3 Linear regression basics

3.1 Introduction

Linear regression analysis is often the starting point of an empirical investigation. Because of its relative simplicity, it is useful for illustrating the different steps of a typical modeling cycle that involves an initial specification of the model followed by estimation, diagnostic checks, and model respecification. The purpose of such a linear regression analysis may be to summarize the data, generate conditional predictions, or test and evaluate the role of specific regressors. We will illustrate these aspects using a specific data example.

This chapter is limited to basic regression analysis on cross-section data of a continuous dependent variable. The setup is for a single equation and exogenous regressors. Some standard complications of linear regression, such as misspecification of the conditional mean and model errors that are heteroskedastic, will be considered. In particular, we model the natural logarithm of medical expenditures instead of the level. We will ignore other various aspects of the data that can lead to more sophisticated nonlinear models presented in later chapters.

3.2 Data and data summary

The first step is to decide what dataset will be used. In turn, this decision depends on the population of interest and the research question itself. We discussed how to convert a raw dataset to a form amenable to regression analysis in chapter 2. In this section, we present ways to summarize and gain some understanding of the data, a necessary step before any regression analysis.

3.2.1 Data description

We analyze medical expenditures of individuals 65 years and older who qualify for health care under the U.S. Medicare program. The original data source is the Medical Expenditure Panel Survey (MEPS).

Medicare does not cover all medical expenses. For example, copayments for medical services and expenses of prescribed pharmaceutical drugs were not covered for the time period studied here. About half of eligible individuals therefore purchase supplementary insurance in the private market that provides insurance coverage against various out-of-pocket expenses.

In this chapter, we consider the impact of this supplementary insurance on total an-
medical expenditures of an individual, measured in dollars. A formal investigation
must control for the influence of other factors that also determine individual medical
expenditure, notably, sociodemographic factors such as age, gender, education and in-
come, geographical location, and health-status measures such as self-assessed health
presence of chronic or limiting conditions. In this chapter, as in other chapters,
instead deliberately use a short list of regressors. This permits shorter output and
simpler discussion of the results, an advantage because our intention is to simply explain
methods and tools available in Stata.
the

3.2.2 Variable description

Given the Stata dataset for analysis, we begin by using the `describe` command to list
various features of the variables to be used in the linear regression. The command with-
a variable list describes all the variables in the dataset. Here we restrict attention
the variables used in this chapter.
to

```
. * Variable description for medical expenditure dataset
. use mus03data.dta
. describe totexp ltotexp posexp suppins phylim actlim totchr age female income
```

variable name	storage type	display format	value label	variable label
totexp	double	%12.0g		Total medical expenditure
ltotexp	float	%9.0g		ln(totexp) if totexp > 0
posexp	float	%9.0g		=1 if total expenditure > 0
suppins	float	%9.0g		=1 if has supp priv insurance
phylim	double	%12.0g		=1 if has functional limitation
actlim	double	%12.0g		=1 if has activity limitation
totchr	double	%12.0g		# of chronic problems
age	double	%12.0g		Age
female	double	%12.0g		=1 if female
income	double	%12.0g		annual household income/1000

variable types and format columns indicate that all the data are numeric. In this
case, some variables are stored in single precision (`float`) and some in double precision
The
ble). From the variable labels, we expect `totexp` to be nonnegative; `ltotexp` to
missing if `totexp` equals zero; `posexp`, `suppins`, `phylim`, `actlim`, and `female` to
(
0 or 1; `totchr` to be a nonnegative integer; `age` to be positive; and `income` to be
negative or positive. Note that the integer variables could have been stored much more
compactly as `integer` or `byte`. The variable labels provide a short description that is
helpful but may not fully describe the variable. For example, the key regressor `suppins`
created by aggregating across several types of private supplementary insurance. No
labels for the values taken by the categorical variables have been provided.
was

3.2.3 Summary statistics

It is essential in any data analysis to first check the data by using the summarize command.

```
. * Summary statistics for medical expenditure dataset
. summarize totexp ltotexp posexp suppins phylim actlim totchr age female income
```

Variable	Obs	Mean	Std. Dev.	Min	Max
totexp	3064	7030.889	11852.75	0	125610
ltotexp	2955	8.059866	1.367592	1.098612	11.74094
posexp	3064	.9644256	.1852568	0	1
suppins	3064	.5812663	.4934321	0	1
phylim	3064	.4255875	.4945125	0	1
actlim	3064	.2836162	.4508263	0	1
totchr	3064	1.754243	1.307197	0	7
age	3064	74.17167	6.372938	65	90
female	3064	.5796345	.4936982	0	1
income	3064	22.47472	22.53491	-1	312.46

On average, 96% of individuals incur medical expenditures during a year; 58% have supplementary insurance; 43% have functional limitations; 28% have activity limitations; and 58% are female, as the elderly population is disproportionately female because of the greater longevity of women. The only variable to have missing data is ltotexp, the natural logarithm of totexp, which is missing for the $(3064 - 2955) = 109$ observations with totexp $= 0$.

All variables have the expected range, except that income is negative. To see how many observations on income are negative, we use the tabulate command, restricting attention to nonpositive observations to limit output.

```
. * Tabulate variable
. tabulate income if income <= 0
```

annual household income/1000	Freq.	Percent	Cum.
-1	1	1.14	1.14
0	87	98.86	100.00
Total	88	100.00	

Only one observation is negative, and negative income is possible for income from self-employment or investment. We include the observation in the analysis here, though checking the original data source may be warranted.

Much of the subsequent regression analysis will drop the 109 observations with zero medical expenditures, so in a research paper, it would be best to report summary statistics without these observations.

3.2.4 More-detailed summary statistics

Additional descriptive analysis of key variables, especially the dependent variable, is useful. For totexp, the level of medical expenditures, summarize, detail yields

```
* Detailed summary statistics of a single variable
summarize totexp, detail
                    Total medical expenditure
-------------------------------------------------------------
      Percentiles      Smallest
 1%           0             0
 5%         112             0
10%         393             0       Obs                3064
25%        1271             0       Sum of Wgt.        3064
50%      3134.5                     Mean             7030.889
                        Largest     Std. Dev.        11852.75
75%        7151        104823
90%       17050        108256       Variance         1.40e+08
95%       27367        123611       Skewness         4.165058
99%       62346        125610       Kurtosis         26.26796
```

Medical expenditures vary greatly across individuals, with a standard deviation of 11,853, which is almost twice the mean. The median of 3,134 is much smaller than the mean of 7,031, reflecting the skewness of the data. For variable x, the skewness statistic is a scale-free measure of skewness that estimates $E\{(x-\mu)^3\}/\sigma^{3/2}$, the third central moment standardized by the second central moment. The skewness is zero for symmetrically distributed data. The value here of 4.16 indicates considerable right skewness. The kurtosis statistic is an estimate of $E\{(x-\mu)^4\}/\sigma^4$, the fourth central moment standardized by the second central moment. The reference value is 3, the value for normally distributed data. The much higher value here of 26.26 indicates that the tails are much thicker than those of a normal distribution. You can obtain additional summary statistics by using the centile command to obtain other percentiles and by using the table command, which is explained in section 3.2.5.

We conclude that the distribution of the dependent variable is considerably skewed and has thick tails. These complications often arise for commonly studied individual-level economic variables such as expenditures, income, earnings, wages, and house prices. It is possible that including regressors will eliminate the skewness, but in practice, much of the variation in the data will be left unexplained ($R^2 < 0.3$ is common for individual-level data) and skewness and excess kurtosis will remain.

Such skewed, thick-tailed data suggest a model with multiplicative errors instead of additive errors. A standard solution is to transform the dependent variable by taking the natural logarithm. Here this is complicated by the presence of 109 zero-valued observations. We take the expedient approach of dropping the zero observations from analysis in either logs or levels. This should make little difference here because only 3.6% of the sample is then dropped. A better approach, using two-part or selection models, is covered in chapter 16.

The output for tabstat in section 3.2.5 reveals that taking the natural logarithm for these data essentially eliminates the skewness and excess kurtosis.

The user-written **fsum** command (Wolfe 2002) is an enhancement of **summarize** that enables formatting the output and including additional information such as percentiles and variable labels. The user-written **outsum** command (Papps 2006) produces a text file of means and standard deviations for one or more subsets of the data, e.g., one column for the full sample, one for a male subsample, and one for a female subsample.

3.2.5 Tables for data

One-way tables can be created by using the `table` command, which produces just frequencies, or the `tabulate` command, which additionally produces percentages and cumulative percentages; an example was given in section 3.2.3.

Two-way tables can also be created by using these commands. For frequencies, only `table` produces clean output. For example,

```
* Two-way table of frequencies
table female totchr
```

=1 if female	# of chronic problems							
	0	1	2	3	4	5	6	7
0	239	415	323	201	82	23	4	1
1	313	466	493	305	140	46	11	2

provides frequencies for a two-way tabulation of gender against the number of chronic conditions. The `tabulate` command is much richer. For example,

```
* Two-way table with row and column percentages and Pearson chi-squared
tabulate female suppins, row col chi2
```

Key
frequency *row percentage* *column percentage*

=1 if female	=1 if has supp priv insurance		Total
	0	1	
0	488	800	1,288
	37.89	62.11	100.00
	38.04	44.92	42.04
1	795	981	1,776
	44.76	55.24	100.00
	61.96	55.08	57.96
Total	1,283	1,781	3,064
	41.87	58.13	100.00
	100.00	100.00	100.00

Pearson chi2(1) = 14.4991 Pr = 0.000

Comparing the row percentages for this sample, we see that while a woman is more likely to have supplemental insurance than not, the probability that a woman in this sample has purchased supplemental insurance is lower than the probability that a man in this sample has purchased supplemental insurance. Although we do not have the information to draw these inferences for the population, the results for Pearson's chi-squared test soundly reject the null hypothesis that these variables are independent. Other tests of association are available. The related command `tab2` will produce all possible two-way tables that can be obtained from a list of several variables.

For multiway tables, it is best to use `table`. For the example at hand, we have

```
. * Three-way table of frequencies
. table female totchr suppins
```

=1 if female	=1 if has supp priv insurance and # of chronic problems, 0							
	0	1	2	3	4	5	6	7
0	102	165	121	68	25	6	1	
1	135	212	233	134	56	22	1	2

=1 if female	=1 if has supp priv insurance and # of chronic problems, 1							
	0	1	2	3	4	5	6	7
0	137	250	202	133	57	17	3	1
1	178	254	260	171	84	24	10	

An alternative is to use `tabulate` with the `by` prefix, but the results are not as neat as those from `table`.

The preceding tabulations will produce voluminous output if one of the variables being tabulated takes on many values. Then it is much better to use `table` with the `contents()` option to present tables that give key summary statistics for that variable, such as the mean and standard deviation. Such tabulations can be useful even when variables take on few values. For example, when summarizing the number of chronic problems by gender, `table` yields

```
. * One-way table of summary statistics
. table female, contents(N totchr mean totchr sd totchr p50 totchr)
```

=1 if female	N(totchr)	mean(totchr)	sd(totchr)	med(totchr)
0	1,288	1.659937888	1.261175	1
1	1,776	1.822635135	1.335776	2

Women on average have more chronic problems (1.82 versus 1.66 for men). The option `contents()` can produce many other statistics, including the minimum, maximum, and key percentiles.

