Chapter 7
Specification of Regression Variables

So far we assumed we know what are the variables that needed to be in our regression model. However, what happens if we include in the regression model a variable that should not be there? What if we leave out a variable that should be included? Can we a proxy for a variable that we do not observe? These are the main question this chapter will address.

7.1 Model specification

What happens in practice is that it is difficult to be sure about the correct specification of the regression model. While theory may help, it usually depends on simplifying assumptions that may not necessarily hold. The properties of the regression estimates depend crucially on the validity of the specification of the model. The following is a quick summary of the consequences of misspecifying the regression model:

1. If you leave out a variable that should be included. The regression estimates are potentially biased. The standard errors of the coefficients and the corresponding t and F tests are in general invalid.
2. If you include a variable that should not be in the model. The coefficients will not be biased, but they are potentially inefficient.

7.2 Omitting a variable

7.2.1 The bias problem

Suppose that the true regression model that we should be estimated is given by

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u. \tag{7.1}$$

However, we do not have the variable X_3 or maybe we have it but we do not include it in the model. Hence, we estimate the following model

$$Y = \beta_1 + \beta_2 X_2 + u. \tag{7.2}$$

Then the predicted or fitted values are

$$\hat{Y} = b_1 + b_2 X_2 \tag{7.3}$$

Recall from previous chapters that the formula to estimate b_2 is given by

$$b_2 = \frac{\sum_{i=1}^n (X_{2i} - \bar{X}_2)(Y_i - \bar{Y})}{\sum_{i=1}^n (X_{2i} - \bar{X}_2)^2} \tag{7.4}$$

We say that b_2 is unbiased if its expected value is equal to the true population parameter β_2. If we plug Equation 7.1 into Equation 7.4 and take expectations we obtain

$$\begin{aligned} E[b_2] &= E\Big[\frac{\sum_{i=1}^n (X_{2i} - \bar{X}_2)(Y_i - \bar{Y})}{\sum_{i=1}^n (X_{2i} - \bar{X}_2)^2}\Big] \\ &= \beta_2 + \beta_3 \frac{\sum_{i=1}^n (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum_{i=1}^n (X_{2i} - \bar{X}_2)^2}. \end{aligned} \tag{7.5}$$

For b_2 to be unbiased we need that the second term on the right-hand side be equal to zero. This term is known as the *omitted variable bias* and it will be zero if $\beta_3 = 0$ or if $\sum_{i=1}^n (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3) / \sum_{i=1}^n (X_{2i} - \bar{X}_2)^2$ is equal to zero. Then the conditions for b_2 to be unbiased in the estimation of Equation 7.2 are:

1. That X_3 does not affect Y. That is, $\beta_3 = 0$.
2. That X_2 and X_3 are linearly uncorrelated. That is, the slope coefficient when we regress X_3 on X_2 os zero, $\sum_{i=1}^n (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)] / [\sum_{i=1}^n (X_{2i} - \bar{X}_2)^2 = 0$.

7.2.2 Invalid statistical tests

When a variable is omitted from the model, the standard errors of the coefficients and the texts statistics are generally invalid. This means that the t and F tests cannot be used.

7.2.3 Example

Consider the case where the true model to explain wages is given by

$$\texttt{wage} = \beta_1 + \beta_2\texttt{educ} + \beta_3\texttt{ability} + u. \tag{7.6}$$

That is, your `wage` is determined by your number of years of formal education (`educ`) and your `ability`. The problem in this equation is that actually it is very difficult to measure `ability`. Hence, we decide omit it and estimate the following model

$$\texttt{wage} = \beta_1 + \beta_2\texttt{educ} + u. \tag{7.7}$$

What is the problem with the estimate of β_2 if we use Equation 7.7? Is is biased! To get an idea of the size of the bias we will proxy `ability` with another variable, `IQ`. Equation 7.5 becomes

$$E[b_2] = \beta_2 + \beta_3 \frac{\sum_{i=1}^{n}(\texttt{educ}_i - \overline{\texttt{educ}})(\texttt{IQ}_i - \overline{\texttt{IQ}})}{\sum_{i=1}^{n}(\texttt{educ}_i - \overline{\texttt{educ}})^2}. \tag{7.8}$$

