9 Regression models for categorical dependent variables

Researchers in the social sciences often deal with categorical dependent variables, whose values may be dichotomous (e.g., rented apartment, yes or no), nominal (party identification: CDU, SPD, or 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 by describing a procedure for dealing with dichotomous dependent variables: logistic regression.

Logistic regression is, for the most part, similar to linear regression, so we will explain it as an analogy to the previous chapter. If you have no previous experience or knowledge of linear regression, we would advise you to first read chapter 8 up to page 186.

As in linear regression, in logistic regression the dependent variables are predicted by a linear combination of independent variables:

$$b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_{K-1} x_{K-1,i}$$

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 $b_1, b_2, \ldots, b_{K-1}$ represent the weights assigned to the variables.

Notice, however, that we did not say that the dependent variable y_i is equal to that linear combination. In contrast to linear regression, in logistic regression you must consider a particular transformation of the dependent variable. Why such a transformation is required and why linear regression is inappropriate are explained in section 9.1, while the transformation itself is explained in section 9.2.1.

Section 9.2.2 explains the method by which the logistic regression coefficients are determined. As this explanation is slightly more difficult and is not required to understand logistic regression, you can skip it for now.

Calculating a logistic regression with Stata is explained in section 9.3. Then we discuss methods of verifying the basic assumptions of the model in section 9.4. The procedure for verifying the joint significance of the coefficients is discussed in section 9.5, while section 9.6 demonstrates a few possibilities for refining the modeling of correlations.

For an overview of further procedures, in particular procedures for categorical variables with more than two values, see section 9.7.

As with linear regression (chapter 8), you will need to do some additional reading if you want to understand the techniques we describe fully. We suggest that you read Hosmer and Lemeshow (2000) and Long (1997).

9.1 The linear probability model

Why is linear regression not suitable for categorical dependent variables? Imagine that you are employed by an international ship safety regulatory agency and are assigned 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 this investigation, we have provided you with data on the sinking of the Titanic.[1] Open the file by typing[2]

```
. use titanic2, clear
```

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

```
. describe
. tab1 _all
```

You will discover that the file contains details on the age (`age2`), gender (`sex`), and passenger class (`class`) of the Titanic's passengers, as well as whether or not they survived the catastrophe (`survived`).

In order to clarify the disadvantages of using linear regression with 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 look like where the Y variable represents the variable for survival and X variable represents age? You may want to sketch this scatterplot yourself.

Note 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, the number of points on the 0-line should increase in relation to those on the 1-line the further to the right you go. To check whether your chart is correct, type

[1] The data provided are real. The dataset and its exact description can be found at *http://amstat.org/publications/jse/archive.htm*. For teaching purposes, 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. Among the files you installed in section 0.2 is the do-file we used to change the dataset, (`crtitanic2.do`), as well as the original dataset (`titanic.dta`).

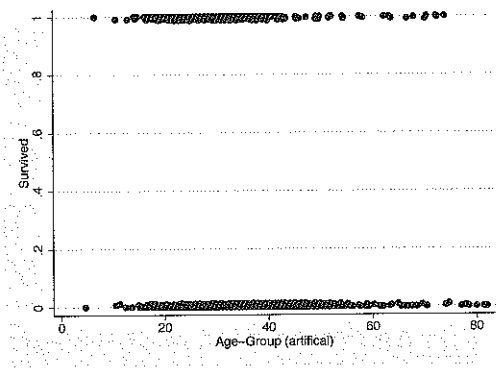
[2] Please make sure that your working directory is `c:\data\kk`; see page 9.

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

This diagram is not particularly informative, as the plot symbols are often directly marked on top of each other, hiding the number of data points.

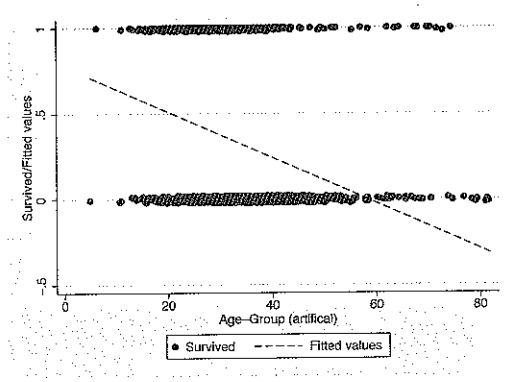
With the help of the `scatter` option `jitter()`, you can produce a more informative diagram. `jitter()` adds a small random number to each data point, thus showing points that were previously hidden under other points. Within the brackets is a number between 1 and 30 that controls the size of the random number; you should generally use small numbers if possible.

