Exercise – Understanding the omitted variable bias using simulation

1. Aim of the exercise

In regression analysis, we often encounter situations where we cannot observe all the independent variables that are potentially correlated with the main variable of interest and hence should be controlled for. When this happens, the zero conditional mean assumption is violated. Econometrics classes demonstrate the resulting bias in the coefficient estimate of the main variable of interest theoretically, which makes the understanding of this bias somewhat abstract. This exercise illustrates the omitted variable bias through simulation. It shows how the sampling distribution of the OLS estimator, in particular its mean, is affected when a variable relevant to the main variable of interest is omitted from the regression.

2. Theory

Consider the linear regression model

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \varepsilon_i.$$

where for simplicity we do not consider a constant. Suppose that

$$\mathrm{E}\left[\varepsilon_i \mid x_{i1}\right] = 0,$$

and

$$\mathrm{E}\left[\varepsilon_i \mid x_{i2}\right] = 0.$$

Suppose that we do not observe x_{i2} so that it enters the error. The model becomes

$$y_i = x_{i1}\beta_1 + \varepsilon_i^*$$

where

$$\varepsilon_i^* = x_{i2}\beta_2 + \varepsilon_i.$$

Then,

$$\begin{aligned}
\mathrm{E}\left[\varepsilon_i^* \mid x_{i1}\right] &= \mathrm{E}\left[x_{i2}\beta_2 \mid x_{i1}\right] + \mathrm{E}\left[\varepsilon_i \mid x_{i1}\right] \\
&= \beta_2 \mathrm{E}\left[x_{i2} \mid x_{i1}\right] + 0 \\
&\neq 0
\end{aligned}$$

if $\beta_2 \neq 0$ and $\mathrm{E}\left[x_{i2} \mid x_{i1}\right] \neq 0$. $\beta_2 \neq 0$ means that x_{i2} should enter the model. $\mathrm{E}\left[x_{i2} \mid x_{i1}\right] \neq 0$ means that x_{i1} and x_{i2} are correlated. The zero conditional mean assumption in this case is violated for ε_i^*.

What is the implication of

$$\mathrm{E}\left[\varepsilon_i^* \mid x_{i1}\right] \neq 0$$

for the OLS estimator $\hat{\beta}_1$? The OLS estimator $\hat{\beta}_1$ when x_{i2} is omitted from the regression is given by

$$\begin{aligned}
\hat{\beta}_1 &= (\boldsymbol{x}_1'\boldsymbol{x}_1)^{-1}\boldsymbol{x}_1'\boldsymbol{y} \\
&= (\boldsymbol{x}_1'\boldsymbol{x}_1)^{-1}\boldsymbol{x}_1'(\boldsymbol{x}_1\beta_1 + \boldsymbol{x}_2\beta_2 + \boldsymbol{\varepsilon}) \\
&= \beta_1 + (\boldsymbol{x}_1'\boldsymbol{x}_1)^{-1}\boldsymbol{x}_1'\boldsymbol{x}_2\beta_2 + (\boldsymbol{x}_1'\boldsymbol{x}_1)^{-1}\boldsymbol{x}_1'\boldsymbol{\varepsilon}.
\end{aligned}$$

$\hat{\beta}_1$ represents the impact of x_1 on y where y is in fact driven not only by x_1 but also x_2 according to the true DGP. This means that we explain y only with x_1 whereas we should explain it also with x_2.

Taking the expectation conditional on $\boldsymbol{X}$, we have

$$\mathrm{E}\left[\hat{\beta}_1 \mid \boldsymbol{X}\right] = \beta_1 + (\boldsymbol{x}_1'\boldsymbol{x}_1)^{-1}\boldsymbol{x}_1'\boldsymbol{x}_2\beta_2$$

since $\mathrm{E}\left[\boldsymbol{\varepsilon} \mid \boldsymbol{X}\right] = \mathbf{0}$ in the true model. In two cases $\hat{\beta}_1$ is an unbiased estimator. First, if

$$(\boldsymbol{x}_1'\boldsymbol{x}_1)^{-1}\boldsymbol{x}_1'\boldsymbol{x}_2 = 0,$$

meaning that there is no correlation between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ in the sample. Realize that the stated expression is the OLS estimate of the coefficient of $\boldsymbol{x}_1$ from the regression of $\boldsymbol{x}_2$ on $\boldsymbol{x}_1$. Second, if

$$\beta_2 = 0,$$

meaning that $\boldsymbol{x}_2$ does not enter the true model. Otherwise $\hat{\beta}_1$ is subject to the omitted variable bias.