Notice that we can actually analyze if the bias is positive or negative based on the signs of the second part on the right-hand size. It seems that β_3 should be positive because higher ability (or IQ) should be correlated positively with wages. Moreover, the part that multiplies β_3 should also be positive because education and ability (or IQ) seem to be positively correlated. Hence, the whole second part on the right-hand side is positive, implying that β_2 is biased upwards. This means that on average we will be getting a larger coefficient (by estimating Equation 7.7) than the true coefficient (if we were estimating the true Equation 7.6).

Let's look at this empirically by estimating Equations 7.6 and 7.7 with real data (where we use `IQ` in place of `ability`):

```
Model 1: OLS, using observations 1-935
Dependent variable: wage

             coefficient   std. error   t-ratio    p-value
  --------------------------------------------------------
  const       146.952       77.7150      1.891    0.0589    *
  educ         60.2143       5.69498    10.57     9.35e-025 ***

Mean dependent var   957.9455   S.D. dependent var   404.3608
Sum squared resid    1.36e+08   S.E. of regression   382.3203
R-squared            0.107000   Adjusted R-squared   0.106043
F(1, 933)            111.7929   P-value(F)           9.35e-25
Log-likelihood      -6885.458   Akaike criterion     13774.92
Schwarz criterion    13784.60   Hannan-Quinn         13778.61
```

$$\widehat{\texttt{wage}} = \underset{(77.715)}{146.952} + \underset{(5.6950)}{60.2143}\,\texttt{educ}$$

$$N = 935 \quad \bar{R}^2 = 0.1060 \quad F(1,933) = 111.79 \quad \hat{\sigma} = 382.32$$

(standard errors in parentheses)

```
Model 2: OLS, using observations 1-935
Dependent variable: wage

             coefficient   std. error   t-ratio    p-value
```

```
  -------------------------------------------------------
  const      -128.890      92.1823      -1.398    0.1624
  educ         42.0576      6.54984      6.421    2.15e-010 ***
  IQ            5.13796     0.955827     5.375    9.66e-08  ***

Mean dependent var   957.9455   S.D. dependent var   404.3608
Sum squared resid    1.32e+08   S.E. of regression   376.7300
R-squared            0.133853   Adjusted R-squared   0.131995
F(2, 932)            72.01515   P-value(F)           8.27e-30
Log-likelihood      -6871.185   Akaike criterion     13748.37
Schwarz criterion    13762.89   Hannan-Quinn         13753.91
```

$$\widehat{\texttt{wage}} = \underset{(92.182)}{-128.890} + \underset{(6.5498)}{42.0576}\,\texttt{educ} + \underset{(0.95583)}{5.13796}\,\texttt{IQ}$$

$$N = 935 \quad \bar{R}^2 = 0.1320 \quad F(2,932) = 72.015 \quad \hat{\sigma} = 376.73$$

(standard errors in parentheses)

The empirical results are consistent with our theoretical analysis. The estimate of β_2 in Equation 7.7 is too large (upward biased). The bias can be obtained separately by estimating a regression of `IQ` on `educ` and then plugging the results in Equation 7.8

```
Model 3: OLS, using observations 1-935
Dependent variable: IQ

             coefficient   std. error   t-ratio    p-value
  -------------------------------------------------------
  const       53.6872       2.62293      20.47    3.36e-077 ***
  educ         3.53383      0.192210     18.39    1.16e-064 ***

Mean dependent var   101.2824   S.D. dependent var   15.05264
Sum squared resid    155346.5   S.E. of regression   12.90357
R-squared            0.265943   Adjusted R-squared   0.265157
F(1, 933)            338.0192   P-value(F)           1.16e-64
Log-likelihood      -3716.973   Akaike criterion     7437.946
Schwarz criterion    7447.627   Hannan-Quinn         7441.637
```

$$\widehat{\texttt{IQ}} = \underset{(2.6229)}{53.6872} + \underset{(0.19221)}{3.53383}\,\texttt{educ}$$

$$N = 935 \quad \bar{R}^2 = 0.2652 \quad F(1,933) = 338.02 \quad \hat{\sigma} = 12.904$$