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



On examining the chart, you might get the impression that there is 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 8.1.2):

```
. 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, you might say, for example, that the survival of a 5-year-old averages around 0.7. This is a less-than-convincing interpretation if you considers that passengers can only survive or not survive; they cannot survive just a little bit.

However, the predicted value of the dichotomous dependent variable can also be interpreted in a different way. You need to understand what the arithmetic mean of a dichotomous variable with the values of 0 and 1 signifies. The mean of the variable `survived`, for example, is 0.3230. This reflects the share of passengers survived.[3] So, we see that 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 is 0.32. In general, the predicted values of the linear regression are estimates of the conditional mean of the dependent variable. Thus you can 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 LPM (Aldrich and Nelson 1984).

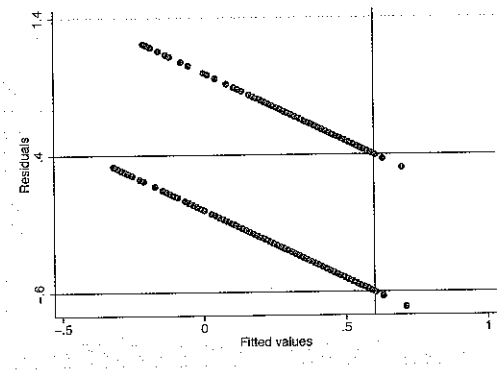
How can you interpret the negative predicted values for passengers over 60 with the help of the probability interpretation? In fact, you cannot, as according the mathematical definition of probabilities, they should be between 0 and 1. Given sufficient small or large values of the X variable, a model 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.[4]

The second problem affects the homoskedastic assumption of linear regression that we introduced in section 8.3.2. According to this assumption, the variance of errors for all values of $\widehat{Y}$ should be constant. We suggested that the scatterplot of the residuals against the predicted values indicated a possible violation of this assumption. You can achieve a graph such as this for our linear probability model by typing

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

[3] You can confirm this for yourself by typing `tab survived`.

[4] In practice, this is a problem of little importance when predicted values of over 1 or under 0 do not appear for real values of the independent variables. However, using a model that would prevent such impossible probabilities from the start seems sensible.



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 her age, she 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 $1-\widehat{y}_i$ or $-\widehat{y}_i$. The variance of the residuals is $\widehat{y}_i \times (1-\widehat{y}_i)$ and is therefore larger as the predicted values approach 0.5. The residuals of the linear probability model are therefore by definition heteroskedastic, so that the standard errors of the coefficients will be wrong.

In conclusion, although a linear regression with a dichotomous dependent variable is possible, it leads to two problems. First, not all predicted values can be interpreted, and second, this model does not allow for correct statistical inference. To avoid these problems, you need a model that 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.

9.2 Basic concepts

9.2.1 Odds, log odds, and 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 because we attempted to represent values between 0 and 1 with a straight line. The values calculated with a linear regression[5] 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, which we will explain using the Titanic data from the previous section.

[5] We showed you a linear combination like this on page 245.

We previously received indications that children had a higher chance of survival than adults did. Now we want to investigate whether women were more likely to survive than men. You can obtain an 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
frequency *row percentage*

Gender	Survived no	yes	Total
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 7.2.1, we interpreted tables like this using row or column percentages. By using the available row percentages, we 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, you can do a similar comparison by dividing the number of survivors by the number of dead. For the women this would be 344:126.

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

You will get the same number[6] if you divide the proportional values (in this case, the row percentages)

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

You can interpret these ratios as follows: for women, the probability of surviving is almost 3 times as high as the probability of dying. In other words, the probability of dying is around one-third ($1:2.73 = 0.366$) the probability of surviving. In practice, we would say that the odds of surviving are generally around 2.73 to 1, while the odds of dying lie at around 1 to 2.73.

[6]The deviations are due to roundoff error.

In general, this relationship can be written as

$$\text{odds}_{\text{surviving}} = \frac{\text{Probability}_{\text{surviving}}}{\text{Probability}_{\text{dying}}} \tag{9.1}$$

or slightly shorter by using symbols instead of text:

$$\text{odds} = \frac{P(Y = 1)}{1 - P(Y = 1)} \tag{9.2}$$

The probabilities of survival $P(Y = 1)$ and dying $P(Y = 0)$ can be found, respectively, in the numerator and the denominator. Since the only two alternatives are surviving or dying, their probabilities sum to one, so we replace $P(Y = 0)$ with $1 - P(Y = 1)$.

You can also calculate the chance of survival for men: their odds of survival are considerably lower than those of the women: $367/1364 = .269$. This means that for men the probability that they will survive stands at $0.269 : 1$; 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 using a measured value. For instance, you can compare the chance of survival for men with that of women by dividing the odds for men by the odds for women:

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

This relationship is called the "odds ratio".

In this case, we would say that the odds of survival for a man are 0.099 times, or ten times lower, than those of a woman. Apparently, the principle of "women and children first" appears to have been adhered to. Whether this *appearance* actually holds is something that we will investigate in more detail on page 278.

However, first we should consider the suitability of using odds for our statistical model. In the previous section, we looked at the probabilities of surviving the Titanic catastrophe by passenger age. We found 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?

(*Continued on next page*)

Table 9.1: Probabilities, odds, and logits

$P(Y=1)$	odds $= \frac{P(Y=1)}{1-P(Y=1)}$		ln(odds)
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 9.1, we list 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. If we presume that the values represent the chance of survival for passengers on the Titanic of different ages, 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. Using (9.2), you can calculate the odds that an individual within each of these 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, one person out of one hundred should have survived: in other terms, a ratio of one to ninety-nine (1:99). If you calculate 1/99, you get the value 0.010101. You can perform this calculation for each row in the table. The values of the odds lie between 0 and $+\infty$; odds of 0 occur if there are no survivors within a specific group, while odds $+\infty$ occur when nearly everyone in a large group survives. If the number of survivors is equal to the number of victims, we get odds of 1.

Odds are therefore *slightly* better suited than probabilities for use as dependent variables in a regression model; 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. You can avoid this problem by using the (natural) logarithm of the odds. These values, called *logits*, are displayed 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 that stretch to $-\infty$

as you approach 0. The logarithm of numbers over 1 stretches towards $+\infty$. Note also the symmetry of the values. At a probability of 0.5 the odds lie at 1:1 or 50:50. The logarithmic value lies at 0. If you look at the probabilities above and below 0.5, 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 can therefore be represented by a linear combination of variables and hence is better suited for use as a dependent variable. Unfortunately, the logit is not always easy to interpret. Your employers are unlikely to understand you if you tell 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, by simply transforming (9.2), you can convert the values of the logits back into probabilities

$$P(Y = 1) = \frac{e^L}{1 + e^L} \tag{9.3}$$

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

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

In this graph, we see another interesting characteristic of logits: while the range of values of the logits has 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 you approach the boundary values of the probabilities, the less the probabilities change. In other words, the probabilities asymptotically approach the values 0 and 1, but they *never* go over the boundaries. From this we can deduce that on the basis of a linear combination, predicted logits can always be converted into probabilities within the permitted boundaries of 0 and 1.

In the introduction we said that a logistic regression uses a linear combination of variables to predict the outcome, but we did not specify what the dependent variable was. However, now we are in a position to do so:

$$\widehat{L}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_{K-1} x_{K-1,i} \tag{9.4}$$

This is called the 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) increase by b units.

Before we use logistic regression, let's examine the procedure for determining the b coefficients of (9.4). In the case of the linear regression, we used the OLS procedure for estimation. For the logistic regression, we instead use the process of maximum likelihood. The logic of this process is somewhat more complex than that of OLS, even

though the basic principle is similar: you look for the b coefficients that are optimal in a certain respect. We will explain this process in detail in the following section. However, you do not need to work through the example to understand the section that follows it!

9.2.2 Excursion: The maximum likelihood principle

In discussing linear regression, we explained the OLS process used to determine the b coefficients. In principle, you could calculate the logarithmic odds for each combination of the independent variables and use these in an OLS regression model. Nevertheless, for reasons we will not explain here, a procedure such as this is not as *efficient* as the process of estimation applied in logistic regression: the maximum likelihood principle.[7] Using this technique, you can determine the b coefficients so that the proportionate values you observed become maximally probable. What does this mean? Before we can answer this question, we need to make a little detour:

On page 248, we informed you that 32.3 percent of the Titanic passengers survived. Suppose that you had determined this figure from a sample of the passengers. In this case, you could ask yourself how likely such a percentage may be, when the *true* number of the survivors amounts to, say, 60 percent, of the passengers? To answer this question, imagine that you have selected a single passenger from the population. If 60 percent of the passengers survived, the probability that this passenger will be a survivor is 0.6 and the probability that he or she will be a victim is 0.4. Now select a second person from the population. Whether this person is a survivor or a victim, the probabilities remain the same (sampling with replacement).

In figure 9.1, we have conducted all possible samples with three observations. We obtained $2^n = 2^3 = 8$ samples with a size of $n = 3$. In the first sample, we only observed survivors (S). The probability that a sample randomly selects 3 survivors is $0.6 \times 0.6 \times 0.6 = 0.6^3 = 0.216$. In the second, third, and fifth samples, we observed two survivors and one victim (V). Each of these samples has probability $0.6 \times 0.6 \times 0.4 = 0.6^2 \times 0.4^1 = 0.144$. In total, the probability of such a sample is $0.144 \times 3 = 0.432$.

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

[7] Andreß, Hagenaars, and Kühnel (1997, 40–45) introduction to the maximum likelihood principle served as a model for the following section.

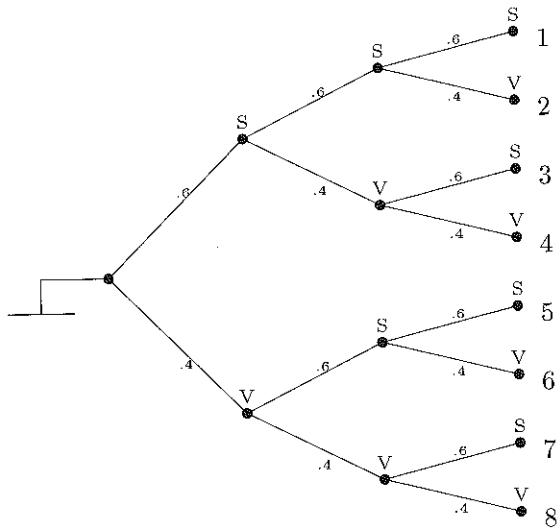


Figure 9.1: Sample of a dichotomous characteristic with the size of 3

Generally, the probability of observing h successes in a sample of size n is

$$P(h|\pi, n) = \binom{n}{h}\pi^h(1-\pi)^{n-h} \tag{9.5}$$

where π defines the probability of a positive outcome in the population. The term $\binom{n}{h}$ stands for $n!/h!(n-h)!$. It enables us to calculate the number of potential samples in which the dichotomous characteristic appears n times. In Stata, the probability of 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, we are usually not interested in this figure; instead our attention is on π, the characteristic's share in the population. Although π is unknown, we can consider what value of π would make the given sample most probable. For this, we can use various values for π in the (9.5) and then select the value that results in the highest probability. Formally, this means that we are searching for the value of π for which the likelihood

$$\mathcal{L}(\pi|h, n) = \binom{n}{h}\pi^h(1-\pi)^{n-h} \tag{9.6}$$

is maximized. We can forgo a calculation of $\binom{n}{h}$, as this terms remains constant for all values of π. Note that the likelihood is calculated with the same formula as in (9.5). If (9.5) is evaluated for all possible values of h, the probabilities sum to one, but this is not

the case for the values of $\mathcal{L}$ and all possible values of π. Therefore, we must differentiate between likelihood and probability.

You can do this for sample 2 from figure 9.1 (2 survivors and 1 victim) by creating an artificial dataset with 100 observations:

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

Now generate the variable `pi` by rendering a series of possible values for π:

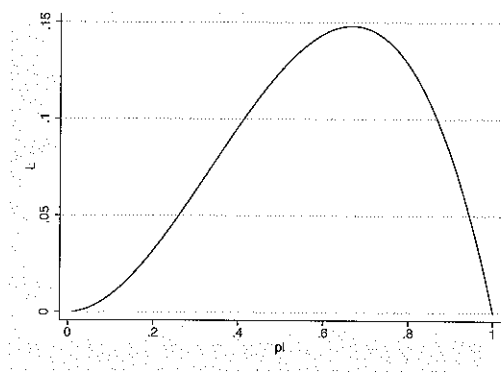
```
. generate pi = _n/100
```

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

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

With the help of a graph, you can then analyze which π 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 can you estimate the b coefficients of our regression model with the maximum likelihood principle from the (9.4)? The answer is simple. Instead of directly inserting the values for π, you can calculate π with the help of our regression model. Now insert (9.4) in (9.3)

$$\pi = \widehat{P}(Y = 1) = \frac{e^{b_0 + b_1 x_{1i} \ldots + b_{K-1} x_{K-1,i}}}{1 + e^{b_0 + b_1 x_{1i} \ldots + b_{K-1} x_{K-1,i}}} \tag{9.7}$$

and again in (9.6):

$$\mathcal{L}(b_k|f,n,m) = \widehat{P}(Y=1)^h \times \left\{1-\widehat{P}(Y=1)\right\}^{n-h}$$
$$= \left(\frac{e^{b_0+b_1x_{1i}+\cdots+b_{K-1}x_{K-1,i}}}{1+e^{b_0+b_1x_{1i}+\cdots+b_{K-1}x_{K-1,i}}}\right)^h \times \left(1-\frac{e^{b_0+b_1x_{1i}+\cdots+b_{K-1}x_{K-1,i}}}{1+e^{b_0+b_1x_{1i}+\cdots+b_{K-1}x_{K-1,i}}}\right)^{n-h} \quad (9.8)$$

After doing 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. You will not find an analytical solution with this model, unlike linear OLS regression. For this reason, iterative algorithms are used to maximize the log likelihood.

We have introduced the maximum likelihood principle for logistic regression with a dichotomous dependent variable. In principle, we can apply it to many different models by adapting (9.6) to reflect the distributional assumptions we wish to make. The resulting likelihood function is then maximized using a mathematical algorithm. Stata has a command called `ml` to do this, which is described in detail in Gould, Pitblado, and Sribney (2003).

9.3 Logistic regression with Stata

Let's briefly set aside our Titanic example in favor of an alternative. Say that you assumed that when the age and household income of a surveyed individual increases, the probability of living in an apartment or house they own also increases. In addition, you expect that the share of individuals who own their own residence[8] to be higher in West Germany than it is in East Germany.

Now let's load our dataset `data1.dta`.

