8 Introduction to linear regression

In the last chapter, we described the distributions of a few different variables for various subgroups. For example, we compared the distributions of income and political party affiliation for men and women using a number of techniques. One of those techniques was cross-tabulation, which we used to examine the relative frequencies of votes cast for different groups formed by the values of a second variable—`gender`, in this case (page 142). Applying a different technique to the `income` variable, we compared the distribution of income for men and women using statistics such as means, quantiles, and standard deviations (page 156). In other words, we looked at how income *depends* on gender. Therefore, `income` was our dependent variable, and our *independent* variable was `gender`.

The techniques described in chapter 7 provide a reasonably good representation of your data if you want to compare the distribution of one variable for a few different subgroups formed by a second variable. However, if you are interested in the relationship between two variables with many categories, a "scatterplot" may be more useful. A scatterplot is a graphical representation of the joint distribution of two variables. When you draw the scatterplot, each observation is plotted in two-dimensional space (in other words, along two axes). The coordinates of each point are the values of the variables for that particular observation. The values of the independent variable are graphed on the x-axis, while the values of the dependent variable are graphed on the y-axis.

Three different examples[1] of scatterplots can be seen in figure 8.1.

[1] The data for these examples are taken from the WHO and UNICEF web sites. Detailed information are provided as notes in the data file (see footnote 8 on page 97). These data are included as Stata datasets in the data package provided with this book.

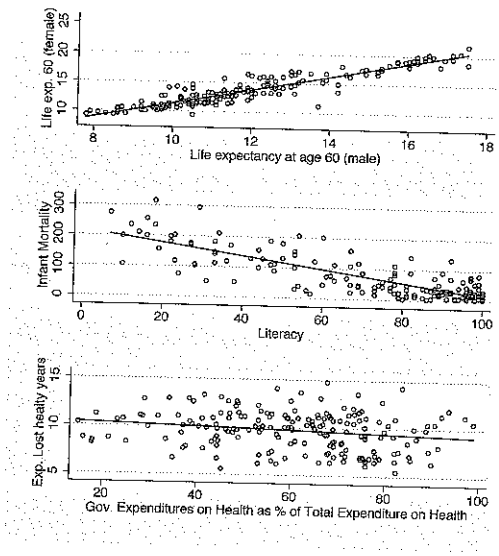


Figure 8.1: Three different scatterplots

`grscatter.do`

The first scatterplot shows data from 188 nations on the life expectancy at age 60 for females plotted against the life expectancy at age 60 for males. The dots are distributed from the lower left-hand corner to the upper right-hand corner. This suggests that high life expectancies for males go along with high life expectancies for females. Cases such as these are called "positive" relationships.

The second scatterplot depicts the relationship between infant mortality and female literacy. There we find the data points for 162 nations spreading out from the upper left-hand corner to the lower right-hand corner. This means that the higher the female literacy rate in a country is, the lower is the observed infant mortality rate. This is called a "negative" relationship.

The third scatterplot shows the relationship between the expected lost healthy years at birth for males and the governmental health expenditures as a percentage of total health expenditures. In this case, the observations from the 190 different countries are distributed fairly evenly over the entire diagram. The relationship between health expenditures and lost healthy years is therefore not obvious. We can, at best, find a weak relationship.

All three graphs contain a solid straight line that summarizes the relationship between the two variables and is called a "regression line". In the first scatterplot example, the dots are close to the regression line; there we have a *strong* correlation. In contrast, balloonlike clouds of dots, as in the third example, indicate a *weak* correlation. One way to measure the strength of the correlation is Pearson's correlation coefficient r. A Pearson's correlation coefficient of 0 means that no relationship can be observed between the

two variables. Both -1 and $+1$ represent the strongest possible observed relationships, but -1 indicates a negative relationship and $+1$ indicates a positive relationship.

Regardless of the strengths of correlation, there is not necessarily a causal relationship between the variables. The life expectancy of women is most likely not caused by the life expectancy of men. You can instead think of a common cause for both of them. You could hypothesize on the causal link between literacy and infant mortality, but neither scatterplots nor regression lines can test such an assumption (King, Keohane, and Verba 1994 and Berk 2004).

Creating scatterplots for different values of r is useful for getting an idea of the relationship. You can practice doing so by using a small demonstration we wrote for Stata.[2] Type

```
. do cplot 0.5
```

and you will see a scatterplot of two variables whose correlation coefficient is $r = 0.5$. You can vary the strength of the relationship by changing the number you enter for r after the `do cplot` command.

A simple linear regression analysis aims to characterize the relationship between one dependent variable and one independent variable with a straight line. A straightforward generalization of this is multiple linear regression analysis, which characterizes the relationship between one dependent and *more than one* independent variables. Note that the term "multivariate regression" is reserved for a technique for more then one *dependent* variables.

We begin by outlining the basic principle behind simple linear regression in section 8.1 and then discuss the relationship between the estimated parameters and the true population parameters in section 8.5. We then extend the model to deal with multiple independent variables in section 8.2. Linear regression analysis requires us to make several assumptions, and section 8.3 introduces several techniques to check those assumptions. Refinements of the basic model are the subject of section 8.4. We then discuss alternative methods of computing standard errors and other extensions of the linear regression model in section 8.6.

While we will explain some of the statistical background, our main purpose is to show you how to perform regression analysis with Stata. You will need to do additional reading to gain a full understanding of regression analysis. Books that work well with our approach are Hamilton (1992) and Fox (1997). We also highly recommend that you read Berk (2004) for a critical discussion of common mistakes.

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

8.1 Simple linear regression

8.1.1 The basic principle

In this section, we will introduce terms such as "OLS", "RSS", "predicted values", and "regression". If you are already familiar with these terms, you may skip this section.

The basic principle of all regression models is straightforward. To describe the relationship between your variables, you are looking for an equation that allows you to predict the values of a *dependent* variable as well as possible with the help of one or more *independent* variables. As an example, consider the following situation.

You have a hunch that the size of someone's dwelling is determined by his or her net income. You believe that the higher someone's income is, the larger the home will be. At the same time, you know that all apartments have a certain minimum size. You could formalize your suspicion about the relationship between income and home size with the aid of a simple equation:

$$\widehat{y}_{\text{Lopez}} = b_0 + b_1 x_{\text{Lopez}} \qquad \text{with} \quad b_0, b_1 > 0 \tag{8.1}$$

The Lopez family's predicted home size ($\widehat{y}_{\text{Lopez}}$) is calculated by assuming that all homes are at least b_0 square feet and adding to that a fraction b_1 of the family's net household income.[3] The term b_1 accounts for the fact that income is measured in, say, dollars, while home size is measured in square feet. The parameters b_0 and b_1 are assumed to be the same for all households and are called the "regression parameters" or "regression coefficients".

Now you might argue that income is not the only variable that affects home size. For example, family size or the ages of family members might play a role, as well. You may, in fact, come up with any number of factors that might affect home size. If you do not know all the factors, the estimated home size $\widehat{y}_i$ you calculate using the above equation will always deviate from the observed values. This deviation is called the "residual". In general, you might represent the relationship between an individual's actual home size y_i, the predicted size of her home $\widehat{y}_i$, and the residual e_i in the following way:

$$y_i = \widehat{y}_i + e_i \tag{8.2}$$

or, using your first hunch,

$$y_i = \underbrace{b_0 + b_1 x_i}_{\widehat{y}_i} + e_i \tag{8.3}$$

[3] The symbol $\widehat{y}_i$ is always used for the predicted values of the dependent variable. So $\widehat{y}_{\text{Lopez}}$ is the predicted value for the Lopez family.

In (8.3), the value of y for observation i is described as a linear combination of x_i and some noise. The coefficients b_0 and b_1 are constants and are independent of the individuals i; e_i represents noise due to factors not accounted for in the equation for each individual i.

The goal of your analysis is to find values for b_0 and b_1. You can use a number of techniques to find the regression parameters. Here we will limit ourselves to the one that is the simplest and in the widest use: ordinary least squares (OLS). The point of this technique is to make the difference between the predicted values and the observed values as small as possible.

To understand what this means, look at the scatterplot in figure 8.2. Try to find a straight line that depicts the relationship between the two variables. You will find that not all points lie on one single line. You might try to draw a straight line among the points in such a way that the distances between the points and the line are as small as possible. To find these distances, you might use a ruler.

The goal was to minimize the differences across all points, so looking at one of the distances will not provide you with enough information to choose the best line. What else can you do? You could try adding the distances up for all the points. If you did this, you would notice that negative and positive distances could cancel each other out. To find a way around this problem, you might use the squared distances instead.

If you drew several straight lines and measured the distances between the points and every new line, the straight line with the smallest sum of squared distances would be the one that reflects the relationship the best. This search for the line with the best fit is the idea behind the OLS estimation technique: it attempts to minimize the sum of squared residuals (e_i^2). The points on the straight line represent the predicted values of ($\widehat{y}_i$) for all values of X. If your model fits the data well, all points will be close to the straight line and the sum of the squared residuals will be small. If your model does not fit the data well, the points will be spread out and the sum of the squared residuals will be large.

(*Continued on next page*)

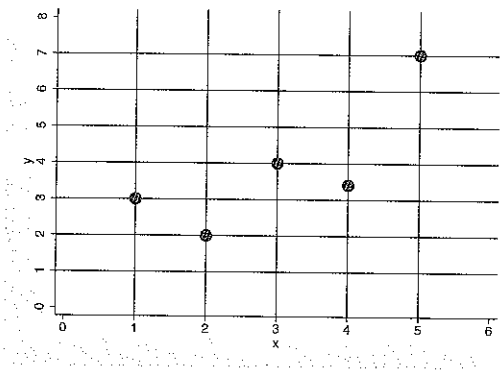


Figure 8.2: Exercise for the OLS principle

grreg1.do

We have prepared a small demonstration of the OLS solution to the regression problem in figure 8.2. Typing

```
. do grreg2.do
```

causes figure 8.2 to be displayed with the regression line.

We can also present the OLS principle a bit more formally. We are looking for those parameters (b_0, b_1) in (8.3) for which the sum of the squared residuals (the residual sum of squares, abbreviated RSS) is at a minimum. Those parameters are the y-axis intercepts and the slopes of the lines we drew. A search for the best fit using a trial-and-error technique like the one described above would be very time consuming. Using mathematical techniques to minimize the RSS is an easier way to find our parameters that more reliably leads to the correct solution. Mathematically, the RSS can be written as the difference between the observed and predicted values:

$$\text{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \tag{8.4}$$

Substituting $\widehat{y}_i$, we can write the above equation as

$$\text{RSS} = \sum e_i^2 = \sum (y_i - b_0 - b_1 x_i)^2 \tag{8.5}$$

Now that we have defined the RSS mathematically, we can use the OLS technique to minimize it.[4] This means that we must find values for b_0 and b_1 for which (8.5) is

[4]The exact mathematical procedure for this technique has been presented in a number of different ways. For fans of a graphic interpretation, we recommend Cook and Weisberg (1999) or Hamilton (1992) to start with.

as small as possible. To do this, we can take the first partial derivatives of (8.5) with respect to b_0 and b_1, set them equal to zero, and solve for b_0 and b_1. At this point, it is not particularly important that you be able to take the derivative yourself. You should, however, be aware that the entire technique is nothing more than a search for the minimum of a function with two unknowns.

If, on the other hand, you wish to review the high school and college math necessary for taking partial derivatives, you can find a helpful review in Hagle (1996, 38–58).[5]

Before we continue with the mathematics, we will show you how to compute a regression with Stata. The next subsection explains how to do this, and you will see how easy and helpful it is to use statistical packages for these kinds of computations. But be careful: despite the simplicity of the computational work, you must always think carefully about what exactly you are doing. During the course of the chapter, we will look at substantive problems caused by naively applying these regression technique.

8.1.2 Linear regression using Stata

In this subsection, we will explain how to calculate a regression with Stata. In the previous subsection, we voiced a suspicion that home size is influenced by net household income. You might now be interested in a specification of this relationship. A good place to begin would be with a linear regression of home size (`sqfeet`) on net household income (`hhinc`). The Stata command you will need to perform your regression is pretty simple:

[5] To reconstruct the transformations used in finding values for b_0 and b_1 for which RSS is at a minimum, you can do so as follows:

$$\frac{\partial \text{RSS}}{\partial b_0} = -2\sum y_i + 2nb_0 + 2nb_1 \sum x_i \tag{8.6}$$

If you set this partial derivative equal to zero and solve for b_0, you will get

$$b_0 = \overline{y} - b_1\overline{x} \tag{8.7}$$

Following the same principle, you can find the first partial derivative with respect to b_1:

$$\frac{\partial \text{RSS}}{\partial b_1} = -2\sum y_i x_i + 2b_0 \sum x_i + 2b_1 \sum x_i^2 = 0 \tag{8.8}$$

Now you replace b_0 with $\overline{y} - b_1\overline{x}$. After a few transformations, you end up with

$$b_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} \tag{8.9}$$

You can find a more detailed presentation of this derivation in Hamilton (1992, 33).

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

Source	SS	df	MS	Number of obs =	3126
Model	114301950	1	114301950	F(1, 3124) =	694.26
Residual	514327350	3124	164637.436	Prob > F =	0.0000
				R-squared =	0.1818
Total	628629300	3125	201161.376	Adj R-squared =	0.1816
				Root MSE =	405.76

sqfeet	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hhinc	.1786114	.0067787	26.35	0.000	.1653202	.1919025
_cons	600.2684	14.91459	40.25	0.000	571.025	629.5118

As you can see, the command consists of the `regress` statement and a list of variables. The first variable is the dependent variable, and the second is the independent variable. The output contains three different sections: the "table of ANOVA results" in the upper left-hand corner, the "model fit table" in the upper right-hand corner, and the "table of coefficients" in the bottom half of the output. We will now explain what is contained in each of these sections, beginning with the table of coefficients.

The table of coefficients

At the bottom of the table in the column labeled `Coef.`, you will find the regression coefficients, that is, the values for b_0 and b_1, from (8.3).

To the right of the column of regression coefficients are several other statistics used to measure the accuracy with which those coefficients have been estimated. We discuss those after we show how to interpret the coefficients and make predictions.

At this point, you may be asking yourself what all the numbers represent. The value for b_0 is written in the regression-output row labeled `_cons`. b_0 is 600.2684 in this example. According to this model, every family has a home that is at least 600 ft^2, regardless of whether the family has any income. The value for b_1 is stored in the row that begins with `hhinc` and is about 0.1786114. This means that according to the regression model, the home size will increase by about 0.18 ft^2 with every additional dollar of income.

Assuming that the Lopez family has a net monthly income of \$1,748 at its disposal, you can use (8.1) to estimate how big the family's home might be:

$$\widehat{y}_{\text{Lopez}} = 600.2684 + 0.1786114 \times \$1,748$$

You can calculate this amount directly within Stata using the `display` command, much as you would use a pocket calculator. Type

```
. display 600.2684 + 0.1786114 * 1748
912.48113
```

If you use the numbers displayed in the table of coefficients, you must deal with two problems: First, typing numbers by hand often leads to mistakes. In addition, the figures in the output have been rounded. For computations like the one above, we recommend using the results saved internally by Stata (see chapter 4). Commands that fit regression models are considered to be e-class in Stata, so you can look at the saved results with the command `ereturn list`. If you do this, you might find yourself searching in vain for the regression coefficients. This is because all the regression coefficients are stored in a matrix named `e(b)`. The easiest way to access the values contained in this matrix is to use the construction `_b[`*varname*`]`, where *varname* is replaced by the name of either an independent variable or the constant (`_cons`).

The computation for the Lopez family would then look like this:

```
. display _b[_cons]+_b[hhinc]*1748
912.48107
```

This number differs a bit from the number we computed above because the results saved by Stata are accurate to about the 16^{th} decimal place. You can see the effect of raising income by \$1 on home size. If you enter \$1749 instead of \$1748 as the value for income, you will see that the predicted value for home size increases by $b_1 = b_{\texttt{hhinc}} = 0.1786114\,\text{ft}^2$.

You might be interested not in an estimated home size for a family with a certain income, but in the actual home sizes of all families in our data who have that income. To see the sizes of the homes of all the families with a net household income of \$1748, you could use the following command:

```
. list sqfeet hhinc if hhinc==1748
```

As you can see here, the predicted home size of $912\,\text{ft}^2$ is not displayed, but rather various values between $474\,\text{ft}^2$ and $1227\,\text{ft}^2$ appear instead. The observed values of y_i differ from the predicted values $\widehat{y}_i$. These differences are the residuals.

If you want to compute the predicted values for every household in your dataset, you could use the saved regression coefficients.[6] To compute the predicted values this way, you would type[7]

```
. generate sqfeethat=_b[_cons]+_b[hhinc]*hhinc
```

This is the same principle that was used in the previous `display` command, except that the home size is not only predicted for the Lopez family, but for all families. The result of this computation is stored in the `sqfeethat` variable. We use the suffix `hat` to indicate that this is a "predicted" variable.[8]

[6] You can use the saved regression coefficients anywhere Stata expects an *expression*; see section 3.1.5.