(standard errors in parentheses)

Replacing the valued in Equation 7.8

$$\begin{aligned} E[b_2] &= \beta_2 + \beta_3 \frac{\sum_{i=1}^{n}(\texttt{educ}_i - \overline{\texttt{educ}})(\texttt{IQ}_i - \overline{\texttt{IQ}})}{\sum_{i=1}^{n}(\texttt{educ}_i - \overline{\texttt{educ}})^2}. \qquad (7.9) \\ &= \beta_2 + 5.13796 \times 3.53383 \\ &= \beta_2 + 18.15667 \end{aligned}$$

That is exactly the difference between the coefficients in Equations 7.6 and 7.7, 60.2143 - 42.0576 = 18.15667.

7.3 Including a variable that should not be included

Suppose that the true population model is given by

$$Y = \beta_1 + \beta_2 X_2 + u. \tag{7.10}$$

However, for some season you include X_3 and end up estimating the following model

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u. \tag{7.11}$$

In a regression model like Equation 7.11 with two variables (X_2 and X_3) the OLS estimator for b_2 is given by

$$\begin{aligned} b_2 = & \frac{\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)(Y_i-\bar{Y})\sum_{i=1}^{n}(X_{3i}-\bar{X}_3)^2}{\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)^2\sum_{i=1}^{n}(X_{3i}-\bar{X}_3)^2-\left(\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)(X_{3i}-\bar{X}_3)\right)^2} \\ & - \frac{\sum_{i=1}^{n}(X_{3i}-\bar{X}_3)(Y_i-\bar{Y})\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)(X_{3i}-\bar{X}_3)}{\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)^2\sum_{i=1}^{n}(X_{3i}-\bar{X}_3)^2-\left(\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)(X_{3i}-\bar{X}_3)\right)^2} \end{aligned} \tag{7.12}$$

Which is certainly different from the OLS estimator for b_2 in Equation 7.10,

$$b_2 = \frac{\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)^2} \tag{7.13}$$

Interestingly, b_2 in both Equations (7.12 and 7.13) is unbiased, $E(b_2) = \beta_2$. Hence, estimating the effect of X_2 on Y will yield unbiased estimates even if we include irrelevant variables. Then, what is the problem? Including irrelevant variables will inflate the standard errors of the coefficients. This means that the estimate b_2 from Equation 7.11 will be inefficient. The implied population variance of b_2 in Equation 7.11 is

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)^2} \cdot \frac{1}{(1-r_{X_2X_3}^2)} \tag{7.14}$$

where $r_{X_2X_3}^2$ is the correlation coefficient between X_2 and X_3, while the population variance of b_2 in Equation 7.10 is

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)^2}. \tag{7.15}$$

Notice that because $0 \leq r_{X_2X_3}^2 \leq 1$, the population variance in Equation 7.15 is larger than the implied population variance in Equation 7.14. Actually, they will be equal if $r_{X_2X_3}^2 = 0$, that is, if X_2 and X_3 are linearly uncorrelated. Moreover, when linearly uncorrelated $\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)(X_{3i}-\bar{X}_3)=0$, then Equation 7.12 reduces to 7.13, meaning that including X_3 in the equation will not affect the estimation of β_2. While the population variances are the same, the estimated (sample) variances will still differ due to a reduction in the degrees of freedom.

