5 Specifying the functional form

5.1 Introduction

A key assumption maintained in the previous chapter is that the functional form was correctly specified. Here we discuss some methods for checking the validity of this assumption. If the *zero-conditional-mean* assumption

$$E[u \mid x_1, x_2, \ldots, x_k] = 0 \tag{5.1}$$

is violated, the coefficient estimates are inconsistent.

The three main problems that cause the zero-conditional-mean assumption to fail in a regression model are

- improper specification of the model;
- endogeneity of one or more regressors; or
- measurement error of one or more regressors.

The *specification* of a regression model may be flawed in its list of included regressors or in the functional form specified for the estimated relationship. *Endogeneity* means that one or more regressors may be correlated with the error term, a condition that often arises when those regressors are simultaneously determined with the response variable. *Measurement error* of a regressor implies that the underlying behavioral relationship includes one or more variables that the econometrician cannot accurately measure. This chapter discusses specification issues, whereas chapter 8 addresses endogeneity and measurement errors.

5.2 Specification error

The consistency of the linear regression estimator requires that the sample regression function correspond to the underlying population regression function or true model for the response variable y:

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

Specifying a regression model often involves making a sequence of decisions about the model's contents. Economic theory often provides some guidance in model specification but may not explicitly indicate how a specific variable should enter the model, identify

the functional form, or spell out precisely how the stochastic elements (u_i) enter the model. Comparative static results that provide expected signs for derivatives do not indicate which functional specification to use for the model. Should it be estimated in levels; as a log-linear structure; as a polynomial in one or more of the regressors; or in logarithms, implying a constant-elasticity relationship? Theory is often silent on such specifics, yet we must choose a specific functional form to proceed with empirical research.[1]

Let us assume that the empirical specification may differ from the population regression function in one of two ways (which both might be encountered in the same fitted model). Given the dependent variable y, we may omit relevant variables from the model, or we may include irrelevant variables in the model, making the fitted model "short" or "long", respectively, relative to the true model.

5.2.1 Omitting relevant variables from the model

Suppose that the true model is

$$y = \mathbf{x}_1\boldsymbol{\beta}_1 + \mathbf{x}_2\boldsymbol{\beta}_2 + u$$

with k_1 and k_2 regressors in the two subsets, but that we regress y on just the $\mathbf{x}_1$ variables:

$$y = \mathbf{x}_1\boldsymbol{\beta}_1 + u$$

This step yields the least-squares solution

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_1 &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} && (5.2) \\ &= \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{u} && (5.3) \end{aligned}$$

Unless $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$ or $\boldsymbol{\beta}_2 = 0$, the estimate $\widehat{\boldsymbol{\beta}}_1$ is biased, since

$$E[\widehat{\boldsymbol{\beta}}_1|\mathbf{X}] = \boldsymbol{\beta}_1 + \mathbf{P}_{1\cdot 2}\,\boldsymbol{\beta}_2$$

where $\mathbf{P}_{1\cdot 2} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2$ is the $k_1 \times k_2$ matrix reflecting the regression of each column of $\mathbf{X}_2$ on the columns of $\mathbf{X}_1$. If $k_1 = k_2 = 1$ and the single variable in $\mathbf{X}_2$ is correlated with the single variable in $\mathbf{X}_1$, we can derive the direction of bias. Generally, with multiple variables in each set, we can make no statements about the nature of the bias of the $\widehat{\boldsymbol{\beta}}_1$ coefficients.

We may conclude that the cost of omitting relevant variables is high. If $E[\mathbf{x}_1'\mathbf{x}_2] \neq \mathbf{0}$, (5.3) would have showed that the estimator was inconsistent. If the population correlations between elements of $\mathbf{x}_1$ and $\mathbf{x}_2$ are zero, regression estimates would be consistent but probably biased in finite samples. In economic research, a variable mistakenly excluded from a model is unlikely to be uncorrelated in the population or in the sample with the regressors.

1. This requirement holds unless one chooses to use nonparametric methods that are beyond the scope of this book. See Hardle (1990) for an introduction to nonparametric methods.

Specifying dynamics in time-series regression models

A related concern arises in models for time-series data, in which theory rarely fully specifies the *time form* of the dynamic relationship. For instance, consumer theory may specify the ultimate response of an individual's saving to a change in her after-tax income. However, theory may fail to indicate how rapidly the individual will adjust her saving to a permanent change in her salary. Will that adjustment take place within one, two, three, or more biweekly pay periods? From our analysis of the asymmetry of specification error, we know that the advice to the modeler should be "do not underfit the dynamics." If we do not know the time form of a dynamic relationship with certainty, we should include several lagged values of the regressor. We can then use the "test down" strategy discussed below to determine whether the longer lags are necessary. Moreover, omitting higher-order dynamic terms may cause apparent nonindependence of the regression errors, as signaled by residual independence tests.