3. Application

3.1. Clear the memory

Clear the memory from possible calculations from an earlier session.

```
% 3.1. Clear the memory
clear;
```

3.2. Set the number of simulations

Set the number of simulations to be carried out.

```
% 3.2. Set the number of simulations
N_sim = 1000;
```

3.3. Set the sample size

Assume a linear regression model with `N_obs` observations available for the variables of this model.

```
% 3.3. Set the sample size
N_obs = 1000;
```

3.4. Set true values for the coefficients of the model

Assume that the liner regression model contains two independent variables and, for simplicity, no constant. Assume also that this is the true DGP and that we know the true values of the coefficients.

```
20 % 3.4. Set true values for the coefficients
21 B_true = [0.5 0.75]';
```

3.5. Create the constant term

Create the constant term.

```
23 % 3.5. Create the constant term
24 x_0 = ones(N_obs,1);
```

3.6. Create a vector of covariances between two independent variables

A problem related to the systematic part of the DGP is omitting a relevant independent variable that is part of the true DGP. One potential reason is that the researcher does not know that a particular variable belongs in the specification or cannot collect data on that variable. If the omitted independent variable is uncorrelated with all of the independent variables that are included in the regression model, leaving it out will not bias the estimated coefficients. The omitted variable could, however, still explain some part of the variation in the dependent variable. If the omitted independent variable is correlated with one or more of the independent variables, this will bias the coefficient estimates of those variables.

Here we examine the omitted variable problem across a range of correlations between an included independent variable and the omitted independent variable. We define correlation levels using 10 different covariance values, ranging from 0 to 0.99. Line 27 creates a vector array containing these values. We assume that each variable has a variance of 1. Line 28 defines the column dimension of this vector. We will use this variable when iterating over different correlation scenarios in our simulation.

```
26 % 3.6. Create a vector of covariances between two independent variables
27 sigma_x_1_x_2 = 0:0.11:0.99;
28 N_sig = size(sigma_x_1_x_2,2);
```

3.7. Preallocate a matrix to store the simulated OLS coefficient estimates

Preallocate a matrix that will store the coefficient estimates from repeated sampling, simulated at different correlation levels between the included and omitted independent variables. The matrix is `N_sim` × `N_sim` because we will simulate `N_sim` samples, at `N_sig` different levels of correlation.

```
30 % 3.7. Preallocate a matrix for storing OLS estimates from all samples
31 B_hat_1_sim = NaN(N_sim,N_sig);
```

3.8. Define an input argument for the multivariate normal random number generator

In the next section, we will draw random numbers from the multivariate normal distribution to create two independent variables that are correlated with each other. To do this, we will make use of the built-in MATLAB `mvnrnd` function. The function accepts three input arguments. The first input argument is the mean vector of the distribution. The second input argument is the covariance matrix of the distribution. The third input argument specifies the number of observations to be drawn for each random variable of the distribution. The third input argument is defined above. Here we define the first input argument. The second input argument will be defined within the simulation in the next section because in each iteration of the simulation the covariance matrix will be updated in accordance with the different correlation levels between two two variables.

```
34 % 3.8. Define mean vector for the multivariate random number generator
35 mu = [0 0];
```

3.9. Create the sampling distribution of the OLS estimator at different correlation levels between the included and omitted independent variables

Here we consider a nested loop structure, with one outer loop and one inner loop. The inner for loop simulates `N_sim` coefficient estimates from repeated sampling. The outer for loop repeats this simulation for 10 different correlation levels between the included and the omitted variables. Consider the inner for loop. Line 37 defines the index of the for loop. Line 39 defines the covariance matrix of the included and omitted variables at given covariance values. Variances of each independent variable are set to 1. In line 41 we supply the `mvnrnd` function with the defined covariance matrix, and with two other input arguments defined in the previous section. The function generates random values for the included and omitted variables from the multivariate normal distribution so that the variables are correlated.

In lines 42 and 43 we define the included (`x_1`) and omitted (`x_2`) variables.

In line 45 we generate random values for the error term. Note that we rule out heteroskedasticity by setting the standard deviation of the error term to 1. In line 46 we generate values for the dependent variable using both `x_1` and `x_2` following the true DGP. In line 47 we use the function `exercisefunction` to estimate the coefficient of `x_1` using this dependent variable but `x_1` as the only explanatory variable, incorrectly ignoring `x_2`. In line 48 we collect the simulated coefficient estimates at different levels of correlation between the included and the omitted variable in the matrix array `B_hat_1_sim(i,j)`.