7.3.1 Example

Consider the following model where we want to see how age affects the likelihood of being married. Are older people more likely to be married? Well, let's estimate the exact response of `married` to `age`,[1]

$$\texttt{married} = \beta_1 + \beta_2 \texttt{age} + \varepsilon \tag{7.16}$$

The estimation results from Gretl are

```
Model 1: OLS, using observations 1-935
Dependent variable: married

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const       0.540935     0.107608      5.027    5.98e-07 ***
  age         0.0106442    0.00323870    3.287    0.0011   ***

Mean dependent var   0.893048   S.D. dependent var   0.309217
Sum squared resid    88.28274   S.E. of regression   0.307608
R-squared            0.011445   Adjusted R-squared   0.010385
F(1, 933)            10.80160   P-value(F)           0.001052
Log-likelihood      -223.4066   Akaike criterion     450.8133
Schwarz criterion    460.4944   Hannan-Quinn         454.5047
```

$$\widehat{\texttt{married}} = \underset{(0.10761)}{0.540935} + \underset{(0.0032387)}{0.0106442}\,\texttt{age}$$

$$N = 935 \quad \bar{R}^2 = 0.0104 \quad F(1,933) = 10.802 \quad \hat{\sigma} = 0.30761$$

(standard errors in parentheses)

If the average `age` in the sample is 33 years of age, the predicted value for `married` is 89.2 ($\widehat{\texttt{married}} = 0.5409 + 0.0106 \times 33$). This means that if you are 33 years old, the probability that you are married is 89.2%. In addition, every year you get older, the probability that you are married increases by 0.011 or about 1.%. For some reason you think that only fools get married and then you decide to wrongly estimate the model

$$\texttt{married} = \beta_1 + \beta_2 \texttt{age} + \beta_3 \texttt{IQ} + \varepsilon \tag{7.17}$$

[1] Because `married` is actually a dummy variable this is a linear probability model, a type of model that we will see in detail in Chapter 9.

where the variable IQ is X_3 in Equation 7.11 and should not be in the model. The estimation results from Gretl are

```
Model 2: OLS, using observations 1-935
Dependent variable: married

             coefficient    std. error    t-ratio   p-value
  ---------------------------------------------------------
  const       0.563197      0.129804       4.339    1.59e-05 ***
  age         0.0106007     0.00324337     3.268    0.0011   ***
  IQ         -0.000205573   0.000669635   -0.3070   0.7589

Mean dependent var   0.893048   S.D. dependent var   0.309217
Sum squared resid    88.27381   S.E. of regression   0.307757
R-squared            0.011545   Adjusted R-squared   0.009424
F(2, 932)            5.442677   P-value(F)           0.004467
Log-likelihood      -223.3594   Akaike criterion     452.7187
Schwarz criterion    467.2404   Hannan-Quinn         458.2559
```

$$\widehat{\texttt{married}} = \underset{(0.12980)}{0.563197} + \underset{(0.0032434)}{0.0106007}\,\texttt{age} - \underset{(0.00066963)}{0.000205573}\,\texttt{IQ}$$

$$N = 935 \quad \bar{R}^2 = 0.0094 \quad F(2,932) = 5.4427 \quad \hat{\sigma} = 0.30776$$

(standard errors in parentheses)

Not surprisingly, the effect of IQ on married is not significant. This means that fools are not more likely to be married. However, the results do not necessarily support the conjecture that higher IQ is associated with married people either. Nevertheless, including IQ does not seems to help in the estimation of β_2. As we have seen theoretically, the estimate of the second equation is less efficient as can be appreciated from its larger standard error ($0.003243 > 0.003239$).