```
. use data1, clear
```

To check your assumption, you can calculate a logistic regression model of residence ownership against the independent variables of age, household income, and an East–West variable.

Stata has two commands for fitting logistic regression models, `logit` and `logistic`. The commands differ in how they report the estimated coefficients. `logit` reports the actual bs in (9.4), while `logistic` reports the odds ratios discussed previously. Because we have emphasized using a linear combination of variables to explain the dependent variable, we will focus on `logit` and show you how to obtain odds ratios after estimation. Some researchers, particularly biostatisticians and others in the medical field,

[8] In the following, we will refer to living in an apartment or house that the individual owns as *residence ownership*. In this respect, children may also be considered to "own" housing. For household income, we will use the word "income".

focus almost exclusively on odds ratios and therefore typically use `logistic` instead. Regardless of how the coefficients are reported, both commands fit the same underlying statistical model.

Note that at least one category of the dependent variable must be 0, as `logit` takes a value of 0 to represent failure and any other value as representing success. Normally you use a dependent variable with the values 0 and 1, where the category assigned the value of 1 means success. For our example, the variable `owner` should be generated with the values of 1 for house owner and 0 for tenant as follows:[9]

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

We generate the East–West variable analogously as we did previously for our linear regression model (page 193):

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

It would also 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, we recommend that you center the two metrically independent variables `age` and `hhinc`, e.g., deduct the mean of the variable from each value. The mean of centered variables is zero, making it easier to interpret regression models at various points (Aiken and West 1991). You can center `age` and `hhinc` with the following commands:[10]

```
. 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 < .
```

Now we are ready to fit the logistic regression model:

[9] For details on this command, see page 78. The command `label list` determines the assignment of the values to labels (section 5.5).

[10] The Stata commands used here are explained in chapter 4.

```
. 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
Logistic regression                                   Number of obs   =       3200
                                                      LR chi2(3)      =     327.83
                                                      Prob > chi2     =     0.0000
Log likelihood = -1927.5979                           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 one from linear regression. At the bottom of the output is the coefficient block, which contains the coefficients for the dependent variables and the constants along with their standard errors, significance tests, and confidence intervals. At the top left is the iterations block with some results that are related to the maximum likelihood calculation, and at the top right we see block describing the model fit. In the following sections, we will discuss each of these blocks along the lines of our explanation of linear regression.

9.3.1 The coefficients block

The b coefficients can be found in the first column of the coefficient block.[11] The b coefficients formally indicate how the predicted values change when the corresponding independent variables increase by one unit, just like in linear regression, although here the predicted values are the logarithmic odds of success, not the mean of the dependent variable. For example, you would interpret the regression coefficient of `cage` as follows: the logarithmic odds of residence ownership rises on average by 0.0189758 if age increases by one year. You would interpret the regression coefficient of `chhinc` the same way. Form the regression coefficient of `east`, we can say that 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, we might instead say that 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 0. Due to centering, this means that the logarithmic chance of residence ownership for West German individuals with a mean age and mean income lies at −0.6023514.

[11] In the second column, you will find the standard errors of the regression coefficients, which will help you calculate significance tests, as well as confidence interval limits. You can interpret these figures corresponds the same way you did the figures in the linear regression (section 8.5). Note that in logistic regression, you usually evaluate the significance of the coefficients using a likelihood-ratio test (section 9.5).

Because changes in the logarithm of the odds of a positive outcome are not very easy to interpret, we will discuss their interpretation in more detail.

Sign interpretation

As a first step, consider just the signs and relative sizes 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 it decreases. In our example, the probability of house ownership increases with age and with income. The probability of home ownership is lower in the East than it is in the West.

Interpretation with odds ratios

Using the model equation, we want to calculate the predicted 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. This also is true for the centered age variable (`cage`), where individuals with the mean age have the value 0. Finally, West Germans surveyed are assigned 0 for the variable `east`. Thus the predicted logit for a West German of mean age and mean income is simply equal to the constant term.

By calculating the exponential of the regression function, you can convert the logarithmic odds to odds:[12]

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

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

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

An older person's probability of owning their residence is therefore slightly larger than that of those with average age. We can use the *odds ratio* (page 251) to compare the outcomes of the two 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, a person is 1.02 times as likely to own her residence. Increasing age by two years increases the likelihood of owning a residence by $1.02 \times 1.02 \approx 1.04$. Notice that odds ratios work in a multiplicative fashion.

[12] We covered working with the saved coefficients in detail in section 9.1.

You can reduce the complexity of calculating odds ratios if you consider that, in order to determine the odds ratios, you must first calculate the odds for a particular value of X and then for the value $X + 1$. After that, you divide both results by each other, which can be presented as follows:

$$\text{odds ratio} = \frac{e^{b_0+b_1(X+1)}}{e^{b_0+b_1X}} = \frac{e^{b_0+b_1X}e^{b_1}}{e^{b_0+b_1X}} = e^{b_1} \tag{9.9}$$

You can therefore obtain the odds ratio simply by computing the exponential of the corresponding b coefficient.

Many logistic regression users prefer the interpretation of results in the form of the odds ratios. For this reason, Stata also has the command `logistic` that directly reports the odds ratios. If you have already fit your model using `logit`, typing in `logistic` will redisplay the output in terms of odds ratios.

Probability interpretation

The third possibility for interpreting the coefficient is provided by (9.3), which we used to show you how to convert logits into probabilities. For example, to compute the probability of residence ownership for West Germans with mean ages and incomes, you could type

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

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

The `predict` command enables you to generate a new variable that contains the predicted probability for every observation in the sample. You just enter the `predict` command along with the name of the variable you want to contain the predicted probabilities.

```
. predict Phat
```

Here the name `Phat` indicates that this variable deals with predicted probabilities. You can also calculate the predicted logits with the `xb` option of the `predict` command.