5.2.2 Graphically analyzing regression data

With Stata's graphics, you can easily perform exploratory data analysis on the regression sample, even with large datasets. In specification analysis, we may want to examine the simple bivariate relationships between y and the regressors in `x`. Although multiple linear regression coefficients are complicated functions of the various bivariate regression coefficients among these variables, we still often find it useful to examine a set of bivariate plots. We use `graph matrix` to generate a set of plots illustrating the bivariate relationships underlying our regression model of median housing prices:

(*Continued on next page*)

```
. graph matrix lprice lnox ldist rooms stratio, ms(Oh) msize(tiny)
```

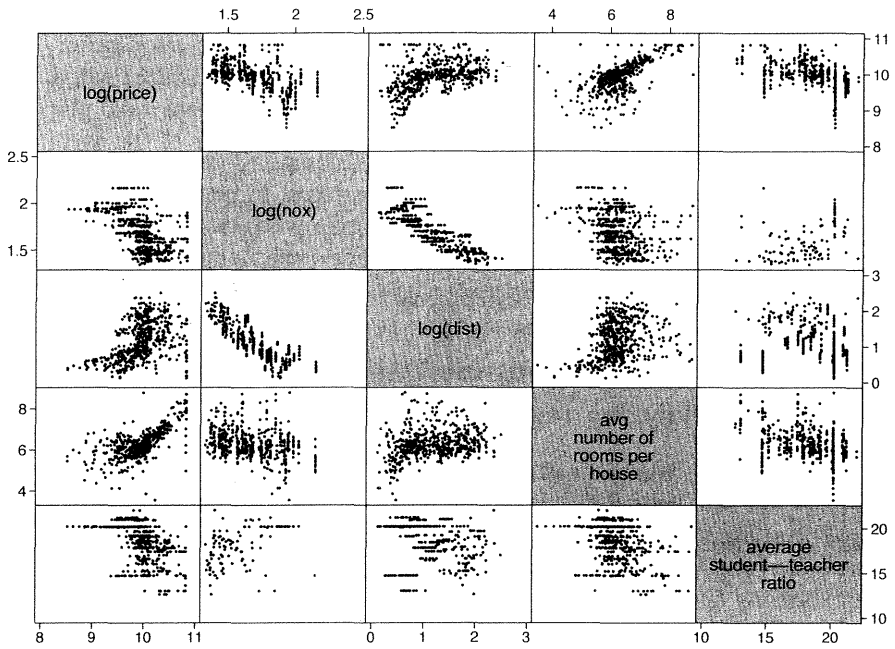

Figure 5.1: `graph matrix` of regression variables

The first row (or column) of the plot matrix in figure 5.1 illustrates the relationships between the variable to be explained (the log of median housing price) and the four causal factors. These plots are the $y - x$ planes, in which a simple regression of log housing price on each of these factors in turn would determine the line of best fit.

The other bivariate graphs below the main diagonal are also illustrative. If any of these relationships could be fitted well by a straight line, the intercorrelation among those regressors would be high, and we would expect collinearity problems. For instance, the scatter of points between `lnox` and `ldist` appears to be compact and linear. The `correlate` command shows that those two variables have a simple correlation of -0.86.

5.2.3 Added-variable plots

The added-variable plot identifies the important variables in a relationship by decomposing the multivariate relationship into a set of two-dimensional plots.[2] Taking each regressor in turn, the added-variable plot is based on two residual series. The first series, e_1, contains the residuals from the regression of $\mathbf{x}_g$ on all other $\mathbf{x}$, whereas the second series, e_2, contains the residuals from the regression of y on all x variables except $\mathbf{x}_g$. That is, e_1 represents the part of $\mathbf{x}_g$ that cannot be linearly related to those other regressors, whereas e_2 represents the information in y that is not explained by all other regressors (excluding $\mathbf{x}_g$). The added-variable plot for $\mathbf{x}_g$ is then the scatter of e_2 (on the y-axis) versus e_1 (on the x-axis). Two polar cases (as discussed by Cook and Weisberg [1994, 194]) are of interest. If most points are clustered around a horizontal line at ordinate zero in the added-variable plot, $\mathbf{x}_g$ is irrelevant. On the other hand, if most points are clustered around a vertical line with abscissa zero, the plot would indicate near-perfect collinearity. Here as well, adding $\mathbf{x}_g$ to the model would not be helpful.

The strength of a linear relationship between e_1 and e_2 (that is, the slope of a least-squares line through this scatter of points) represents the marginal value of $\mathbf{x}_g$ in the full model. If the slope is significantly different from zero, $\mathbf{x}_g$ makes an important contribution to the model beyond that of the other regressors. The more closely the points are grouped around a straight line in the plot, the more important is the contribution of $\mathbf{x}_g$ at the margin. As an added check, if the specification of the full model (including $\mathbf{x}_g$) is correct, the plot of e_1 versus e_2 must exhibit linearity. Significant departures from linearity in the plot cast doubt on the appropriate specification of $\mathbf{x}_g$ in the model.

After `regress`, the command to generate an added-variable plot is given as

`avplot` *varname*

where *varname* is the variable on which the plot is based, which can be a regressor or a variable not included in the regression model. Alternatively,

`avplots`

produces one graph with all added-variable plots from the last regression, as we now illustrate.