The `table` command with the `contents()` option can additionally produce two-way and multiway tables of summary statistics. As an example,

```
. * Two-way table of summary statistics
. table female suppins, contents(N totchr mean totchr)
```

=1 if female	=1 if has supp priv insurance 0	1
0	488	800
	1.530737705	1.73875
1	795	981
	1.803773585	1.837920489

shows that those with supplementary insurance on average have more chronic problems. This is especially so for males (1.74 versus 1.53).

The `tabulate, summarize()` command can be used to produce one-way and two-way tables with means, standard deviations, and frequencies. This is a small subset of the statistics that can be produced using `table`, so we might as well use `table`.

The `tabstat` command provides a table of summary statistics that permits more flexibility than `summarize`. The following output presents summary statistics on medical expenditures and the natural logarithm of expenditures that are useful in determining skewness and kurtosis.

```
. * Summary statistics obtained using command tabstat
. tabstat totexp ltotexp, stat (count mean p50 sd skew kurt) col(stat)
```

variable	N	mean	p50	sd	skewness	kurtosis
totexp	3064	7030.889	3134.5	11852.75	4.165058	26.26796
ltotexp	2955	8.059866	8.111928	1.367592	-.3857887	3.842263

This reproduces information given in section 3.2.4 and shows that taking the natural logarithm eliminates most skewness and kurtosis. The `col(stat)` option presents the results with summary statistics given in the columns and each variable being given in a separate row. Without this option, we would have summary statistics in rows and variables in the columns. A two-way table of summary statistics can be obtained by using the `by()` option.

(*Continued on next page*)

3.2.6 Statistical tests

The ttest command can be used to test hypotheses about the population mean of a single variable (H_0: $\mu = \mu^*$ for specified value μ^*) and to test the equality of means (H_0: $\mu_1 = \mu_2$). For more general analysis of variance and analysis of covariance, the oneway and anova commands can be used, and several other tests exist for more specialized examples such as testing the equality of proportions. These commands are rarely used in microeconometrics because they can be recast as a special case of regression with an intercept and appropriate indicator variables. Furthermore, regression has the advantage of reliance on less restrictive distributional assumptions, provided samples are large enough for asymptotic theory to provide a good approximation.

For example, consider testing the equality of mean medical expenditures for those with and without supplementary health insurance. The ttest totexp, by(suppins) unequal command performs the test but makes the restrictive assumption of a common variance for all those with suppins=0 and a (possibly different) common variance for all those with suppins=1. An alternative method is to perform ordinary least-squares (OLS) regression of totexp on an intercept and suppins and then test whether suppins has coefficient zero. Using this latter method, we can permit all observations to have a different variance by using the vce(robust) option for regress to obtain heteroskedastic-consistent standard errors; see section 3.3.4.

3.2.7 Data plots

It is useful to plot a histogram or a density estimate of the dependent variable. Here we use the kdensity command, which provides a kernel estimate of the density.

The data are highly skewed, with a 97th percentile of approximately $40,000 and a maximum of $1,000,000. The kdensity totexp command will therefore bunch 97% of the density in the first 4% of the x axis. One possibility is to type kdensity totexp if totexp < 40000, but this produces a kernel density estimate assuming the data are truncated at $40,000. Instead, we use command kdensity totexp, we save the evaluation points in kx1 and the kernel density estimates in kd1, and then we line-plot kd1 against kx1.

We do this for both the level and the natural logarithm of medical expenditures, and we use graph combine to produce a figure that includes both density graphs (shown in figure 3.1). We have

```
* Kernel density plots with adjustment for highly skewed data
kdensity totexp if posexp==1, generate (kx1 kd1) n(500)
graph twoway (line kd1 kx1) if kx1 < 40000, name(levels)
kdensity ltotexp if posexp==1, generate (kx2 kd2) n(500)
graph twoway (line kd2 kx2) if kx2 < ln(40000), name(logs)
graph combine levels logs, iscale(1.0)
```

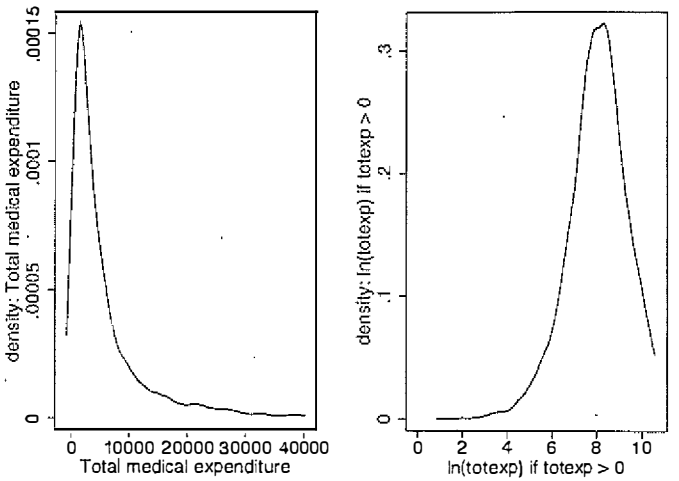

Figure 3.1. Comparison of densities of level and natural logarithm of medical expenditures

Only positive expenditures are considered, and for graph readability, the very long right tail of totexp has been truncated at \$40,000. In figure 3.1, the distribution of totexp is very right-skewed, whereas that of ltotexp is fairly symmetric.

3.3 Regression in levels and logs

We present the linear regression model, first in levels and then for a transformed dependent variable, here in logs.

3.3.1 Basic regression theory

We begin by introducing terminology used throughout the rest of this book. Let θ denote the vector of parameters to be estimated, and let $\widehat{\theta}$ denote an estimator of θ. Ideally, the distribution of $\widehat{\theta}$ is centered on θ with small variance, for precision, and a known distribution, to permit statistical inference. We restrict analysis to estimators that are consistent for θ, meaning that in infinitely large samples, $\widehat{\theta}$ equals θ aside from negligible random variation. This is denoted by $\widehat{\theta} \xrightarrow{p} \theta$ or more formally by $\widehat{\theta} \xrightarrow{p} \theta_0$, where θ_0 denotes the unknown "true" parameter value. A necessary condition for consistency is correct model specification or, in some leading cases, correct specification of key components of the model, most notably the conditional mean.

Under additional assumptions, the estimators considered in this book are asymptotically normally distributed, meaning that their distribution is well approximated by the multivariate normal in large samples. This is denoted by

$$\widehat{\theta} \overset{a}{\sim} N\{\theta,\ \text{Var}(\widehat{\theta})\}$$

where $\text{Var}(\widehat{\theta})$ denotes the (asymptotic) variance–covariance matrix of the estimator (VCE). More efficient estimators have smaller VCEs. The VCE depends on unknown parameters, so we use an estimate of the VCE, denoted by $\widehat{V}(\widehat{\theta})$. Standard errors of the parameter estimates are obtained as the square root of diagonal entries in $\widehat{V}(\widehat{\theta})$. Different assumptions about the data-generating process (DGP), such as heteroskedasticity, can lead to different estimates of the VCE.

Test statistics based on asymptotic normal results lead to the use of the standard normal distribution and chi-squared distribution to compute critical values and p-values. For some estimators, notably, the OLS estimator, tests are instead based on the t distribution and the F distribution. This makes essentially no difference in large samples with, say, degrees of freedom greater than 100, but it may provide a better approximation in smaller samples.

3.3.2 OLS regression and matrix algebra

The goal of linear regression is to estimate the parameters of the linear conditional mean

$$E(y|\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta} = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K \tag{3.1}$$

where usually an intercept is included so that $x_1 = 1$. Here $\mathbf{x}$ is a $K \times 1$ column vector with the jth entry—the jth regressor x_j—and $\boldsymbol{\beta}$ is a $K \times 1$ column vector with the jth entry β_j.

Sometimes $E(y|\mathbf{x})$ is of direct interest for prediction. More often, however, econometrics studies are interested in one or more of the associated marginal effects (MEs),

$$\frac{\partial E(y|\mathbf{x})}{\partial x_j} = \beta_j$$

for the jth regressor. For example, we are interested in the marginal effect of supplementary private health insurance on medical expenditures. An attraction of the linear model is that estimated MEs are given directly by estimates of the slope coefficients.

The linear regression model specifies an additive error so that, for the typical ith observation,

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i, \quad i = 1, \ldots, N$$

The OLS estimator minimizes the sum of squared errors, $\sum_{i=1}^{N}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$.

Matrix notation provides a compact way to represent the estimator and variance matrix formulas that involve sums of products and cross products. We define the $N \times 1$

column vector y to have the ith entry y_i, and we define the $N \times K$ regressor matrix X to have the ith row $\mathbf{x}_i'$. Then the OLS estimator can be written in several ways, with

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i=1}^{N}\mathbf{x}_i y_i \\ &= \begin{bmatrix} \sum_{i=1}^{N}x_{1i}^2 & \sum_{i=1}^{N}x_{1i}x_{2i} & \cdots & \sum_{i=1}^{N}x_{1i}x_{Ki} \\ \sum_{i=1}^{N}x_{2i}x_{1i} & \sum_{i=1}^{N}x_{2i}^2 & & \vdots \\ & & \ddots & \\ \sum_{i=1}^{N}x_{Ki}x_{1i} & & \cdots & \sum_{i=1}^{N}x_{Ki}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{N}x_{1i}y_i \\ \sum_{i=1}^{N}x_{2i}y_i \\ \vdots \\ \sum_{i=1}^{N}x_{Ki}y_i \end{bmatrix}\end{aligned}$$

We define all vectors as column vectors, with a transpose if row vectors are desired. By contrast, Stata commands and Mata commands define vectors as row vectors, so in parts of Stata and Mata code, we need to take a transpose to conform to the notation in the book.

3.3.3 Properties of the OLS estimator

The properties of any estimator vary with the assumptions made about the DGP. For the linear regression model, this reduces to assumptions about the regression error u_i.

The starting point for analysis is to assume that u_i satisfies the following classical conditions:

1. $E(u_i|\mathbf{x}_i) = 0$ (exogeneity of regressors)
2. $E(u_i^2|\mathbf{x}_i) = \sigma^2$ (conditional homoskedasticity)
3. $E(u_iu_j|\mathbf{x}_i,\mathbf{x}_j) = 0,\ i \neq j$, (conditionally uncorrelated observations)

Assumption 1 is essential for consistent estimation of $\boldsymbol{\beta}$ and implies that the conditional mean given in (3.1) is correctly specified. This means that the conditional mean is linear and that all relevant variables have been included in the regression. Assumption 1 is relaxed in chapter 6.

Assumptions 2 and 3 determine the form of the VCE of $\widehat{\boldsymbol{\beta}}$. Assumptions 1–3 lead to $\widehat{\boldsymbol{\beta}}$ being asymptotically normally distributed with the default estimator of the VCE

$$\widehat{\mathbf{V}}_{\text{default}}(\widehat{\boldsymbol{\beta}}) = s^2(\mathbf{X}'\mathbf{X})^{-1}$$

where

$$s^2 = (N-k)^{-1}\sum_i \widehat{u}_i^2 \tag{3.2}$$

and $\widehat{u}_i = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$. Under assumptions 1–3, the OLS estimator is fully efficient. If, additionally, u_i is normally distributed, then "t statistics" are exactly t distributed. This

fourth assumption is not made, but it is common to continue to use the t distribution in the hope that it provides a better approximation than the standard normal in finite samples.

When assumptions 2 and 3 are relaxed, OLS is no longer fully efficient. In chapter 5, we present examples of more-efficient feasible generalized least-squares (FGLS) estimation. In the current chapter, we continue to use the OLS estimator, as is often done in practice, but we use alternative estimates of the VCE that are valid when assumption 2, assumption 3, or both are relaxed.