[7] After entering this command, you will get a warning that some missing values have been generated. Those missing values are for all the families for whom the dataset contains no income information.

[8] You may have noticed that we placed a "hat" (circumflex) on y in the above equations to indicate a predicted value ($\widehat{y}$), as opposed to a value actually measured for a certain family (y).

If the last command seems like too much work, there is an easier and better way to get the same result: the `predict` command computes the predicted values after each regression command and stores them in a variable. If you enter

```
. predict yhat1
```

Stata will store the predicted values into the new variable `yhat1`,[9] which contains the same values as the `sqfeethat` variable. If you want to convince yourself that this is the case, type `list sqfeethat yhat1`. Because it is used after estimation, the `predict` command is called a "postestimation" command.

If you have already calculated the predicted values, it is easy to calculate the values of the residuals. They are just the differences between the observed and predicted values:

```
. generate resid1=sqfeet-sqfeethat
```

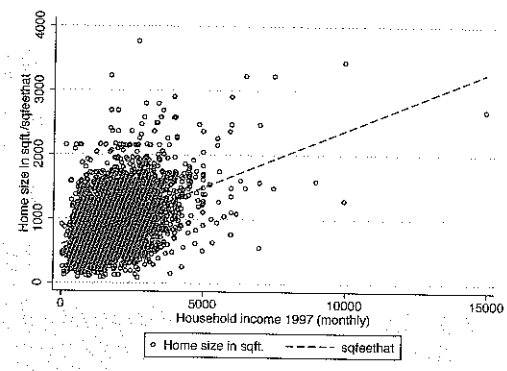
This difference is nothing more than the distance you measured between each point and the straight line in the figure on page 182.

You can also compute the residuals by using the `predict` postestimation command with the `residuals` option and specifying a variable name (here, `resid2`):[10]

```
. predict resid2, resid
```

Let's use a graph to look at the results. We might want to draw a scatterplot with `sqfeet` against `hhinc`, overlaid by a line plot of the predicted values (`yhat1`) against `hhinc`.

```
. graph twoway (scatter sqfeet hhinc, msymbol(oh)) (line sqfeethat hhinc, sort)
```



[9] In each subsection of this chapter, we perform a separate computation of the predicted values, meaning that we compute variables with predicted values a number of times. We use the name `yhat` with a running number for each of these variables. This running number has no meaning but is simply used to denote a new computation.

[10] Please resist the temptation to set `e` as a name for the residuals. The name `e` may, in principle, be a valid variable name, but using it might lead to confusion if scientific notation is used for numbers. See section 5.1.1 for a list of variable names you should avoid.

Standard errors

Thus far, we have treated the coefficients of our regression model as if they have been determined without any level of uncertainty. However, typically your dataset is only a random sample from a large population. The coefficients that you obtain from `regress` are therefore only estimates of the values that describe the entire population. If you were able to collect a second random sample from the same population, you would obtain different estimates. We therefore need a way to describe the variability that we would obtain if we were to apply our estimator to many different samples. Said slightly differently, we need a way to determine the "precision" with which the coefficients we obtained from our sample estimate the population parameters. Standard errors, which are really nothing more than sample standard deviations associated with estimated coefficients, are our solution.

Standard errors allow us to perform statistical inference; that is, they allow us to tests hypotheses about the underlying population parameters. For example, we might want to test whether the parameter on a given independent variable is zero, which means that the variable has no impact on the dependent variable.

Note that techniques for statistical inference are frequently misused. These techniques are based on a series of assumptions, which tend to be fulfilled only in high-quality samples. In the context of linear regression, the default statistical inference techniques assume a simple random sample from a large population, as well as uncorrelated errors and homoskedasticity. The confidence intervals are, in addition, based on the assumption of normally distributed errors. You will have to evaluate whether these assumptions hold or not. In section 8.3, we present several techniques that can be used to check those assumptions.

Here we give a few illustrations of the techniques for statistical inference. To begin with, compute the following regression model to analyze home size:

```
. regress sqfeet hhinc
```

On the right side of the table of coefficients, you see for each coefficient its 95% confidence interval boundaries; for household income, these are 0.165 and 0.192. When you think of confidence intervals, remember that if we were to draw many random samples out of the population and compute for each one the regression coefficient and the corresponding confidence intervals around this coefficient, then 95% of all intervals would contain the "true" coefficient for the population.[11]

For a quick check, you can determine whether the value zero is included in the confidence limits. If so, you can assume that the corresponding independent variable has no influence on the dependent variable in the population. Often you will see the t value used to determine the significance of a coefficient, meaning statistical significance, not substantive significance. With the help of the t distribution, this significance test tells you how likely it is that you would observe a value at least as extreme as the

[11] This does *not* mean that the true value is between the interval limits with a probability of 0.95.

particular coefficient you observed under the assumption that the 'true' coefficient in the population is *zero* (null hypothesis). The probability of observing a given t value with a given sampling size under the null hypothesis is shown in column $P > |t|$. A small value (e.g., smaller than 0.05) in this column tells you only that it is unlikely that you will observe a value like the one you would compute if the true coefficient in the population were *zero*. This means that your test is based on the hypothesis that household income has *no* influence at all on home size—a statement that you probably would not dare to make. We therefore recommend that you keep track of the confidence intervals and the effect size itself.

To compute the 95% confidence interval, you add or subtract roughly 1.96 times the standard error to the regression coefficient.[12] You can obtain the standard error of a particular explanatory variable k through dividing the standard error of the residuals (s_{e_i}) by the sum of the squared residuals resulting from a regression of k on all other independent variables.

The table of ANOVA results

ANOVA is short for "analysis of variance". We use the term "table of ANOVA results" to describe the upper left-hand section of the Stata regression output, where you will find the variation in the dependent variable divided into an explained portion and an unexplained portion. For handy reference, we reproduce here the table of ANOVA results that you have already seen on page 184:

Source	SS	df	MS
Model	114301950	1	114301950
Residual	514327350	3124	164637.436
Total	628629300	3125	201161.376

We can learn a bit more about the table of ANOVA results using a fictional example. Say that you are asked to predict the size of an apartment belonging to a student named Paul. If you do not know anything about Paul, you might answer that his apartment is as big as the average student apartment. In this case, your guess would be reasonable because the average apartment size is the value with which you get the smallest squared error. In other words, the mean apartment size is the OLS estimate of apartment size.

In table 8.1, we have listed the apartment sizes and household sizes of three students in a hypothetical city. The average student apartment size in that city is 590 ft^2, which we calculated using data for all the students in the city, not just the ones listed here.[13]

[12] The exact value varies with the sample size; however, 1.96 is a good approximation for a sample size above 30.

[13] We got the idea of using use a table like this one from Hair et al. (1995).

If you use $590\,\text{ft}^2$ to estimate the size of Paul's apartment, you end up with a number that is $160\,\text{ft}^2$ too high.[14] If you also use the mean to estimate the other students' apartment sizes, then in one case you make a correct prediction and in the other case you underestimate the student's apartment size by $270\,\text{ft}^2$. If you take the squares of these differences and sum them, the result is a total squared deviation of $98,500\,\text{ft}^4$. This number is usually called the "total sum of squares" (TSS). In general,

$$\text{TSS} = \sum (y_i - \overline{y})^2 \tag{8.10}$$

This corresponds to the expression you find in the numerator of the formula for the variance (s^2). The TSS is therefore sometimes also called the "variation".

Maybe you should not make your prediction using only the mean. You might wish to make use of other information you have about the students. After all, it is reasonable to assume that the size of the apartment increases with the number of people living there. If all the students you know have bedrooms that are about $160\,\text{ft}^2$, you might think this number holds true for most other students. So the apartment would have to have at least $160\,\text{ft}^2$ for each of the students living there, but it is likely to be even larger. An apartment usually has at least one bathroom and a kitchen, and you might think that they take up about $320\,\text{ft}^2$ combined. You might describe this hypothesis using the equation below:

$$y_i = 320 + 160x_i \tag{8.11}$$

You could use that model to compute an apartment size for each household size. If you did this, you would calculate the difference between the actual apartment size and the apartment size you predicted with your model; this is the amount displayed in the last column of the table. To compare these differences with the TSS we calculated above, you would have to square these deviations and sum them. If you did this, you would have calculated the "residual sum of squares" (RSS) we introduced in section 8.1.1. For your hypothesis, the value of RSS is 8,600.

Table 8.1: Apartment and household size

	Apt. Size	City	Diff.	HH Size	Estim.	Residual
Paul	430	590	−160	1	480	−50
John	590	590	0	2	640	−50
Ringo	860	590	+270	3	800	+60

[14] In table 8.1, the difference between the observed value and the predicted mean is calculated as follows: $430 - 590 = -160$.

If you subtract the RSS from the TSS, you get the model sum of squares (MSS), which indicates how much you have been able to improve your estimation by using your hypothesis:

TSS	=	98,500
−RSS	=	8,600
= MSS	=	89,900

The squared residuals that you get when you use household size to predict apartment size are about 89,900 smaller than the ones you got without taking this knowledge into account. That means that the actual apartment sizes are much closer to your predicted values when you use household size in making your prediction.

The MSS therefore can be regarded as a baseline to measure the quality of our model. The higher the MSS, the better are your predictions compared with the prediction based solely on the mean. The mean can be regarded as the standard against which to judge the quality of your prediction.

In the ANOVA part of the regression output, you will find information about the MSS, RSS, and TSS in the column labeled `SS`. The first row of numbers (`Model`) describes the MSS, the second (`Residual`) describes the RSS, and the third (`Total`) describes the TSS. If you look at the output on page 188, you will see that our RSS is 514327350. The sum of the squared residuals taking the mean as the estimate (TSS) is 628629300, and the difference between these two quantities (MSS) is 114301950. These and all other numbers in the table of ANOVA results are relevant for the model fit table.

The column labeled `df` contains the number of degrees of freedom.[15] The number of degrees of freedom equals the number of unknowns that can vary freely. For the MSS, the number of degrees of freedom is just the number of independent variables included in the model, that is, $k-1$, where k is the number of regression coefficients (the constant and all independent variables). The number of degrees of freedom for the RSS is $n-k$, where n is the number of observations. The number of degrees of freedom for the TSS is $n-1$. The last column contains the average sum of squares (MS). You may want to compute these numbers yourself by dividing the first column by the second column (the number of degrees of freedom).

The model fit table

Here, once again, is the model fit table from page 184:

```
Number of obs =    3126
F(  1,  3124) =  694.26
Prob > F      =  0.0000
R-squared     =  0.1818
Adj R-squared =  0.1816
Root MSE      =  405.76
```

[15] For a very well-written explanation of the concept of degrees of freedom, see Howell (1997, 53).

In the previous section, we showed that the MSS tells you how much the sum of the squared residuals decreases when you add independent variables to the model. If you were looking at models with different independent variables, you might want to compare the explanatory power of those models using the MSS. You could not, however, use the *absolute* value of the MSS to do so. That value depends not only on the quality of the model, but on how much influence the squared residuals (TSS) had there in the first place.

To compare models, you must look at how much the model reduces the squared residuals relative to the total amount of squared residuals. You can do this using the "coefficient of determination", or R^2:

$$R^2 = \frac{\text{MSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum e^2}{\sum (y_i - \overline{y})^2} \tag{8.12}$$

R^2 represents the squared residuals that are explained by the model as a share of the total squared residuals. When we say that the model explains a portion of the residuals, we mean that portion of the residuals of the model without independent variables that disappears when we use a model *with* independent variables. For this reason, R^2 is called the "explained variation" or the "explained variance". You will find this statistic in the model fit table of the Stata output, where it is called `R-squared`.

In our example, $R^2 = 0.1818$, meaning that household size (the independent variable in our model) "explains" 18 percent of the variation in apartment size.

R^2 is a useful indicator of a model's explanatory power, but it should not be considered in isolation. Unfortunately, often people evaluate the quality of a regression model only by looking at the size of R^2, which is not only invalid but very dangerous. In section 8.3, we will show you why.

One alternative to R^2 is the root MSE, which is the square root of the average residual of the model from the table of ANOVA results:

$$\text{root MSE} = \sqrt{\frac{\text{RSS}}{n-k}} \tag{8.13}$$

This statistic is easy to interpret, as it has the same units as the dependent variable. In our example, a root MSE of 405.76 can be interpreted as showing that we are, on average for our data, about 406 ft^2 "off the mark" in predicting a respondent's apartment size with our model. (This interpretation is not completely correct since it is not a literal average. After all, $\sqrt{\sum e_i^2} \neq \sum e_i$. But the above interpretation seems justified to us.)

There are two rows of the model fit table that we still haven't talked about: the rows labeled "$F(1, 3124)$" and "Prob > F". The values in these rows are included because we are using a sample to test our regression model and therefore want some measure of its significance.[16] The F value is calculated using the following equation:

[16] For more about the technical term "significance", see section 8.5.

$$F = \frac{\text{MSS}/k - 1}{\text{RSS}/n - k} \tag{8.14}$$

This F statistic is the ratio of the two values in the third column of the ANOVA table. It is F distributed and forms the basis of a significance test for R^2. In other words, the value of F is used to test the hypothesis that the R^2 we calculated with our sample data is significantly different from the population value of zero.[17] Said another way, you want to estimate the probability of observing the reduction in RSS in the model if, in fact, the independent variables in the model have no explanatory power.[18] The value listed for "Prob > F" gives the probability that the R^2 we calculated with our sample data will be observed if the value of R^2 in the population is actually equal to zero.

8.2 Multiple regression

Load `data1.dta` into working memory:

```
. use data1, clear
```

In the previous section, we introduced linear regression with one independent variable. A multiple regression is an extension of the simple linear regression presented in that section. Unlike in the simple regression, you can use several independent variables in a multiple regression. Analogous to (8.3), the model equation of the multiple linear regression is

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_{K-1} x_{K-1,i} + e_i \tag{8.15}$$

The equation for the simple linear regression has been extended with additional X variables and the attendant regression coefficients. You might want to use a model like this for two reasons.

In section 8.1.2, you calculated a simple linear regression of the dwelling size on household income. You were able to explain 18 percent of the variation in apartment size with this regression, and the average error in predicting apartment size was 406 ft^2. If you want to maximize the predictive power of our model, there is no reason to be satisfied with the performance of this simple model. You could improve the predictive power of our model by including other variables. This would be the primary reason for using a regression with more than one independent variable.

A second reason is a bit more complicated. In one of the previous sections, we used household income as an independent variable. Suppose that you also wanted to allow for the effect of household size. You have, however, reason to assume that household income

[17] A description of this relationship can be found in Gujarati (1995, 244–250).

[18] This F test is often called a test of the null hypothesis—that all coefficients but the constant are zero Gujarati (1995,.247). Incidentally, the confidence intervals might *not* contain the value zero, but the overall model may nevertheless not be significant.

is related to the size of the household, as additional family members might contribute to the overall household income. At the same time, it is reasonable to assume that households with more members need more space than those with fewer members. Thus the regression coefficient that we computed for household income may already include the effect of household size. In cases like this, the regression coefficient on household income is said to be "biased". You might try to combat this bias by including additional variables in the model.

In the next section, we are going to show you how to calculate a multiple linear regression model in Stata and then interpret the regression coefficients. After that, we will present some computations that are specific to this kind of regression. Finally, we will illustrate what is meant by the formal phrase "controlling for" when it is used for the interpretation of regression coefficients in multiple-regression models (section 8.2.3).

8.2.1 Multiple regression using Stata

The Stata command for computing a multiple regression is similar to that for simple linear regression. You enter additional variables to the list of variables; the order in which you enter them does not matter. You can apply the general rules for lists of variables (page 47), but remember that the dependent variable is always the first one in your list.

The output for the multiple linear regression resembles the one for a simple linear regression, except that, for each additional independent variable, you get one more row for the corresponding coefficient. Finally, you obtain the predicted values using the `predict` command as you did above.

For example, say that you want to compute a regression model of dwelling size that contains not only household size and household income, but a location variable for the difference between East and West Germany and an ownership variable indicating owned and rented living space. To do this, you will need to recode some of the variables:[19]

```
. generate owner = renttype == 1 if renttype < .
. generate east = state >=11 & state<=16 if state < .
```

Now you can compute the regression model:

(*Continued on next page*)

[19] See chapter 5 if you have any problems with these commands.

```
. regress sqfeet hhinc hhsize east owner
```

Source	SS	df	MS	Number of obs =	3125
				F(4, 3120) =	442.09
Model	227333588	4	56833397	Prob > F =	0.0000
Residual	401094539	3120	128555.942	R-squared =	0.3617
				Adj R-squared =	0.3609
Total	628428127	3124	201161.372	Root MSE =	358.55

sqfeet	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hhinc	.1168247	.0064601	18.08	0.000	.1041582	.1294912
hhsize	32.76697	5.185336	6.32	0.000	22.59996	42.93399
east	-99.99731	14.2274	-7.03	0.000	-127.8933	-72.1013
owner	383.6225	13.89444	27.61	0.000	356.3793	410.8657
_cons	524.1375	17.36074	30.19	0.000	490.0979	558.1771

Note that the number of observations has decreased from 3,126 to 3,125 because of missing values for the state variable. Observations that have a missing value in any of the variables are dropped when you fit the model. This is called "casewise deletion" of missing values. Type `search impute` to learn about other ways of dealing with missing values.