2. For details about the added-variable plot, see Cook and Weisberg (1994, 191–194). See [R] **regress postestimation** for more details about its implementation in Stata.

```
. use http://www.stata-press.com/data/imeus/hprice2a, clear
(Housing price data for Boston-area communities)
. generate rooms2 = rooms^2
. regress lprice lnox ldist rooms rooms2 stratio lproptax
```

Source	SS	df	MS	Number of obs =	506
				F(6, 499) =	138.41
Model	52.8357813	6	8.80596356	Prob > F =	0.0000
Residual	31.7464896	499	.06362022	R-squared =	0.6247
				Adj R-squared =	0.6202
Total	84.5822709	505	.167489645	Root MSE =	.25223

lprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnox	-.6615694	.1201606	-5.51	0.000	-.8976524	-.4254864
ldist	-.095087	.0421435	-2.26	0.024	-.1778875	-.0122864
rooms	-.5625662	.1610315	-3.49	0.001	-.8789496	-.2461829
rooms2	.0634347	.0124621	5.09	0.000	.0389501	.0879193
stratio	-.0362928	.0060699	-5.98	0.000	-.0482185	-.0243671
lproptax	-.2211125	.0410202	-5.39	0.000	-.301706	-.1405189
_cons	14.15454	.5693846	24.86	0.000	13.03585	15.27323

```
. avplots, ms(Oh) msize(small) col(2)
```

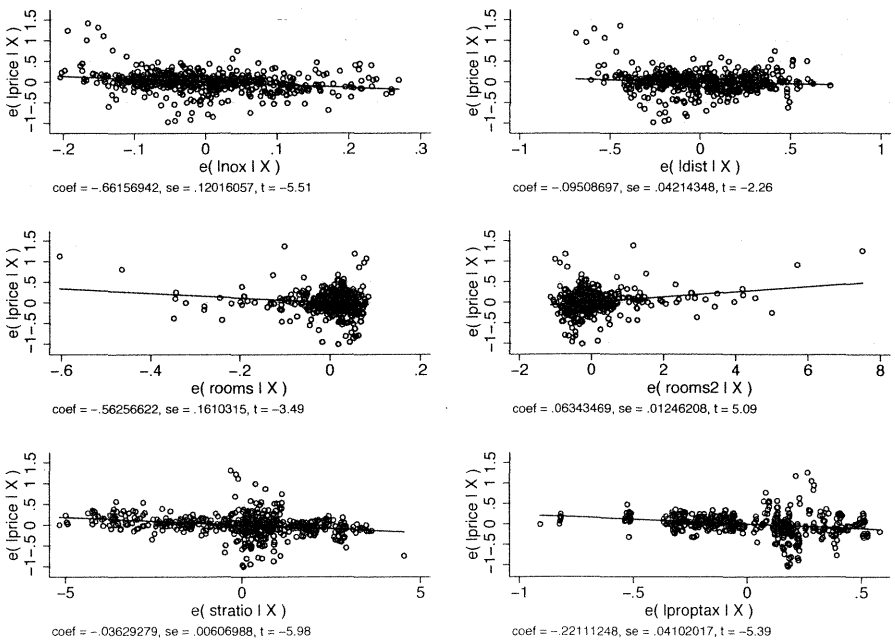

Figure 5.2: Added-variable plots

In each pane of figure 5.2, we see several observations that are far from the straight line linking the response variable and that regressor. The outlying values are particularly evident in the graphs for `lnox` and `ldist`, where low values of $E[\texttt{lnox}|X]$ and $E[\texttt{ldist}|X]$ are associated with prices much higher than those predicted by the model. The t statistics shown in each panel test the hypothesis that the least-squares line has a slope significantly different from zero. These test statistics are identical to those of the original regression, shown above.

5.2.4 Including irrelevant variables in the model

Including irrelevant regressors does not violate the zero-conditional-mean assumption. Since their population coefficients are zero, including them in the regressor list does not cause the conditional mean of the u process to differ from zero. Suppose that the true model is

$$y = \mathbf{x}_1\boldsymbol{\beta}_1 + u \tag{5.4}$$

but we mistakenly include several $\mathbf{x}_2$ variables in our regression model. In that case, we fail to impose the restrictions that $\boldsymbol{\beta}_2 = 0$. Since $\boldsymbol{\beta}_2 = 0$ in the population, including $\mathbf{x}_2$ leaves our estimates of $\boldsymbol{\beta}_1$ unbiased and consistent, as is the estimate of σ_u^2. Overfitting the model and including the additional variables causes a loss of *efficiency* (see section 4.2.3). By ignoring that the $\mathbf{x}_2$ variables do not belong in the model, our estimates of $\boldsymbol{\beta}_1$ are less precise than they would be with the correctly specified model, and the estimated standard errors of $\boldsymbol{\beta}_1$ will be larger than those fitted from the correct model of (5.4). This property is especially apparent if we have $k_1 = k_2 = 1$ and the correlation between x_1 and x_2 is high. Mistakenly including $\mathbf{x}_2$ will lead to imprecise estimates of $\boldsymbol{\beta}_1$.