3.3.4 Heteroskedasticity-robust standard errors

Given assumptions 1 and 3, but not 2, we have heteroskedastic uncorrelated errors. Then a robust estimator, or more precisely a heteroskedasticity-robust estimator, of the VCE of the OLS estimator is

$$\widehat{\mathbf{V}}_{\text{robust}}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\left(\frac{N}{N-k}\sum_i \widehat{u}_i^2\mathbf{x}_i\mathbf{x}_i'\right)(\mathbf{X}'\mathbf{X})^{-1} \tag{3.3}$$

For cross-section data that are independent, this estimator, introduced by White (1980), has supplanted the default variance matrix estimate in most applied work because heteroskedasticity is the norm, and in that case, the default estimate of the VCE is incorrect.

In Stata, a robust estimate of the VCE is obtained by using the `vce(robust)` option of the `regress` command, as illustrated in section 3.4.2. Related options are `vce(hc2)` and `vce(hc3)`, which may provide better heteroskedasticity-robust estimates of the VCE when the sample size is small; see [R] **regress**. The robust estimator of the VCE has been extended to other estimators and models, and a feature of Stata is the `vce(robust)` option, which is applicable for many estimation commands. Some user-written commands use `robust` in place of `vce(robust)`.

3.3.5 Cluster–robust standard errors

When errors for different observations are correlated, assumption 3 is violated. Then both default and robust estimates of the VCE are invalid. For time-series data, this is the case if errors are serially correlated, and the `newey` command should be used. For cross-section data, this can arise when errors are clustered.

Clustered or grouped errors are errors that are correlated within a cluster or group and are uncorrelated across clusters. A simple example of clustering arises when sampling is of independent units but errors for individuals within the unit are correlated. For example, 100 independent villages may be sampled, with several people from each village surveyed. Then, if a regression model overpredicts y for one village member, it is likely to overpredict for other members of the same village, indicating positive correlation. Similar comments apply when sampling is of households with several individuals in each household. Another leading example is panel data with independence over individuals but with correlation over time for a given individual.

Given assumption 1, but not 2 or 3, a cluster–robust estimator of the VCE of the OLS estimator is

$$\widehat{V}_{\text{cluster}}[\widehat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1}\left(\frac{G}{G-1}\frac{N-1}{N-k}\sum_g \mathbf{X}_g\widehat{\mathbf{u}}_g\widehat{\mathbf{u}}_g'\mathbf{X}_g'\right)(\mathbf{X}'\mathbf{X})^{-1}$$

where $g = 1, \ldots, G$ denotes the cluster (such as village), $\widehat{\mathbf{u}}_g$ is the vector of residuals for the observations in the gth cluster, and $\mathbf{X}_g$ is a matrix of the regressors for the observations in the gth cluster. The key assumptions made are error independence across clusters and that the number of clusters $G \to \infty$.

Cluster–robust standard errors can be computed by using the `vce(cluster clustvar)` option in Stata, where clusters are defined by the different values taken by the *clustvar* variable. The estimate of the VCE is in fact heteroskedasticity-robust and cluster–robust, because there is no restriction on $\text{Cov}(u_{gi}, u_{gj})$. The cluster VCE estimate can be applied to many estimators and models; see section 9.6.

Cluster–robust standard errors must be used when data are clustered. For a scalar regressor x, a rule of thumb is that cluster–robust standard errors are $\sqrt{1 + \rho_x\rho_u(M-1)}$ times the incorrect default standard errors, where ρ_x is the within-cluster correlation coefficient of the regressor, ρ_u is the within-cluster correlation coefficient of the error, and M is the average cluster size.

It can be necessary to use cluster–robust standard errors even where it is not immediately obvious. This is particularly the case when a regressor is an aggregated or macro variable, because then $\rho_x = 1$. For example, suppose we use data from the U.S. Current Population Survey and regress individual earnings on individual characteristics and a state-level regressor that does not vary within a state. Then, if there are many individuals in each state so M is large, even slight error correlation for individuals in the same state can lead to great downward bias in default standard errors and in heteroskedasticity–robust standard errors. Clustering can also be induced by the design of sample surveys. This topic is pursued in section 5.5.

3.3.6 Regression in logs

The medical expenditure data are very right-skewed. Then a linear model in levels can provide very poor predictions because it restricts the effects of regressors to be additive. For example, aging 10 years is assumed to increase medical expenditures by the same amount regardless of observed health status. Instead, it is more reasonable to assume that aging 10 years has a multiplicative effect. For example, it may increase medical expenditures by 20%.

We begin with an exponential mean model for positive expenditures, with error that is also multiplicative, so $y_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})\varepsilon_i$. Defining $\varepsilon_i = \exp(u_i)$, we have $y_i = \exp(\mathbf{x}_i'\boldsymbol{\beta} + u_i)$, and taking the natural logarithm, we fit the log-linear model

$$\ln y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$$

by OLS regression of $\ln y$ on x. The conditional mean of $\ln y$ is being modeled, rather than the conditional mean of y. In particular,

$$E(\ln y|\mathbf{x}) = \mathbf{x}'\beta$$

assuming u_i is independent with conditional mean zero.

Parameter interpretation requires care. For regression of $\ln y$ on x, the coefficient β_j measures the effect of a change in regressor x_j on $E(\ln y|\mathbf{x})$, but ultimate interest lies instead on the effect on $E(y|\mathbf{x})$. Some algebra shows that β_j measures the proportionate change in $E(y|\mathbf{x})$ as x_j changes, called a semielasticity, rather than the level of change in $E(y|\mathbf{x})$. For example, if $\beta_j = 0.02$, then a one-unit change in x_j is associated with a proportionate increase of 0.02, or 2%, in $E(y|\mathbf{x})$.

Prediction of $E(y|\mathbf{x})$ is substantially more difficult because it can be shown that $E(\ln y|\mathbf{x}) \neq \exp(\mathbf{x}'\beta)$. This is pursued in section 3.6.3.

3.4 Basic regression analysis

We use `regress` to run an OLS regression of the natural logarithm of medical expenditures, `ltotexp`, on `suppins` and several demographic and health-status measures. Using $\ln y$ rather than y as the dependent variable leads to no change in the implementation of OLS but, as already noted, will change the interpretation of coefficients and predictions.

Many of the details we provide in this section are applicable to all Stata estimation commands, not just to `regress`.

3.4.1 Correlations

Before regression, it can be useful to investigate pairwise correlations of the dependent variables and key regressor variables by using `correlate`. We have

```
  * Pairwise correlations for dependent variable and regressor variables
. correlate ltotexp suppins phylim actlim totchr age female income
(obs=2955)
```

	ltotexp	suppins	phylim	actlim	totchr	age
ltotexp	1.0000					
suppins	0.0941	1.0000				
phylim	0.2924	-0.0243	1.0000			
actlim	0.2888	-0.0675	0.5904	1.0000		
totchr	0.4283	0.0124	0.3334	0.3260	1.0000	
age	0.0858	-0.1226	0.2538	0.2394	0.0904	1.0000
female	-0.0058	-0.0796	0.0943	0.0499	0.0557	0.0774
income	0.0023	0.1943	-0.1142	-0.1483	-0.0816	-0.1542

	female	income
female	1.0000	
income	-0.1312	1.0000

Medical expenditures are most highly correlated with the health-status measures phylim, actlim, and totchr. The regressors are only weakly correlated with each other, aside from the health-status measures. Note that correlate restricts analysis to the 2,955 observations where data are available for all variables in the variable list. The related command pwcorr, not demonstrated, with the sig option gives the statistical significance of the correlations.

3.4.2 The regress command

The regress command performs OLS regression and yields an analysis-of-variance table, goodness-of-fit statistics, coefficient estimates, standard errors, t statistics, p-values, and confidence intervals. The syntax of the command is

regress *depvar* [*indepvars*] [*if*] [*in*] [*weight*] [, *options*]

Other Stata estimation commands have similar syntaxes. The output from regress is similar to that from many linear regression packages.

For independent cross-section data, the standard approach is to use the vce(robust) option, which gives standard errors that are valid even if model errors are heteroskedastic; see section 3.3.4. In that case, the analysis-of-variance table, based on the assumption of homoskedasticity, is dropped from the output. We obtain

```
. * OLS regression with heteroskedasticity-robust standard errors
. regress ltotexp suppins phylim actlim totchr age female income, vce(robust)
Linear regression                               Number of obs =    2955
                                                F(  7,  2947) =  126.97
                                                Prob > F      =  0.0000
                                                R-squared     =  0.2289
                                                Root MSE      =  1.2023
```

ltotexp	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
suppins	.2556428	.0465982	5.49	0.000	.1642744	.3470112
phylim	.3020598	.057705	5.23	0.000	.1889136	.415206
actlim	.3560054	.0634066	5.61	0.000	.2316797	.4803311
totchr	.3758201	.0187185	20.08	0.000	.3391175	.4125228
age	.0038016	.0037028	1.03	0.305	-.0034587	.011062
female	-.0843275	.045654	-1.85	0.065	-.1738444	.0051894
income	.0025498	.0010468	2.44	0.015	.0004973	.0046023
cons	6.703737	.2825751	23.72	0.000	6.149673	7.257802

The regressors are jointly statistically significant, because the overall F statistic of 126.97 has a p-value of 0.000. At the same time, much of the variation is unexplained with $R^2 = 0.2289$. The root MSE statistic reports s, the standard error of the regression, defined in (3.2). By using a two-sided test at level 0.05, all regressors are individually statistically significant because $p < 0.05$, aside from age and female. The strong statistical insignificance of age may be due to sample restriction to elderly people and the inclusion of several health-status measures that capture well the health effect of age.

Statistical significance of coefficients is easily established. More important is the economic significance of coefficients, meaning the measured impact of regressors on medical expenditures. This is straightforward for regression in levels, because we can directly use the estimated coefficients. But here the regression is in logs. From section 3.3.6, in the log-linear model, parameters need to be interpreted as semielasticities. For example, the coefficient on suppins is 0.256. This means that private supplementary insurance is associated with a 0.256 proportionate rise, or a 25.6% rise, in medical expenditures. Similarly, large effects are obtained for the health-status measures, whereas health expenditures for women are 8.4% lower than those for men after controlling for other characteristics. The income coefficient of 0.0025 suggests a very small effect, but this is misleading. The standard deviation of income is 22, so a 1–standard deviation in income leads to a 0.055 proportionate rise, or 5.5% rise, in medical expenditures.

MEs in nonlinear models are discussed in more detail in section 10.6. The preceding interpretations are based on calculus methods that consider very small changes in the regressor. For larger changes in the regressor, the finite-difference method is more appropriate. Then the interpretation in the log-linear model is similar to that for the exponential conditional mean model; see section 10.6.4. For example, the estimated effect of going from no supplementary insurance (suppins=0) to having supplementary insurance (suppins=1) is more precisely a $100 \times (e^{0.256} - 1)$, or 29.2%, rise.

The regress command provides additional results that are not listed. In particular, the estimate of the VCE is stored in the matrix e(V). Ways to access this and other stored results from regression have been given in section 1.6. Various postestimation commands enable prediction, computation of residuals, hypothesis testing, and model specification tests. Many of these are illustrated in subsequent sections. Two useful commands are

```
. * Display stored results and list available postestimation commands
. ereturn list
  (output omitted)
. help regress postestimation
  (output omitted)
```

3.4.3 Hypothesis tests

The test command performs hypothesis tests using the Wald test procedure that uses the estimated model coefficients and VCE. We present some leading examples here, with a more extensive discussion deferred to section 12.3. The F statistic version of the Wald test is used after regress, whereas for many other estimators the chi-squared version is instead used.

