272
- `kdensity`, 175–180
- Kendall’s tau-b, 151
- kernel density estimator, 175–180
- key variable, 321–322
- `label data`, 333–334
- labels
  - and values, 107
  - datasets, 333–334
  - display, 107
  - values, 21–22, 105–106
  - variables, 21, 105
- legend, 114, 135–136
- leverage, 218, 281
- `lfit` (plotype), 137
- likelihood, 263–264
- likelihood ratio test, 284–286, 289
- Likelihood-Ratio  $\chi^2$ , 273
- limits, 9–10
- `line` (plotype), 115, 119–122
- linear combination, 189
- linear probability model, *see* regression (LPM)
- linear regression, *see* regression
- linearity assumption, 209–212, 277–281
- `list`, 13–14
- `local`, 79, 341–346
- local macros, *see* macros
- local mean regression, 278
- loess, *see* LOWESS
- `log()` (function), 83–84
- `log` (scale suboption), 125–126
- log-files
  - finish recording, 42
  - interrupt recording, 35
  - log commands, 33–36
  - smcl, 41
  - start recording, 41–42
- logarithm, 83–84
- logical expressions, *see* expressions
- `logistic`, 269
- logistic regression
  - coefficients, 267–271
  - command, 266–267
  - dependent variable, 258–262
  - diagnostic, 277–284
  - estimation, 262–265
  - fit, 272–277
  - marginal effect, 296
- `logit`, 266–267
- logit-model, *see* logistic regression
- logits, 260–261
- loops
  - foreach, 66–69
  - forvalues, 69–70
- `lower()` (function), 97
- LOWESS, 211, 278–279
- `lowess` (plotype), 279
- `lowess` (statistical graph), 279
- LPM, *see* regression (LPM)
- macro
  - extended macro functions, 367–368
  - global, 343
  - local, 79–80, 341–346
- manuals, 4–5
- `margin()` (graph option), 123–124
- marker
  - colors, 118
  - labels, 113–114, 127–129
  - options, 116
  - sizes, 119
  - symbols, 113, 116–118
- master data, 321
- match, *see* datasets (combine)
- `matrix` (command), 280
- `matrix` (graph type), 111, 209–210, 212–213
- maximum, 16, 164–165
- maximum likelihood
  - principle, 262–265
  - search domain, 272
- `mbands` (plot-type), 210
- mean, *see* average
- median, 164–165
- median regression, 220, 239–240
- median-trace, 210

- memory, *see* RAM
- merge**, 320–330
- meta data, 326
- minimum, 16, 164–165
- missing
  - encode, 104
- missing** (tabulate-option), 148
- missing values, *see* missings
- missings
  - coding, 316
  - definition, 13
  - in expressions, 58–59
  - set, 18, 103–104
- ML, *see* maximum-likelihood
- mlabel()** (scatter option), 127–129
- mlabposition()** (scatter option), 128–129
- mlabsize()** (scatter option), 128–129
- mlabvposition()** (scatter option), 129
- MLE, *see* maximum-likelihood
- mlogit**, 295
- more off, 40–41
- MSS, 196–197
- multicollinearity, 221, 226
- multinomial logistic regression, 293–297
- mvdecode**, 18, 103–104
- mvencode**, 104
  
- neqany()** (egen-function), 94
- net install**, 377–378
- NetCourses, 373–374
- newlist, 68
- non-linear relationships, 232, 287–288
- normal distribution
  - density, 291
  - density function, 291–292
- note()** (graph option), 136–137
- notes**, 104
- null model, 272
- numlabel**, 107
- numlist**, 62, 68
  
- observations
  - definition, 12
  - list, 13–14
  
- odds, 258–260
- odds-ratio, 259–260, 268–269
- odds-ratio interpretation, 268–269
- OLS, 189–191
- operators, 59–61
- options, 19–20, 54–55
- order**, 333
- ordered logistic regression, *see* proportional odds model
- ordinal logit model, *see* proportional odds model
- ordinary least squares, *see* OLS
  