Clearly, overfitting the model costs much less than underfitting, as discussed earlier. The long model delivers unbiased and consistent estimates of all its parameters, including those of the irrelevant regressors, which tend to zero.

5.2.5 The asymmetry of specification error

We may conclude that the costs of these two types of specification error are asymmetric. We would much rather err on the side of caution (including additional variables) to avoid the inconsistent estimates that would result from underfitting the model. Given this conclusion, a model selection strategy that starts with a simple specification and seeks to refine it by adding variables is flawed. The opposite approach, starting with a general specification and seeking to refine it by imposing appropriate restrictions, has much more to recommend it.[3] Although a general specification may be plagued by collinearity, a recursive simplification strategy is much more likely to yield a usable model at the end of the specification search. Ideally, we would not need to search for a specification. We would merely write down the fitted model that theory propounds, run one regression,

3. The *general-to-specific* approach proposed by econometrician David Hendry in several of his works implements such a refinement strategy. See http://ideas.repec.org/e/phe33.html for more information.

and evaluate its results. Unfortunately, most applied work is not that straightforward. Most empirical investigations contain some amount of specification searching.

In considering such a research strategy, we also must be aware of the limits of statistical inference. We might run 20 regressions in which the regressors do not appear in the true model, but at the 5% level, we would expect one of those 20 estimates to erroneously show a relationship between the response variable and regressors. Many articles in the economic literature decry "data mining" or "fishing for results". The rationale for fitting a variety of models in search of the true model is to avoid using statistical inference to erroneously reject a theory because we have misspecified the relationship. If we write down one model that bears little resemblance to the true model, fit that model, and conclude that the data reject the theory, we are resting our judgment on the *maintained hypothesis* that we have correctly specified the population model. But if we used a transformation of that model, or added omitted variables to the model, our inference might reach a different conclusion.

5.2.6 Misspecification of the functional form

A model that includes the appropriate regressors may be misspecified because the model may not reflect the algebraic form of the relationship between the response variable and those regressors. For instance, suppose that the true model specifies a nonlinear relationship between y and x_j—such as a polynomial relationship—and we omit the squared term.[4] Doing so would be misspecifying the functional form. Likewise, if the true model expresses a constant-elasticity relationship, the model fitted to logarithms of y and $\mathbf{x}$ could render conclusions different from those of a model fitted to levels of the variables. In one sense, this problem may be easier to deal with than the omission of relevant variables. In a misspecification of the functional form, we have all the appropriate variables at hand and only have to choose the appropriate form in which they enter the regression function. Ramsey's omitted-variable regression specification error test (RESET) implemented by Stata's `estat ovtest` may be useful in this context.

5.2.7 Ramsey's RESET

Linear regression of y on the levels of various x's restricts the effects of each x_j to be strictly linear. If the functional relationship linking y to x_j is nonlinear, a linear function may work reasonably well for some values of x_j but will eventually break down. Ramsey's RESET is based on this simple notion. RESET runs an augmented regression that includes the original regressors, powers of the predicted values from the original regression, and powers of the original regressors. Under the null hypothesis of no misspecification, the coefficients on these additional regressors are zero. RESET is simply a Wald test of this null hypothesis. The test works well because polynomials in $\widehat{y}$ and x_j can approximate a variety of functional relationships between y and the regressors $\mathbf{x}$.

4. Distinguish between a model linear in the *parameters* and a nonlinear relationship between y and x. $y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + u_i$ is linear in the $\boldsymbol{\beta}$ parameters but defines a nonlinear function, $E[y|x] = f(x)$.

As discussed in [R] **regress postestimation**, to compute the RESET after `regress`, we use the following command syntax:[5]

```
estat ovtest [ , rhs ]
```

The parsimonious flavor of the test, computed by default, augments the regression with the second, third, and fourth powers of the $\widehat{y}$ series. With the `rhs` option, powers of the individual regressors themselves are used. This option may considerably reduce the power of the test in small samples because it will include many regressors. For example, if we perform RESET after our regression model of log housing prices,

```
. quietly regress lprice lnox ldist rooms stratio
. estat ovtest
Ramsey RESET test using powers of the fitted values of lprice
       Ho:  model has no omitted variables
                 F(3, 498) =      9.69
                  Prob > F =      0.0000
. estat ovtest, rhs
Ramsey RESET test using powers of the independent variables
       Ho:  model has no omitted variables
                F(12, 489) =     11.79
                  Prob > F =      0.0000
```