One difficulty in interpreting probabilities is that they do not increase at the same rate for each unit increase in an independent variable. For example, 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
```

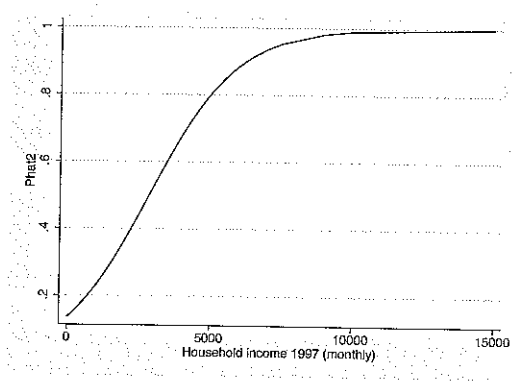
Comparing West Germans with the mean age with those 10 years older, the probability of residence ownership *increases* by around $0.3983 - 0.3538 = 0.0445$. If age increases by 10 more years, the probability increases by $0.4445 - 0.3983 = 0.0462$ and then by $0.4917 - 0.4445 = 0.0472$. We see that an increase in age of 10 years at a time does not lead to a constant change in the predicted probability.

One solution would be to show the probabilities in a conditional-effects plot. Similar to the graphs shown in section 8.4.2, this plot graphs the predicted values for various characteristics of the independent variables. Thus you could, for example, generate a variable with income-dependent predicted probabilities of West Germans with the mean age:[13]

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

and display this in a graph:

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



The graph shows that the increase in probabilities is not constant over all income values. Depending on income, the probability of home ownership will rise either rapidly or slowly.

9.3.2 The iteration block

In the upper-left part of the `logit` output (see page 259) are a number of rows beginning with the word *iteration*. This sort of output is typical for models whose coefficients are determined by maximum likelihood. As we mentioned in our discussion of this procedure, when you use the maximum likelihood principle, there is typically no closed-form mathematical equation that can be solved to obtain the b coefficients. Instead, an iterative procedure must be used that tries a sequence of different coefficient values. As

[13] At the mean age, the value of the centered age variable is 0, so you can omit age from the equation. The variable `east` is zero for West Germans, so you can omit it from the calculations as well.

the algorithm gets closer to the solution, the value of the likelihood function changes by less and less.

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 (see section 8.1.2), which contained figures for TSS, RSS, and MSS. TSS was the sum of the squared residuals from predicting 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 thus represents how many fewer errors 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 coefficients cannot be interpreted in the same way as with linear regression. Two values of the likelihood function are of particular 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 the constant term equal 0 ($\mathcal{L}_0$).[14] The last likelihood represents the maximized value. The larger the difference between the first and last values of the log likelihood, the stronger is the advantage of the model with independent variables compared to the null model. In this sense, you can consider TSS analogous to $\mathcal{L}_0$, RSS to $\mathcal{L}_K$, and MSS to $\mathcal{L}_0 - \mathcal{L}_K$.

Other than the first and last log likelihoods, the rest of the figures in the iteration block are of little interest, with one exception. Under certain circumstances, the maximum likelihood process delivers a solution for the b coefficients that is not optimal. This may occur if the domain where you are searching for the coefficients is not concave or "flat". This is a somewhat technical issue, and we do not wish to delve any further. However, note especially that a large number of iterations may indicate a difficult function to maximize, though it is difficult to say how many iterations are too many. You should generally expect more iterations the larger the number of independent variables in your model.

In general, the logistic regression model's likelihood function is "well-behaved", meaning that it is relatively easy to maximize. However, if you do have problems obtaining convergence, you may want to remove a few independent variables from your specification and try again.

9.3.3 The model fit block

R^2 was used to assess the fit of a linear regression model. The reason R^2 is so commonly used is that it has, on one hand, clear boundaries of 0 and 1, and on the other, a clear interpretation of the *share of explained variance*. There is no comparable generally accepted measured value for logistic regression. Instead, many different statistics have been suggested, some of which we will introduce here.

[14] This is true for Stata's `logit` and `logistic` commands. However, other maximum likelihood commands use alternative starting values, and in those cases the iteration-zero log likelihood is *not* the value obtained when all the slope parameters are set to zero.

One such measure of fit is reported in the model fit block of `logit`: the pseudo-R^2 (p^2). Nevertheless, it is already a mistake to speak of *the* pseudo-R^2. There are various definitions for pseudo-R^2 (Veall and Zimmermann 1994 and Long and Freese 2003). Therefore, you should always indicate which pseudo-R^2 you are referring to. The one reported by Stata is the one suggested by McFadden (1973), which is why we refer to it as p^2_{MF}.

McFadden's p^2_{MF} is defined in a way that is clearly analogous to the R^2 in linear. (Recall that $R^2 = \text{MSS}/\text{TSS} = 1 - \text{RSS}/\text{TSS}$). p^2_{MF} is defined as

$$p^2_{\text{MF}} = \frac{\ln \mathcal{L}_0 - \ln \mathcal{L}_K}{\ln \mathcal{L}_0} = 1 - \frac{\ln \mathcal{L}_K}{\ln \mathcal{L}_0} \tag{9.10}$$

where $\mathcal{L}_0$ is the likelihood from the model with just a constant term and $\mathcal{L}_K$ is the likelihood of the full model. As is the case in R^2, p^2_{MF} lies within the boundaries of 0 and 1; however, interpreting the content is disproportionately more difficult. "The higher, the better" is pretty much the only thing that you can say of p^2_{MF}. In our example (page 259), the value of p^2_{MF} at around 0.08 is what most people would agree is rather small.

Besides McFadden's Pseudo-R^2, the likelihood-ratio χ^2 value ($\chi^2_{\mathcal{L}}$) is another indicator of the quality of the overall model. It, too, is based on the difference between the likelihood functions for the full and constant-only models. However, unlike p^2_{MF}, this difference is not standardized to lie between 0 and 1. It is defined as

$$\chi^2_{\mathcal{L}} = -2(\ln \mathcal{L}_0 - \ln \mathcal{L}_K) \tag{9.11}$$

$\chi^2_{\mathcal{L}}$ follows a χ^2 distribution, and as with the F value in linear regression, you can use $\chi^2_{\mathcal{L}}$ to investigate the hypothesis that the independent variables do not have any explanatory power or, equivalently, that all the coefficients other than the constant are all zero. The probability of this hypothesis being true is reported in the line that reads "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 of the linear regression F test, rejection of this null hypothesis is not sufficient for us to be satisfied with the results.

As with linear regression, you should not judge a model's suitability purely by the measured values within the model fit block, especially in logistic regression, as there is no single generally accepted measured value for doing so. For this reason, we will discuss some alternative measures that are not reported in the output.

Classification tables

The fit of the linear regression model was primarily assessed on the basis of the residuals $(y - \widehat{y})$. In logistic regression, one way to assess fit is with a classification table, in which every observation is assigned one of the two outcomes of the dependent variable. The

positive outcome is normally assigned when the model predicts a probability of over 0.5, whereas an observation is assigned a negative outcome if a probability of under 0.5 is predicted. For example, you could do this manually assuming you have already created the variable `Phat` containing the predicted probabilities by typing

```
. 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 containing the classified values and the original values:

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

Key
frequency *column percentage* *cell percentage*

ownerhat	owner 0	1	Total
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

The sensitivity and the specificity of the model are commonly used by people in the medical profession. Sensitivity is the share of observations classified as residence owners within the observations who actually do own their residences. Specificity is the share of observations classified as tenants among those who are actual tenants. In the example above, the sensitivity is 28.10%, and the specificity is 90.77%.

The count R^2 is commonly used in the social sciences. It deals with the share of overall correctly predicted observations, which you can determine by adding the overall shares in the main diagonal of the above-generated table. However, it is easier to use the `estat classification` command, which you can use to generate the table in a different order, as well as derive the sensitivity, specificity, count, R^2, and other figures:

This means that when predicting with a model that includes independent variables, our error rate drops by 12 percent compared to prediction based solely on the marginal distribution of the dependent variable. You can receive the adjusted count R^2, as well as other model fit measured values through Scott Long and Jeremy Freese's ado package `fitstat`, available from the SSC archive (see section 12.3.2). Two common fit statistics, the AIC and BIC, are available with the `estat ic` command.

Pearson chi-squared

A second group of fit statistics is based on the Pearson residuals. For you to understand these, we must explain the term *covariate pattern*, which 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, where j indexes each covariate pattern that occurs. By typing

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

you can view the index number representing the covariate pattern of each observation.

The Pearson residuals are obtained by comparing the number of successes with covariate pattern j (y_j) with the predicted number of successes with that covariate pattern ($m_j\widehat{P}_j$, where $\widehat{P}_j$ is the predicted probability of success for covariate pattern j). The Pearson residual is defined as

$$r_{P(j)} = \frac{(y_j - m_j\widehat{P}_j)}{\sqrt{m_j\widehat{P}_j(1-\widehat{P}_j)}} \tag{9.13}$$

(Multiplying $\widehat{P}_j$ by the number of cases with that covariate pattern results in the predicted number of successes in pattern j.) Notice that unlike residuals in linear regression, which are in general different for each observation, the Pearson residuals for two observations differ only if those observations do not have the same covariate pattern. Typing

```
. predict pres, resid
```

generates a variable containing the Pearson residuals. The sum of the square of this variable over all covariate patterns produces the Pearson chi-squared statistic. You can obtain this statistic by typing

```
. estat gof
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
```

This test is for the hypothesis of the conformity of predicted and observed frequencies across covariate patterns. A small χ^2 value (high p-value) indicates small differences between the observed and the estimated frequencies, while a large χ^2 value (low p-value) suggests that the difference between observed and estimated values cannot be explained by a random process. Be careful when interpreting the p-value as a true "significance" level: a p-value under 0.05 may indicate that the model does not represent reality, but values over 5 percent do not necessarily mean that the model fits the data well. A p-value of, say, 6 percent is still fairly small, even though you cannot formally reject the null hypothesis that the difference between observed and estimated values is completely random at significance levels below 6%.

The χ^2 test is not very suitable when the number of covariate patterns (here 3,183) is close to the number of observations in the model (here: 3,200). Hosmer and Lemeshow (2000, 140–145) have therefore suggested modifying the test by sorting the data by the predicted probabilities and dividing them into g approximately equal-sized groups. They then suggest comparing the frequency of the observed successes in each group with the frequency estimated by the model. A large p-value indicates a small difference between the observed and the estimated frequencies.