You interpret the coefficients in a multiple regression model in exactly the same way you do in the simple linear regression. The only difference is that the b coefficients are now calculated *controlling for* the effect of all the other independent variables. We will discuss the meaning of that phrase in section 8.2.3. At this point, we will confine ourselves to once again illustrating the formal interpretation of the coefficients.

The regression coefficients reflect the average change in the size of the dwelling as the independent variable in question increases by one unit. The coefficient might, for example, be interpreted as saying that "with each additional dollar of household income, the size of the dwelling increases by an average of about $0.117\,\text{ft}^2$". Similarly, the dwelling size increases by an average of about $32.77\,\text{ft}^2$ for each additional person in the household.

The variables `east` and `owner` are dummy variables, or variables that have only two categories, denoted by the values 0 and 1.[20] In principle, you interpret these variables the same way that you interpret all the other variables. For example, let's look at the `owner` variable, which has a value of 0 for all renters and 1 for all owners: for each unit by which the `owner` variable increases, the dwelling increases by an average of about $384\,\text{ft}^2$. Since a dummy variable can be increased by one unit only once, we could also say, "Owners live in dwellings that are, on average, about $384\,\text{ft}^2$ larger than the ones in which renters live." In the same way, the dwellings in East Germany are, on average, around $100\,\text{ft}^2$ *smaller* than the dwellings in West Germany.

The regression constant indicates how large a dwelling is whose observation has a value of 0 for all variables included in the model. This value would refer to dwelling

[20] There are other possibilities for coding binary variables (Aiken and West 1991, 127–130).

size for western households with no household income and no household members. This is a fairly uninteresting piece of information, as there is no person with a household size of zero. For this reason, it often makes sense to *center* continuous independent variables in a regression model, which means subtracting the mean from each individual value. For example, people with a value of 0 for a centered `income` variable would have the mean income, and the regression constant would therefore be the mean predicted income value. This procedure allows you to interpret the constant and the coefficients when interaction terms are used in the regression model. A method for centering a variable is described in chapter 4.

8.2.2 Additional computations

Adjusted R^2

In adding the two dummy variables and the household size variable to our regression model, you have increased R^2 from 18 to 36 percent. This is an obvious improvement in the explanatory power of our model, but you need to put this improvement in perspective: R^2 almost always increases if you add variables to the model.[21] The effect of these additional variables on R^2 is offset by the effect of additional observations. You can therefore safeguard against misleading increases in R^2 by making sure that you have enough observations to test your model. In the example above, the ratio between observations and independent variables that was used in the model is quite favorable. However, if you intend to work with only a small number of observations (e.g., if your dataset comprises country-level information for European countries) and you use many independent variables, R^2 will quickly become an unreliable measure.[22]

Perhaps it will be easier to understand why a small number of observations leads to a higher R^2 if you imagine a scatterplot with two points. These two points can be easily connected by a straight line, which is the regression line. Now you have "explained" all the variance, as there are no distances left between either of the points and the line. But does this mean that the two variables for which you made the scatterplot are really related to each other? Not necessarily. Imagine, for example, that you plotted the gross national products of Great Britain and Germany against the lengths of their coasts and drew a regression line. You would be able to explain the difference between the gross national products of Germany and Great Britain "perfectly"; at the same time, you would be forced to leave the scientific community.

Given the effects of the number of observations and the number of independent variables on R^2, you may want a more meaningful measure of your model's explanatory power. The adjusted R^2 (`Adj R-squared`) results from a correction that accounts for the number of model parameters k (everything on the right-hand side of your equation) and the number of observations (Greene 2003, 35)

[21] The only situation in which R^2 does not increase is when the coefficient of the additional variable is exactly equal to zero. In practice, this case is almost never observed.

[22] You will find a list of problems related to the use of R^2 in Kennedy (1997, 26–28).

$$R_a^2 = 1 - \frac{n-1}{n-k}(1 - R^2) \tag{8.16}$$

where k is the number of parameters and n is the number of observations. As long as the number of observations is sufficiently large, the adjusted R^2 will be close to R^2.

Standardized regression coefficients

In our regression model, the coefficient for household size is much larger than the one for household income. If you look only at the absolute size of the coefficients, you might be tempted to assume that the household size has a larger influence on dwelling size than does household income. But you will recognize that the coefficients reflect how much a dependent variable changes if the independent variable is changed by one unit. In comparing the two coefficients, you are comparing the change in dwelling size if household income increases by one dollar with the change in dwelling size if the size of the household increases by one person.

To compare the effects of variables measured in different units, you will often use standardized regression coefficients (b_k^*), which are calculated as follows

$$b_k^* = b_k \frac{s_{X_k}}{s_Y} \tag{8.17}$$

where b_k is the coefficient of the kth variable, s_Y is the standard deviation of the dependent variable, and s_{X_k} is the standard deviation of the kth independent variable.

The standardized regression coefficients are often called beta coefficients, which is why you use the `beta` option to look at them. If you want to reexamine your coefficients, note that Stata displays the results of the last model (with no recalculation) if you type `regress` without a list of variables. If you do this, you end up with values for `beta` in the rightmost column of the table of coefficients[23]:

```
. regress, beta noheader
```

sqfeet	Coef.	Std. Err.	t	P>|t|	Beta
hhinc	.1168247	.0064601	18.08	0.000	.2789243
hhsize	32.76697	5.185336	6.32	0.000	.0942143
east	-99.99731	14.2274	-7.03	0.000	-.1009768
owner	383.6225	13.89444	27.61	0.000	.4113329
_cons	524.1375	17.36074	30.19	0.000	.

The beta coefficients are interpreted in terms of the effect of standardized units. For example, as household income increases by one standard deviation, the size of the dwelling increases by about 0.28 standard deviations. In contrast, a one-standard-deviation increase in household size leads to an increase in dwelling size of about 0.09

[23] The `noheader` option suppresses the output of the ANOVA table and the model fit table.

standard deviations. If you look at the beta coefficients, household income has a stronger effect on dwelling size than does the size of the household.

Understandably, using the standardized regression coefficients to compare the effect sizes of the different variables in a regression model is quite popular. But people often overlook some important points in doing so:

- You cannot use standardized regression coefficients for binary variables. Because the standard deviation of a dichotomous variable is a function of its skewness, the standardized regression coefficient gets smaller as the skewness of the variable gets larger.[24]
- If interaction terms are used (see section 8.4.2), calculating b_k^* using (8.17) is invalid; if interactions are included in your model, you cannot interpret the beta coefficients provided by Stata. If you want to study effect sizes with beta coefficients that are appropriate for interactions, you must transform all the variables that are part of the interaction term in advance, using a z standardization (Aiken and West 1991, 28–48).
- You should not compare standardized regression coefficients computed with different datasets, as the variances of the variables will likely differ among those datasets (Berk 2004, 28–31).

8.2.3 What does "under control" mean?

The b coefficients from any regression model show how much the predicted value of the dependent variable changes with a one-unit increase in the independent variable. In a multiple-regression model, this increase is calculated *controlling for* the effects of all the other variables. In other words, we see the effect of changing one variable by one unit while holding all other variables constant. In this section, we will explain this concept in greater detail. We will do so using a simpler version of the regression model used above. Here only the regression coefficients are of interest to us:

```
. regress sqfeet hhsize hhinc, noheader
```

sqfeet	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hhsize	40.96258	5.809275	7.05	0.000	29.5722	52.35297
hhinc	.1650224	.0069971	23.58	0.000	.1513031	.1787418
_cons	520.5754	18.6216	27.96	0.000	484.0635	557.0872

[24] To make this point clear, we wrote a small do-file demonstration: `anbeta.do`. This program computes 1,000 regressions with a dichotomous independent variable that takes on the values 0 and 1. In the first regression, no observation has a value of 1 for the independent variable. In each additional regression, the number of observations where $X = 1$ increases by one, until the last regression, where all cases have the value 1 for the independent variable. A figure is drawn with the beta coefficients from each of those 1,000 regressions.

Look for a moment at the coefficient for household income, which differs from the coefficients we calculated both for the simple model (page 184) and for the multiple model (page 194): What is the reason for this change? To find an answer, you need to compute the coefficient for household income in a slightly different way: To begin, compute the residuals of the regression of dwelling size on household size:

```
. regress sqfeet hhsize, noheader
```

sqfeet	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hhsize	79.88333	5.946378	13.43	0.000	68.22431	91.54235
_cons	738.6072	17.16877	43.02	0.000	704.9445	772.2699

```
. predict e_fs, resid
(81 missing values generated)
```

When you do this, you create a new variable that stores the residuals: `e_fs`. Before continuing, you should give some serious thought to the meaning of those residuals.

We suggest that the residuals reflect the size of the dwelling adjusted for household size. In other words, the residuals reflect that part of the dwelling size that has nothing to do with household size. You could also say that they are that part of the information about dwelling size that cannot already be found in the information about the household size.

Now compute the residuals for a regression of household income on household size:

```
. regress hhinc hhsize, noheader
```

hhinc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hhsize	227.0473	14.14031	16.06	0.000	199.3223	254.7723
_cons	1335.758	40.74527	32.78	0.000	1255.869	1415.648

```
. predict e_hh, resid
(139 missing values generated)
```

These residuals also have a substantive interpretation. If we apply the above logic, they reflect that part of household income that has nothing to do with household size. They therefore represent household income *adjusted for* household size.

Now compute a linear regression of `e_fs` on `e_hh`.

```
. regress e_fs e_hh, noheader
```

e_fs	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
e_hh	.1650187	.006996	23.59	0.000	.1513014	.178736
_cons	-1.354776	7.200201	-0.19	0.851	-15.47238	12.76283

Take a close look at the b coefficient for `e_hh`, which corresponds to the coefficient in the multiple-regression model you calculated above.[25] If you interpreted this coefficient the same as one from a simple linear regression, you might say that dwelling size, adjusted for household size, increases about $0.165\,\text{ft}^2$ with each additional dollar of household income, adjusted for household size. The same interpretation holds true for the coefficients in the multiple-regression model. The regression coefficients in the multiple-regression model therefore reflect the effect of the independent variable in question on the dependent variable, adjusted for the effect of all other independent variables. This is what "controlling for" means.

8.3 Regression diagnostics

It is so easy to compute a multiple regression model using modern statistical software packages that people tend to forget that there are several assumptions behind a multiple regression; if they do not hold true, these assumptions can lead to questionable results. These assumptions are called "Gauss–Markov assumptions".[26]

To illustrate the importance of the underlying assumptions, open the data file `anscombe.dta` , and compute the following regression models:[27]

```
. use anscombe, clear
. regress y1 x1
. regress y2 x2
. regress y3 x3
. regress y4 x4
```

Note the estimated results for each regression model: the estimated coefficients, the variance of the residuals (RSS), and the explained variance R^2. Evidently you cannot find any difference between these four models just by looking at the numbers: you got an R^2 of 0.67 in all four models. The constant (or intercept) is 3, and the slope of the regression line is 0.5. If you did not know about the regression assumptions or regression diagnostics, you would probably stop your analysis at this point, supposing that you had a good fit for all models.

Now draw a scatterplot for each of these variable combinations, and then consider which model convinces you and which one does not; you can do this by typing the commands `scatter y1 x1`, `scatter y2 x2`, etc., one after each other. We actually used `granscomb1.do` to produce the graphs. But we have put them on page 200, so as not to spoil the surprise.

The scatterplots in figure 8.3 show, without a doubt, that there is good reason to be cautious in interpreting regression results. Looking at just the R^2 or coefficients can be very misleading!

[25] Differences are due to rounding errors.

[26] If you are already familiar with the Gauss–Markov assumptions and how to check them, you might want to get a quick overview of regression diagnostics within Stata by typing `help regress postestimation`.

[27] The data file was created by Anscombe (1973).

Now we want to show you how to check the Gauss–Markov conditions and correct any violations of them. Most of the diagnostic techniques we present are graphical, so you will need to understand the basics of the Stata `graph` command (see chapter 6). For an overview of various graphical diagnostic techniques, see Cook and Weisberg (1994). See Berk (2004, chapter 9) for a discussion on the limitations and potential hazards of using regression diagnostics.

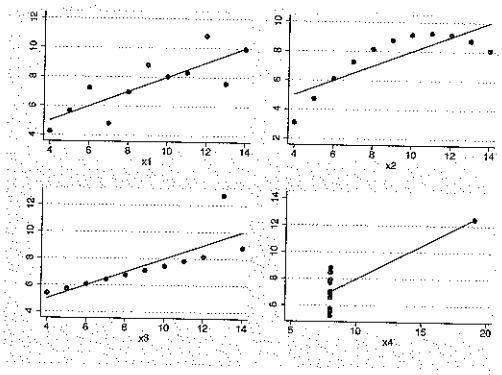


Figure 8.3: The Anscombe quartet

`granscomb1.do`

8.3.1 Violation of $E(\epsilon_i) = 0$

The unobserved influence on the dependent variable for each observation is called the error. By assumption, the average across observations of these errors is zero. This assumption may be violated if

1. the relationship between the dependent and independent variables is nonlinear,
2. some outliers have a strong effect on the regression coefficients,
3. some influential factors have been omitted that in fact are correlated with the included independent variables.

Violating $E(\epsilon_i) = 0$ results in biased regression coefficients, so it is therefore very important to verify that this holds. All the problems that showed up in the Anscombe quartet are due to violations of one or more of these assumptions.

There are special techniques for testing each of the problems named above. You can see all three possible causes using a residual-versus-fitted plot, which is a scatterplot of the residuals of a linear regression against the predicted values. For the last computed regression, you build the plot by typing

```
. regress y4 x4
. predict yhat
. predict resid, resid
. scatter resid yhat
```

or by using the `rvfplot` command, which generates one of the specialized statistical graphs mentioned in section 6.2.

```
. rvfplot
```

With `rvfplot`, you can use all the graphic options that are available for scatterplots. Figure 8.4 shows the residual-versus-fitted plots for all regressions in the Anscombe example. Note that in these graphs, the mean of the residuals is always equal to zero. This is true by definition. In a regression model, the regression coefficients are calculated so that the mean of the sample residuals is equal to zero. To fulfill the assumption that $E(\epsilon_i) = 0$, not only must the overall mean of the residuals be zero, but the mean of the residuals must be zero *locally*. This means that the mean of the residuals is zero for any slice of the x-axis. This is true only for the first and the last regression model.

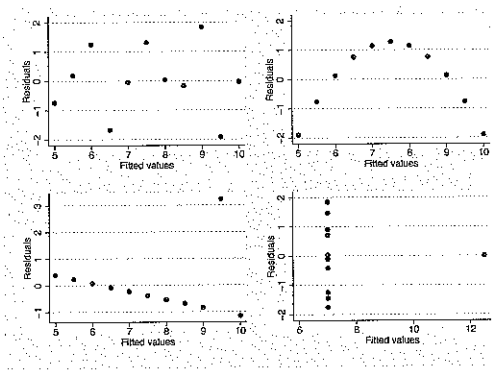


Figure 8.4: Residual-versus-fitted plots of the Anscombe quartet

`granscomb2.do`

Note that in a regression with only one independent variable, violations of the regression assumptions can be seen with a simple scatterplot of the dependent variable against the independent variable. The advantage of the residual-versus-fitted plot is that it also applies to multiple regression models.

In practice, a violation of $E(\epsilon_i) = 0$ is usually not as obvious as it is in the Anscombe data. For this reason, we will now introduce some special diagnostic tools for determining which of the three possibilities might be causing the violation of this assumption.

Linearity

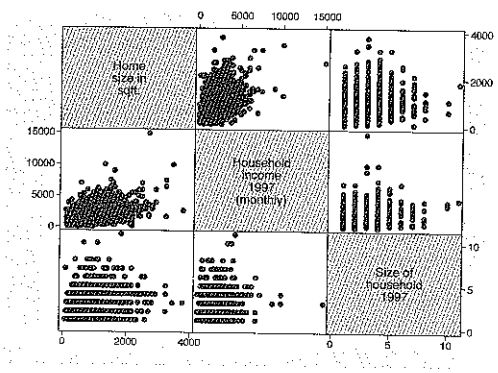
To understand the following examples, you might want to start with a regression of home size on household income and household size using the GSOEP data:

```
. use data1, clear
. regress sqfeet hhinc hhsize
```

One of the most important requirements for a linear regression is that the dependent variable can indeed be described as a linear function of the independent variables. To examine the functional form of the relation, you should use nonparametric techniques, where you try to have as few prior assumptions as possible. A good example is a scatterplot reflecting only some general underlying assumptions derived from perception theory.