we can reject RESET's null hypothesis of no omitted variables for the model using either form of the test. We respecify the equation to include the square of `rooms` and include another factor, `lproptax`, the log of property taxes in the community:

```
. regress lprice lnox ldist rooms rooms2 stratio lproptax
      Source |       SS       df       MS              Number of obs =     506
-------------+------------------------------           F(  6,   499) =  138.41
       Model |  52.8357813     6  8.80596356            Prob > F      =  0.0000
    Residual |  31.7464896   499   .06362022            R-squared     =  0.6247
-------------+------------------------------           Adj R-squared =  0.6202
       Total |  84.5822709   505  .167489645            Root MSE      =  .25223

------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lnox |  -.6615694   .1201606    -5.51   0.000    -.8976524   -.4254864
       ldist |   -.095087   .0421435    -2.26   0.024    -.1778875   -.0122864
       rooms |  -.5625662   .1610315    -3.49   0.001    -.8789496   -.2461829
      rooms2 |   .0634347   .0124621     5.09   0.000     .0389501    .0879193
     stratio |  -.0362928   .0060699    -5.98   0.000    -.0482185   -.0243671
    lproptax |  -.2211125   .0410202    -5.39   0.000     -.301706   -.1405189
       _cons |   14.15454   .5693846    24.86   0.000     13.03585    15.27323
------------------------------------------------------------------------------
. estat ovtest
Ramsey RESET test using powers of the fitted values of lprice
       Ho:  model has no omitted variables
                 F(3, 496) =      1.64
                  Prob > F =      0.1798
```

5. A more general command that implements several flavors of the RESET, and may be applied after instrumental-variables estimation, is Mark Schaffer's `ivreset`, available from `ssc`.

This model's predicted values no longer reject the RESET. The relationship between `rooms` and housing values appears to be nonlinear (although the pattern of signs on the `rooms` and `rooms2` coefficients is not that suggested by theory). But as theory suggests, communities with higher property tax burdens have lower housing values, ceteris paribus.

5.2.8 Specification plots

Many plots based on the residuals have been developed to help you evaluate the specification of the model because certain patterns in the residuals indicate misspecification. We can graph the residuals versus the predicted values with `rvfplot` (residual-versus-fitted plot) or plot them against a specific regressor with `rvpplot` (residual-versus-predictor plot) by using the regression model above:

```
. quietly regress lprice lnox ldist rooms rooms2 stratio lproptax
. rvpplot ldist, ms(Oh) yline(0)
```

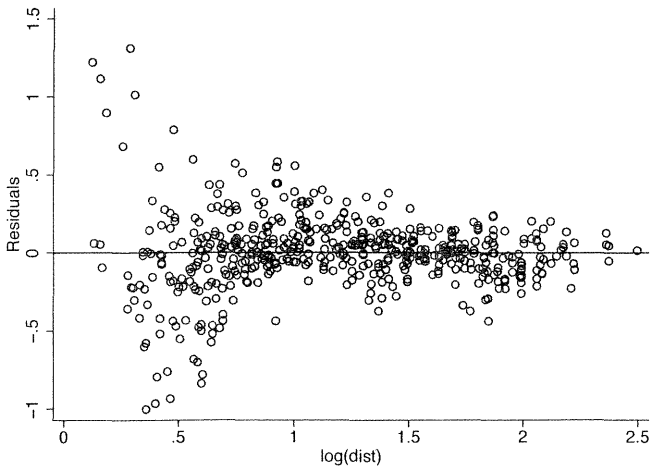

Figure 5.3: Residual-versus-predictor plot

The latter plot is displayed in figure 5.3. Any pattern in this graph indicates a problem with the model. For instance, the residuals appear much more variable for low levels versus high levels of log of distance (`ldist`), so the assumption of homoskedasticity (a constant variance of the disturbance process) is untenable.

A variety of other graphical techniques for identifying specification problems have been proposed, and several are implemented in Stata; see [R] **regress postestimation**.

5.2.9 Specification and interaction terms

We might also encounter specification error with respect to interactions among the regressors. If the true model implies that $\partial y/\partial x_j$ is a function of x_ℓ, we should fit the model

$$y = \beta_1 + \beta_2 X_2 + \cdots + \beta_j x_j + \beta_\ell X_\ell + \beta_p (x_j \cdot x_\ell) + \cdots + u \tag{5.5}$$

in which the regressor $(x_j \cdot x_\ell)$ is an *interaction term*. With this term added to the model, we find that $\partial y/\partial x_j = \beta_j + \beta_p x_\ell$. The effect of x_j then depends on x_ℓ. For example, in a regression of housing prices on attributes of the dwelling, the effect of adding a bedroom to the house may depend on the house's square footage.[6] If the coefficient β_p is constrained to equal zero [that is, if we estimate (5.5) without interaction effects], the partial derivatives of both x_j and x_ℓ are constrained to be constant rather than varying, as they would be for the equation including the interaction term. If the interaction term or terms are irrelevant, their t statistics will indicate that you can safely omit them. But here the correct specification of the model requires that you enter the regressors in the proper form in the fitted model.