You can obtain the Hosmer–Lemeshow test by using `estat gof` together with the `group()` option. Enter the number of groups into which you want to divide the data between the brackets. $g = 10$ is often used.

```
. estat gof, group(10)
```

9.4 Logistic regression diagnostics

We now discuss two methods to test the specification of a logistic regression model. First, logistic regression assumes a linear relationship between the logarithmic odds of success and the independent variables. Thus you should test the validity of this assumption before interpreting the results.

Second, you need to deal with influential observations, meaning small observations that have a strong influence on the results of a statistical procedure. Occasionally these outliers, as they are also known, turn out to be the result of incorrectly entered data, but usually they indicate that variables are missing from the model.

9.4.1 Linearity

We used graphical analyses to discover nonlinear relationships in the linear regression model, and we used smoothing techniques to make the relationship more visible. You can also use certain scatterplots for logistic regression, but you should consider two issues. First, the median trace used in linear regression as a scatterplot smoother does not work for dichotomous variables because the median can only take values of 0 and 1.[15]

[15] The value 0.5 can occur if there is an equal number of 0 and 1 values.

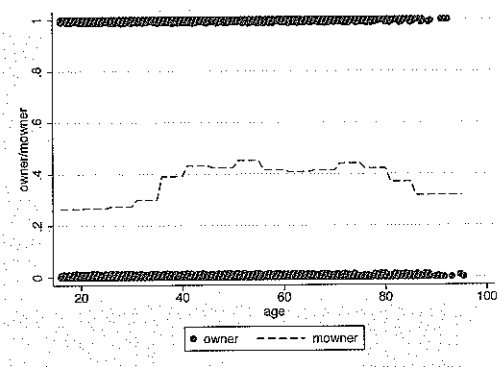
Second, the functional form of the scatterplot does not have to be linear, as linearity is only assumed with respect to the logits. The functional form between the probabilities and the independent variable has the shape of an S (see the graph on page 253).

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

Ideally, the graph should show the local mean regression to have an S-shaped curve like the illustration on page 9.2.1. However, the graph often only depicts a small section of the S-shape, so if the band means range only from about 0.2 to 0.8, the mean regression should be almost linear. U-shaped, reverse U-shaped, and other noncontinuous curves represent potential problems.

Stata does not have a specific command for simple local mean regression, but you can do it easily nonetheless:[16]

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



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

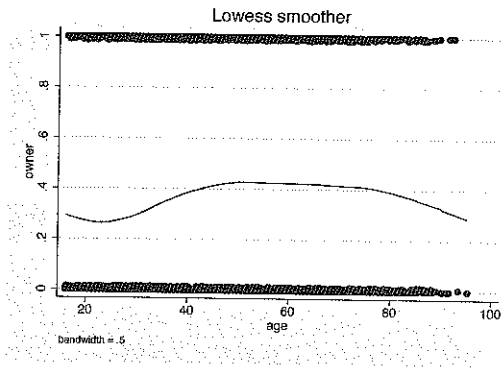
Cleveland's (1979) locally weighted scatterplot smoother (LOWESS)[17] is an alternative that is often better for investigating functional forms. You can use this smoother with the `twoway` plottype `lowess` or the statistical graph command `lowess`. We will not discuss the calculation of this smoother but refer you to the excellent explanation of the logic behind LOWESS in Cleveland (1994). Note, however, that you can adjust

[16] For the function `autocode()`, see page 152. For the command egen, see section 5.2.2 on page 87.

[17] 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.

the level of smoothing by specifying a value between 0 and 1 in the `bwidth()` option, with higher numbers specifying increased smoothing. Also note that LOWESS is a computationally intensive process, so it may take some time to display the following graph on your screen:

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



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, are likely to live in their own houses or apartments.

Both graphs show a correlation that contradicts the S-shaped correlation required by logistic regression. As with linear regression, U-shaped relationships can be modeled through the generation of polynomials. Nevertheless, before you do this, check if the U-shaped relationship is still visible after controlling for household income. You can do this using a technique similar to the local mean regression discussed in the previous chapter for linear regression.[18] In this process, you replace the age variable of your regression model with a set of dummy variables (see page 278 and section 8.4.1.) for the grouped version of the age variable on page 269:

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

Fitting this model yields a total of 14 b coefficients for the age variables. Each b coefficient 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, the age b coefficients should increase continuously and steadily. This does not appear to

[18] For the following process, see Hosmer and Lemeshow (2000, 90). Alternatively, scatterplots with smoothers of the dependent variables can be made against one independent variable controlling for various combinations of the other independent variables (Schnell 1994, 253). Fox (1997) demonstrates a process related to the component-plus-residual plot (page 203).

be the case for the coefficients in front of us. You can easily evaluate this by graphically depicting the rise of the b coefficients with the following commands:

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

The symbol ' in the `matrix` command is a simple quotation mark and can be found next to the ";" key on English keyboards. In this case, it stands for the transpose of a matrix, which we will explain later.

Explanation: Stata stores the coefficients of statistical models in a row vector called `e(b)`. These are nothing more than the stored results to which we have repeatedly referred . Matrices and vectors are of particular interest, as they contain numerous saved results. The vector `e(b)`, for example, contains the estimated coefficients of the regression model. By typing

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

you can take a closer look at `e(b)`.

As with results saved in `r()` macros, you can also perform calculations on matrices saved in `e()`-class macros using `matrix` commands.[19] In the commands above, we used a `matrix` command to transpose the row vector `e(b)` into the column vector `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 `e(b)`. In contrast, in the newly generated column vector, the coefficients are under one other.

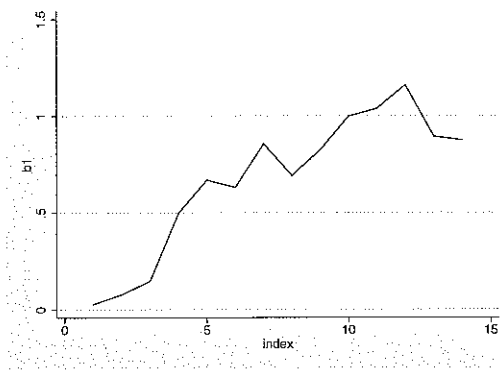
The column vector is therefore nothing more than a list of numbers. The `svmat` command writes these numbers as a variable in our dataset; specifying the `names` option gives the variable a name. Note that Stata automatically adds the character 1 onto the name you choose. Stata does this because the command can also save matrices with a number of columns as variables, whereby every column in the matrix becomes a variable.

After you use the `svmat` command, your dataset will contain the new variable `b1`, which contains the 14 age b coefficients, the b coefficient for household income, and the b coefficient for the constants. The first 14 numbers in the variable $b1$ are the age coefficients. In our case, the coefficients for the age dummies will appear first in the dataset, and so we can graph them by typing

(*Continued on next page*)

[19] For an overview of the `matrix` commands, type `help matrix`.

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



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. Including 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 9.5 on page 275 and on page 278.

9.4.2 Influential cases

Influential data points are observations that heavily influence the b coefficients of a regression model. Said another way, if we were to remove an influential data point and then refit our model, our coefficient estimates would change by more than a trivial amount. As explained on page 209, 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.

This concept is somewhat more problematic in logistic regression than it is in linear regression, as you can only measure the approximate leverage and discrepancy (Fox 1997, 459). In Stata, you can approximate the leverage values by typing

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

Note that you must refit the original model with `cage` and `chhinc` as independent variables, since `predict` always refers to the last model fitted, and we just fit a model including dummy variables to control for age. After getting the predicted leverage values, you can obtain the standardized residuals as an approximation to discrepancy by typing

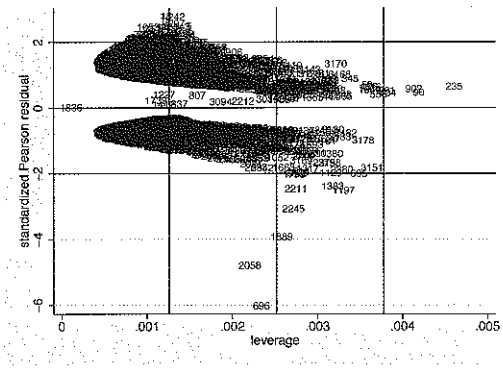
```
. predict spres, rstandard
```

In logistic regression, the standardized residuals for observations having the same covariate pattern are identical. The same also applies for the leverage values. To isolate those covariate patterns having high leverage and discrepancy values, you can produce a graph that compares the standardized residuals with the leverage values. Fox (1997, 461) uses a diagram with vertical lines at the mean of the leverage values and at two and three times the mean. To produce 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 4) to later use them as vertical lines in the graph.