- package description, 378–379
- panel data, *see* datasets (panel data)
- partial correlation, *see* regression (standard. coefficient)
- partial regression plot, *see* added-variable plot
- partial residual plot, *see* comp.-plus-residual plot
- PDF, 142–143
- Pearson residual, 276–277
- Pearson- $\chi^2$ , 276–277
- percentiles, *see* quantiles
- PICT, 142–143
- pie** (graph type), 111, 156–157, 168
- pie charts, 156–157, 168
- plot region, 113, 123–124
- plotregion()** (graph option), 123–124
- PNG, 142–143
- pocket calculator, 59
- portable document format, *see* PDF
- PostScript, *see* EPS
- ppfad.dta**, 326
- predict**, 194
- predicted values, 193–194
- Pregibons  $\delta\beta$ , *see*  $\delta\beta$
- preserve**, 71
- probability interpretation, 269, 271
- probit**, 293
- probit model, 291–293
- program define**, 346–351
- program drop**, 348
- programs

- and do-files, 347–348
- debugging, 349, 359
- define, 346–347
- in do-files, 349–351
- naming, 348
- redefine, 348
- syntax, 356–360, 363–365
- syntax checks, 365–367
- proportional odds model, 298–299
- PS, 142–143
- pseudo  $r^2$ , 272–273
- PSID, 318, 330
- pwd, 28, 63
- pweight (weighting type), 74–75
- Q-Q plots, 182–183
- quantile plot, 180–182
- quantile regression, 239–240
- quantiles, 163–165
- quartiles, 164–165
- quietly, 362
- r, see correlation coefficient
- r() (saved results), 77–79
- r(max) (saved result), 77–78
- r(mean) (saved result), 77–78
- r(mean) (saved result), 69
- r(min) (saved result), 77
- r(N) (saved result), 77–78
- r(sd) (saved result), 77–78
- r(sum) (saved result), 77
- r(sum\_w) (saved result), 77
- r(Var) (saved result), 77–78
- r-class, see commands (r-class)
- r2 =  $r^2$ , 197–198
- RAM, 9–10, 334–336
- random effects model, 250–251
- random numbers, 68, 84
- range() (scale suboption), 124–125
- RAW, 305
- raw data, see RAW
- recode, see variables (replace)
- recode() (function), 160–161
- recode, 92–93
- reference lines, 113–114, 126–127
- regress, 26, 191–192
- regression
  - ANOVA-Table, 195–197
  - autocorrelation, 224
  - coefficient, 192–194, 201–202
  - command, 191–192, 200–201
  - control, 204–206
  - diagnostic, 206–224
  - fit, 197–199
  - homoscedasticity, 222–224
  - linear, 26, 185–252
  - LPM, 254–257
  - multiple, 199–200
  - non-linear relationships, 231–234
  - omitted variables, 221
  - panel data, 240–251
  - residuals, 194–195
  - simple, 188–191
  - standard error, 236
  - standard. coefficient, 203–204
  - with heteroscedasticity, 234–235
- replace, 25, 81–92
- reshape, 243–245
- resid (predict-option), 194–195
- residual
  - definition, 188
  - sum, 190–191, 196
- Residual Sum of Squares, see RSS
- residual-vs.-fitted plot, 208–209, 222–223, 234–235
- response variable, see variables (dependent)
- restore, 71
- results window, see windows (result)
- return list, 78
- reverse (scale suboption), 125–126
- review, see windows (review-window)
- RGB, 118
- rmiss() (egen-function), 94
- robust, 223–224
- Root MSE, 198
- round() (function), 223
- RSS, 196
- rstud (predict option), 223
- running counter, 88–89