You can use a scatterplot matrix to look at the relationships between all variables of a regression model. Scatterplot matrices draw scatterplots between all variables of a specified variable list. Here is an example:

```
. graph matrix sqfeet hhinc hhsize
```



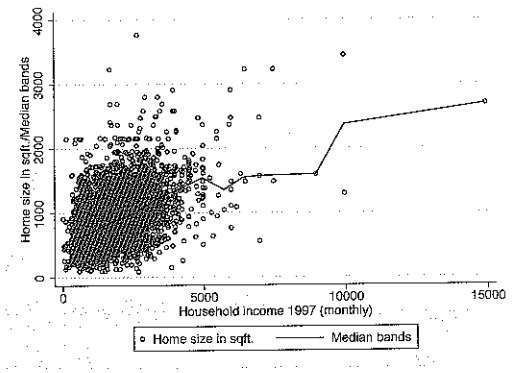
In each plot, the variable to the side of the graph is used as the Y variable, and the variable above or below the graph is used as the X variable. In the very first line of the figure are scatterplots of home size against all the independent variables of the regression model.

However, scatterplots often only show the functional form of a relationship for small sample sizes. If you deal with larger sample sizes, you will need more information to improve the scatterplot. For this purpose, Stata allows you to overlay scatterplots with a scatterplot smoother (Fox 2000).

One example of a scatterplot smoother is the "median trace". To construct a median trace, you divide the variable plotted on the x-axis of a twoway plot into strips and calculate the median for each strip. Then the medians are connected with straight lines. In Stata, you get the median trace as plottype `mband` of twoway graphs. The

`bands(k)` option of this plottype is used to decide the number of strips into which the x-axis should be divided. The smaller the number of bands, the smoother is the line.

```
. twoway (scatter sqfeet hhinc, ms(oh)) (mband sqfeet hhinc, bands(20) clp(solid))
```



The figure shows a linear trend for the majority of the data. The figure also shows that there are many outliers on both variables: income and home size.[28] Even if you can establish a linear relationship between two variables, that relationship may change when you include other variables in the regression model. That is, the functional form of a relation between two variables may change under the influence of other variables.

One clue about the relation between one independent variable (e.g., household income) and the dependent variable (home size) if you control for other independent variables (such as household size) is given by plotting the residuals against the independent variables.[29] But plotting the residuals against one of the independent variables does not indicate the exact shape of any curvilinearity. For example, a U-shaped relation and a logarithmic relation might produce the same plot under certain circumstances (Berk and Both 1995). [30]

The component-plus-residual plots—also known as partial residual plots—are a modification of the plot just described: they allow the determination of the functional form of the relation. Within the component-plus-residual plots, instead of using the residual, the product of the residual and the linear part of the independent variable is plotted against the other independent variables. What this means is shown in the following example.

To examine the linearity between home size and household size in a multiple regression model, you can first compute the regression of home size on household income and household size and save the residuals as `e1`:

[28] These might be observations that will heavily influence the regression result. In the next section, you will find out more about this issue.

[29] When the sample size becomes large, it is reasonable to use a scatterplot smoother.

[30] You must distinguish between these two kinds of relations: if there is a U-shaped relation, you must insert a quadratic term, whereas it might be sufficient to transform the dependent variable if there is a logarithmic relation (see section 8.4.3).

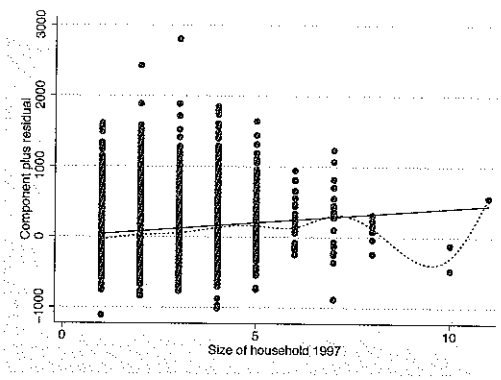
```
. regress sqfeet hhinc hhsize
. predict e1, resid
```

Then you can add the linear part of household size to the saved residuals and plot the resulting number against household size:

```
. generate e1plus= e1 + _b[hhsize]*hhsize
. twoway (scatter e1plus hhsize) (mband e1plus hhsize, bands(20))
```

You would end up with the same result if you used the wrapper `cprplot`, which is implemented in Stata; this wrapper will run the same procedure for any independent variable of your choice.[31] After `cprplot`, you enter the name of the independent variable for which you want to create the variable. The straight line in the resulting graph is equivalent to the regression line. We have also added a median spline, which is similar to the median trace but uses curves to connect the different medians.

```
. cprplot hhsize, mspline msopts(bands(20))
```



You might infer from the graph that home size decreases for a household size of seven or more. In this case, however, this is probably an effect of the unstable median computation within the upper bands of household size since there are few households with seven or more members.

Potential solutions

In our example, the relations seem to be linear. In the presence of nonlinear relations, you need to transform the independent variables involved or include additional quadratic terms in the equation; see section 8.4.3.

[31] You will also find the augmented component-plus-residual plot from Mallows (1986): `acprplot`. Also, instead of the median trace used here, you could use the locally weighted scatterplot smoother (LOWESS) (Cleveland 1994, 168). Then you would use the option `lowess`.

Influential cases

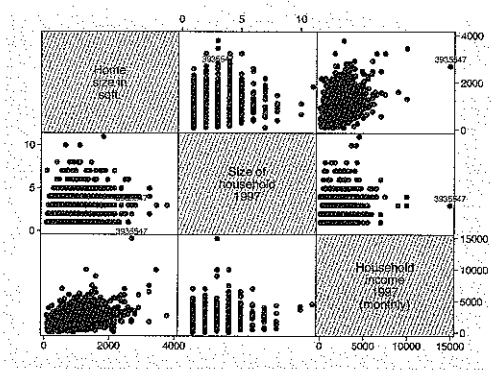
"Influential" cases are observations that heavily influence the results of a regression model. Mostly, these are observations that have unusual combinations of the regression variables included in the model (multivariate outliers). As an example, think of a person with a huge income living in a very small home.

It may not be possible to detect multivariate outliers in a bivariate scatterplot. Observations that show up as outliers in one scatterplot might in fact turn out to be normal if you controlled for other variables.

If, for example, the person mentioned had been interviewed in her secondary residence, the small home size is less surprising. Thus it is often possible to *explain* multivariate outliers. In these cases, the solution for this problem is to include a variable in the regression model that captures the explanation. In our example, you would have to include in the regression model a variable that indicates whether this is the primary or secondary residence.

You can find signs of influential cases using a scatterplot matrix that is built from the variables included in the regression model. As each data point of one of these scatterplots lies on the same row or column as that of the other scatterplot, you can locate conspicuous observations over the entire set of scatterplots (Cleveland 1993, 275). Our example illustrates this with the help of one observation, which we have highlighted.

```
. gen str label = string(persnr) if hhinc == 14925
. graph matrix sqfeet hhsize hhinc, mlab(label) mlabpos(6)
```



A more formal way to discover influential cases is to use DFBETAs. The computation of DFBETAs has a simple logic: First, you compute a regression model, and second compute it again with one observation deleted. Then you compare the two results. If there is a big difference in the resulting coefficients, the observation that was excluded in the second computation has a big influence on the coefficients. You then repeat this technique for each observation to determine its influence on the regression coefficients. You compute this for each of the k regression coefficients separately. More formally, the equation for computing the influence of the ith case on the kth regression coefficient is

$$\text{DFBETA}_{ik} = \frac{b_k - b_{k(i)}}{s_{e(i)}/\sqrt{\text{RSS}_k}} \tag{8.18}$$

where b_k is the coefficient of variable k and $b_{k(i)}$ is the corresponding coefficient without observation i; $s_{e(i)}$ is the standard deviation of the residuals without observation i. The ratio in the denominator standardizes the difference so that the influences on the coefficients are comparable (Hamilton 1992, 125).

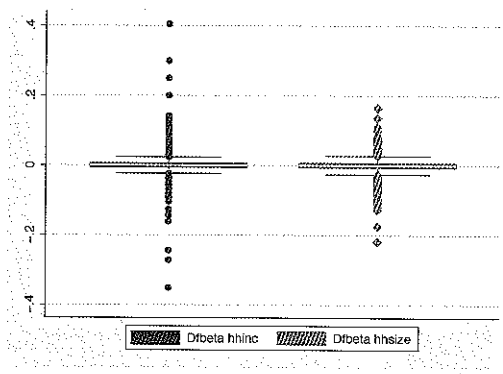
In Stata, you compute values for DFBETA_{ik} using the `dfbeta` command. You enter this command after the regression command, with a variable list in which you specify the coefficients for which you want to view the change. If you do not specify a variable list, all coefficients are used. The results of the command `dfbeta` are stored in variables whose names begin with "DF".

Typing

```
. regress sqfeet hhinc hhsize
. dfbeta
```

generates two variables: `DFhhinc` and `DFhhsize`. Both variables contain, for each observation, its influence on the regression coefficient. If there are indeed influential cases in your data file, you can detect this using

```
. graph box DF*
```



Values of $|\text{DFBETA}| > 2/\sqrt{n}$ are considered large (Belsley et al. 1980, 28). [32] In our model, several observations exceed this boundary value. With

```
. foreach var of varlist DF* {
. list persnr 'var' if (abs('var') > 2/sqrt(e(N))) & 'var' < .
. }
```

you obtain a list of these observations.[33]

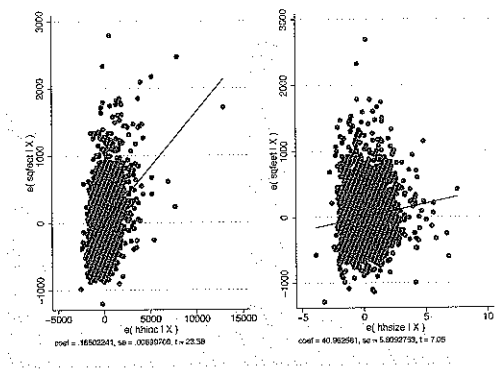
[32] Other authors use 1 as the boundary value for DFBETA (Bollen and Jackman 1990, 267).

[33] The command `foreach` is explained in section 3.2.2. The expression `abs()` is a general Stata function that returns the absolute value of the argument included in the parentheses (see section 3.1.6). Finally, `e(N)` is the number of observations included in the last regression model fitted (see chapter 4).

Another way to detect outliers is to use the added-variable plot (partial regression plot). To create the added-variable plot of the variable X_1, you first compute a regression of Y on all independent variables besides X_1. Then you compute a regression of X_1 on all remaining independent variables. You then save the residuals of both regressions and plot them against each other.[34]

In Stata, you can also create added-variable plots by using the `avplot` or `avplots` commands. `avplot` creates the added-variable plot for one explicitly named independent variable, whereas `avplots` shows all possible plots in one graph:

```
. regress sqfeet hhinc hhsize
. avplots
```



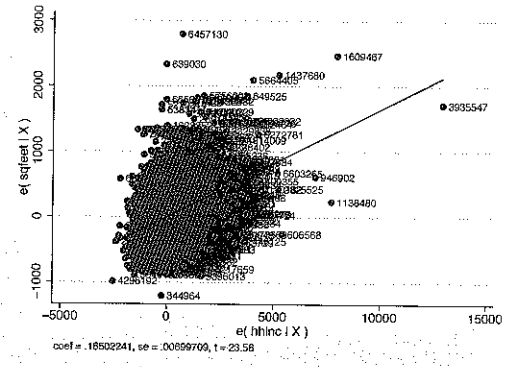
In these plots, points that are far from the regression line are "multivariate outliers". These kinds of observations have more potential to influence the regression results. In the figure above, some observations are conspicuous in that household income is higher than you would assume by looking at the values of the remaining variables. In the plot for household income, one observation in particular is cause for concern—the one with the largest house.

You can type

```
. avplot hhinc, mlabel(persnr)
```

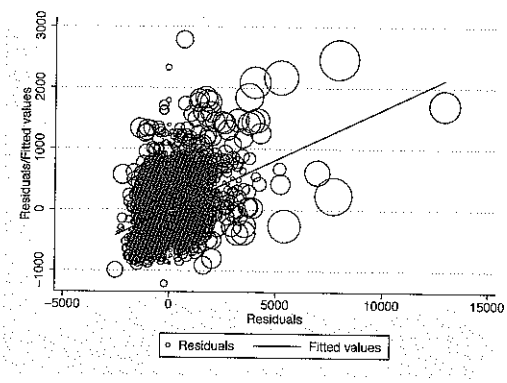
to identify the personal identification number of this observation:

[34]The logic behind added-variable plots corresponds to the way the b coefficients are interpreted in a multiple-regression model (see section 8.2.3). A scatterplot of the residuals that were created there would be an added-variable plot.



Hamilton (1992, 128–129, 141) recommends using an added-variable plot where the size of the plot symbol is proportional to DFBETA. To do this, you must create the plot yourself. In the multiple linear regression we used above, you would create such a plot for household income as follows:[35]

```
. regress sqfeet hhsize
. predict esqfeet, resid
. regress hhinc hhsize
. predict ehhinc, resid
. generate absDF = abs(DFhhinc)
. twoway (scatter esqfeet ehhinc [weight = absDF], msymbol(oh))
> (lfit esqfeet ehhinc, clp(solid))
```



In Stata graphs, you can control the size of the plot symbol using *weights*. Here it is not important which kind of weights (`fweights` or `aweights`, for example) you use. In this example, you need to pay attention to possible negative values of DFBETA. So you can compute the absolute values of DFBETA first and use these values for weighting.[36]

[35] For this example, we use the variable `DFhhinc`, which we created on page 206. The axes of this graph are both labeled as "residuals" automatically by the command `predict`. If you want to change these labels, see section 6.3.4.

[36] You will find some general remarks about weights in section 3.3.

The previous figure shows that the multivariate outlier has an appreciable influence on the regression line. Even stronger are two observations more to the left, but they cancel each other out. Altogether, we seem to find the influential cases mainly in the upper region of income, regardless of the other variables. Those few observations with high income have a disproportionately strong influence on the regression result.

So far, the impact of single observations has been examined separately for the different coefficients. If you have many independent variables, you will find it more complicated to interpret the numerous DFBETA values. With "Cook's D", you have a statistic available that estimates the effect of one observation on all regression coefficients simultaneously (Fox 1991, 84) and hence the influence of one observation on the entire regression model. You get this statistic by entering `predict` after the regression command:

```
. predict cook, cooksd
```

The idea behind this statistic is that the influence of one observation on the regression model is composed of two aspects: the value of the dependent variable and the combination of independent variables. An influential case has an unusual value on Y *and* an unusual combination of values on the Xs. Only if both aspects are present will the coefficients be strongly affected by this observation. The graphs in figure 8.5 make this clearer. The graphs present scatterplots of the home size against the income of five Englishmen in 1965, 1967, and 1971.

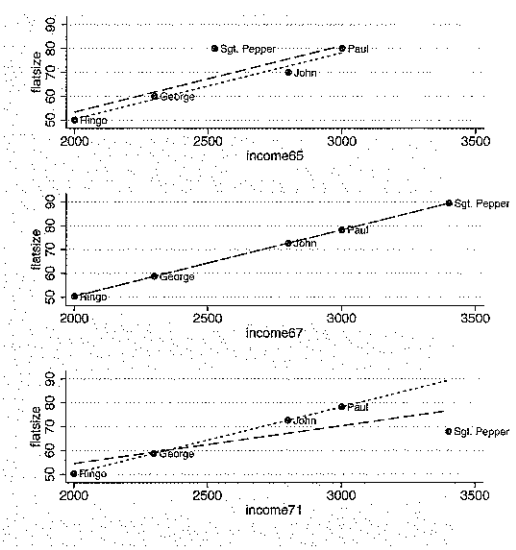


Figure 8.5: Scatterplots to picture leverage and discrepancy

`grbeatles.do`

In the first scatterplot, which shows the year 1965, Sgt. Pepper has an extraordinarily large home given his income. Sgt. Pepper's income is, however, anything but extraordinary: it is equal to the mean net income of the five Englishmen. We draw two regression lines in this picture. The dotted line is the regression line that results from a regression without Sgt. Pepper. When Sgt. Pepper is included in the regression, the regression line is shifted upwards. There is no change in the slope of the line (the b coefficient of income).

In the scatterplot for 1967, Sgt. Pepper has an extraordinarily high income. The size of his home corresponds, however, exactly to the square footage we would expect from our model. Sgt. Pepper has therefore an extraordinarily large value of X but, given this value for X, a quite common Y value. The regression lines that result from the regressions with and without Sgt. Pepper are identical in this case.