As an example of misspecification due to interaction terms, we include `taxschl`—an interaction term between `lproptax`, the logarithm of average property taxes in the community, and `stratio`, the student–teacher ratio—in our housing-price model.[7] Both are negative factors, in the sense that buyers would prefer to pay lower taxes and enjoy schools with larger staff and would not be willing to pay as much for a house in a community with high values for either attribute.

```
. generate taxschl = lproptax * stratio
. regress lprice lnox ldist lproptax stratio taxschl
```

Source	SS	df	MS	Number of obs =	506
Model	38.7301562	5	7.74603123	F(5, 500) =	84.47
Residual	45.8521148	500	.09170423	Prob > F =	0.0000
				R-squared =	0.4579
Total	84.5822709	505	.167489645	Adj R-squared =	0.4525
				Root MSE =	.30283

lprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnox	-.9041103	.1441253	-6.27	0.000	-1.187276	-.6209444
ldist	-.1430541	.0501831	-2.85	0.005	-.2416499	-.0444583
lproptax	-1.48103	.5163117	-2.87	0.004	-2.495438	-.4666219
stratio	-.4388722	.1538321	-2.85	0.005	-.7411093	-.1366351
taxschl	.0641648	.026406	2.43	0.015	.0122843	.1160452
_cons	21.47905	2.952307	7.28	0.000	15.6786	27.27951

The interaction term is evidently significant, so a model excluding that term can be considered misspecified for the reasons discussed in section 5.2.1, although the omitted variable is algebraically related to the included regressors. The interaction term has

6. And vice versa: in (5.5), $\partial y/\partial x_\ell$ is a function of the level of x_j.
7. We exclude the `rooms` and `rooms2` regressors from this example for illustration.

a positive sign, so the negative partial derivative of `lprice` with respect to `lproptax` (`stratio`) becomes less negative (closer to zero) for higher levels of `stratio` (`lproptax`).

5.2.10 Outlier statistics and measures of leverage

To evaluate the adequacy of the specification of an fitted model, we must also consider evidence relating to the model's robustness to *influential data*. The OLS estimator is designed to fit the regression sample as well as possible. However, our objective in fitting the model often includes inference about the population from which the sample was drawn or computing out-of-sample forecasts. Evidence that the model's coefficients have been strongly influenced by a few data points or of structural instability over subsamples casts doubt on the fitted model's worth in any broader context. For this reason, we consider tests and plots designed to identify influential data.

A variety of statistics have been designed to evaluate influential data and the relationship between those data and the fitted model. A pioneering work in this field is Belsley, Kuh, and Welsch (1980) and the later version, Belsley (1991). An *outlier* in a regression relationship is a data point with an unusual value, such as a value of housing price twice as high as any other or a community with a student–teacher ratio 3 standard deviations below the mean. An outlier may be an observation associated with a large residual (in absolute terms), a data point that the model fits poorly.

On the other hand, an unusual data point that is far from the center of mass of the x_j distribution may also be an outlier, although the residual associated with that data point will often be small because the least-squares process attaches a squared penalty to the residual in forming the least-squares criterion. Just as the arithmetic mean (a least-squares estimator) is sensitive to extreme values (relative to the sample median), the least-squares regression fit will attempt to prevent such an unusual data point from generating a sizable residual. We say that this unusual point has a high degree of *leverage* on the estimates because including it in the sample alters the estimated coefficients by a sizable amount. Data points with large residuals may also have high leverage. Those with low leverage may still have a large effect on the regression estimates. Measures of influence and the identification of influential data points take their leverage into account.

You can calculate a measure of each data point's leverage after `regress` with

```
. predict double lev if e(sample), leverage
```

These leverage values are computed from the diagonal elements of the "hat matrix", $h_j = \mathbf{x}_j(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_j'$, where $\mathbf{x}_j$ is the jth row of the regressor matrix.[8] You can use `lvr2plot` to graphically display leverage values versus the (normalized) squared residuals. Points with very high leverage or very large squared residuals may deserve scrutiny. We can also examine those statistics directly. Consider our housing-price regression model, for which we compute leverage and squared residuals. The `town` variable identifies the community.

8. The formulas for `predict` options are presented in [R] **regress postestimation**.

```
. quietly regress lprice lnox ldist rooms rooms2 stratio lproptax
. generate town = _n
. predict double lev if e(sample), leverage
. predict double eps if e(sample), res
. generate double eps2 = eps^2
. summarize price lprice
```

Variable	Obs	Mean	Std. Dev.	Min	Max
price	506	22511.51	9208.856	5000	50001
lprice	506	9.941057	.409255	8.517193	10.8198

We then list the five largest values of the leverage measure, using `gsort` to produce the descending-sort order:

```
. gsort -lev
. list town price lprice lev eps2 in 1/5
```

	town	price	lprice	lev	eps2
1.	366	27499	10.2219	.17039262	.61813718
2.	368	23100	10.04759	.11272637	.30022048
3.	365	21900	9.994242	.10947853	.33088957
4.	258	50001	10.8198	.08036068	.06047061
5.	226	50001	10.8198	.0799096	.03382768

We can also examine the towns with the largest squared residuals:

```
. gsort -eps2
. list town price lprice lev eps2 in 1/5
```

	town	price	lprice	lev	eps2
1.	369	50001	10.8198	.02250047	1.7181195
2.	373	50001	10.8198	.01609848	1.4894088
3.	372	50001	10.8198	.02056901	1.2421055
4.	370	50001	10.8198	.0172083	1.0224558
5.	406	5000	8.517193	.00854955	1.0063662

As these results show, a large value of leverage does not imply a large squared residual, and vice versa. Several of the largest values of leverage or the squared residuals correspond to the extreme values of median housing prices recorded in the dataset, which range from $5,000 to $50,001. These data may have been coded with observations outside that range equal to that minimum or maximum value, respectively.