A common test is one of equality of coefficients. For example, consider testing that having a functional limitation has the same impact on medical expenditures as having an activity limitation. The test of $H_0: \beta_{\texttt{phylim}} = \beta_{\texttt{actlim}}$ against $H_a: \beta_{\texttt{phylim}} \neq \beta_{\texttt{actlim}}$ is implemented as

```
  * Wald test of equality of coefficients
  quietly regress ltotexp suppins phylim actlim totchr age female
> income, vce(robust)
  test phylim = actlim
 ( 1)  phylim - actlim = 0
       F(  1,  2947) =    0.27
            Prob > F =    0.6054
```

Because $p = 0.61 > 0.05$, we do not reject the null hypothesis at the 5% significance level. There is no statistically significant difference between the coefficients of the two variables.

The model can also be fitted subject to constraints. For example, to obtain the least-squares estimates subject to $\beta_{\texttt{phylim}} = \beta_{\texttt{actlim}}$, we define the constraint using `constraint define` and then fit the model using `cnsreg` for constrained regression with the `constraints()` option. See exercise 2 at the end of this chapter for an example.

Another common test is one of the joint statistical significance of a subset of the regressors. A test of the joint significance of the health-status measures is one of H_0: $\beta_{\texttt{phylim}} = 0$, $\beta_{\texttt{actlim}} = 0$, $\beta_{\texttt{totchr}} = 0$ against H_a: at least one is nonzero. This is implemented as

```
. * Joint test of statistical significance of several variables
. test phylim actlim totchr
 ( 1)  phylim = 0
 ( 2)  actlim = 0
 ( 3)  totchr = 0
       F(  3,  2947) =  272.36
            Prob > F =    0.0000
```

These three variables are jointly statistically significant at the 0.0.5 level because $p = 0.000 < 0.05$.

3.4.4 Tables of output from several regressions

It is very useful to be able to tabulate key results from multiple regressions for both one's own analysis and final report writing.

The `estimates store` command after regression leads to results in `e()` being associated with a user-provided model name and preserved even if subsequent models are fitted. Given one or more such sets of stored estimates, `estimates table` presents a table of regression coefficients (the default) and, optionally, additional results. The `estimates stats` command lists the sample size and several likelihood-based statistics.

We compare the original regression model with a variant that replaces `income` with `educyr`. The example uses several of the available options for `estimates table`.

```
. * Store and then tabulate results from multiple regressions
. quietly regress ltotexp suppins phylim actlim totchr age female income,
> vce(robust)
. estimates store REG1
. quietly regress ltotexp suppins phylim actlim totchr age female educyr,
> vce(robust)
. estimates store REG2
. estimates table REG1 REG2, b(%9.4f) se stats(N r2 F ll)
> keep(suppins income educyr)
```

Variable	REG1	REG2
suppins	0.2556	0.2063
	0.0466	0.0471
income	0.0025	
	0.0010	
educyr		0.0480
		0.0070
N	2955.0000	2955.0000
r2	0.2289	0.2406
F	126.9723	132.5337
ll	-4.73e+03	-4.71e+03

legend: b/se

This table presents coefficients (b) and standard errors (se), with other available options including t statistics (t) and p-values (p). The statistics given are the sample size, the R^2, the overall F statistic (based on the robust estimate of the VCE), and the log likelihood (based on the strong assumption of normal homoskedastic errors). The keep() option, like the drop() option, provides a way to tabulate results for just the key regressors of interest. Here educyr is a much stronger predictor than income, because it is more highly statistically significant and R^2 is higher, and there is considerable change in the coefficient of suppins.

3.4.5 Even better tables of regression output

The preceding table is very useful for model comparison but has several limitations. It would be more readable if the standard errors appeared in parentheses. It would be beneficial to be able to report a p-value for the overall F statistic. Also some work may be needed to import the table into a table format in external software such as Excel, Word, or LaTeX.

The user-written esttab command (Jann 2007) provides a way to do this, following the estimates store command. A cleaner version of the previous table is given by

```
. * Tabulate results using user-written command esttab to produce cleaner output
. esttab REG1 REG2, b(%10.4f) se scalars(N r2 F ll) mtitles
> keep(suppins income educyr) title("Model comparison of REG1-REG2")
```

Model comparison of REG1-REG2

	(1) REG1	(2) REG2
suppins	0.2556***	0.2063***
	(0.0466)	(0.0471)
income	0.0025*	
	(0.0010)	
educyr		0.0480***
		(0.0070)
N	2955	2955
r2	0.2289	0.2406
F	126.9723	132.5337
ll	-4733.4476	-4710.9578

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Now standard errors are in parentheses, the strength of statistical significance is given using stars that can be suppressed by using the `nostar` option, and a title is added.

The table can be written to a file that, for example, creates a table in LaTeX.

```
  * Write tabulated results to a file in latex table format
  quietly esttab REG1 REG2 using mus03table.tex, replace b(%10.4f) se
> scalars(N r2 F ll) mtitles keep(suppins age income educyr _cons)
> title("Model comparison of REG1-REG2")
```

Other formats include `.rtf` for rich text format (Word), `.csv` for comma-separated values, and `.txt` for fixed and tab-delimited text.

As mentioned earlier, this table would be better if the p-value for the overall F statistic were provided. This is not stored in `e()`. However, it is possible to calculate the p-value given other variables in `e()`. The user-written `estadd` command (Jann 2005) allows adding this computed p-value to stored results that can then be tabulated with `esttab`. We demonstrate this for a smaller table to minimize output.

```
  * Add a user-calculated statistic to the table
  estimates drop REG1 REG2
  quietly regress ltotexp suppins phylim actlim totchr age female income,
> vce(robust)
  estadd scalar pvalue = Ftail(e(df_r),e(df_m),e(F))
  (output omitted)
  estimates store REG1
  quietly regress ltotexp suppins phylim actlim totchr age female educyr,
> vce(robust)
```

```
. estadd scalar pvalue = Ftail(e(df_r),e(df_m),e(F))
(output omitted)
. estimates store REG2
. esttab REG1 REG2, b(%10.4f) se scalars(F pvalue) mtitles keep(suppins)
```

	(1) REG1	(2) REG2
suppins	0.2556***	0.2063***
	(0.0466)	(0.0471)
N	2955	2955
F	126.9723	132.5337
pvalue	0.0000	0.0000

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

The `estimates drop` command saves memory by dropping stored estimates that are no longer needed. In particular, for large samples the sample inclusion indicator `e(sample)` can take up much memory.

Related user-written commands by Jann (2005, 2007) are `estout`, a richer but more complicated version of `esttab`, and `eststo`, which extends `estimates store`. Several earlier user-written commands, notably, `outreg`, also create tables of regression output but are generally no longer being updated by their authors. The user-written `reformat` command (Brady 2002) allows formatting of the usual table of output from a single estimation command.

3.5 Specification analysis

The fitted model has $R^2 = 0.23$, which is reasonable for cross-section data, and most regressors are highly statistically significant with the expected coefficient signs. Therefore, it is tempting to begin interpreting the results.

However, before doing so, it is useful to subject this regression to some additional scrutiny because a badly misspecified model may lead to erroneous inferences. We consider several specification tests, with the notable exception of testing for regressor exogeneity, which is deferred to chapter 6.

3.5.1 Specification tests and model diagnostics

In microeconometrics, the most common approach to deciding on the adequacy of a model is a Wald-test approach that fits a richer model and determines whether the data support the need for a richer model. For example, we may add additional regressors to the model and test whether they have a zero coefficient.

Stata also presents the user with an impressive and bewildering menu of choices of diagnostic checks for the currently fitted regression; see [R] **regress postestimation**. Some are specific to OLS regression, whereas others apply to most regression models. Some are visual aids such as plots of residuals against fitted values. Some are diagnostic statistics such as influence statistics that indicate the relative importance of individual observations. And some are formal tests that test for the failure of one or more assumptions of the model. We briefly present plots and diagnostic statistics, before giving a lengthier treatment of specification tests.

3.5.2 Residual diagnostic plots

Diagnostic plots are used less in microeconometrics than in some other branches of statistics, for several reasons. First, economic theory and previous research provide a lot of guidance as to the likely key regressors and functional form for a model. Studies rely on this and shy away from excessive data mining. Secondly, microeconometric studies typically use large datasets and regressions with many variables. Many variables potentially lead to many diagnostic plots, and many observations make it less likely that any single observation will be very influential, unless data for that observation are seriously miscoded.

We consider various residual plots that can aid in outlier detection, where an outlier is an observation poorly predicted by the model. One way to do this is to plot actual values against fitted values of the dependent variable. The postestimation command `rvfplot` gives a transformation of this, plotting the residuals $\widehat{u}_i = y_i - \widehat{y}_i$ against the fitted values $\widehat{y}_i = \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$. We have

```
. * Plot of residuals against fitted values
. quietly regress ltotexp suppins phylim actlim totchr age female income,
> vce(robust)
. rvfplot
```

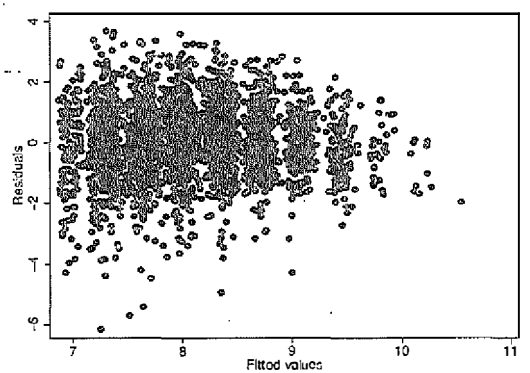

Figure 3.2. Residuals plotted against fitted values after OLS regression

Figure 3.2 does not indicate any extreme outliers, though the three observations with a residual less than -5 may be worth investigating. To do so, we need to generate $\widehat{u}$ by using the `predict` command, detailed in section 3.6, and we need to list some details on those observations with $\widehat{u} < -5$. We have

```
* Details on the outlier residuals
predict uhat, residual
predict yhat, xb
list totexp ltotexp yhat uhat if uhat < -5, clean
```

	totexp	ltotexp	yhat	uhat
1.	3	1.098612	7.254341	-6.155728
2.	6	1.791759	7.513358	-5.721598
3.	9	2.197225	7.631211	-5.433987

The three outlying residuals are for three observations with the very smallest total annual medical expenditures of, respectively, \$3, \$6, and \$9. The model evidently greatly overpredicts for these observations, with the predicted logarithm of total expenditures (`yhat`) much greater than `ltotexp`.

Stata provides several other residual plots. The `rvpplot` postestimation command plots residuals against an individual regressor. The `avplot` command provides an added-variable plot, or partial regression plot, that is a useful visual aid to outlier detection. Other commands give component-plus-residual plots that aid detection of nonlinearities and leverage plots. For details and additional references, see [R] **regress postestimation**.

3.5.3 Influential observations

Some observations may have unusual influence in determining parameter estimates and resulting model predictions.

Influential observations can be detected using one of several measures that are large if the residual is large, the leverage measure is large, or both. The leverage measure of the ith observation, denoted by h_i, equals the ith diagonal entry in the so-called hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$. If h_i is large, then y_i has a big influence on its OLS prediction $\widehat{y}_i$ because $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$. Different measures, including h_i, can be obtained by using different options of `predict`.