In the scatterplot for 1971, Sgt. Pepper has an extraordinarily high income and, for this income, an extraordinarily small home. Here both aspects mentioned above are present. Accordingly, the regression line changes.[37]

The idea that the effect of a certain point is determined by the extreme values of X and Y can be described mathematically as

$$\text{influence} = \text{leverage} \times \text{discrepancy} \tag{8.19}$$

where the leverage signifies how extraordinary the combination of the X values is (as in the second scatterplot) and the discrepancy signifies how extraordinary the Y value is (as in the first scatterplot). As leverage and discrepancy are multiplied, the influence of any given observation is equal to 0 if one or both aspects are missing.

To compute the influence as shown in (8.19), you need some measures of the leverage and the discrepancy. First, look at a regression model with only one independent variable. In this case, the leverage of a specific observation increases with its distance from the mean of the independent variable. Therefore, a measure of the leverage would be the ratio of that distance to the sum of the distances of all observations.[38]

When there are several independent variables, the distance between any given observation and the centroid of the independent variables is used, controlling for the correlation and variance structure of the independent variables (also see Fox 1997, 271). In Stata, you obtain the leverage value for every observation by using the `predict lev, leverage` command after the corresponding regression. When you type that command, Stata saves the leverage value of every observation in a variable called `lev`.

[37] Think of the regression line as a seesaw, with the support at the mean of the independent variable. Points that are far away from the support and from the regression line are the most influential points.

[38] Specifically,

$$h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{j=1}^{n} (x_j - \overline{x})^2} \tag{8.20}$$

To measure the discrepancy, it seems at first obvious that you should use the residuals of the regression model. But this is not in fact reasonable. Points with a high leverage pull the regression line in their direction, and therefore that they may have small residuals. If you used residuals as a measure of discrepancy in (8.19), you might compute small values for the influence of an observation, although the observation changed the regression results markedly.[39]

Hence, to determine the discrepancy you need a statistic that is adjusted for the leverage. The standardized residual e_i' is such a statistic. You can obtain the values of the standardized residuals by using the `predict varname, rstandard` command, which you can enter after a regression.[40]

After finding a statistic for both discrepancy and leverage, you can multiply the two statistics together in accordance with (8.19). But you should provide an appropriate weight for each value to be multiplied. We leave that task to the statisticians, one of whom—Cook (1977)—suggested the following computation

$$D_i = \underbrace{\frac{h_i}{1-h_i}}_{\text{leverage}} \times \underbrace{\frac{e_i'^2}{k+1}}_{\text{discrepancy}} \tag{8.21}$$

where e_i' is the standardized residual and h_i is the leverage of the ith observation.[41] Values of Cook's D that are higher than 1 or $4/n$ are considered large. Schnell (1994, 225) recommends using a graph to determine influential cases. In this graph, the value of Cook's D for each observation is plotted against its serial number within the data file, and the threshold is marked by a horizontal line.

To construct this graph, you must first compute the values for Cook's D after the corresponding regression:

```
. regress sqfeet hhsize hhinc
. predict cooksd, cooksd
```

Then you save the threshold in a local macro (`max`) using the number of observations in the last regression model, which is stored by Stata as an internal result in `e(N)` (see chapter 4 and section 11.2.1):

```
. local max = 4/e(N)
```

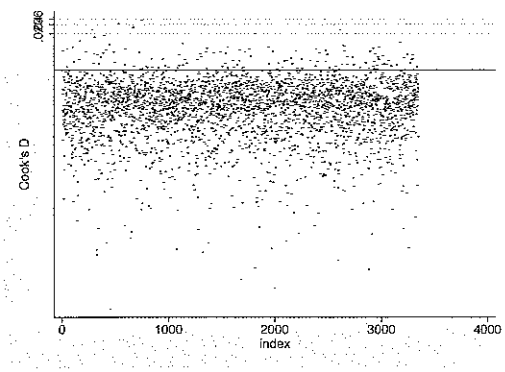
[39] This can be demonstrated with the fourth graph in the Anscombe quartet (page 200). If you computed the influence of this outlier with (8.19) and thereby used the residuals as a statistic for discrepancy, the influence of this outlier would be equal to 0.

[40] You can choose any variable name for *varname*.

[41] There is a very useful teaching tool you can see by typing the command `regpt`, which is taken from an ado-file programmed by the Academic Technology Services of the University of California, Los Angeles. To learn more about ado-files, see chapter 11; to learn more about ado-files provided over the Internet, see chapter 12.

Now you build a variable `index`, which contains the serial observation number, and use this variable as the x-axis on our graph. Next construct the graph with a logarithmic y-axis:

```
. generate index = _n
. graph twoway scatter cooksd index, yline('max') msymbol(p) yscale(log)
```



The figure shows a large number of observations that are above the critical value, especially those with a comparatively high income:

```
. generate bigcook = cooksd > 'max'
. tabulate bigcook, summarize(hhinc)
```

	Summary of Household income 1997 (monthly)		
bigcook	Mean	Std. Dev.	Freq.
0	1860	945	2975
1	2723	1931	226
Total	1921	1068	3201

In summary, the analyses you ran in this section show a clear finding: using these diagnostic techniques, you found a few observations with high incomes to be conspicuous. The results of the model are much more strongly affected by these few observations than by all the other observations with low, medium, or high (but not very high) income. In particular, the added-variable plot on page 208 shows that these influential observations are not problematic for the example because they influence the regression model only in the bivariate case. If you excluded all the observations with a very high income, the coefficients would stay pretty much as they are.

Potential solutions

You may wonder what to do when influential observations are present. If an influential case can be attributed unquestionably to a measurement error, you should either correct the error or delete the observation from the file. If influential observations result

from extreme values of the dependent variable, it is reasonable to use median regression (section 8.6.1).

Almost always, however, influential observations result from an incompletely specified model. Exceptional cases are in this case exceptional only because our theory explains them insufficiently. As in our example, where observations with a high income influence the regression extraordinarily, you should ask if another factor influences home size that is typically related to high (or to low) income. With *right-skewed* distributions, such as that of income, you may want to change the model to use the logarithm of household income instead of household income itself. In the current context, this means that household income is supposed to be in a logarithmic relation to home size: the higher the household income gets, the smaller is the change in home size with each additional dollar of household income.

Omitted variables

Variables are called "omitted variables" or omitted factors if they influence the dependent variable and are at the same time correlated with one or more of the independent variables of the regression model. Strictly speaking, nonlinear relations and influential cases are omitted factors, too. In the first case, you may have overlooked the fact that an independent variable does not have the same influence on the dependent variable throughout the whole range of the dependent variable. In the second case, you may have neglected to model an explicit error theory or overlooked a mechanism that would explain the outliers.

To figure out which variables have been omitted, you can begin by graphing the residuals against all variables that are not included in the model. But this is obviously only possible for those variables that are included in the data file. Even if these graphs do not show any distinctive features, there still may be a problem. This diagnostic tool is therefore necessary but not sufficient.

Identifying omitted factors is, first of all, a theoretical problem. Thus we warn against blindly using tools to identify omitted variables.

In trying to include all important influential factors in the model, there is an additional risk called *multicollinearity*. We will introduce an extreme case of multicollinearity in section 8.4.1 when we discuss how to include categorical independent variables in regression models. If there is a perfect linear relation between two variables of the regression model,[42] Stata will exclude one of them when calculating the model.

But even if the two variables are not a perfect linear combination of each other, some problems can arise: The standard errors of the coefficients might increase, and there might be an unexpected change in the size of the coefficients or their signs. You should therefore avoid including variables in the regression model haphazardly. If your model fits the data well based on R^2 but nevertheless has few significant coefficients, then multicollinearity may be a problem.

[42] For example, $x_1 = 2 + x_2$.

Finally, you can use the `vif` command to detect multicollinearity after regression. This command gives you what is called a variance inflation factor for each independent variable. See, for example, Fox (1997, 338) for an interpretation and explanation of this tool.

8.3.2 Violation of Var$(\epsilon_i) = \sigma^2$

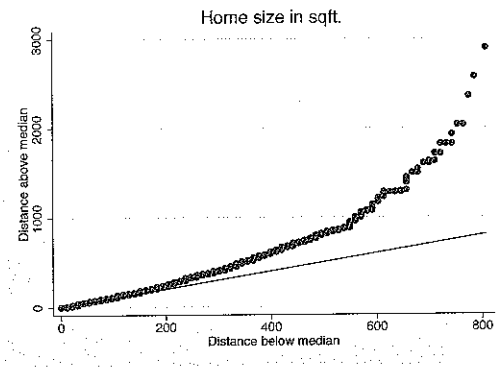
The assumption that $\text{Var}(\epsilon_i) = \sigma^2$ requires that the variance of the errors be the same for all values of X. This assumption is called homoskedasticity, and its violation is called heteroskedasticity. Unlike the violation of $E(\epsilon_i) = 0$, heteroskedasticity does not lead to biased coefficients. But when the homoskedasticity assumption is violated, the coefficients of a regression model are not efficient. With inefficient estimation, there is an increasing probability that a particular regression coefficient deviates from the true value for the population. Said another way, heteroskedasticity causes the standard errors of the coefficients to be incorrect, and that obviously has an impact on any statistical inference that you perform.

There are many possible reasons for heteroskedasticity. Frequently you find heteroskedasticity if the dependent variable of your regression model is not symmetric. To test the symmetry of variables, you will find the graphical techniques described in section 7.3.3 to be very useful.

Stata has a special technique for checking the symmetry of a distribution, called a "symmetry plot" (Chambers et al. 1983, 29). To construct a symmetry plot, you first determine the median. Then you compute the distances between the observations next in size and the median. In a symmetry plot, you plot these two quantities against each other. You do the same with the next observation, and so on. If all distances are the same, the plot symbols will lie on the diagonal. If the distances of the observations above the median are larger then those below, the distribution is right-skewed. If the reverse is true, the distribution is left-skewed.

In Stata, the `symplot` command graphs a symmetry plot of a given variable. Here we graph the symmetry plot for home size:

```
. symplot sqfeet
```



The figure shows an obviously right-skewed distribution of the variable home size. With this kind of distribution, there is risk of violating the homoskedasticity assumption.

The residual-versus-fitted plot (Cleveland 1994, 126) is the standard technique for examining the homoskedasticity assumption. We want to introduce one variation of this plot, which emphasizes the variance of the residuals. You therefore divide the x-axis into k groups with the same number of observations and then draw a box plot of the studentized residuals for each group.

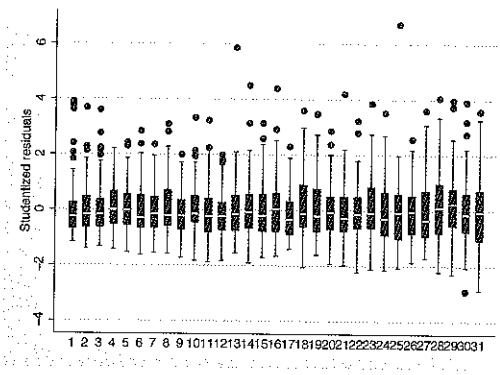
To do this, you again compute the regression model, the predicted values, and the studentized residuals:

```
. regress sqfeet hhinc hhsize
. predict yhat3
. predict rstud, rstud
```

For this example, we chose the number of groups used for the x-axis so that each box plot contains roughly 100 observations:[43]

[43] The function `round()` is a general Stata function (see section 3.1.6). The saved result `e(N)` stores the number of observations of the last regression model. The `xtile` command is described in section 7.3.1.

```
. local groups = round(e(N)/100,1)
. xtile groups = yhat3, nq('groups')
. graph box rstud, over(groups)
```



In the figure, you can see that there is a slight increase in the variance of the residuals.

Potential solutions

In many cases, you can simply transform the dependent variable to remove heteroskedasticity. The transformation should end in a symmetric variable. For right-skewed variables, a logarithmic transformation is often sufficient. In addition, the `boxcox` command allows you to transform a variable so that it is as symmetric as possible. You will find additional discussion of the Box–Cox transformation in section 8.4.3.

If transforming the dependent variable does not remove heteroskedasticity in the regression model, you cannot use the standard errors of the coefficients (as they are given in the regression output) for a significance test. If you are nevertheless interested in a significance test, you might want to try the `robust` option in the regression command. When you use this option, the standard errors are computed in such a way that homoskedasticity of the error terms need not be assumed.

8.3.3 Violation of Cov$(\epsilon_i, \epsilon_j) = 0,\ i \neq j$

What $\text{Cov}(\epsilon_i, \epsilon_j) = 0,\ i \neq j$, means is that the errors are not correlated with each other. The violation of this assumption is called "autocorrelation", which results in inefficient estimation of the coefficients.

For example, in the preceding sections you tried to predict home size. Now suppose that you have surveyed home size by letting the interviewers estimate the size instead of asking the respondent. In this case, it is reasonable to assume that some of the interviewers tend to overestimate the sizes of the dwellings, whereas others tend to underestimate them. In this case, all the observations from one interviewer should be

similar in over- or underestimating home size. A similar situation occurs if all people in a household are interviewed. In this case, as well as in the above, there may be factors within the unobserved influences (ϵ_i) that are the same for all members of a household. The same might be true for respondents of a particular sampling unit.

This example shows that you can deal with a violation of the independence assumption, even with cross-sectional data. In more recent years, the literature on complex samples (Lee, Forthofer, and Lorimor 1989; Lehtonen and Pakkinen 1995; Skinner, Holt, and Smith 1989) as well in the multilevel literature (Kreft and de Leeuw 1998), shows ways to handle the violations of the independence assumption we mentioned in the examples. We will give a more detailed discussion of this problem later.

Autocorrelation is a key concept, especially in time-series analysis, as successive observations tend to be more similar than observations separated by a large time span (serial autocorrelation). The Durbin–Watson test statistic has been developed for time-series analysis, and in Stata it is available using the `estat dwatson` command after regression. However, you must define the data as a time series beforehand.[44]

8.4 Model extensions

In this section, we will introduce three extensions to the linear model you have seen so far. These extensions are used for categorical independent variables, interaction terms, and modeling curvilinear relationships. Interpreting refined models can sometimes be rather difficult, so we will introduce conditional-effects-plots as a graphical way to display regression results.

8.4.1 Categorical independent variables

You need to be cautious when you include a categorical variable with more than two categories in the regression model. Take, for example, marital status. The variable `marital` has six categories, namely married, separated, unmarried, divorced, widowed, and grass-widowed (partner is living abroad):

```
. tabulate marital
```

Marital status 1997	Freq.	Percent	Cum.
married	1,860	55.69	55.69
separate	83	2.49	58.17
unmarried	800	23.95	82.13
divorce	270	8.08	90.21
widowed	312	9.34	99.55
grasswid	15	0.45	100.00
Total	3,340	100.00	

[44] As we do not discuss time-series analysis in this book, we refer here to the online help `tsset` and to the manual entry [U] **26.13 Models with time-series data**.

It would not make sense to include marital status in the same way as all other independent variables since assuming that going from being married to separated has the same effect on home size as going from divorced to widowed. However, you would assume this implicitly when a categorical variable with several categories is included in a regression model without any changes. What you need instead are contrasts between the individual categories.

Let's say that you want to include a variable that differentiates between married and unmarried respondents. To do so, you can create a dichotomous variable with the response categories 0 for not married and 1 for married.

```
. generate married = marital == 1 if marital < .
```

You can interpret the resulting b coefficient for this variable in the same way as any other dummy variable; accordingly, you could say that married respondents live on average in a space that is b square feet bigger than the one unmarried people live in.

All other contrasts can be build in the same way:

```
. generate separated = marital == 2 if marital < .
. generate unmarried = marital == 3 if marital < .
. generate divorced = marital == 4 if marital < .
. generate widowed = marital == 5 if marital < .
. generate grasswid = marital == 6 if marital < .
```

Each contrast displays the difference between respondents with one particular marital status and all other respondents. Beware that Stata will automatically remove one of the dummy variables if you include all contrasts in the regression model:

```
. regress sqfeet hhinc hhsize married-grasswid, noheader
```

sqfeet	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hhinc	.1644524	.0070193	23.43	0.000	.1506894	.1782153
hhsize	39.28901	6.440657	6.10	0.000	26.66065	51.91736
married	-42.47589	47.98608	-0.89	0.376	-136.5634	51.61162
separated	(dropped)					
unmarried	-73.35993	48.77556	-1.50	0.133	-168.9954	22.27553
divorced	-107.4403	52.90216	-2.03	0.042	-211.1669	-3.713738
widowed	-20.81282	52.21819	-0.40	0.690	-123.1983	81.57269
grasswid	-329.026	120.9142	-2.72	0.007	-566.1055	-91.9465
_cons	579.1672	48.63016	11.91	0.000	483.8168	674.5176

The reason for dropping one of the contrasts is that of six new dummy variables, only five are needed to know the marital status of each individual person. If five dummies indicate a person does not live with a spouse, is not divorced, is not unmarried, is not widowed, and is not grass-widowed, the person must be separated from her spouse. The sixth dummy variable tells you nothing that cannot be gleaned from the other five, since there are six possibilities. Computationally, it is not even possible to estimate coefficients on all six dummies in addition to the constant term because those six dummies sum to one and hence are perfectly correlated with the constant term.