The DFITS statistic

A summary of the leverage values and magnitude of residuals is provided by the DFITS statistic of Welsch and Kuh (1977),

$$\text{DFITS}_j = r_j \sqrt{\frac{h_j}{1-h_j}}$$

where r_j is a studentized (standardized) residual,

$$r_j = \frac{e_j}{s_{(j)}\sqrt{1-h_j}}$$

with $s_{(j)}$ referring to the root mean squared error (s) of the regression equation with the jth observation removed. Working through the algebra shows that either a large value of leverage (h_j) or a large absolute residual (e_j) will generate a large $|\text{DFITS}_j|$. The DFITS measure is a scaled difference between the in-sample and out-of-sample predicted values for the jth observation. DFITS evaluates the result of fitting the regression model including and excluding that observation. Belsley, Kuh, and Welsch (1980) suggest that a cutoff value of $|\text{DFITS}_j| > 2\sqrt{k/N}$ indicates highly influential observations. We now compute DFITS in our housing-price regression model:[9]

```
. predict double dfits if e(sample), dfits
```

We then sort the calculated DFITS statistic in descending order and calculate the recommended cutoff value as an indicator variable, `cutoff`, equal to 1 if the absolute value of DFITS is large and zero otherwise. Consider the values of DFITS for which `cutoff` = 1:

9. See [R] **regress postestimation** for more details.

```
. gsort -dfits
. quietly generate cutoff = abs(dfits) > 2*sqrt((e(df_m)+1)/e(N)) & e(sample)
. list town price lprice dfits if cutoff
```

	town	price	lprice	dfits
1.	366	27499	10.2219	1.5679033
2.	368	23100	10.04759	.82559867
3.	369	50001	10.8198	.8196735
4.	372	50001	10.8198	.65967704
5.	373	50001	10.8198	.63873964
6.	371	50001	10.8198	.55639311
7.	370	50001	10.8198	.54354054
8.	361	24999	10.12659	.32184327
9.	359	22700	10.03012	.31516743
10.	408	27901	10.23642	.31281326
11.	367	21900	9.994242	.31060611
12.	360	22600	10.02571	.28892457
13.	363	20800	9.942708	.27393758
14.	358	21700	9.985067	.24312885
490.	386	7200	8.881836	-.23838749
491.	388	7400	8.909235	-.25909393
492.	491	8100	8.999619	-.26584795
493.	400	6300	8.748305	-.28782824
494.	416	7200	8.881836	-.29288953
495.	402	7200	8.881836	-.29595696
496.	381	10400	9.249561	-.29668364
497.	258	50001	10.8198	-.30053391
498.	385	8800	9.082507	-.302916
499.	420	8400	9.035987	-.30843965
500.	490	7000	8.853665	-.3142718
501.	401	5600	8.630522	-.33273658
502.	417	7500	8.922658	-.34950136
503.	399	5000	8.517193	-.36618139
504.	406	5000	8.517193	-.37661853
505.	415	7012	8.855378	-.43879798
506.	365	21900	9.994242	-.85150064

About 6% of the observations are flagged by the DFITS cutoff criterion. Many of those observations associated with large positive DFITS have the top-coded value of $50,001 for median housing price, and the magnitude of positive DFITS is considerably greater than that of negative DFITS. The identification of top-coded values that represent an arbitrary maximum recorded price suggests that we consider a different estimation technique for this model. The `tobit` regression model, presented in section 10.3.2, can properly account for the *censored* nature of the median housing price.

The DFBETA statistic

We may also want to focus on one regressor and consider its effect on the estimates by computing the DFBETA series with the `dfbeta` command after a regression.[10] The jth observation's DFBETA measure for regressor ℓ may be written as

$$\text{DFBETA}_j = \frac{r_j v_j}{\sqrt{v^2(1-h_j)}}$$

where the v_j are the residuals obtained from the partial regression of x_ℓ on the remaining columns of $\mathbf{X}$, and v^2 is their sum of squares. The DFBETAs for regressor ℓ measure the distance that this regression coefficient would shift when the jth observation is included or excluded from the regression, scaled by the estimated standard error of the coefficient. One rule of thumb suggests that a DFBETA value greater than unity in absolute value might be reason for concern since this observation might shift the estimated coefficient by more than one standard error. Belsley, Kuh, and Welsch (1980) suggest a cutoff of $|\text{DFBETA}_j| > 2/\sqrt{N}$.