A commonly used measure is dfits_i, which can be shown to equal the (scaled) difference between predictions of y_i with and without the ith observation in the OLS regression (so dfits means difference in fits). Large absolute values of dfits indicate an influential data point. One can plot dfits and investigate further observations with outlying values of dfits. A rule of thumb is that observations with $|\text{dfits}| > 2\sqrt{k/N}$ may be worthy of further investigation, though for large datasets this rule can suggest that many observations are influential.

The `dfits` option of `predict` can be used after `regress` provided that regression is with default standard errors because the underlying theory presumes homoskedastic errors. We have

```
. * Compute dfits that combines outliers and leverage
. quietly regress ltotexp suppins phylim actlim totchr age female income
. predict dfits, dfits
. scalar threshold = 2*sqrt((e(df_m)+1)/e(N))
. display "dfits threshold = "  %6.3f threshold
dfits threshold =  0.104
. tabstat dfits, stat (min p1 p5 p95 p99 max) format(%9.3f) col(stat)
    variable |       min        p1        p5       p95       p99       max
-------------+--------------------------------------------------------------
       dfits |    -0.421    -0.147    -0.083     0.085     0.127     0.221
----------------------------------------------------------------------------

. list dfits totexp ltotexp yhat uhat if abs(dfits) > 2*threshold & e(sample),
> clean

          dfits   totexp    ltotexp       yhat        uhat
  1.  -.2319179       3   1.098612   7.254341   -6.155728
  2.  -.3002994       6   1.791759   7.513358   -5.721598
  3.  -.2765266       9   2.197225   7.631211   -5.433987
 10.  -.2170063      30   3.401197   8.348724   -4.947527
 42.  -.2612321     103   4.634729    7.57982   -2.945091
 44.  -.4212185     110    4.70048   8.993904   -4.293423
108.  -.2326284     228   5.429346   7.971406    -2.54206
114.  -.2447627     239   5.476463   7.946239   -2.469776
137.  -.2177336     283   5.645447   7.929719   -2.284273
211.   -.211344     415   6.028278   8.028338    -2.00006
2925.  .2207284   62346   11.04045   8.660131    2.380323
```

Here over 2% of the sample has |dfits| greater than the suggested threshold of 0.104. But only 11 observations have |dfits| greater than two times the threshold. These correspond to observations with relatively low expenditures, or in one case, relatively high expenditures. We conclude that no observation has unusual influence.

3.5.4 Specification tests

Formal model-specification tests have two limitations. First, a test for the failure of a specific model assumption may not be robust with respect to the failure of another assumption that is not under test. For example, the rejection of the null hypothesis of homoskedasticity may be due to a misspecified functional form for the conditional mean. An example is given in section 3.5.5. Second, with a very large sample, even trivial deviations from the null hypothesis of correct specification will cause the test to reject the null hypothesis. For example, if a previously omitted regressor has a very small coefficient, say, 0.000001, then with an infinitely large sample the estimate will be sufficiently precise that we will always reject the null of zero coefficient.

Test of omitted variables

The most common specification test is to include additional regressors and test whether they are statistically significant by using a Wald test of the null hypothesis that the coefficient is zero. The additional regressor may be a variable not already included, a transformation of a variable(s) already included such as a quadratic in age, or a quadratic

with interaction terms in age and education. If groups of regressors are included, such as a set of region dummies, `test` can be used after `regress` to perform a joint test of statistical significance.

In some branches of biostatistics, it is common to include only regressors with $p < 0.05$. In microeconometrics, it is common instead to additionally include regressors that are statistically insignificant if economic theory or conventional practice includes the variable as a control. This reduces the likelihood of inconsistent parameter estimation due to omitted-variables bias at the expense of reduced precision in estimation.

Test of the Box–Cox model

A common specification-testing approach is to fit a richer model that tests the current model as a special case and perform a Wald test of the parameter restrictions that lead to the simpler model. The preceding omitted-variable test is an example.

Here we consider a test specific to the current example. We want to decide whether a regression model for medical expenditures is better in logs than in levels. There is no obvious way to compare the two models because they have different dependent variables. However, the Box–Cox transform leads to a richer model that includes the linear and log-linear models as special cases. Specifically, we fit the model with the transformed dependent variable

$$g(y_i, \theta) \equiv \frac{y_i^{\theta} - 1}{\theta} = \mathbf{x}_i'\boldsymbol{\beta} + u_i$$

where θ and $\boldsymbol{\beta}$ are estimated under the assumption that $u_i \sim N(0, \sigma^2)$. Three leading cases are 1) $g(y, \theta) = y - 1$ if $\theta = 1$; 2) $g(y, \theta) = \ln y$ if $\theta = 0$; and 3) $g(y, \theta) = 1 - 1/y$ if $\theta = -1$. The log-linear model is supported if $\widehat{\theta}$ is close to 0, and the linear model is supported if $\widehat{\theta} = 1$.

The Box–Cox transformation introduces a nonlinearity and an additional unknown parameter θ into the model. This moves the modeling exercise into the domain of nonlinear models. The model is straightforward to fit, however, because Stata provides the boxcox command to fit the model. We obtain

```
. * Boxcox model with lhs variable transformed
. boxcox totexp suppins phylim actlim totchr age female income if totexp>0, nolog
Fitting comparison model
Fitting full model

                                                Number of obs   =       2955
                                                LR chi2(7)      =     773.02
Log likelihood = -28518.267                     Prob > chi2     =      0.000

------------------------------------------------------------------------------
      totexp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      /theta |   .0758956   .0096386     7.87   0.000     .0570042    .0947869
------------------------------------------------------------------------------
```

```
Estimates of scale-variant parameters

                    Coef.

Notrans
     suppins     .4459618
      phylim      .577317
     actlim     .6905939
     totchr     .6754338
        age     .0051321
     female    -.1767976
     income     .0044039
      _cons     8.930566

     /sigma     2.189679


   Test        Restricted      LR statistic      P-value
    H0:      log likelihood        chi2        Prob > chi2

theta = -1     -37454.643     17872.75           0.000
theta =  0     -28550.353        64.17           0.000
theta =  1     -31762.809      6489.08           0.000
```

The null hypothesis of $\theta = 0$ is strongly rejected, so the log-linear model is rejected. However, the Box–Cox model with general θ is difficult to interpret and use, and the estimate of $\widehat{\theta} = 0.0759$ gives much greater support for a log-linear model ($\theta = 0$) than the linear model ($\theta = 1$). Thus we prefer to use the log-linear model.

Test of the functional form of the conditional mean

The linear regression model specifies that the conditional mean of the dependent variable (whether measured in levels or in logs) equals $\mathbf{x}_i'\boldsymbol{\beta}$. A standard test that this is the correct specification is a variable augmentation test. A common approach is to add powers of $\widehat{y}_i = \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$, the fitted value of the dependent variable, as regressors and a test for the statistical significance of the powers.

The `estat ovtest` postestimation command provides a RESET test that regresses y on $\mathbf{x}$ and $\widehat{y}^2$, $\widehat{y}^3$, and $\widehat{y}^4$, and jointly tests that the coefficients of $\widehat{y}^2$, $\widehat{y}^3$, and $\widehat{y}^4$ are zero. We have

```
. * Variable augmentation test of conditional mean using estat ovtest
. quietly regress ltotexp suppins phylim actlim totchr age female income,
> vce(robust)

. estat ovtest

Ramsey RESET test using powers of the fitted values of ltotexp
       Ho:  model has no omitted variables
                F(3, 2944) =      9.04
                  Prob > F =      0.0000
```

The model is strongly rejected because $p = 0.000$.

An alternative, simpler test is provided by the `linktest` command. This regresses y on $\widehat{y}$ and $\widehat{y}^2$, where now the original model regressors $\mathbf{x}$ are omitted, and it tests whether the coefficient of $\widehat{y}^2$ is zero. We have

```
. * Link test of functional form of conditional mean
. quietly regress ltotexp suppins phylim actlim totchr age female income,
> vce(robust)

. linktest
```

Source	SS	df	MS	Number of obs =	2955
				F(2, 2952) =	454.81
Model	1301.41696	2	650.708481	Prob > F =	0.0000
Residual	4223.47242	2952	1.43071559	R-squared =	0.2356
				Adj R-squared =	0.2350
Total	5524.88938	2954	1.87030785	Root MSE =	1.1961

ltotexp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_hat	4.429216	.6779517	6.53	0.000	3.09991	5.758522
_hatsq	-.2084091	.0411515	-5.06	0.000	-.2890976	-.1277206
_cons	-14.01127	2.779936	-5.04	0.000	-19.46208	-8.56046

Again the null hypothesis that the conditional mean is correctly specified is rejected. A likely reason is that so few regressors were included in the model, for pedagogical reasons.

The two preceding commands had different formats. The first test used the `estat ovtest` command, where `estat` produces various statistics following estimation and the particular statistics available vary with the previous estimation command. The second test used `linktest`, which is available for a wider range of models.

Heteroskedasticity test

One consequence of heteroskedasticity is that default OLS standard errors are incorrect. This can be readily corrected and guarded against by routinely using heteroskedasticity-robust standard errors.

Nonetheless, there may be interest in formally testing whether heteroskedasticity is present. For example, the retransformation methods for the log-linear model used in section 3.6.3 assume homoskedastic errors. In section 5.3, we present diagnostic plots for heteroskedasticity. Here we instead present a formal test.

A quite general model of heteroskedasticity is

$$\text{Var}(y|\mathbf{x}) = h(\alpha_1 + \mathbf{z}'\boldsymbol{\alpha}_2)$$

where $h(\cdot)$ is a positive monotonic function such as $\exp(\cdot)$ and the variables in $\mathbf{z}$ are functions of the variables in $\mathbf{x}$. Tests for heteroskedasticity are tests of

$$H_0: \boldsymbol{\alpha}_2 = \mathbf{0}$$

and can be shown to be independent of the choice of function $h(\cdot)$. We reject H_0 at the α level if the test statistic exceeds the α critical value of a chi-squared distribution

with degrees of freedom equal to the number of components of **z**. The test is performed by using the `estat hettest` postestimation command. The simplest version is the Breusch–Pagan Lagrange multiplier test, which is equal to N times the uncentered explained sum of squares from the regression of the squared residuals on an intercept and **z**. We use the `iid` option to obtain a different version of the test that relaxes the default assumption that the errors are normally distributed.

Several choices of the components of **z** are possible. By far, the best choice is to use variables that are a priori likely determinants of heteroskedasticity. For example, in regressing the level of earnings on several regressors including years of schooling, it is likely that those with many years of schooling have the greatest variability in earnings. Such candidates rarely exist. Instead, standard choices are to use the OLS fitted value $\widehat{y}$, the default for `estat hettest`, or to use all the regressors so $\mathbf{z} = \mathbf{x}$. White's test for heteroskedasticity is equivalent to letting **z** equal unique terms in the products and cross products of the terms in **x**.

We consider $\mathbf{z} = \widehat{y}$ and $\mathbf{z} = \mathbf{x}$. Then we have

```
. * Heteroskedasticity tests using estat hettest and option iid
. quietly regress ltotexp suppins phylim actlim totchr age female income
. estat hettest, iid
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of ltotexp
         chi2(1)=          32.87
         Prob > chi2  =    0.0000
. estat hettest suppins phylim actlim totchr age female income, iid
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: suppins phylim actlim totchr age female income
         chi2(7)=          93.13
         Prob > chi2  =    0.0000
```

Both versions of the test, with $\mathbf{z} = \widehat{y}$ and with $\mathbf{z} = \mathbf{x}$, have $p = 0.0000$ and strongly reject homoskedasticity.