Note that the constant represents the predicted value for respondents with zero on all covariates, in this case respondents who are separated, since the separated dummy was not included in the model. The predicted home size of persons with a different family status differs by the amount of the according b coefficient. Married respondents have, therefore, on average a home size that is 42.48 ft^2 smaller than those of respondents who are separated. Even smaller are the homes of unmarried respondents—about 73.36 ft^2 smaller than those of separated respondents. All other coefficients are interpreted accordingly.

The coefficients you obtained may be somewhat surprising. It would be more reasonable if married respondents have on average larger home sizes than those who are separated, other things constant; however, according to the model, the opposite is true. Upon closer inspection, though, the results are more reasonable. Separated respondents typically live alone, which means their household size is one. While married couples have an average home size 42 ft^2 less than separated people holding all other factors constant, the `hhsize` variable's coefficient of 39.3 implies that a married couple with no children has, in fact, a home which is an average of just 3 ft^2 less than a separated person. In addition, our dataset has only a few separated respondents.

Instead of creating contrast variables by hand, you can use the `tab` *varname*`, gen(`*newvar*`)` command, where *varname* is the name of the categorical variable you want to use. This command ensures that there will be as many variables—*newvar1* to *newvarK*—as there are categories in the categorical variable.

```
. tabulate egp, gen(egp_)
```

This command creates the variables `egp_1` to `egp_11`. You can include those in your regression model. It is usually a good idea to decide on a contrast that will be left out and used as a comparison category, but which one you use does not affect the results substantively.

The `xi` prefix is another shortcut for creating dummy variables that you can use with all kinds of model commands. You compute the regression of home size on the categorical variable type of household by typing

```
. xi: regress sqfeet i.htyp
```

This command creates eight dummy variables (this is the number of categories in the variable `hhtype`) to be used in the regression model. By default, the smallest category is omitted and used as the reference category. The `xi` prefix is especially helpful if you intend to model interaction effects with those dummy variables. However, using multiple `xi` prefixes for several models takes time, since all dummy variables must be created over and over for each model.

8.4.2 Interaction terms

In order to discuss modeling interaction effects, we will return to our analysis of income inequality between men and women from chapter 1. There we tried to "explain" the gross income by gender and occupational status for all respondents who have some kind of income. The analysis showed that women earn less on average then men and that this difference can only in part be explained by the difference in full-time and part-time occupation between men and women.

Let's begin by reproducing the model using the do-file `anchap1.do`:

```
. do anchap1.do
```

Assume that you hypothesize that income depends not only on the respondents' gender and occupational status, but also on their educational level. A higher educational level—you believe—leads to higher income. In this case, you should include education in your regression model. Say that you also assume that the educational advantage is more significant the older people are. There are three good reasons why this is a reasonable assumption:

- More highly educated people begin their occupational careers later in life. Starting salaries for highly educated people are therefore not necessarily—if at all—higher than their co-workers in the same age group with less education but more occupational experience. However, with increased occupational tenure, the educational advantage increases, as well.
- People with less education have a higher risk of becoming unemployed. Periods of unemployment can lead to difficulties in finding a new job, with an increased likelihood of having part-time or minimum-wage jobs. This would lead in extreme cases to a loss in income as age increases for people with less education.
- Education was for a long time the most important variable for determining income. Nowadays there are professions that do not require professional training or a college degree. Income inequality between educational groups could therefore be seen as an extinguishing phenomena that is only visible for older generations.

All three arguments could lead us to hypothesize that the effect of education on income increases with age. Such effects that vary with values of a third variable are called interaction effects, which you can include in a regression model by multiplying the relevant variables.

Here education and age are relevant variables for the interaction effect. To extend the regression model, include years of education (`yedu`). You can create the variable for age using year of birth (`ybirth`) and the date of the interview:

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

It is advantageous to center continuous variables, such as length of education or age, before they are included in a regression model, for several reasons. This is especially true in the presence of interaction terms (Aiken and West 1991).

To center a variable, subtract the mean from each value. However, note that you need to compute the mean only for those respondents that will be included in the regression model later on. In order to know which respondents these are, you need to know which respondents have valid values on all variables that will be included in the regression model. You can use the `egen` function `rowmiss(`*varlist*`)` to do so. This function counts the number of missing values for the specified variable list.

```
. egen miss = rowmiss(income yedu ybirth fulltime)
```

You generate the `miss` variable using this `egen` command. The variable `miss` takes on the value zero for all respondents that do not have missing values for any of the specified variables. This allows you to specify the correct mean, as we discussed above:[45]

```
. summarize yedu if miss == 0
. generate cyedu = yedu - r(mean) if miss==0
```

and

```
. summarize age if miss==0
. generate cage = age - r(mean) if miss==0
```

To build the interaction term, multiply both variables that are part of the interaction.

```
. generate yeduage = cyedu * cage
```

The linear regression model will be extended by the variables `cage` and `cyedu` and the interaction term `yeduage` that you just created:

```
. regress income men fulltime cage cyedu yeduage
```

Source	SS	df	MS		Number of obs =	1545
					F(5, 1539) =	50.24
Model	359958784	5	71991756.8		Prob > F =	0.0000
Residual	2.2051e+09	1539	1432839.59		R-squared =	0.1403
					Adj R-squared =	0.1375
Total	2.5651e+09	1544	1661333.49		Root MSE =	1197

income	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
men	444.0768	61.78785	7.19	0.000	322.8796	565.2741
fulltime	764.3836	90.45379	8.45	0.000	586.9579	941.8093
cage	-.893062	2.158071	-0.41	0.679	-5.126133	3.340009
cyedu	119.3288	12.50207	9.54	0.000	94.80594	143.8517
yeduage	1.11756	.8829518	1.27	0.206	-.6143558	2.849476
_cons	1007.47	86.07676	11.70	0.000	838.6304	1176.311

[45] The expression `r(mean)` refers to the mean saved by `summarize` (see chapter 4).

Without going into details, we want to explain the graphical representation of the regression model results, in particular, the interpretation of the interaction terms. Even if most people use tables to display the results of a multiple linear regression, we think graphs are more reasonable for complex models.[46] Graphically displaying the results is especially useful for models with interaction terms; conditional-effects plots are commonly used.

To construct a conditional-effects plot, you draw different regression lines for different combinations of independent variables. Here you can create a graph that indicates the correlation between age and income and, according to our hypothesis about the interaction effect, the correlation for an average length of education, as well as for the lowest and highest years of education. This means that you need to compute three different regression lines.

For a simple linear regression, we showed you on page 185 how to compute the regression line according to the model results:

```
. predict yhat1
```

Alternatively, you could get the predicted values by typing

```
. generate yhat2 = _b[_cons] + _b[fulltime]*fulltime + _b[men]*men +
> _b[cage]*cage + _b[cyedu]*cyedu + _b[yeduage]*yeduage
```

Unlike simple linear regression, in a multiple regression the predicted values cannot be displayed on a straight line in a two-dimensional graph. However, you can obtain a line if you fix the values of all variables but one. You can, for example, compute the predicted values for female respondents with part-time employment and an average education. Since you centered the variable for education, you know that 0 represents the average length of education. You can therefore compute the desired values by fixing all variables except `age` to zero:

```
. generate yh_yedu0 = _b[_cons] +  _b[cage] * cage
```

Notice that most of the terms in the equation disappeared because you replaced the variables' values with zeros. The same is true for the interaction effect since multiplying `age` with the average value of the centered education variable is zero, as well.

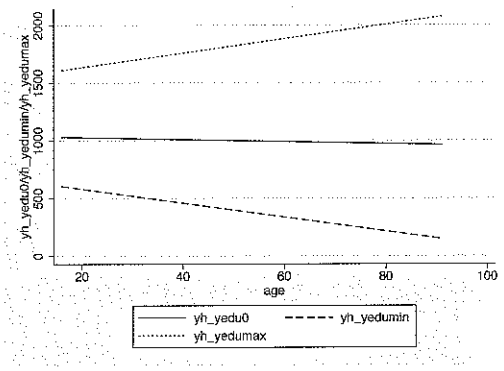
The predicted values for the lowest education group can be computed accordingly. To do so, you first must `summarize` the variable `cyedu` and use the saved result `r(min)`, which contains the minimum value of the variable used in the last `summarize` command. Beware that fixing the interaction term is more complicated, since it should vary with age but not with education. You therefore multiply the interaction term with the minimum value of education *and* `cyedu`. Now you can fix the interaction term accordingly for the highest educational group:

[46] In section 12.3.1, we introduce a command that is usually used to transform Stata regression tables in regression tables usually used for publications.

```
. summarize cyedu
. generate yh_yedumin = _b[_cons] + _b[cage]*cage + _b[cyedu]*r(min) +
> _b[yeduage]*cage*r(min)
. generate yh_yedumax = _b[_cons] + _b[cage]*cage + _b[cyedu]*r(max) +
> _b[yeduage]*cage*r(max)
```

You have thus created three different variables that contain values for three different educational levels for women who work part-time. Each of these three variables is in itself a function of age. The values for each educational level form a line that can be displayed graphically. Note that in this command, the star behind `yh_edu` is used to include all variables beginning with `yh_edu` in the `graph` command (see section 3.1.2):

```
. graph twoway line yh_yedu* age, sort
```



In this graph, the top line represents respondents from the highest educational group, and the bottom line represents respondents from the lowest educational group.[47] The graph shows that age has a different effect for each educational level. The higher the educational level, the greater is the increase of income with increasing age. In models without interaction terms, the lines in a conditional-effects plots would always be parallel.

8.4.3 Regression models using transformed variables

There are two main reasons to use transformed variables in a regression model:

1. the presence of a nonlinear relationship
2. a violation of the homoskedasticity assumption

Depending on the reason you want to transform the variables, there are different ways to proceed. In the presence of a nonlinear relation, you would (normally) transform the

[47] To improve this legend, refer to section 6.3.4.

independent variable, but in the presence of heteroskedasticity, you would transform the *dependent* variable. We will begin by explaining how to model nonlinear relationships and then how to deal with heteroskedasticity (also see Mosteller and Tukey 1977).

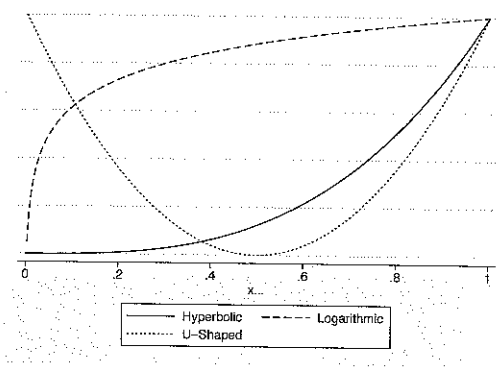
Nonlinear relations

We introduced regression diagnostic techniques for detecting nonlinear relationships in section 8.3.1. However, in many cases, theoretical considerations already provide enough reason to model a nonlinear relationship: think, for example, about the correlation between female literacy rate and birth rate. You would expect a negative correlation for these two variables, and you would also expect the birth rate not to drop linearly towards zero. Rather you would expect birth rate to decrease with an increase in literacy rate to levels of around one or two births per women.

Nonlinear relationships occur quite frequently in cases where income is used as an independent variable. For many relationships, income changes in the lower range of income have a greater effect on the dependent variable than income changes in the upper part of the income distribution. Income triples with a change from \$500 to \$1,500, while the increase is only ten percent for a change from \$10,000 to \$11,000, although the increase is in both cases \$1,000.

When modeling nonlinear relationships, you first need to know or at least hypothesize a functional form of the relationship. Here you need to distinguish among three basic types of nonlinear relations: logarithmic, hyperbolic, and U-shaped. Stylized versions of these relationship can be produced with the twoway plottype `function` (see [G] **graph twoway function**):

```
. twoway (function y = x^3, yaxis(1) yscale(off axis(1)))
> (function y = ln(x), yaxis(2) yscale(off axis(2)))
> (function y = (-1)* x +  x^2, yaxis(3) yscale(off axis(3))),
> legend(label(1 "Hyperbolic") label(2 "Logarithmic") label(3 "U-Shaped"))
```



In logarithmic relationships, the dependent variable increases with increasing values of the independent variable. However, with increasing values of the independent

variable, the increase in the dependent variable levels off. In hyperbolic relationships, the relation is reversed, as the dependent variable increases only moderately at the beginning and increases with increasing values of the independent variable. In U-shaped relationships, the sign of the effect of the independent variable changes. All three basic types can occur in opposite directions. For logarithmic relationships, this would mean that the values decrease rapidly at the beginning and slower later on. For hyperbolic relationships, the values drop slowly at the beginning and rapidly later on. For U-shaped relationships, the values first decrease and increase later on, or vice versa. In practice, logarithmic relationships occur quite frequently.

To model logarithmic relations, you first form the log of the independent variable and replace the original variable in the regression model with this new variable. A strong logarithmic relationship can be found between the countries' gross domestic product and infant mortality rate. The file `uno.dta` contains these data.[48]

```
. use uno.dta, clear
. graph twoway scatter infmort gdp
```

You can model this logarithmic relationship by first creating the log of the X variable

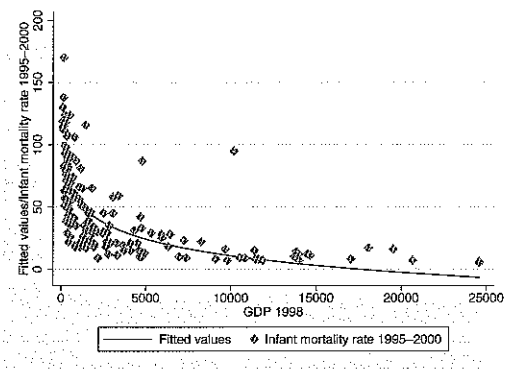
```
. generate loggdp = log(gdp)
```

and then using this variable instead of the original X variable:

```
. regress infmort loggdp
. predict yhat1
```

You can see the logarithmic relationship between the predicted value of the regression model `yhat1` and the untransformed independent variable:

```
. twoway (scatter infmort gdp) (line yhat1 gdp, sort)
```

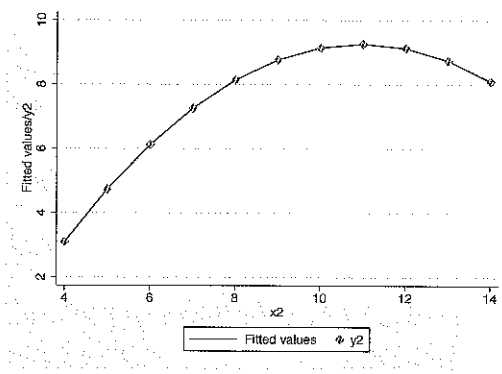


[48] The original example was introduced by Fox (2000), and the values have been updated using information provided by the United Nations (*http://www.un.org/Depts/unsd/social/main.htm*).

You use a similar procedure to model hyperbolic relations, except that now you square the original variable instead of taking its logarithm. Here the original variable is also replaced by the newly transformed variable.[49]

The situation is different when you are modeling a U-shaped relationship. Although you still square the independent variable, the newly generated variable does not replace the original variable. Instead both variables will be used in the regression model. A U-shaped relation is one of the examples in *Anscombe's quartet* on page 200. Including the quadratic term will allow you to model this correlation perfectly:

```
. use anscombe, clear
. generate x2q = x2^2
. regress y2 x2 x2q
. predict yhat
. twoway (line yhat x2, sort) (scatter y2 x2)
```



If you are thinking of using transformations of the independent variables, see Cook and Weisberg (1999, chapter 16) for some precautions for doing so.

Eliminating heteroskedasticity

In section 8.3.2, we discussed skewed dependent variables as one of the possible causes of heteroskedasticity. In this case, you would need to transform the dependent variable in order to remove heteroskedasticity. Note that the interpretation of the regression model changes when you include a transformed variable. Transforming the dependent variable leads to a nonlinear relationship between the dependent and *all* independent variables (Hair et al. 1995, 75).

The aim in transforming a variable is to obtain a fairly symmetric or normal dependent variable. Remember the following rule of thumb: If the distribution is wide, the inverse of the variable is a useful transformation $(1/Y)$. If the distribution is skewed to

[49] Examples for hyperbolic relations are rare in the social sciences. The salary of a Formula 1 race driver could possibly show a hyperbolic relation to the number of Grand-Prix victories.

the right (such as home size in our example), taking the log is reasonable, and you can take the square root if the distribution is skewed to the left (Fox 1997, 59–82).

Besides following these rules, you can use the Stata command `bcskew0` which uses a Box–Cox transformation that will lead to a (nearly) unskewed distribution[50].

```
. use data1,clear
. bcskew0 bcsqfeet = sqfeet
```

The residual-versus-fitted plot (page 200) tells you something about the type of transformation necessary. If the spread of the residuals increases with increasing values in the predicted variable, the inverse of Y is a better dependent variable. If there is a decreasing spread of the residuals with increasing values in the predicted variable, you will want to replace the original dependent variable Y with the square root of Y (Hair et al. 1995, 70).