7.4 Testing a linear restriction

Testing linear restriction on the regression coefficients is sometimes very useful. Consider the following regression model,

$$\log \texttt{wage} = \beta_1 + \beta_2 \texttt{exper} + \beta_3 \texttt{educ} + \varepsilon \tag{7.18}$$

The regression output in Gretl is

```
Model 1: OLS, using observations 1-935
Dependent variable: logwage

             coefficient   std. error   t-ratio    p-value
  --------------------------------------------------------
  const       5.50271      0.112037     49.12     8.13e-261 ***
  educ        0.0777820    0.00657687   11.83     3.62e-030 ***
  exper       0.0197768    0.00330251    5.988    3.02e-09  ***
```

```
Mean dependent var   6.779004   S.D. dependent var   0.421144
Sum squared resid    143.9786   S.E. of regression   0.393044
R-squared            0.130859   Adjusted R-squared   0.128994
F(2, 932)            70.16174   P-value(F)           4.13e-29
Log-likelihood      -452.0704   Akaike criterion     910.1407
Schwarz criterion    924.6624   Hannan-Quinn         915.6779
```

$$\widehat{\texttt{logwage}} = \underset{(0.11204)}{5.50271} + \underset{(0.0065769)}{0.0777820}\,\text{educ} + \underset{(0.0033025)}{0.0197768}\,\text{exper}$$

$$N = 935 \quad \bar{R}^2 = 0.1290 \quad F(2,932) = 70.162 \quad \hat{\sigma} = 0.39304$$

(standard errors in parentheses)

Let's say that we want to text whether the effect of a year on education on wages is the same as the effect of a year of experience of wages. That is, we want to text the following null hypothesis,

$$H_0 : \beta_2 = \beta_3 \tag{7.19}$$

While it may be tempting to just look and compare the regression estimates b_2 and b_3, this approach is not correct. Remember that b_2 and b_3 are just estimates and are not the unknown β_2 and β_3. The statistically correct approach is to run an auxiliary restricted regression where we force $b_2 = b_3$. Then, we have to compare if the regression fit with the *restricted* coefficients is significantly lower that the regression fit with the *unrestricted* (original) regression. To do this we calculate the residual sum of squares from the restricted model (RSS_R) and the residual sum of squares from the unrestricted model (RSS_U) and calculate the following F statistic:

$$F_{r,n-k} = \frac{(RSS_R - RSS_U)/r}{RSS_U/(n-k)} \tag{7.20}$$

where F is distributed with r and $n-k$ degrees of freedom. The number of restrictions r is equal to one in our example.

This is done automatically in Gretl. After you estimate the unrestricted regression model, in the regression output window you have to go to `Tests` → `Linear restrictions` and a new window will open. In the new window you have to type the command `b[educ] - b[exper] = 0` to obtain

```
Restriction:
 b[educ] - b[exper] = 0

Test statistic: F(1, 932) = 97.8892, with p-value = 5.14357e-022

Restricted estimates:

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       6.24122      0.0877816    71.10     0.0000    ***
  educ        0.0214837    0.00346501    6.200    8.46e-010 ***
  exper       0.0214837    0.00346501    6.200    8.46e-010 ***

  Standard error of the regression = 0.412948
```

The calculated F-statistics (that used Equation 7.20) is 97.8892 with an associated p-value that is below 0.05. This means that the fit in the two regression equations is significantly different and we reject the null hypothesis presented in Equation 7.19. We conclude that the effect of education and experience have a significantly different effect on wages.

If you want to test whether education had four times the effect on wages than experience, the null is

$$H_0 : \beta_2 = 4 \times \beta_3 \tag{7.21}$$

The command in Gretl is `b[educ] - 4*b[exper] = 0` to have

```
Restriction:
 b[educ] - 4*b[exper] = 0

Test statistic: F(1, 932) = 0.0126711, with p-value = 0.910399

Restricted estimates:

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       5.50597      0.108181      50.90    1.89e-271 ***
  educ        0.0778171    0.00656598    11.85    2.78e-030 ***
  exper       0.0194543    0.00164150    11.85    2.78e-030 ***

  Standard error of the regression = 0.392836
```

Notice that the F-statistics is fairly small and has a p-value that is now greater than 5%. We do not reject the null hypothesis and conclude that, on average, one year of education has four times the effect on wages than one year of experience.