Omnibus test

An alternative to separate tests of misspecification is an omnibus test, which is a joint test of misspecification in several directions. A leading example is the information matrix (IM) test (see section 12.7), which is a test for correct specification of a fully parametric model based on whether the IM equality holds. For linear regression with normal homoskedastic errors, the IM test can be shown to be a joint test of heteroskedasticity, skewness, and nonnormal kurtosis compared with the null hypothesis of homoskedasticity, symmetry, and kurtosis coefficient of 3; see Hall (1987).

The `estat imtest` postestimation command computes the joint IM test and also splits it into its three components. We obtain

```
* Information matrix test
quietly regress ltotexp suppins phylim actlim totchr age female income
estat imtest
```

Cameron & Trivedi's decomposition of IM-test

Source	chi2	df	p
Heteroskedasticity	139.90	31	0.0000
Skewness	35.11	7	0.0000
Kurtosis	11.96	1	0.0005
Total	186.97	39	0.0000

The overall joint IM test rejects the model assumption that $y \sim N(\mathbf{x}'\boldsymbol{\beta}, \sigma^2\mathbf{I})$, because $p = 0.0000$ in the Total row. The decomposition indicates that all three assumptions of homoskedasticity, symmetry, and normal kurtosis are rejected. Note, however, that the decomposition assumes correct specification of the conditional mean. If instead the mean is misspecified, then that could be the cause of rejection of the model by the IM test.

3.5.5 Tests have power in more than one direction

Tests can have power in more than one direction, so that if a test targeted to a particular type of model misspecification rejects a model, it is not necessarily the case that this particular type of model misspecification is the underlying problem. For example, a test of heteroskedasticity may reject homoskedasticity, even though the underlying cause of rejection is that the conditional mean is misspecified rather than that errors are heteroskedastic.

To illustrate this example, we use the following simulation exercise. The DGP is one with homoskedastic normal errors

$$y_i = \exp(1 + 0.25 \times x_i + 4 \times x_i^2) + u_i,$$
$$x_i \sim U(0,1), \quad u_i \sim N(0,1)$$

We instead fit a model with a misspecified conditional mean function:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + v$$

We consider a simulation with a sample size of 50. We generate the regressors and the dependent variable by using commands detailed in section 4.2. We obtain

```
* Simulation to show tests have power in more than one direction
clear all
set obs 50
obs was 0, now 50
set seed 10101
. generate x = runiform()                    // x ~ uniform(0,1)
```

```
. generate u = rnormal()                    // u ~ N(0,1)
. generate y = exp(1 + 0.25*x + 4*x^2) + u
. generate xsq = x^2
. regress y x xsq
      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  2,    47) =  168.27
       Model |  76293.9057     2  38146.9528            Prob > F      =  0.0000
    Residual |  10654.8492    47  226.698919            R-squared     =  0.8775
-------------+------------------------------           Adj R-squared =  0.8722
       Total |  86948.7549    49  1774.46438            Root MSE      =  15.057

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |  -228.8379   29.3865    -7.79   0.000    -287.9559   -169.7199
         xsq |   342.7992   28.71815   11.94   0.000     285.0258    400.5727
       _cons |   28.68793   6.605434    4.34   0.000     15.39951    41.97635
------------------------------------------------------------------------------
```

The misspecified model seems to fit the data very well with highly statistically significant regressors and an R^2 of 0.88.

Now consider a test for heteroskedasticity:

```
. * Test for heteroskedasticity
. estat hettest
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of y
         chi2(1)      =        22.70
         Prob > chi2  =    0.0000
```

This test strongly suggests that the errors are heteroskedastic because $p = 0.0000$, even though the DGP had homoskedastic errors.

The problem is that the regression function itself was misspecified. A RESET test yields

```
. * Test for misspecified conditional mean
. estat ovtest
Ramsey RESET test using powers of the fitted values of y
       Ho:  model has no omitted variables
                  F(3, 44) =   2702.16
                  Prob > F =      0.0000
```

This strongly rejects correct specification of the conditional mean because $p = 0.0000$.

Going the other way, could misspecification of other features of the model lead to rejection of the conditional mean, even though the conditional mean itself was correctly specified? This is an econometrically subtle question. The answer, in general, is yes. However, for the linear regression model, this is not the case essentially because consistency of the OLS estimator requires only that the conditional mean be correctly specified.

3.6 Prediction

For the linear regression model, the estimator of the conditional mean of y given $\mathbf{x} = \mathbf{x}_p$, $E(y|\mathbf{x}_p) = \mathbf{x}_p'\boldsymbol{\beta}$, is the conditional predictor $\widehat{y} = \mathbf{x}_p'\widehat{\boldsymbol{\beta}}$. We focus here on prediction for each observation in the sample. We begin with prediction from a linear model for medical expenditures, because this is straightforward, before turning to the log-linear model.

Further details on prediction are presented in section 3.7, where weighted average prediction is discussed, and in sections 10.5 and 10.6, where many methods are presented.

3.6.1 In-sample prediction

The most common type of prediction is in-sample, where evaluation is at the observed regressor values for each observation. Then $\widehat{y}_i = \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$ predicts $E(y_i|\mathbf{x}_i)$ for $i = 1, \ldots, N$.

To do this, we use `predict` after `regress`. The syntax for `predict` is

`predict` [*type*] *newvar* [*if*] [*in*] [, *options*]

The user always provides a name for the created variable, *newvar*. The default option is the prediction $\widehat{y}_i$. Other options yield residuals (usual, standardized, and studentized), several leverage and influential observation measures, predicted values, and associated standard errors of prediction. We have already used some of these options in section 3.5. The `predict` command can also be used for out-of-sample prediction. When used for in-sample prediction, it is good practice to add the `if e(sample)` qualifier, because this ensures that prediction is for the same sample as that used in estimation.

We consider prediction based on a linear regression model in levels rather than logs. We begin by reporting the regression results with `totexp` as the dependent variable.

```
. * Change dependent variable to level of positive medical expenditures
. use mus03data.dta, clear
. keep if totexp > 0
(109 observations deleted)
```

```
. regress totexp suppins phylim actlim totchr age female income, vce(robust)
Linear regression                               Number of obs     =       2955
                                                F(  7,  2947)     =      40.58
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1163
                                                Root MSE          =      11285
```

totexp	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
suppins	724.8632	427.3045	1.70	0.090	-112.9824	1562.709
phylim	2389.019	544.3493	4.39	0.000	1321.675	3456.362
actlim	3900.491	705.2244	5.53	0.000	2517.708	5283.273
totchr	1844.377	186.8938	9.87	0.000	1477.921	2210.832
age	-85.36264	37.81868	-2.26	0.024	-159.5163	-11.20892
female	-1383.29	432.4759	-3.20	0.001	-2231.275	-535.3044
income	6.46894	8.570658	0.75	0.450	-10.33614	23.27402
_cons	8358.954	2847.802	2.94	0.003	2775.07	13942.84

We then predict the level of medical expenditures:

```
. * Prediction in model linear in levels
. predict yhatlevels
(option xb assumed; fitted values)
  summarize totexp yhatlevels
```

Variable	Obs	Mean	Std. Dev.	Min	Max
totexp	2955	7290.235	11990.84	3	125610
yhatlevels	2955	7290.235	4089.624	-236.3781	22559

The summary statistics show that on average the predicted value `yhatlevels` equals the dependent variable. This suggests that the predictor does a good job. But this is misleading because this is always the case after OLS regression in a model with an intercept, since then residuals sum to zero implying $\sum y_i = \sum \widehat{y}_i$. The standard deviation of `yhatlevels` is \$4,090, so there is some variation in the predicted values.

For this example, a more discriminating test is to compare the median predicted and actual values. We have

```
* Compare median prediction and median actual value
tabstat totexp yhatlevels, stat (count p50) col(stat)
```

variable	N	p50
totexp	2955	3334
yhatlevels	2955	6464.692

There is considerable difference between the two, a consequence of the right-skewness of the original data, which the linear regression model does not capture.

The `stdp` option provides the standard error of the prediction, and the `stdf` option provides the standard error of the prediction for each sample observation, provided the

original estimation command used the default VCE. We therefore reestimate without vce(robust) and use predict to obtain

```
* Compute standard errors of prediction and forecast with default VCE
quietly regress totexp suppins phylim actlim totchr age female income
predict yhatstdp, stdp
predict yhatstdf, stdf
summarize yhatstdp yhatstdf
```

Variable	Obs	Mean	Std. Dev.	Min	Max
yhatstdp	2955	572.7	129.6575	393.5964	2813.983
yhatstdf	2955	11300.52	10.50946	11292.12	11630.8

The first quantity views $\mathbf{x}_i'\widehat{\boldsymbol{\beta}}$ as an estimate of the conditional mean $\mathbf{x}_i'\boldsymbol{\beta}$ and is quite precisely estimated because the average standard deviation is \$573 compared with an average prediction of \$7,290. The second quantity views $\mathbf{x}_i'\widehat{\boldsymbol{\beta}}$ as an estimate of the actual value y_i and is very imprecisely estimated because $y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$, and the error u_i here has relatively large variance since the levels equation has $s = 11285$.

More generally, microeconometric models predict poorly for a given individual, as evidenced by the typically low values of R^2 obtained from regression on cross-section data. These same models may nonetheless predict the conditional mean well, and it is this latter quantity that is needed for policy analysis that focuses on average behavior.

3.6.2 Marginal effects

The mfx postestimation command calculates MEs and elasticities evaluated at sample means, along with associated standard errors and confidence intervals where relevant. The default is to obtain these for the quantity that is the default for predict. For many estimation commands, including regress, this is the conditional mean. Then mfx computes for each continuous regressor $\partial E(y|\mathbf{x})/\partial x$, and for 0/1 indicator variables $\Delta E(y|\mathbf{x})$, evaluated at $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ and $\mathbf{x} = \bar{\mathbf{x}}$.

For the linear model, the estimated ME of the jth regressor is $\widehat{\beta}_j$, so there is no need to use mfx. But mfx can also be used to compute elasticities and semielasticities. For example, the eyex option computes the elasticity $\partial y/\partial x \times (x/y)$, evaluated at sample means, which equals $\widehat{\beta}_j \times (\bar{x}_j/\bar{y})$ for the linear model. We have

```
. * Compute elasticity for a specified regressor
. quietly regress totexp suppins phylim actlim totchr age female income,
> vce(robust)
. mfx, varlist(totchr) eyex
Elasticities after regress
      y  = Fitted values (predict)
           7290.2352
```

variable	ey/ex	Std. Err.	z	P>|z|	[95% C.I.]		X
totchr	.457613	.04481	10.21	0.000	.369793	.545433	1.8088

A 1% increase in chronic problems is associated with a 0.46% increase in medical expenditures. The `varlist(totchr)` option restricts results to just the regressor `totchr`.

The `predict()` option of `mfx` allows the computation of MEs for the other quantities that can be produced using `predict`.

3.6.3 Prediction in logs: The retransformation problem

Transforming the dependent variable by taking the natural logarithm complicates prediction. It is easy to predict $E(\ln y|\mathbf{x})$, but we are instead interested in $E(y|\mathbf{x})$ because we want to predict the level of medical expenditures rather than the natural logarithm. The obvious procedure of predicting $\ln y$ and taking the exponential is wrong because $\exp\{E(\ln y)\} \neq E(y)$, just as, for example, $\sqrt{E(y^2)} \neq E(y)$.