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

Next we generate the graph with the standardized residuals against the leverage values. To generate vertical lines, we use the `xline()` option. We use the number of covariate patterns as the plot symbol. These patterns are found in the variable `cpatt`, which we generated on page 267:

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



Eight covariate patterns in the graph are particularly conspicuous: both patterns having the lowest standardized residuals, and the six patterns having standardized residuals under -2 *and* leverage values over twice the average. The following command shows that invariably the latter consist of observations from West Germany with comparatively high income but not 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 is determined by Cook's D (see section 8.3.1), which involves multiplying the leverage and discrepancy. An analogous measured value for logistic regression is

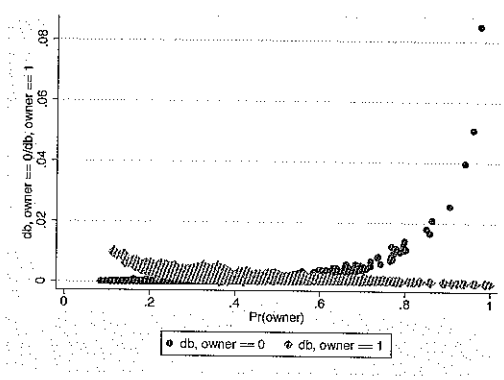
$$\Delta\beta = \underbrace{\frac{r^2_{P(j)}}{(1-h_j)^2}}_{\text{Discrepancy}} \times \underbrace{h_j}_{\text{Leverage}} \tag{9.14}$$

where h_j is the value for the leverage. In Stata, you can obtain this value via

```
. predict db, dbeta
```

as a variable under the name `db`. A scatterplot of $\Delta\beta$ against the predicted probabilities is often used, in which observations with success as the outcome are displayed in a different color or symbol than those of failure. The `separate` command is particularly useful for the latter[20]:

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

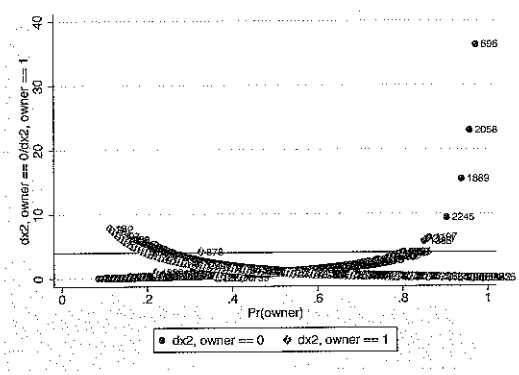


[20] For an explanation of `separate`, type `help separate`. The variable `Phat` was generated on page 261 with `predict Phat`.

The curve from the bottom left to the top right consists of all tenants, while the curve that slopes downward from left to the bottom right consists of all residence owners. Several covariate patterns for tenants are noteworthy, namely those which have a high predicted probability of residence ownership. If you enter the number of covariate patterns into the graph instead of the symbols, you will see that these patterns are the same ones our previous analysis detected.

The Pearson residuals allow for a further test statistic for influential observations. As shown in section 9.3.3, the sum of the squared Pearson residuals is a measure of the deviation of the predicted values from the observed values. The contribution of each covariate pattern to this measure matches the square of the Pearson residual. If you divide this contribution by $1 - h_j$, you get $\Delta\chi^2_{P(j)}$, which indicates the change in the Pearson chi-squared statistic when the covariate pattern j is removed from the dataset. The scatterplot of $\Delta\chi^2_{P(j)}$ against the predicted probabilities is well suited to the discovery of covariate patterns that are hard to predict through the model. Here it would be useful to enter Hosmer and Lemeshow's raw threshold value of $\Delta\chi^2_{P(j)}$ of four into the graph (Hosmer and Lemeshow 2000, 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 you can eliminate data errors, you should determine if a variable important to the model was left out. This could be a subgroup for which the assumed correlation between age, household income, region, and residence ownership does not hold.

9.5 Likelihood-ratio test

In section 9.3.3, we showed you how to calculate χ^2_L. That statistic compares the likelihood of the fitted model with that of a model in which all the coefficients other

than the constant are set to 0. A large value of $\chi^2_{\mathcal{L}}$ indicates that the full model does significantly better at explaining the dependent variable than the constant-only model.

You can apply the same principal to answer the question of whether the addition of more independent variables achieves a significant increase in the explanatory power of our model compared to a null model with fewer independent variables. For example, you can ask whether the fit of a model on residence ownership against household income increases if we include an age variable. To answer this question, you can carry out a calculation that is analogous to the test of the overall model by again using -2 times the difference between the log likelihood of the model without age ($\ln \mathcal{L}_{\text{without}}$) and the log likelihood of the model with age ($\ln \mathcal{L}_{\text{with}}$):

$$\chi^2_{\mathcal{L}(\text{Diff})} = -2(\ln \mathcal{L}_{\text{without}} - \ln \mathcal{L}_{\text{with}}) \tag{9.15}$$

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

You can easily calculate $\chi^2_{\mathcal{L}(\text{Diff})}$ in Stata using the `lrtest` command. In our example, we want to investigate the significance of the age effects. First, we calculate the model with the variable we want to investigate:

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

We store this model internally using the command `estimates store`, and we name the model `full`:

```
. estimates store full
```

Now we calculate the reduced model.

```
. logit owner chhinc east
```

Then you can use `lrtest` to test the difference between this model and the previously stored model. You can simply list the name of the stored model (`full`) and, optionally, the name of the model against which it should be compared. If you do not specify 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 $\chi^2_{\mathcal{L}(\text{Diff})}$ value of 76.89 or higher in our sample is very small when the age coefficient in the population is 0. You can therefore be fairly certain that the age coefficient is not zero. However, this statistic does not reveal anything about the degree of influence of age on residence ownership; for that you need to consider the coefficient on age.

When using the likelihood-ratio test, note that only models that are nested can be compared with one another. This means that the full model must contain all the

variables of the reduced model. Furthermore, both models must be calculated using the same set of observations. The latter may be problematic if, for example, some observations in your full model must be excluded due to missing values, while they may be included in the reduced model if you leave out a variable. In such cases, Stata displays a warning message ("observations differ").

If you wish to compare models not fitted to the same sets of observations, an alternative is to use "information criteria" which are based on the log-likelihood function and are valid even when comparing nonnested models. Two of the most common information criteria are the BIC (Bayesian information criterion) and AIC (Akaike's information criterion), which are obtained through the `estat ic` command mentioned earlier (page 267). An excellent introduction to the statistical foundations of these indices is provided by Raftery (1995).

9.6 Refined models

As with the linear regression model, the logistic regression model can also be expanded in various ways to investigate complicated causal hypotheses, particularly, in three ways: specifying nonlinear relationships, comparing subgroups (categorical variables), and investigating varying correlations between subgroups (interaction effects). Because these procedures for expanding the model are similar to those for linear regression, we will use a few examples from our discussion of linear regression.

Nonlinear relationships

During the diagnosis of our regression model, we saw signs of a U-shaped correlation between age and the logarithmic chance of residence ownership (section 9.4.1). In this respect, U-shaped correlations are only one form of nonlinear relationships. Logarithmic or hyperbolic relationships can also occur. 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-shaped distribution of the logit transformation.

There are many ways to account for nonlinear relationships. If you have an assumption as to why older people are less likely to own a residence than middle-aged people, you should incorporate the variable in question into the regression model. If, for example, you suspect that the observed decline is a consequence of older people moving into nursing homes to receive full-time care, you might want to introduce some type of variable indicating whether a person is in poor health.

An alternative way of controlling for a nonlinear relationship is to categorize the independent variable into several groups and use a set of dummy variables instead of the continuous variable. We discussed a strategy like this on page 270. A more common strategy is to use transformations or polynomials of the independent variables. The

rules for linear regression apply here, too: in the case of hyperbolic relationships, the X variable is squared, and in the case of logarithmic relationships, the logarithm of the X variable is used. For U-shaped relationships, we use the squared X variable in addition to the original X variable.

To model the U-shaped relationship between residence ownership and age, you could proceed as follows:

```
. generate cage2 = cage^2
(140 missing values generated)
. logit owner cage cage2 chhinc east, nolog
Logistic regression                             Number of obs   =       3200
                                                LR chi2(4)      =     332.12
                                                Prob > chi2     =     0.0000
Log likelihood = -1925.4516                     Pseudo R2       =     0.0794

------------------------------------------------------------------------------
       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 in a conditional-effects plot (see section 8.4.2).

Categorical independent variables

Categorical variables are used in logistic regression the same way they are used in linear regression (section 8.4.1). 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*.

Let's continue our investigation into the Titanic catastrophe (see section 9.1). You want to see whether the seafaring principle of women and children first was put into practice or whether, as shown in the film *Titanic*, the first-class gentlemen took their places in the lifeboats at the expense of the third-class women and children.

You have previously established that women and children evidently really did have better chances of survival than men did (and adults, respectively). To look into this more closely, load the original dataset:

```
. use titanic, clear
```

This dataset contains dichotomous variables for survival, age (children are coded zero and adults are coded one), and sex (females are coded as zero and males as one), as well as a categorical variable for first-class passengers (1), second-class passengers (2), third-class passengers (3), and crew (4).

The film *Titanic* assumed that, besides gender and age, passenger class was also a criterion for a place in the lifeboats. You can verify this assumption using a logistic regression model of survival against age, sex, and class. To use the independent variable `class` in the regression model, first transform it into a set of dummy variables. You can do this with the following command:[21]

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

This command generates four dummy variables named `class1` to `class4`, which you can now use in your regression model. As in linear regression, you must choose one of the variables as the reference category and omit it from the model. In this case, you will want to use the first-class passengers as the reference category.