- running sum, 90–91
- rvfplot, 208–209
- sampling-weights, *see* weights
- SAS files, 303–305
- save, 29, 333–334
- saved results, 69, 77–80, 193–194, 280
- scatter (plottype), 111, 115
- scatterplot, 185–186
- scatterplot matrix, 209–210, 212–213
- scatterplot smoother, 210
- sensitivity, 274–275
- separate, 182–183
- sign interpretation, 268
- SJ, 24, 373
- SJ-Ados, 376–378
- SMCL, 41, 370
- SOEP, *see* GSOEP
- sort, 15
- sort (scatter option), 121–122
- specificity, 274–275
- spreadsheet format, 305–307
- SPSS files, 303–305
- ssc install, 378
- SSC-Ados, 378
- SSC-IDEAS, 378
- standard deviation, 16, 162
- Stata Journal, 373
- Stata Technical Bulletin, 373
- Stata-Press, 373
- stata.toc, 378–379
- Statalist, 373
- statistical inference, 235
- STB, 24, 373
- STB-Ados, 376–378
- stereotype model, 297–298
- storage types, 95–96, 107–108
- strings, 308–309, 316
  - display format, 102
  - in expressions, 97
  - replace substrings, 98–99
  - storage type, 95–96
  - to dates, 101
  - to numeric, 96–97
- subinstr() (function), 99
- subscripts, 90–92
- substr() (function), 98–99
- subtitle() (graph option), 136–137
- sum() (function), 90–91
- summarize, 16–17, 164–165
- summarize() (tabulate-option), 166
- summary graphs, 168–170
- summary tables, 166–168
- superposition, 157–158, 169–170
- survey data, 238–239
- svmat, 280
- svy, 238–239
- symmetry plot, 222
- symplot (statistical graph), 222
- syntax, 356–360, 363–365
- syntax diagram, 49
- system files, 303–304
- tab separated values, *see* spreadsheet format
- tab1, 148
- tab2, 151
- table, 166–168
- tabstat, 165
- tabulate, 147–152
- tau-b, *see* Kenndall’s tau-b
- tempvar, 368–370
- text() (twoway option), 129
- textbox options, 134–135
- tick lines, 114, 133
- title() (graph option), 136–137
- total sum of squares, *see* TSS
- total variance, *see* TSS
- trace, 349
- TSS, 195–196
- two-way table, *see* frequency table (two-way)
- twoway (graph type), 111
- U-shaped relationship, 233–234
- unbalanced panel data, 326
- uniform() (function), 68, 84
- update, 374–375
- updating Stata, 374–375
- upper() (function), 97

- use, 11
- using, 62–63
- using data, 321
  
- V, *see* Cramer's V
- value labels, *see* labels (values)
- value**label** (axis-labels sub-option), 132
- variable list, *see* variables (varlist)
- variable window, *see* windows (variables)
- variables
  - \_all, 51
  - allowed names, 83
  - categorical, 225
  - center, 35, 69, 78, 202, 228–229
  - definition, 12
  - delete, 12, 51
  - dependent, 185
  - dummy, 84–85, 155–156, 201, 225–227, 288–289
  - generate, 25, 81–104
  - group, 159–162
  - identifier, 315–316
  - independent, 185
  - multiple codings, 317
  - names, 104–105
  - ordinal, 297
  - replace, 25, 81–104
  - temporary, 368–370
  - transformations, 113, 212, 220–221, 223, 231–235
  - varlist, 14, 50–53
- variance, *see* standard deviation
- variance of residuals, *see* RSS
- variation, 196
- varlist, *see* variables (varlist)
- version**, 40
- view, 41
  
- weights, 70–75
- whisker, 171–172
- wildcards, 52
- windows
  - change, 27
  - command window, 8
  - font sizes, 8
  - graph window, 23
  - preferences, 8
  - result window, 8
  - review window, 8, 14
  - scroll back, 11
  - variables window, 8
- windows metafile, *see* WMF
- WMF, 142–143
- working directory, *see* directory (working directory)
- wstata.exe, 374
  
- xi:, 227
- x**label**() (twoway option), 130–132
- x**line**() (twoway option), 126–127
- x**tick**() (twoway option), 133
- x**scale**() (twoway option), 124–126
- x**size**() (graph option), 123
- xt-commands, 241
- xtgee, 251
- xt**ick**() (twoway option), 133
- xt**itle**() (twoway option), 133–135
- xt**reg**, 248–249, 251
  
- y**label**() (twoway option), 130–132
- y**line**() (twoway option), 126–127
- y**tick**() (twoway option), 133
- y**scale**() (twoway option), 124–126
- y**size**() (graph option), 123
- y**tick**() (twoway option), 133
- y**title**() (twoway option), 133–135
  
- zip archive, 3