We compute DFBETAs for one of the regressors, `lnox`, in our housing-price regression model:

```
. quietly regress lprice lnox ldist rooms rooms2 stratio lproptax
. dfbeta lnox
                          DFlnox:  DFbeta(lnox)
. quietly generate dfcut = abs(DFlnox) > 2/sqrt(e(N)) & e(sample)
. sort DFlnox
. summarize lnox
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        lnox |        506    1.693091    .2014102   1.348073   2.164472
```

10. As discussed in [R] **regress postestimation**, we can calculate one `dfbeta` series with `predict`, whereas you can use one `dfbeta` command to compute one or all of these series and automatically name them.

```
. list town price lprice lnox DFlnox if dfcut
```

	town	price	lprice	lnox	DFlnox
1.	369	50001	10.8198	1.842136	-.4316933
2.	372	50001	10.8198	1.842136	-.4257791
3.	373	50001	10.8198	1.899118	-.3631822
4.	371	50001	10.8198	1.842136	-.2938702
5.	370	50001	10.8198	1.842136	-.2841335
6.	365	21900	9.994242	1.971299	-.2107066
7.	408	27901	10.23642	1.885553	-.1728729
8.	368	23100	10.04759	1.842136	-.1309522
9.	11	15000	9.615806	1.656321	-.1172723
10.	410	27499	10.2219	1.786747	-.1117743
11.	413	17900	9.792556	1.786747	-.0959273
12.	437	9600	9.169518	2.00148	-.0955826
13.	146	13800	9.532424	2.164472	-.0914387
490.	154	19400	9.873029	2.164472	.0910494
491.	463	19500	9.87817	1.964311	.0941472
492.	464	20200	9.913438	1.964311	.0974507
493.	427	10200	9.230143	1.764731	.1007114
494.	406	5000	8.517193	1.93586	.1024767
495.	151	21500	9.975808	2.164472	.1047597
496.	152	19600	9.883285	2.164472	.1120427
497.	460	20000	9.903487	1.964311	.1142668
498.	160	23300	10.05621	2.164472	.1165014
499.	491	8100	8.999619	1.806648	.1222368
500.	362	19900	9.898475	2.04122	.1376445
501.	363	20800	9.942708	2.04122	.1707894
502.	490	7000	8.853665	1.806648	.1791869
503.	358	21700	9.985067	2.04122	.1827834
504.	360	22600	10.02571	2.04122	.2209745
505.	361	24999	10.12659	2.04122	.2422512
506.	359	22700	10.03012	2.04122	.2483543

Compared to the DFITS measure, we see a similar pattern for the DFBETA for `lnox`, with roughly 6% of the sample exhibiting large values of this measure. As with DFITS, the large positive values exceed their negative counterparts in magnitude. Many of the positive values are associated with the top-coded median house value of $50,001. These (presumably wealthy) communities have values of `lnox` well in excess of its minimum or mean. In contrast, many of the communities with large negative DFBETA values have extremely high values (or the maximum recorded value) of that pollutant.

How should we react to this evidence of many data points with a high degree of leverage? For this research project, we might consider that the price data have been improperly coded, particularly on the high end. Any community with a median housing value exceeding $50,000 has been coded as $50,001. These observations in particular have been identified by the DFITS and DFBETA measures.

Removing the bottom-coded and top-coded observations from the sample would remove communities from the sample nonrandomly, affecting the wealthiest and poorest communities. To resolve this problem of *censoring* (or coding of extreme values) we could use the `tobit` model (see section 10.3.2). A version of the tobit model, two-limit tobit, can handle censoring at both lower and upper limits.

5.3 Endogeneity and measurement error

In econometrics, a regressor is endogenous if it violates the zero-conditional-mean assumption $E[u \mid X] = 0$: that is, if the variable is correlated with the error term, it is *endogenous*. I deal with this problem in chapter 8.

We often must deal with measurement error, meaning that the variable that theory tells us belongs in the relationship cannot be precisely measured in the available data. For instance, the exact marginal tax rate that an individual faces will depend on many factors, only some of which we might be able to observe. Even if we knew the individual's income, number of dependents, and homeowner status, we could only approximate the effect of a change in tax law on her tax liability. We are faced with using an approximate measure, including some error of measurement, whenever we try to formulate and implement such a model.

This issue is similar to a proxy variable problem, but here it is not an issue of a latent variable such as aptitude or ability. An observable magnitude does exist, but the econometrician cannot observe it. Why is this measurement error of concern? Because the economic behavior we want to model—that of individuals, firms, or industries—presumably is driven by the *actual* measures, not our mismeasured approximations of those factors. If we fail to capture the actual measure, we may misinterpret the behavioral response.

Mathematically, measurement error (commonly termed errors-in-variables) has the same effect on an OLS regression model as endogeneity of one or more regressors (see chapter 8).

Exercises

1. Using the `lifeexpW` dataset, regress `lifeexp` on `popgrowth` and `lgnppc`. Generate an added-value plot by using `avplot safewater`. What do you conclude about the regression estimates?
2. Refit the model, including `safewater`. Use Ramsey's RESET to evaluate the specification. What do you conclude?
3. Generate the `dfits` series from the regression, and list the five countries with the largest absolute value of the DFITS measure. Which of these countries stand out?
4. Refit the model, omitting Haiti, and apply the RESET. What do you conclude about the model's specification?