8.5 More on standard errors

We have mentioned that the standard errors reported by `regress` require you to make a host of assumptions in order for them to be valid. Multicollinearity, for example, typically causes standard errors to be inflated, making coefficients appear insignificant. Heteroskedasticity also affects standard errors. In this section, we briefly introduce two alternative methods of obtaining standard errors and confidence intervals.

Bootstrap techniques

Confidence intervals are computed under the assumption that the regression coefficients are normally distributed. Based on this assumption, you can multiply the standard errors with the critical value.[51] However, with finite samples the coefficients may not have a normal distribution. Therefore, we want to introduce a technique you can use to compute confidence intervals in a different way. The technique is called the "bootstrap", suggesting that we have to help ourselves using the information at hand instead of relying on distribution assumptions.[52]

The bootstrap technique is based on the assumption that all the information you have about your population is contained within the sample data, meaning you use *only* the sample data to assess the population. Let's step back for a second. Assume that you could draw as many samples from the population as you wanted, and you compute confidence intervals for the regression coefficient for each of these samples; about 95 percent of all 95% confidence intervals computed in this way would include the true

[50] Make sure that the variable used in `bcskew0` does not include negative values or the value zero.

[51] The critical value for a 95% confidence interval for models with more than 120 degrees of freedom is 1.96. Remember that the degrees of freedom are the number of cases minus number of coefficients including the constant b_0.

[52] See Mooney and Duval (1993, 42) for a brief overview over the different bootstrap techniques. A detailed explanation is given by Efron and Tibshirani (1993) and a pedagogical introduction by Stine (1990).

value of the coefficient. Unfortunately, you cannot replicate the sample that many times, but you can use the bootstrap technique. Applying the bootstrap technique, you draw a large number of samples out of your sample, where each sample has a similar number of observations as the original sample. In other words, you treat your original sample as if it were the population, then repeatedly draw samples from it. This may sound strange, but all that is required is sampling with replacement. Some observations will probably appear twice or three times in one of the new samples.

Now you compute the statistic (e.g., the mean of a variable, or in this case the regression coefficient b) for each of the samples that had been drawn out of the original sample. The distribution of all the estimation results will enable you to compute confidence interval boundaries for the statistic of interest to you. According to the "percentile method" the 95 percent confidence interval is formed by the value above which there are 2.5 percent of all estimation results, and the value below which there are 2.5 percent of all estimation results.

In Stata, the `bootstrap` prefix command is used to perform the bootstrap.[53] All you need to do is prefix your estimation command with the keyword `bootstrap:`. You can optionally specify the number of samples to draw. If you want to interpret the confidence intervals, we suggest drawing at least 1,000 samples. Do not be surprised if this takes a while on your machine. Finally, note that the entire procedure is based on a random process, so unless you reset the random-number seed each time, you will get slightly different results every time you run your do-file again. To obtain the percentile-based confidence intervals we just discussed, use the `estat bootstrap` command afterwards. `bootstrap` itself reports confidence intervals that assume that the coefficients are normally distributed, which we have just argued is not always appropriate.

```
. use data1, clear
(SOEP'97 (Kohler/Kreuter))
. set seed 731
. bootstrap: regress sqfeet income hhsize
(running regress on estimation sample)
  (output omitted)
. estat bootstrap, percentile
Linear regression                               Number of obs     =       2960
                                                Replications      =         50

------------------------------------------------------------------------------
             |   Observed               Bootstrap
      sqfeet |      Coef.       Bias    Std. Err.  [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .06513619   .0014398   .0074741    .0529878   .0781351  (P)
      hhsize |   83.435215   1.849459   6.4308584   73.36475   96.36674  (P)
       _cons |   644.0526   -8.150171   19.503386   593.7859   667.9052  (P)
------------------------------------------------------------------------------
(P)     percentile confidence interval
```

[53] Prior to Stata 9, the `bootstrap` command was used, which follows a different syntax. Type `help bootstrap` if you are using an older version of Stata.

To apply the bootstrap technique, you need to know how your sample is drawn from the population because this sampling process must be followed during the bootstrap sampling. To introduce the bootstrap, we described the most simple case of a simple random sample. Applying bootstrap techniques to most national survey data is therefore more complex than what we have introduced here since you need to replicate the sampling of, for example, a cluster sample. You can find ways to do so by typing `help bootstrap`.

Confidence intervals in cluster samples

Many surveys are not based on simple random sampling but on multistage clustered samples. One example for a two-stage cluster sample is the selection of hospitals and the selection of patients within each selected hospital. A multistage cluster sample example would be the selection of schools, the selection of classrooms within each school, and within each classroom the selection of students (Levy and Lemeshow 1999, 227). The GSOEP, which is the basis of the dataset `data1.dta` used here, is also a multistage cluster sample (Pannenberg et al. 1998). At the first stage, election districts were sampled from a nationwide register, and afterwards respondents were drawn using what is called a random walk.

The decision to use a clustered sample (and not a simple random sample) can be made for organizational or financial reasons. For example, the absence of a general population register in many countries reduces many researchers' ability to use simple random samples. Sampling in several sampling stages, one of them at the level of small geographical clusters, allows for the selection of respondents without the aid of register data. However, there is a drawback to using a clustered sample: the standard errors can be larger then they would be under simple random sampling (Kish 1965). One reason for this is the relative homogeneity of observations within a cluster. For example, respondents within the same neighborhood can probably afford similar housing and belong to a similar income group, therefore giving similar answers regarding income. Another reason for larger standard errors may be the way data were collected, where each interviewer might influence the responses (Schnell and Kreuter 2005).

In addition, samples are often stratified; that is, population elements are divided into exhaustive subgroups and sampling takes place within each of these subgroups. Unlike clustering, stratification can reduce the standard errors if the stratification variable is correlated with the outcome of interest. However, you will need to take sampling design information into account.

Stata provides special commands to estimate the correct standard errors for complex samples[54]. For example, if you want to take the potentially homogenizing effects of interviewers into account for our example, you can use the following command sequence.

[54] See [SVY] **survey** for an overview of Stata's survey estimation commands.

```
. svyset, psu(intnr)
      pweight: <none>
          VCE: linearized
    Strata 1: <one>
        SU 1: intnr
       FPC 1: <zero>
. svy: regress sqfeet income hhsize
(running regress on estimation sample)
Survey: Linear regression
Number of strata   =         1                  Number of obs      =      2562
Number of PSUs     =       429                  Population size    =      2779
                                                Design df          =       428
                                                F(   2,    427)    =     93.27
                                                Prob > F           =    0.0000
                                                R-squared          =    0.0820
```

sqfeet	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	.0552464	.0079756	6.93	0.000	.0395701	.0709227
hhsize	82.46072	7.144441	11.54	0.000	68.41816	96.50328
_cons	651.5778	22.78619	28.60	0.000	606.7911	696.3646

The `svyset, psu(intnr)` command specifies the name of a variable (here `intnr`) that contains identifiers for the primary sampling units (clusters). You can parse this specification to the regression by putting the `svy` prefix in front of the regression command. The same technique applies to other statistical models as well.

You now obtain results for 2,562 observations, who had been interviewed by 429 interviewers.[55] In this example, the difference between the correct confidence intervals and confidence intervals estimated assuming a simple random sample is negligible.[56] Make sure that your data file includes variables that contain sampling design information so that you have information about the sampling point and about the interviewer) so that you can check your results for design effects. An excellent resource for the application of advanced techniques in Stata is the book by Rabe-Hesketh and Everitt (2004).

8.6 Advanced techniques

In addition to multiple linear regression, there are a number of related models that can be estimated in Stata. There is not enough room in this book to explain all of them in detail. However, a few of these models are so common that we want to describe the general ideas behind them. Each model is explained in detail in the *Stata Reference Manual*, where you will also find selected literature on the model.

[55] The interviewer information is missing for 561 respondents in this data file.

[56] This looks quite different if you look at the confidence intervals for the mean of the variables indicating respondents' concerns, for example np9506 with `svy: mean np9506`. With `svy: mean np9506` followed by `estat effects` you get a design effect of 1.45, which means that the confidence intervals are about 1.45 times higher than the ones estimated assuming simple random sampling.

8.6.1 Median regression

A median regression is quite similar to the ordinary least squares regression we talked about earlier. While the sum of the squared residuals $\sum(\widehat{y}_i - y_i)^2$ is minimized in OLS regression, the sum of the absolute residuals $\sum|\widehat{y}_i - y_i|$ is minimized when applying median regression. Squaring residuals in OLS means that large residuals are more heavily weighted than small residuals. This property is lost in median regression, so it is less sensitive to outliers than OLS regression.

Median regression takes its name from its predicted values, which are estimates of the median of the dependent variable conditional on the values of the independent variables. In OLS, the predicted values are estimates of the conditional means of the dependent variable. The predicted values of both regression techniques describe therefore a measure of a certain property—the central tendency—of the dependent variable.

Stata treats median regression as a special case of a quantile regression. In quantile regression, the coefficient is estimated so that the sum of the weighted (i.e., multiplied by a the factor w_i) absolute residuals is minimized.

$$\sum(|y_i - \widehat{y}_i| \times w_i) = \min \tag{8.22}$$

Weights can be different for positive and negative residuals. If positive and negative residuals are weighted the same way, you get a median regression. If positive residuals are weighted by the factor 1.5 and negative residuals are weighted by the factor 0.5, you get a "3rd quartile regression", etc.

In Stata, you compute quantile regressions using the `qreg` command. Just as in any other Stata model command, the dependent variable follows the command, then you specify the list of independent variables; the default is a median regression.

For this, use the dataset `data2agg.dta`, which contains the mean life satisfaction and the mean income data from the German population from 1984 to 2002.[57]

```
. use data2agg, clear
```

First, take a look at a scatterplot with the regression line of the mean life satisfaction on the mean income:

```
. twoway (lfit lsat inc) (sc lsat inc, mlab(wave)), leg(order(1) lab(1 "OLS"))
```

Note in this figure that the data for 1984 might influence the regression results more than any other data point. Now compute a median regression

```
. qreg lsat inc
```

and compare the predicted values of the median regression with the standard linear fit of the OLS regression:

[57] We used this small dataset to exemplify the effect of median regression. Beware that working with aggregate data is prone for ecological fallacy (Freedman 2004).

```
. predict medhat
. twoway (lfit lsat inc) (sc lsat inc, mlab(wave)) (line medhat inc, sort),
> leg(order(1 3) lab(1 "OLS") lab(3 "Median"))
```

Note that the regression line of the median regression is not as steep as the standard regression line. The reason is that the median regression is more robust to extreme data points, such as those from 1984.

8.6.2 Regression models for panel data

Panel data, or cross-sectional time-series data, contain repeated measures of the same individuals over time. An example of a panel data is the German Socio-Economic Panel (GSOEP). In the GSOEP, about 12,000 persons have been asked identical questions every year since 1984. That is, the GSOEP measures the *same* variables for the *same* respondents at *different* points in time. It should be clear, however, that panel data does not only arise from such "panel surveys". The same data structure is also present if you have collected certain macroeconomic indices in many different countries over time, or even data about certain features of political parties over time. Really, what defines panel data is that the *same* entities are observed at different times. In the remaining section, we will use the term "individuals" for these entities.

In Stata, all the commands that deal with panel data begin with the letters `xt`, and these commands are described in the *Longitudinal/Panel Data Reference Manual* ([XT]). A list of the `xt` commands can be found by typing `help xt`. Among the `xt` commands are some of the more complex models in the statistical universe, which we will not describe here. Instead we will help you understand the thinking behind the major approaches to analyzing panel data, together with examples of how you can use these approaches in Stata.[58]

Before we describe the statistical models, we need to say a word about data management. All Stata commands for panel analysis require a panel dataset that is in long format, so the next section describes how to put your data in this format. Then we will explain fixed-effects models and error-components models.

From wide to long format

Panel data can be stored in wide format or in long format. In wide format, the observations of the dataset are the individuals observed, and the variables are their characteristics at the respective time points. For example, if we ask four specific individuals, say John, Paul, George, and Ringo, about their life satisfaction in 1968, 1969, and 1970, we can store their answers in wide format by making a dataset with four observations, namely John, Paul, George, and Ringo, and three variables reflecting life satisfaction in 1968, 1969, and 1970, respectively (see table 8.2). However, the same information can

[58] For more information, see (Baltagi 1995; Hardin and Hilbe 2003; Diggle, Liang, and Zeger 1994; Wooldridge 2002) and the literature cited in [XT] **xtreg**.

also be stored in long format, where the observations are the individuals *at a specific point in time* and the variables are the observed characteristics. Hence, in our example, there would be three observations for John—one for 1968, one for 1969, and one for 1970—three observations for Paul, etc. The information on life satisfaction would be in one single variable. To keep the information about the timing, we would need a new variable for the year of observation.

Table 8.2: Ways to store panel data

Wide Format

i	X_{1968}	X_{1969}	X_{1970}
John	7	8	5
Paul	5	2	2
George	4	3	1
Ringo	8	8	6

Long Format

i	year	X
John	1968	7
John	1969	8
John	1970	5
Paul	1969	5
Paul	1969	2
Paul	1970	2
George	1968	4
⋮	⋮	⋮
Ringo	1970	6

Stata's `xt` commands generally expect panel data in long format. It is however more common for dataset providers to distribute panel data in wide format.[59] You will often need to reshape your dataset from wide to long.

An example of a panel data in wide format is `data2w.dta`. Please load this dataset to follow our example of changing from wide format to long format:

```
. use data2w, clear
```

This file contains information on year of birth, gender, life satisfaction, marital status, individual labor earnings, and annual work hours of 1,761 respondents (individuals) from the German Socio-Economic Panel (GSOEP). The individuals were observed every year between 1984 and 2002. Therefore, with the exception of the time-invariate variables gender and year of birth, there are 19 variables for each observed characteristic. If you look at the file with

[59] For very large panel studies, such as the German Socio-Economic Panel (GSOEP), the American Panel Study of Income Dynamics (PSID) or the British Household Panel Study (BHPS), the situation tends to be even more complicated. These data are often distributed in more than one file. In this case, you need to first combine these files into one single file. In section 10.4, we show you how to do this using an example from the GSOEP, resulting in a dataset in wide format.

```
. describe
Contains data from data2w.dta
  obs:         1,761                          GSOEP 1984-2002 randomized
                                                (Kohler/Kreuter)
 vars:            80                          1 Sep 2004 09:24
 size:       294,087 (96.9% of memory free)   (_dta has notes)
```

variable name	storage type	display format	value label		variable label
hhnr	long	%12.0g			Fix Household Number
persnr	long	%12.0g			Person ID (n)
sex	byte	%13.0g	sex		Gender (n)
gebjahr	int	%8.0g			Year of birth (n)
lsat1984	byte	%45.0g	sat		General Life Satisfaction
mar1984	byte	%20.0g	d1110484	*	Marital Status of Individual (n)
hour1984	int	%12.0g			Annual Work Hours of Individual (n)
inc1984	float	%9.0g		*	Individual Labor Earnings (n)
lsat1985	byte	%45.0g	sat		General Life Satisfaction
mar1985	byte	%20.0g	d1110485	*	Marital Status of Individual (n)
hour1985	int	%12.0g			Annual Work Hours of Individual (n)
inc1985	float	%9.0g		*	Individual Labor Earnings (n)
lsat1986	byte	%45.0g	sat		General Life Satisfaction

(*output omitted*)

you will see that the variable names of the file have a specific structure. The first part of the variable names, namely `lsat`, `mar`, `hour`, and `inc`, refers to the content of the variable, whereas the second part refers to the year in which the variable has been observed. Using this type of naming convention makes it easy to reshape data from wide to long.

Unfortunately, in practice variable names rarely follow this naming scheme. Even the variables in the GSOEP do not. For your convenience, we have renamed all the variables in the dataset beforehand, but generally you will need to do this on your own using the `rename` and `renpfix` commands. Renaming all the variables of panel data in wide format can be quite cumbersome. In the do-file `crdata2.do`, we have therefore constructed some loops, which use concepts explained in sections 3.2.2 and 11.2.1. If you need to rename many variables, you should review these concepts.[60]

The command for changing data between wide and long is `reshape`. `reshape long` changes a dataset from wide to long, and `reshape wide` does the same in the other direction. Stata needs to know three pieces of information to reshape data:

1. The variable that identifies the individuals in the data (i.e., the respondents),
2. the characteristics that are under observation, and
3. the times when the characteristics were observed.

[60] The user-written Stata command `soepren` makes it easier to rename GSOEP variables. The command is available on the SSC archive; for information about the SSC archive and installing user-written commands, see chapter 12.

The first piece of information is easy to obtain. In our example data, it is simply the variable `persnr`, which uniquely identifies each individual of the GSOEP. If there is no such variable, you can simply generate a variable containing the running number of each observation (see section 5.1.3).