```
. logit survived age sex class2-class4, nolog
Logistic regression                             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 signs on the age dummy, it appears that the survival chance for adults was smaller than that of the children and that the survival chance for the men was smaller than that of the women. So far, this supports the principle of "women and children first". However, it also becomes apparent that the first-class passengers have the largest chance of survival compared with the rest. The third-class passengers had the smallest chances of survival; in fact, their chances of survival were even smaller than those of the crew. In conclusion, you can state that women and children were indeed favored for rescue, but apparently passenger class also played a role.

To test formally whether the class played a role in determining survival, you need to test whether the coefficients on the class dummies are *jointly* significantly different from zero, and for that you can use the likelihood-ratio test mentioned above. First, we save the full model you just calculated:

```
. estimates store full
```

and then you calculate the model without the class dummies:

```
. logit survived age sex
```

[21] An alternative would be the command `xi`; see section 8.4.1.

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. Thus passenger class did have an impact on survival.

```
. 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 a person's sex plays the same role for adults and children. However, with the principle of "women and children first", 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 both boys and girls are treated equally, the coefficient for sex should be smaller for children than it is for adults; in fact, it should be zero. In other words, the effect of sex on survival varies with age. Effects of independent variables that vary between subgroups are called interaction effects.

You can model interaction effects in logistic regressions the same way as in linear regression models. Multiplying the variables involved in the interaction effect generates interaction terms. Note that here we do not recenter any variables, because both age and sex are dichotomous.

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

After you do this, you can incorporate the interaction terms into the model:

```
. logit survived sex age menage class2-class4, nolog
Logistic regression                                   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
------------------------------------------------------------------------------
```

Consider first the case in which the interaction term is zero, which happens for observations on children or females. The coefficient on `sex` then 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. In this case the coefficient on age indicates that 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, this would yield a $-.72 + (-1.90) = -2.62$ smaller log-chance of survival for a male adult compared with a female adult. Therefore, the survival chance of men was only around one-fourteenth ($e^{-2.62} = 0.07 = 1/14$) that of the women.

9.7 Advanced techniques

Stata allows you to fit numerous related models in addition to the logistic regression we have described above. Unfortunately, there is not enough space in this book in order to show them in detail. However, in this section we will describe the fundamental ideas behind some of the most important procedures. For further information, we will specifically refer you to the entry in the *Stata Reference Manual* corresponding to each command. There you will also find references to the literature.

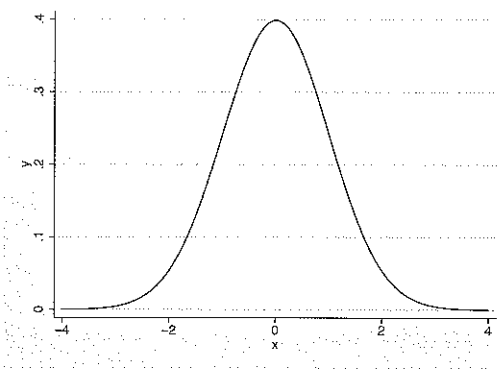
9.7.1 Probit models

In the logistic regression model, we attempted to predict the probability of a success through a linear combination of one or more independent variables. To ensure that the predicted probabilities remained between the limits of 0 and 1, the probability of the success underwent a logit transformation. However, using the logit transformation is not the only way to achieve this. An alternative is the probit transformation used in probit models.

To get some idea of this transformation, visualize the density function of the standard normal distribution:

(*Continued on next page*)

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



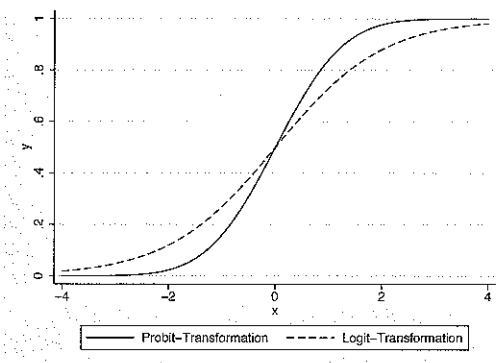
You can interpret this graph in the same way as a histogram or a kernel density estimator (see section 7.3.3); i.e., for this variable, the values around 0 occur most often, and the larger or smaller they become, the rarer they occur.

Suppose that 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 probable. If you want to know the exact answer, you can determine the probability through distribution-function tables for standard normal distribution or through the Stata command

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

The probability of selecting an observation with a value of less than or equal to -2 from a standard normal variate is therefore 0.023. You can repeat the same calculation for any value of X and then enter the calculated probabilities against the values of X in a scatterplot. This results in the distribution function for the standard normal distribution Φ depicted in the following graph:

```
. twoway (function y = normal(x), range(-4 4))
> (function y = exp(x)/(1+exp(x)), range(-4 4)),
> legend(lab(1 "Probit-Transformation") lab(2 "Logit-Transformation"))
```



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

As with the logit transform you used with logistic regression, the normal distribution function can also be used to transform values from $-\infty$ and $+\infty$ into values between 0 and 1. Correspondingly, the inverse of the distribution function for the standard normal distribution (Φ^{-1}) converts probabilities between 0 and 1 for a success ($P(Y = 1)$) into values between $-\infty$ and $+\infty$. The values of this probit transformation are thus also suitable as dependent variables for a linear model. This yields the probit model:

$$\Phi^{-1}\{\widehat{P}(Y = 1)\} = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_{K-1} x_{K-1,i} \tag{9.16}$$

You can estimate the b coefficients of this model through maximum likelihood. You can interpret the coefficients the same way as in logistic regression, except that now the value of the inverse distribution function of the standard normal distribution increases by b units instead of the log-odds ratio increasing by b units for each one-unit change in the corresponding independent variable. Using the distribution function for the standard normal distribution, you can then calculate probabilities of success. Usually, the predicted probabilities of probit models are nearly identical to those of logistic models, and the coefficients are often about 0.58 times the value of those of the logit models (Long 1997, 49).

The Stata command used to calculate probit models is `probit`. For example, you can refit the previous model (see page 280) using `probit` instead of `logit`:

```
. probit survived sex age menage class2-class4, nolog
Probit regression                                 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

See [R] **probit** for more information on this model.

9.7.2 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 all other parties.

The main problem with using multinomial logistic regression is in the interpretation of the coefficients, so this will be the focus point of this section. Nevertheless, in order to understand this problem, you must at least intuitively grasp the statistical fundamentals of the process. These fundamentals will be discussed shortly (Long 1997, cf.).

In multinomial logistic regression, you predict the probability for every value of the dependent variable. You could initially calculate a binary[22] logistic regression for every value of the dependent variable. In our example, you could calculate three separate logistic regressions: one with the dependent variable CDU against non-CDU, one with the dependent variable SPD against non-SPD, and finally one with the dependent variable for the other parties against the CDU and SPD together:

$$\ln \frac{P(Y = \text{CDU})}{P(Y = \text{not-CDU})} = b_0^{(1)} + b_1^{(1)} x_{1i} + b_2^{(1)} x_{2i} + \cdots + b_{K-1}^{(1)} x_{K-1,i}$$

$$\ln \frac{P(Y = \text{SPD})}{P(Y = \text{not-SPD})} = b_0^{(2)} + b_1^{(2)} x_{1i} + b_2^{(2)} x_{2i} + \cdots + b_{K-1}^{(2)} x_{K-1,i}$$

$$\ln \frac{P(Y = \text{other})}{P(Y = \text{not-other})} = b_0^{(3)} + b_1^{(3)} x_{1i} + b_2^{(3)} x_{2i} + \cdots + b_{K-1}^{(3)} x_{K-1,i} \qquad (9.17)$$

[22] In order to differentiate it from multinomial logistic regression, we call the logistic regression of a dichotomous dependent variable a binary logistic regression.

The superscript in parentheses 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 $b_1^{(1)} \ldots b_{K-1}^{(1)}$ 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 the predicted probability of every value of the dependent variable. Note that these predicted probabilities do not all add up to 1. However, they should, as one of the three possibilities—SPD, CDU or Other—must[23] occur.

For this reason, it would appear sensible to jointly estimate $\mathbf{b}^{(1)}$, $\mathbf{b}^{(2)}$, and $\mathbf{b}^{(3)}$ and to adhere to the rule that the predicted probabilities must add up to 1. However, it is not possible to estimate all three sets of coefficients. In order to do so, you must constrain one of the coefficient vectors to be equal to a fixed value, zero being by far the most common choice. After making such a normalization, the remaining coefficients can be estimated using the maximum likelihood principle. Which one of the three sets of coefficients you constrain to be zero does not matter. By default, Stata's `mlogit` command constrains the coefficient vector corresponding to the most frequent outcome.