The log-linear model $\ln y = \mathbf{x}'\boldsymbol{\beta} + u$ implies that $y = \exp(\mathbf{x}'\boldsymbol{\beta})\exp(u)$. It follows that

$$E(y_i|\mathbf{x}_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta})E\{\exp(u_i)\}$$

The simplest prediction is $\exp(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})$, but this is wrong because it ignores the multiple $E\{\exp(u_i)\}$. If it is assumed that $u_i \sim N(0, \sigma^2)$, then it can be shown that $E\{\exp(u_i)\} = \exp(0.5\sigma^2)$, which can be estimated by $\exp(0.5\widehat{\sigma}^2)$, where $\widehat{\sigma}^2$ is an unbiased estimator of the log-linear regression model error. A weaker assumption is to assume that u_i is independent and identically distributed, in which case we can consistently estimate $E\{\exp(u_i)\}$ by the sample average $N^{-1}\sum_{j=1}^{N}\exp(\widehat{u}_j)$; see Duan (1983).

Applying these methods to the medical expenditure data yields

```
* Prediction in levels from a logarithmic model
quietly regress ltotexp suppins phylim actlim totchr age female income
quietly predict lyhat
generate yhatwrong = exp(lyhat)
generate yhatnormal = exp(lyhat)*exp(0.5*e(rmse)^2)
quietly predict uhat, residual
generate expuhat = exp(uhat)
quietly summarize expuhat
generate yhatduan = r(mean)*exp(lyhat)
summarize totexp yhatwrong yhatnormal yhatduan yhatlevels
```

Variable	Obs	Mean	Std. Dev.	Min	Max
totexp	2955	7290.235	11990.84	3	125610
yhatwrong	2955	4004.453	3303.555	959.5991	37726.22
yhatnormal	2955	8249.927	6805.945	1976.955	77723.13
yhatduan	2955	8005.522	6604.318	1918.387	75420.57
yhatlevels	2955	7290.235	4089.624	-236.3781	22559

Ignoring the retransformation bias leads to a very poor prediction, because `yhatwrong` has a mean of \$4,004 compared with the sample mean of \$7,290. The two alternative methods yield much closer average values of \$8,250 and \$8,006. Furthermore, the predictions from log regression, compared with those in levels, have the desirable fea-

ture of always being positive and have greater variability. The standard deviation of yhatnormal, for example, is \$6,806 compared with \$4,090 from the levels model.

3.6.4 Prediction exercise

There are several ways that predictions can be used to simulate the effects of a policy experiment. We consider the effect of a binary treatment, whether a person has supplementary insurance, on medical expenditure. Here we base our predictions on estimates that assume supplementary insurance is exogenous. A more thorough analysis could instead use methods that more realistically permit insurance to be endogenous. As we discuss in section 6.2.1, a variable is endogenous if it is related to the error term. Our analysis here assumes that supplementary insurance is not related to the error term.

An obvious comparison is to compare the difference in sample means $(\overline{y}_1 - \overline{y}_0)$, where the subscript 1 denotes those with supplementary insurance and the subscript 0 denotes those without supplementary insurance. This measure does not control for individual characteristics. A measure that does control for individual characteristics is the difference in mean predictions $(\overline{\widehat{y}}_1 - \overline{\widehat{y}}_0)$, where, for example, $\overline{\widehat{y}}_1$ denotes the average prediction for those with health insurance.

We implement the first two approaches for the complete sample based on OLS regression in levels and in logs. We obtain

```
. * Predicted effect of supplementary insurance: methods 1 and 2
. bysort suppins: summarize totexp yhatlevels yhatduan
```

```
-> suppins = 0
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      totexp |      1207    6824.303    11425.94         9     104823
  yhatlevels |      1207    6824.303    4077.064  -236.3781   20131.43
    yhatduan |      1207    6745.959    5365.255   1918.387   54981.73
```

```
-> suppins = 1
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      totexp |      1748    7611.963    12358.83         3     125610
  yhatlevels |      1748    7611.963    4068.397   502.9237      22559
    yhatduan |      1748    8875.255    7212.993   2518.538   75420.57
```

The average difference is \$788 (from $7612 - 6824$) using either the difference in sample means or the difference in fitted values from the linear model. Equality of the two is a consequence of OLS regression and prediction using the estimation sample. The log-linear model, using the prediction based on Duan's method, gives a larger average difference of \$2,129 (from $8875 - 6746$).

A third measure is the difference between the mean predictions, one with suppins set to 1 for all observations and one with suppins $= 0$. For the linear model, this is simply the estimated coefficient of suppins, which is \$725.

For the log-linear model, we need to make separate predictions for each individual with suppins set to 1 and with suppins set to 0. For simplicity, we make predictions in levels from the log-linear model assuming normally distributed errors. To make these changes and after the analysis have suppins returned to its original sample values, we use preserve and restore (see section 2.5.2). We obtain

```
. * Predicted effect of supplementary insurance: method 3 for log-linear model
. quietly regress ltotexp suppins phylim actlim totchr age female income
. preserve
. quietly replace suppins = 1
. quietly predict lyhat1
. generate yhatnormal1 = exp(lyhat1)*exp(0.5*e(rmse)^2)
. quietly replace suppins = 0
. quietly predict lyhat0
. generate yhatnormal0 = exp(lyhat0)*exp(0.5*e(rmse)^2)
. generate treateffect = yhatnormal1 - yhatnormal0
. summarize yhatnormal1 yhatnormal0 treateffect
```

Variable	Obs	Mean	Std. Dev.	Min	Max
yhatnormal1	2955	9077.072	7313.963	2552.825	77723.13
yhatnormal0	2955	7029.453	5664.069	1976.955	60190.23
treateffect	2955	2047.619	1649.894	575.8701	17532.91

```
. restore
```

While the average treatment effect of $2,048 is considerably larger than that obtained by using the difference in sample means of the linear model, it is comparable to the estimate produced by Duan's method.

3.7 Sampling weights

The analysis to date has presumed simple random sampling, where sample observations have been drawn from the population with equal probability. In practice, however, many microeconometric studies use data from surveys that are not representative of the population. Instead, groups of key interest to policy makers that would have too few observations in a purely random sample are oversampled, with other groups then undersampled. Examples are individuals from racial minorities or those with low income or living in sparsely populated states.

As explained below, weights should be used for estimation of population means and for postregression prediction and computation of MEs. However, in most cases, the regression itself can be fitted without weights, as is the norm in microeconometrics. If weighted analysis is desired, it can be done using standard commands with a weighting option, which is the approach of this section and the standard approach in microeconometrics. Alternatively, one can use survey commands as detailed in section 5.5.

3.7.1 Weights

Sampling weights are provided by most survey datasets. These are called probability weights or pweights in Stata, though some others call them inverse-probability weights because they are inversely proportional to the probability of inclusion of the sample. A pweight of 1,400 in a survey of the U.S. population, for example, means that the observation is representative of 1,400 U.S. residents and the probability of this observation being included in the sample is 1/1400.

Most estimation commands allow probability weighted estimators that are obtained by adding [pweight=*weight*], where *weight* is the name of the weighting variable.

To illustrate the use of sampling weights, we create an artificial weighting variable (sampling weights are available for the MEPS data but were not included in the data extract used in this chapter). We manufacture weights that increase the weight given to those with more chronic problems. In practice, such weights might arise if the original sampling framework oversampled people with few chronic problems and undersampled people with many chronic problems. In this section, we analyze levels of expenditures, including expenditures of zero. Specifically,

```
* Create artificial sampling weights
use mus03data.dta, clear
generate swght = totchr^2 + 0.5
summarize swght
```

Variable	Obs	Mean	Std. Dev.	Min	Max
swght	3064	5.285574	6.029423	.5	49.5

What matters in subsequent analysis is the relative values of the sampling weights rather than the absolute values. The sampling weight variable swght takes on values from 0.5 to 49.5, so weighted analysis will give some observations as much as 49.5/0.5 = 99 times the weight given to others.

Stata offers three other types of weights that for most analyses can be ignored. Analytical weights, called aweights, are used for the quite different purpose of compensating for different observations having different variances that are known up to scale; see section 5.3.4. For duplicated observations, fweights provide the number of duplicated observations. So-called importance weights, or iweights, are sometimes used in more advanced programming.

3.7.2 Weighted mean

If an estimate of a population mean is desired, then we should clearly weight. In this example, by oversampling those with few chronic problems, we will have oversampled people who on average have low medical expenditures, so that the unweighted sample mean will understate population mean medical expenditures.

Let w_i be the population weight for individual i. Then, by defining $W = \sum_{i=1}^{N} w_i$ to be the sum of the weights, the weighted mean $\overline{y}_W$ is

$$\overline{y}_W = \frac{1}{W}\sum_{i=1}^{N} w_i y_i$$

with variance estimator (assuming independent observations) $\widehat{V}(\overline{y}_W) = \{1/W(W-1)\}$ $\sum_{i=1}^{N} w_i(y_i - \overline{y}_W)^2$. These formulas reduce to those for the unweighted mean if equal weights are used.

The weighted mean downweights oversampled observations because they will have a value of pweights (and hence w_i) that is smaller than that for most observations. We have

```
. * Calculate the weighted mean
. mean totexp [pweight=swght]
Mean estimation                   Number of obs   =      3064
```

	Mean	Std. Err.	[95% Conf.	Interval]
totexp	10670.83	428.5148	9830.62	11511.03

The weighted mean of $10,671 is much larger than the unweighted mean of $7,031 (see section 3.2.4) because the unweighted mean does not adjust for the oversampling of individuals with few chronic problems.

3.7.3 Weighted regression

The weighted least-squares estimator for the regression of y_i on $\mathbf{x}_i$ with the weights w_i is given by

$$\widehat{\boldsymbol{\beta}}_W = \left(\sum_{i=1}^{N} w_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i=1}^{N} w_i \mathbf{x}_i y_i$$

The OLS estimator is the special case of equal weights with $w_i = w_j$ for all i and j. The default estimator of the VCE is a weighted version of the heteroskedasticity-robust version in (3.3), which assumes independent observations. If observations are clustered, then the option vce(cluster *clustvar*) should be used.

Although the weighted estimator is easily obtained, for legitimate reasons many microeconometric analyses do not use weighted regression even where sampling weights are available. We provide a brief explanation of this conceptually difficult issue. For a more complete discussion, see Cameron and Trivedi (2005, 818–821).

Weighted regression should be used if a census parameter estimate is desired. For example, suppose we want to obtain an estimate for the U.S. population of the average change in earnings associated with one more year of schooling. Then, if disadvantaged minorities are oversampled, we most likely will understate the earnings increase, because

disadvantaged minorities are likely to have earnings that are lower than average for their given level of schooling. A second example is when aggregate state-level data are used in a natural experiment setting, where the goal is to measure the effect of an exogenous policy change that affects some states and not other states. Intuitively, the impact on more populous states should be given more weight. Note that these estimates are being given a correlative rather than a causal interpretation.

Weighted regression is not needed if we make the stronger assumptions that the DGP is the specified model $y_i = \mathbf{x}_i'\beta + u_i$ and sufficient controls are assumed to be added so that the error $E(u_i|\mathbf{x}_i) = 0$. This approach, called a control-function approach or a model approach, is the approach usually taken in microeconometric studies that emphasize a causal interpretation of regression. Under the assumption that $E(u_i|\mathbf{x}_i) = 0$, the weighted least-squares estimator will be consistent for β for any choice of weights including equal weights, and if u_i is homoskedastic, the most efficient estimator is the OLS estimator, which uses equal weights. For the assumption that $E(u_i|\mathbf{x}_i) = 0$ to be reasonable, the determinants of the sampling frame should be included in the controls $\mathbf{x}$ and should not be directly determined by the dependent variable y.