The two latter pieces of information are coded in the variable names. As we have seen, the first part of the variable names contains the characteristic under observation, and the second part contains the time of observation. We therefore need to tell Stata where the first part of the variable names ends and the second part starts. This information is passed to Stata by listing the variable name stubs that refer to the characteristic under observation. Let us show you how this works for our example:

```
. reshape long inc lsat mar hour, i(persnr) j(wave)
```

First, note the option `i()`, which is required. It is used to specify the variable for the individuals of the dataset. Second, look at what we have specified after `reshape long`. Note that we have listed neither variable names nor a *varlist*. Instead we have specified the name stubs that refer to the characteristic under observation. The remaining part of the variable names is then interpreted by Stata as being information about the time point of the observation. When running the command, Stata strips off the year from the variables that begin with the specified name stub and stores this information in a *new* variable. In our case, the new variable is named `wave`, as we specified this name in the option `j()`. If we had not specified that option, Stata would have used the variable name `_j`.

Now let's take a look at the new dataset.

```
. describe
Contains data
  obs:        33,459                          GSOEP 1984-2002 randomized
                                                (Kohler/Kreuter)
 vars:             9
 size:       836,475 (91.1% of memory free)    (_dta has notes)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
persnr          long   %12.0g                 Person ID (n)
wave            int    %9.0g
hhnr            long   %12.0g                 Fix Household Number
sex             byte   %13.0g      sex        Gender (n)
gebjahr         int    %8.0g                  Year of birth (n)
lsat            byte   %20.0g      sat
mar             byte   %20.0g      d1110402 *
hour            int    %12.0g
inc             float  %9.0g                *
                                           * indicated variables have notes
-------------------------------------------------------------------------------
Sorted by:  persnr  wave
     Note:  dataset has changed since last saved
```

```
. list persnr wave lsat
```

	persnr	wave	lsat
1.	76	1984	8
2.	76	1985	8
3.	76	1986	7
4.	76	1987	8
5.	76	1988	8

(output omitted)

The dataset now has 9 variables instead of 80. Clearly there are still 1,761 *individuals* in the dataset, but since the observations made on the severals occasions for each individual are stacked beneath each other, we end up with 33,459 *observations*. Hence, the data is in long format, as it must be to use the commands for panel data. And working with the `xt` commands is even more convenient if you declare the data to be panel data. You can do this with the `tsset` command by specifying the variable that identifies individuals followed by the variable that indicates time:

```
. tsset persnr wave
        panel variable:  persnr, 76 to 1392841
         time variable:  wave, 1984 to 2002
```

Finally, note that after reshaping the data once, reshaping from long to wide and vice versa is easy:

```
. reshape wide
. reshape long
```

Fixed-effects models

If the data is in long format, you can now run a simple ordinary least squares regression. For example, if you want to find out, whether aging has an effect on general life satisfaction, you might want to run the following regression:

```
. gen age = wave - gebjahr
. replace lsat = .a if lsat < 0
(82 real changes made, 82 to missing)
. regress lsat age
```

Source	SS	df	MS
Model	10.2027446	1	10.2027446
Residual	106166.227	33375	3.18101053
Total	106176.429	33376	3.18122092

```
Number of obs =   33377
F(  1, 33375) =    3.21
Prob > F      =  0.0733
R-squared     =  0.0001
Adj R-squared =  0.0001
Root MSE      =  1.7835
```

lsat	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	-.0006269	.00035	-1.79	0.073	-.001313	.0000592
_cons	7.195336	.0192932	372.95	0.000	7.15752	7.233151

From this regression model, you learn that with age, life satisfaction tends to decrease, but since the coefficient is not significant, you might also say that age does not affect life satisfaction. However, after having read this chapter, you probably do not want to trust this regression model, particularly because of omitted variables. Should you control the relationship for quantities like gender, education, and the historical time in which the respondents grew up?

Now let's imagine that you include a dummy variable for each individual of the GSOEP. As there are 1,761 individuals in the dataset, this would require a regression model with 1,760 dummy variables, which might be slightly overwhelming to work with. But for small datasets like those presented in table 8.2, this is not a problem. So let us deal with this data for now.

```
. preserve
. use beatles, clear
. describe
Contains data from beatles.dta
  obs:            12
 vars:             4                          7 Jul 2004 15:36
 size:           108 (99.9% of memory free)
```

variable name	storage type	display format	value label	variable label
persnr	byte	%9.0g	name	Person
time	int	%9.0g		Year of observation
lsat	byte	%9.0g		Life Satisfaction (fictive)
age	byte	%9.0g		Age in Years

```
Sorted by:
```

This dataset contains the age and (artificial) life satisfaction of four Englishmen at three points in time in long format. The command

```
. regress lsat age
      Source |       SS       df       MS              Number of obs =      12
-------------+------------------------------           F(  1,    10) =    1.97
       Model |  13.460177     1  13.460177             Prob > F      =  0.1904
    Residual |  68.2064897    10 6.82064897            R-squared     =  0.1648
-------------+------------------------------           Adj R-squared =  0.0813
       Total |  81.6666667    11 7.42424242            Root MSE      =  2.6116

------------------------------------------------------------------------------
        lsat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .6902655   .4913643     1.40   0.190    -.4045625    1.785093
       _cons |  -14.32153   13.65619    -1.05   0.319    -44.74941    16.10635
------------------------------------------------------------------------------
```

mirrors the regression analysis from above, showing a slight insignificant, positive effect of age on life satisfaction. Incorporating dummy variables for each individual of the dataset into this regression is straightforward.

```
. tab persnr, gen(d)
```

Person	Freq.	Percent	Cum.
John	3	25.00	25.00
Paul	3	25.00	50.00
George	3	25.00	75.00
Ringo	3	25.00	100.00
Total	12	100.00	

```
. regress lsat age d2-d4
```

Source	SS	df	MS
Model	80.125	4	20.03125
Residual	1.54166667	7	.220238095
Total	81.6666667	11	7.42424242

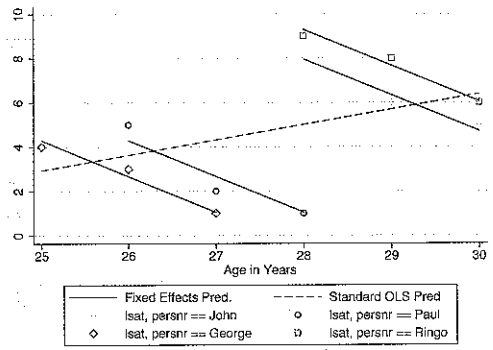
```
Number of obs =      12
F(  4,     7) =   90.95
Prob > F      =  0.0000
R-squared     =  0.9811
Adj R-squared =  0.9703
Root MSE      =   .4693
```

lsat	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	-1.625	.165921	-9.79	0.000	-2.017341	-1.232659
d2	-6.916667	.5068969	-13.65	0.000	-8.115287	-5.718046
d3	-8.541667	.6281666	-13.60	0.000	-10.02704	-7.056289
d4	1.333333	.383178	3.48	0.010	.4272613	2.239405
_cons	53.45833	4.81933	11.09	0.000	42.06243	64.85424

Now it appears that age has a strong negative effect on life satisfaction. The sign of the age effect has reversed, and we will see in a minute why this has happened. But let us first say something about the individual dummies. The coefficients of the individual dummies reflect how strongly the life satisfaction of the four Englishmen differs. You can see that persons 1 and 4 have a much higher life satisfaction than persons 2 and 3. You do not know *why* these people differ in their life satisfaction, the differences are not surprising since different people perceive life differently. Maybe they live in different neighborhoods, have different family backgrounds, grew up under different circumstances, or just have different habits about answering odd questions in population surveys. What is important here is that, since you put individual dummies into the regression model, you have reliably controlled for any differences between the persons. In this sense, the coefficient for age cannot be biased because we omitted stable characteristics of these persons. It is a pure aging effect, which rests solely on the development of the life satisfaction during the aging process of these four men.

This interpretation of the age coefficient can be illustrated with the following plot.

```
. predict yhat
. separate lsat, by(persnr)
. separate yhat, by(persnr)
. twoway (line yhat1-yhat4 age, clstyle(p1 p1 p1 p1))
> (lfit lsat age, clpattern(dash)) (scatter lsat1-lsat4 age),
> legend(order(1 5 6 7 8 9) lab(1 "Fixed Effects Pred.")
> lab(5 "Standard OLS Pred"))
```



The plot is an overlay of a standard scatterplot with different markers for each person in the dataset (`scatter lsat1-lsat4 age`), a conditional effects plot of the regression model with the person dummies (`line yhat1-yhat4 age`) and a simple regression line for all the data (`lfit lsat age`). If you look at the markers for each person separately, you will find that the life satisfaction decreases as the person gets older. At the same time, however, Ringo and John, the two oldest people in the dataset, have a higher life satisfaction than Paul and George. If we do not control for this, differences among people contribute to the age effect. The age effect of the simple OLS regression just shows that the older people have a higher life satisfaction than the younger ones. After we control for the personal differences, the only variation left is that which is within each person, and then age effect reflects the change in life satisfaction as each person gets older.

As the regression model with person dummies restricts itself on the variation within each person, the model is sometimes called the *within estimator*, covariance model, individual dummy-variable model, or *fixed-effects model*.

While the derivation of the fixed-effects model is straightforward, the technical calculation of the model in huge datasets is not. The problem arises because the number of independent variables in regression models is restricted to 800 in Intercooled Stata and to 11,000 in Stata/SE. Therefore, you cannot calculate a fixed-effects model by incorporating individual dummies in datasets with more than 800 or 11,000 individuals respectively.

Fortunately, in the linear regression model, you can use an algebraic trick to calculate the fixed-effects model, anyway. And even more fortunately, Stata has a command that does this algebraic trick for you: `xtreg` with the option `fe`. Here you can use the command for our small example data

```
. xtreg lsat age, fe i(persnr)
```

which reproduces the age coefficient of the model with dummy variables exactly. Note that you do not need to list the dummy variables in this command. Instead you specify the name of the variable, which identifies the individuals in the option `i()`.

The same logic applies if you want to calculate the fixed-effects model for larger datasets. Therefore, you can use the same command also in our previously constructed dataset. As you have already used the command `tsset` above (see page 236), you do not need to specify the `i()` option.

```
. restore
. xtreg lsat age, fe
```

Note that the values of the coefficients for the 1,761 dummy variables are not shown in the output and were not calculated. But the coefficient for age in the model is calculated as if the dummy variables were present. The fixed-effects model controls for all time-invariant differences between the individuals, so the coefficients of the fixed-effects models cannot be biased due to omitted time-invariant characteristics. This feature makes the fixed-effects model particularly attractive.

One side effect of the features of fixed-effects models is that they cannot be used to investigate time-invariant causes of the dependent variables. From a technical point of view, this is because time-invariant characteristics of the individuals are perfectly collinear with the person dummies. From a substantive point of view, this is because fixed-effect models are designed to study the causes of changes within a person. A time-invariant characteristic cannot cause such a change.

8.6.3 Error-components models

Let us begin our description of error-components models with the simple ordinary least squares regression:

```
. regress lsat age
```

This model ignores the panel structure of the data and treats data as cross-sectional. From a statistical point of view, this model violates an underlying assumption of ordinary least squares regression, namely the assumption that all observations are independent of each other. In panel data, you can generally assume that observations from the same individual are more similar to each other than observations from different individuals.

In observing the similarity of the observations from one individual, you might say that the residuals of the above regression are correlated. That is, an individual with a high positive residual at the first time of observation should also have a high positive residual at the second time point, etc.

Let us show you that the residuals of the above regression model are in fact correlated. First, calculate the residuals from the above regression model:

```
. predict res, resid
```

and then we change the dataset to the wide format. Since you have generated a new variable in the long format since last using `reshape`, you cannot just type `reshape wide`; instead, you need to use the full syntax:

```
. reshape wide lsat mar hour inc age res, i(persnr) j(wave)
```

You end up with 19 variables containing the residuals for each individual for every year. These variables can be used to construct a correlation matrix of the residuals. To save some space, we will display this correlation matrix only for the residuals from the eighties:

```
. corr res198?
(obs=1741)
```

	res1984	res1985	res1986	res1987	res1988	res1989
res1984	1.0000					
res1985	0.4406	1.0000				
res1986	0.3861	0.4570	1.0000			
res1987	0.3705	0.4200	0.5006	1.0000		
res1988	0.3310	0.3590	0.4276	0.5241	1.0000	
res1989	0.3159	0.3411	0.4164	0.4755	0.5098	1.0000

As you can see, the residuals are in fact highly correlated. Let us now define this correlation matrix as $\mathbf{R}_{t,s}$

$$\mathbf{R}_{t,s} = \begin{pmatrix} 1 & & & \\ r_{e_{i2},e_{i1}} & 1 & & \\ \vdots & \vdots & \ddots & \\ r_{e_{iT},e_{i1}} & r_{e_{iT},e_{i2}} & \dots & 1 \end{pmatrix} \tag{8.23}$$

As we said, in computing the simple OLS regression on panel data, you assume—among other things—that all correlations of this correlation matrix are 0, or more formally

$$\mathbf{R}_{t,s} = \begin{cases} 1 & \text{for } t = s \\ 0 & \text{otherwise} \end{cases} \tag{8.24}$$

As we have seen, this assumption is not fulfilled in our example regression. Hence, the model is not correctly specified. This is almost always the case for panel data. With panel data, you should expect correlated errors. In error-components models, you can therefore hypothesize about the structure of $\mathbf{R}_{t,s}$. Probably the simplest model after the simple regression model is the random-effects model:

$$\mathbf{R}_{t,s} = \begin{cases} 1 & \text{for } t = s \\ \rho & \text{otherwise} \end{cases} \tag{8.25}$$

Here the hypothetical structure of $\mathbf{R}_t, s$ is that observational units are more similar to each other over time than observations across observational units. The Stata command for random-effects models is `xtreg` with the option `re`.

```
. reshape long
. xtreg lsat age, re
```

Another reasonable assumption for the correlation structure of the residuals might be that the similarity between observations within each observational units is greater the shorter the elapsed time between the observations. This structure can be imposed using an AR(1) correlation matrix:[61]

$$\mathbf{R}_{t,s} = \begin{cases} 1 & \text{for } t = s \\ \rho^{|t-s|} & \text{otherwise} \end{cases} \tag{8.26}$$

Different structures for the correlation matrix allow for a nearly infinite number of model variations. All these variations can be calculated using the `xtgee` command with the `corr()` option for specifying predefined or customized correlation structures. Typing

```
. xtgee lsat age, corr(exchangeable)
```

specifies the random-effects model and produces results similar to those from `xtreg, re`.[62]

```
. xtgee lsat age, corr(ar1)
```

which produces a model with an AR(1) correlation matrix. Typing

```
. xtgee lsat age, corr(independent)
```

produces the standard OLS regression model described at the beginning of this section.

You can interpret the coefficients of error-components models just like the coefficients of a simple OLS regression model. But unlike in the simple OLS model, in an error-components model, if the error-structure is correctly specified, the coefficients are more correct. As the coefficients are based on variations within and between the individuals, you should have no problem investigating the effects of time-invariant independent variables on the dependent variable. Unlike in the fixed-effects model, the coefficients can be biased due to omitted time-invariant covariates.

8.7 Summary

The Stata syntax for computing a regression is always the same, no matter what kind of regression you want. First, you compute the model, save the predicted values and other results, and then test the model assumption, as well as the quality of your model, using the saved results.

[61] AR is short for autoregression.

[62] `xtgee, corr(exchangeable)` and `xtreg, re` produce slightly different results because of implementation details that are too technical to discuss here. In practice, the results are usually quite similar.

`regress` *depvar indepvars* computes an OLS regression of the dependent variable *depvar* on the independent variables *indepvars*.

`predict` *newvar* [, *option*] saves predicted values after regression. You can optionally specify *option* to indicate which results will be saved under the name *newvar*.

`predict yhat` is an example of using `predict` to save the predicted values for each observation in a variable with the name `yhat`. You can use any variable name.

`predict error, resid` saves under the variable name `error` the residuals (the difference between observed and predicted) for each observation.

`regress y x1 x2 x3` computes a multiple regression of the dependent variable `y` on the independent variables `x1`, `x2` and `x3`.

`regress y x1 x2 x3, beta` computes a multiple regression of the dependent variable `y` on the independent variables `x1`, `x2` and `x3`. Standardized residuals will appear in the output.

Additional options for `predict` are

`rstudent` for studentized residuals.

`rstandard` for standardized residuals.

`cooksd` for Cook's D as a measure of the influence each observation has on the entire model.

`dfbeta` for DFBETA as measure of the influence each observation has on an individual regression coefficient.

`stdr` for the standard error for residuals.

The most important commands in regression diagnostics are

`graph twoway (lfit y x) (scatter y x)`, which draws a scatterplot of `y` and `x` with the regression line.

`avplots`, which shows added-variable plots.

`rvfplot`, which delivers a residual-versus-fitted plot.

`cprplot`, which delivers a component-plus residual plot.

`help regress postestimation`, which provides more information on regression diagnostics.