Let us show you an example of interpreting coefficients. Please load `data1.dta`:

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

Now generate a new variable for party 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
(party: 1375 changes made)
. label define party 1 "CDU" 2   "SPD" 3 "Other"
. label value  party party
```

This creates the variable `party` with the values 1 for the CDU/CSU, 2 for the SPD, and 3 for the other parties. Respondents 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 for all estimation commands; i.e., the dependent variable follows the command and is in turn followed by the list of independent variables. With the `baseoutcome()` option, you can select the equation for which the b coefficients are set to 0.

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

[23] In this case, you disregard the possibility of no party preference. If you did not, you would have to calculate a further regression model for this alternative. The predicted probabilities for the four regression should then add up to 1.

```
. mlogit party yedu ybirth, base(1) nolog
Multinomial logistic regression                 Number of obs   =       1360
                                                LR chi2(4)      =      92.69
                                                Prob > chi2     =     0.0000
Log likelihood = -1379.4301                     Pseudo R2       =     0.0325
```

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

(party==CDU is the base outcome)

In contrast to binary logistic regression, the coefficient table 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 $\mathbf{b}^{(\text{CDU})} = 0$, you 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.

Interpreting the coefficients for a multinomial logistic regression is not as easy as for binary logistic regression, as 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, we can demonstrate this with the coefficient for the variable `ybirth` from the equation for the SPD. Writing the probability of preferring the SPD as P_{SPD} and the probability of preferring the CDU as P_{CDU}, the b coefficient for `ybirth` in the SPD equation can be written as

$$b_{ybirth}^{(\text{SPD})} = \ln\left(\frac{\widehat{P}_{\text{SPD}|\text{ybirth}+1}}{\widehat{P}_{\text{CDU}|\text{ybirth}+1}}\right) - \ln\left(\frac{\widehat{P}_{\text{SPD}|\text{ybirth}}}{\widehat{P}_{\text{CDU}|\text{ybirth}}}\right)$$

$$\ln\left(\frac{\widehat{P}_{\text{SPD}|\text{ybirth}+1}}{\widehat{P}_{\text{SPD}|\text{ybirth}}} \times \frac{\widehat{P}_{\text{CDU}|\text{ybirth}}}{\widehat{P}_{\text{CDU}|\text{ybirth}+1}}\right) \quad (9.18)$$

The b coefficient for year of birth in the equation for the SPD, on the one hand, depends on the change in probability of SPD preference with the year of birth. On the other hand, it also depends 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 does not completely depend on the change in the probability of SPD preference. In this respect, the b coefficient can be solely, mainly, or partly dependent on the probability relationship in the base category.

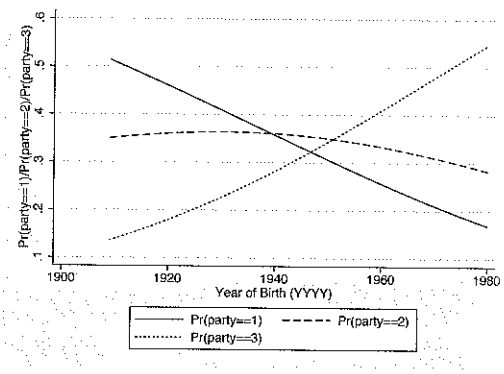
To avoid misinterpreting the multinomial logistic regression, we recommend that you use the conditional-effects plot for the predicted probabilities.[24] To create this plot, first generate the predicted probabilities of the model with `predict`. As there is a predicted probability for every value of the dependent variable, you will have to provide three variable names for the predicted probabilities.

```
. predict PCDU PSPD POther
```

To illustrate the effect of the year of birth, we plot these variables against the year of birth, holding the length of education 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 coefficient for the year of birth.[25]

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



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, so fixing the values of the independent variables may mean that there are too few observations remaining for a sensible plot. The method of recycled predictions described in [R] **mlogit** appears to be good solution to this problem. An even more powerful possibility for producing

[24] One alternative is the *method of recycled predictions* which is described in [R] **mlogit**. A further alternative is calculating marginal effects with the command `mfx compute`.

[25] In the legend, `Pr(party==1)` stands for the CDU, `Pr(party==2)` stands for the SPD and `Pr(party==3)` stands for the other parties.

conditional effects plots is the `prgen` command, which you can download from the SSC archive (see section 12.3.2) and which is more fully described by Long and Freese (2003).

9.7.3 Models for ordinal data

Models for ordinal data are used when the dependent variable has more than two values that can be ranked. An example would be the question regarding concerns about the increase of crime, which respondents could answer with "no concerns", "moderate concerns", or "strong concerns". In a dataset, these could be assigned the values of 0, 1, and 2, or equivalently, 0, 10, and 12. Note that the difference between two consecutive categories is immaterial—all that matters is that the outcomes can be ordered.

In principle, there are two strategies available for modeling ordinal dependent variables. The first uses multinomial logistic regression, whereby certain constraints are imposed upon the coefficients (stereotype model). The second strategy generalizes binary logistic regression for variables with more than two values (proportional odds model). Anderson (1984) discusses the implementation prerequisites for both models.

The logic behind the stereotype model is simple. In multinomial logistic regression, every value of the dependent variable has its own set of coefficients. The length of education in the regression model on page 286 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, you would normally not expect a directional change in the effects. For example, consider the variable for concerns about increasing crime (`np9506`), which contains the values 1 for no concerns, 2 for moderate concerns and 3 for strong concerns. First, 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. This should come as little surprise, since, if education reduces the chance of having moderate concerns (and not of having no concerns), it should also reduce the chance of having strong concerns (and not of having no concerns). However, if you calculate a multinomial logistic regression, this assumption is ignored. Nevertheless, you can include such assumptions in the model by imposing constraints on the b coefficients.

Using constraints, you can impose certain structures for the b coefficients before calculating a model. You could, for example, require that education reduces the chance of having moderate concerns (and not of having no concerns) to the same extent that it does for having strong concerns (and not of having 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 constraint number 1, which states that the coefficient of the variable `yedu` in the third equation be twice as large as the coefficient of the variable `yedu` in the second equation. You impose the constraint by specifying 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 that 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. See [R] **slogit** for more information about this model.

A different approach is followed by the proportional odds model. In the proportional odds model, the value of the ordinal variable is understood as the result of categorizing an underlying metric variable. In our example, you could assume that answers in the `worries` variable only provide a rough indication of the attitudes towards the increase in crime. The attitudes of people probably vary between having infinitely numerous concerns and no concerns whatsoever, so they might take any value in between; that is, attitude is actually a continuous variable E. Instead of observing E, however, all you see are the answers reported on the survey—no concerns, moderate concerns, or strong concerns. Since you have three outcomes in the model, there must also exist two points κ_1 and κ_2 that partition the range of E into the three reported answers. That is, if $E < \kappa_1$, then the person reported no concerns, if $\kappa_1 \leq E \geq \kappa_2$, the person reported moderate concerns, and if $E > \kappa_2$ the person reported strong concerns.

Remember the predicted values ($\widehat{L}$) of the binary logistic regression. These values can take on any values from $-\infty$ to $+\infty$. In this respect, you could interpret these predicted values as the unknown metric attitude E. If you knew the value of κ_1 and κ_2 by assuming a specific distribution for the difference between E and $\widehat{L}$, you could determine the probability that each person reported each of the three levels of concern. The proportional odds model estimates the b's in the linear combination of independent variables as well as the cut points needed to partition the range of E into discrete categories.

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. We will calculate the same model as above:

```
. ologit worries yedu
```

The predicted value of this model for respondents with ten years of education is $S_{10} = -0.068 \times 10 = -0.68$. The value for κ_1 and κ_2 are provided underneath the coefficient block. The probability that respondents with a predicted value of $-.68$ are classified as individuals with moderate concerns matches the probability of $-.68 + u_j \leq -1.196$, or, in other words, the probability that $u_j \leq -1.128$. If you assume that the error term follows the logistic distribution, the probability is $1/(1 + e^{-1.128}) = 0.76$.

For more information on ordered logistic regression in Stata, see [R] **ologit**.

9.8 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` is 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 defined by the user.

`predict statvar, statistic` saves the values of a selected statistic in the new variable `statvar`. The name of the new variable is defined by the user.

`estat gof` calculates the Pearson χ^2 test.

`estat classification` calculates the classification table.

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

`xb` calculates predicted logits.

`deviance` calculates deviance residuals.

`residuals` calculates Pearson residuals.

`rstandard` calculates standardized Pearson residuals.

`dx2` calculates the Hosmer–Lemeshow goodness-of-fit statistic.

`dbeta` calculates the Pregibon Delta–Beta goodness-of-fit statistic.

`number` numbers the covariate patterns.