These points carry over directly to nonlinear regression models. In most cases, microeconometric analyses take on a model approach. In that case, unweighted estimation is appropriate, with any weighting based on efficiency grounds. If a census-parameter approach is being taken, however, then it is necessary to weight.

For our data example, we obtain

```
. * Perform weighted regression
. regress totexp suppins phylim actlim totchr age female income [pweight=swgbt]
(sum of wgt is   1.6195e+04)
Linear regression                               Number of obs =    3064
                                                F(  7,  3056) =   14.08
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0977
                                                Root MSE      =   13824
```

totexp	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
suppins	278.1578	825.6959	0.34	0.736	-1340.818	1897.133
phylim	2484.52	933.7116	2.66	0.008	653.7541	4315.286
actlim	4271.154	1024.686	4.17	0.000	2262.011	6280.296
totchr	1819.929	349.2234	5.21	0.000	1135.193	2504.666
age	-59.3125	68.01237	-0.87	0.383	-192.6671	74.04212
female	-2654.432	911.6422	-2.91	0.004	-4441.926	-866.9381
income	5.042348	16.6509	0.30	0.762	-27.60575	37.69045
_cons	7336.758	5263.377	1.39	0.163	-2983.359	17656.87

The estimated coefficients of all statistically significant variables aside from female are within 10% of those from unweighted regression (not given for brevity). Big differences between weighted and unweighted regression would indicate that $E(u_i|\mathbf{x}_i) \neq 0$ because of model misspecification. Note that robust standard errors are reported by default.

3.7.4 Weighted prediction and MEs

After regression, unweighted prediction will provide an estimate of the sample-average value of the dependent variable. We may instead want to estimate the population-mean value of the dependent variable. Then sampling weights should be used in forming an average prediction.

This point is particularly easy to see for OLS regression. Because $1/N \sum_i (y_i - \widehat{y}_i) = 0$, since in-sample residuals sum to zero if an intercept is included, the average prediction $1/N \sum_i \widehat{y}_i$ equals the sample mean $\overline{y}$. But given an unrepresentative sample, the unweighted sample mean $\overline{y}$ may be a poor estimate of the population mean. Instead, we should use the weighted average prediction $1/N \sum_i w_i \widehat{y}_i$, even if $\widehat{y}_i$ is obtained by using unweighted regression.

For this to be useful, however, the prediction should be based on a model that includes as regressors variables that control for the unrepresentative sampling.

For our example, we obtain the weighted prediction by typing

```
* Weighted prediction
quietly predict yhatwols
mean yhatwols [pweight=swght], noheader
```

	Mean	Std. Err.	[95% Conf.	Interval]
yhatwols	10670.83	138.0828	10400.08	10941.57

```
mean yhatwols, noheader      // unweighted prediction
```

	Mean	Std. Err.	[95% Conf.	Interval]
yhatwols	7135.206	78.57376	6981.144	7289.269

The population mean for medical expenditures is predicted to be \$10,671 using weighted prediction, whereas the unweighted prediction gives a much lower value of \$7,135.

Weights similarly should be used in computing average MEs. For the linear model, the standard ME $\partial E(y_i|\mathbf{x}_i)/\partial x_{ij}$ equals β_j for all observations, so weighting will make no difference in computing the marginal effect. Weighting will make a difference for averages of other marginal effects, such as elasticities, and for MEs in nonlinear models.

3.8 OLS using Mata

Stata offers two different ways to perform computations using matrices: Stata `matrix` commands and Mata functions (which are discussed, respectively, in appendices A and B).

Mata, introduced in Stata 9, is much richer. We illustrate the use of Mata by using the same OLS regression as that in section 3.4.2.

The program is written for the dependent variable provided in the local macro y and the regressors in the local macro xlist. We begin by reading in the data and defining the local macros.

```
. * OLS with White robust standard errors using Mata
. use mus03data.dta, clear
. keep if totexp > 0   // Analysis for positive medical expenditures only
(109 observations deleted)
. generate cons = 1
. local y ltotexp
. local xlist suppins phylim actlim totchr age female income cons
```

We then move into Mata. The st_view() Mata function is used to transfer the Stata data variables to Mata matrices y and X, with tokens("") added to convert `xlist´ to a comma-separated list with each entry in double quotes, necessary for st_view().

The key part of the program forms $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $\widehat{V}(\widehat{\boldsymbol{\beta}}) = (N/N-K)(\mathbf{X}'\mathbf{X})^{-1}(\sum_i \widehat{u}_i^2\mathbf{x}_i\mathbf{x}_i')(\mathbf{X}'\mathbf{X})^{-1}$. The cross-product function cross(X,X) is used to form $\mathbf{X}'\mathbf{X}$ because this handles missing values and is more efficient than the more obvious X'X. The matrix inverse is formed by using cholinv() because this is the fastest method in the special case that the matrix is symmetric positive definite. We calculate the $K \times K$ matrix $\sum_i \widehat{u}_i^2\mathbf{x}_i\mathbf{x}_i'$ as $\sum_i(\widehat{u}_i\mathbf{x}_i')'(\widehat{u}_i\mathbf{x}_i') = \mathbf{A}'\mathbf{A}$, where the $N \times K$ matrix $\mathbf{A}$ has an ith row equal to $\widehat{u}_i\mathbf{x}_i'$. Now $\widehat{u}_i\mathbf{x}_i'$ equals the ith row of the $N \times 1$ residual vector $\widehat{\mathbf{u}}$ times the ith row of the $N \times K$ regressor matrix $\mathbf{X}$, so $\mathbf{A}$ can be computed by element-by-element multiplication of $\widehat{\mathbf{u}}$ by $\mathbf{X}$, or (e:*X), where e is $\widehat{\mathbf{u}}$. Alternatively, $\sum_i \widehat{u}_i^2\mathbf{x}_i\mathbf{x}_i' = \mathbf{X}'\mathbf{D}\mathbf{X}$, where $\mathbf{D}$ is an $N \times N$ diagonal matrix with entries $\widehat{u}_i^2$, but the matrix $\mathbf{D}$ becomes exceptionally large, unnecessarily so, for a large N.

The Mata program concludes by using st_matrix() to pass the estimated $\widehat{\boldsymbol{\beta}}$ and $\widehat{V}(\widehat{\boldsymbol{\beta}})$ back to Stata.

```
. mata
------------------------------------------------- mata (type end to exit) -----
:   // Create y vector and X matrix from Stata dataset
:   st_view(y=., ., "`y´")              // y is nx1
:   st_view(X=., ., tokens("`xlist´")) // X is nxk
:   XXinv = cholinv(cross(X,X))        // XXinv is inverse of X´X
:   b = XXinv*cross(X,y)               // b = [(X´X)^-1]*X´y
:   e = y - X*b
:   n = rows(X)
:   k = cols(X)
:   s2 = (e´e)/(n-k)
:   vdef = s2*XXinv                 // default VCE not used here
:   vwhite = XXinv*((e:*X)´(e:*X)*n/(n-k))*XXinv  // robust VCE
:   st_matrix("b",b´)               // pass results from Mata to Stata
:   st_matrix("V",vwhite)           // pass results from Mata to Stata
: end
```

Once back in Stata, we use ereturn to display the results in a format similar to that for built-in commands, first assigning names to the columns and rows of b and V.

```
. * Use Stata ereturn display to present nicely formatted results
. matrix colnames b = `xlist'
. matrix colnames V = `xlist'
. matrix rownames V = `xlist'
. ereturn post b V
. ereturn display
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
suppins	.2556428	.0465982	5.49	0.000	.1643119	.3469736
phylim	.3020598	.057705	5.23	0.000	.18896	.4151595
actlim	.3560054	.0634066	5.61	0.000	.2317308	.48028
totchr	.3758201	.0187185	20.08	0.000	.3391326	.4125077
age	.0038016	.0037028	1.03	0.305	-.0034558	.011059
female	-.0843275	.045654	-1.85	0.065	-.1738076	.0051526
income	.0025498	.0010468	2.44	0.015	.0004981	.0046015
cons	6.703737	.2825751	23.72	0.000	6.1499	7.257575

The results are exactly the same as those given in section 3.4.2, when we used regress with the vce(robust) option.

3.9 Stata resources

The key Stata references are [U] *User's Guide* and [R] **regress**, [R] **regress postestimation**, [R] **estimates**, [R] **predict**, and [R] **test**. A useful user-written command is estout. The material in this chapter appears in many econometrics texts, such as Greene (2008).

3.10 Exercises

1. Fit the model in section 3.4 using only the first 100 observations. Compute standard errors in three ways: default, heteroskedastic, and cluster–robust where clustering is on the number of chronic problems. Use estimates to produce a table with three sets of coefficients and standard errors, and comment on any appreciable differences in the standard errors. Construct a similar table for three alternative sets of heteroskedasticity-robust standard errors, obtained by using the vce(robust), vce(hc2), and vce(hc3) options, and comment on any differences between the different estimates of the standard errors.

2. Fit the model in section 3.4 with robust standard errors reported. Test at 5% the joint significance of the demographic variables age, female, and income. Test the hypothesis that being male (rather than female) has the same impact on medical expenditures as aging 10 years. Fit the model under the constraint that $\beta_{\texttt{phylim}} = \beta_{\texttt{actlim}}$ by first typing constraint 1 phylim = actlim and then by using cnsreg with the constraints(1) option.

3. Fit the model in section 3.5, and implement the RESET test manually by regressing y on x and $\widehat{y}^2$, $\widehat{y}^3$, and $\widehat{y}^4$ and jointly testing that the coefficients of $\widehat{y}^2$, $\widehat{y}^3$, and $\widehat{y}^4$ are zero. To get the same results as `estat ovtest`, do you need to use default or robust estimates of the VCE in this regression? Comment. Similarly, implement `linktest` by regressing y on $\widehat{y}$ and $\widehat{y}^2$ and testing that the coefficient of $\widehat{y}^2$ is zero. To get the same results as `linktest`, do you need to use default or robust estimates of the VCE in this regression? Comment.

4. Fit the model in section 3.5, and perform the standard Lagrange multiplier test for heteroskedasticity by using `estat hettest` with z = x. Then implement the test manually as 0.5 times the explained sum of squares from the regression of y_i^* on an intercept and $\mathbf{z}_i$, where $y_i^* = \{\widehat{u}_i^2/(1/N)\sum_j \widehat{u}_j^2\} - 1$ and $\widehat{u}_i$ is the residual from the original OLS regression. Next use `estat hettest` with the `iid` option and show that this test is obtained as $N \times R^2$, where R^2 is obtained from the regression of $\widehat{u}_i^2$ on an intercept and $\mathbf{z}_i$.

5. Fit the model in section 3.6 on levels, except use all observations rather than those with just positive expenditures, and report robust standard errors. Predict medical expenditures. Use `correlate` to obtain the correlation coefficient between the actual and fitted value and show that, upon squaring, this equals R^2. Show that the linear model `mfx` without options reproduces the OLS coefficients. Now use `mfx` with an appropriate option to obtain the income elasticity of medical expenditures evaluated at sample means.

6. Fit the model in section 3.6 on levels, using the first 2,000 observations. Use these estimates to predict medical expenditures for the remaining 1,064 observations, and compare these with the actual values. Note that the model predicts very poorly in part because the data were ordered by `totexp`.