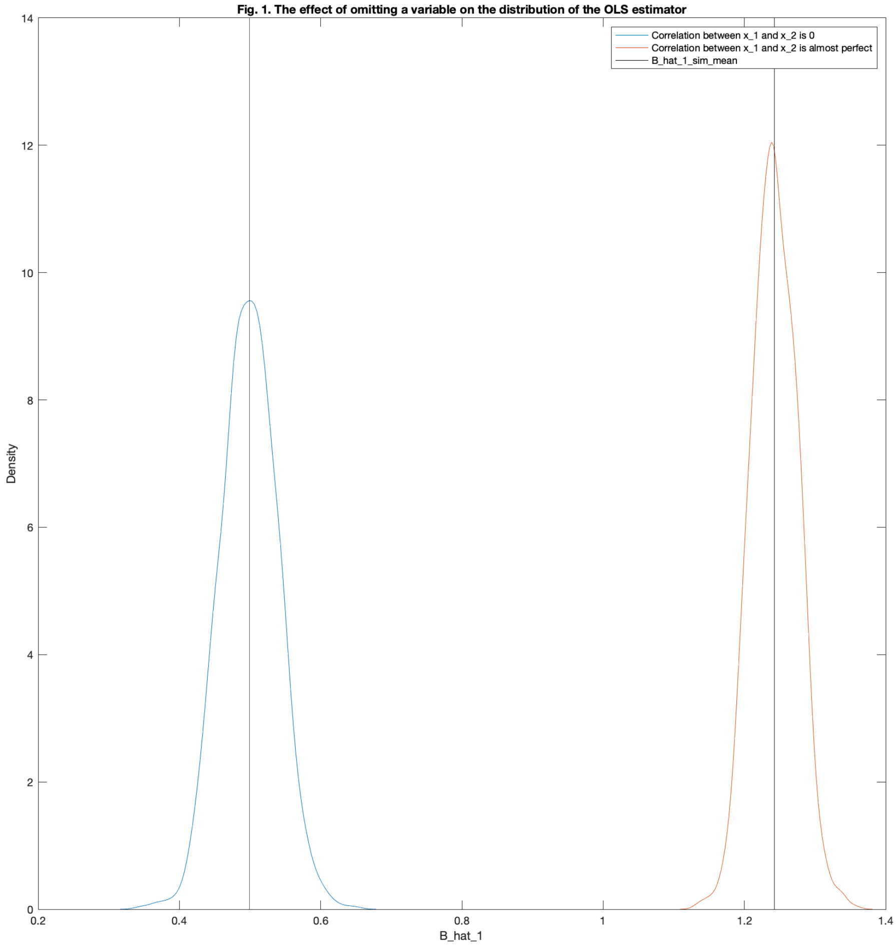
```
36 % 3.9. Create the sampling distribution of the biased OLS estimator
37 for j = 1:N_sig
38     for i = 1:N_sim
39         Sigma = reshape([1 sigma_x_1_x_2(:,j) sigma_x_1_x_2(:,j) 1] ...
40             ,2,2);
41         x_1_x_2_mvn = mvnrnd(mu,Sigma,N_obs);
42         x_1 = x_1_x_2_mvn(:,1);
43         x_2 = x_1_x_2_mvn(:,2);
44         X = [x_1 x_2];
45         u = random('Normal',0,1,[N_obs 1]);
46         y = X*B_true+u;
47         LSS = exercisefunctionlss(y,x_1);
48         B_hat_1_sim(i,j) = LSS.B_hat(1,1);
49     end
50 end
```

4. Plot the sampling distribution of the biased OLS estimator

The plot produced here shows the density estimate of the 1000 OLS coefficient estimates of `x_1`, at a correlation of 0 and 0.99 between `x_1` and `x_2`. The distribution of estimates at correlation 0 is centered right at 0.5, indicating no bias. This demonstrates that omitting a variable that is not correlated with an included variable does not affect parameter estimates in the case of OLS. In contrast, the distribution of estimates when the correlation is 0.99 shows a substantial amount of bias. In fact, recall from above that the coefficient on the omitted variable was set to 0.75. The mean of the distribution with correlation 0.99 is 1.242, which is 0.5 (true coefficient of `x_1`) plus 0.742. Hence, at near-perfect correlation with the omitted variable, almost all of the true effect of that omitted variable is incorrectly attributed to `x_1` through the biased estimate of the coefficient of `x_1`.

```
%% 4. Plot the sampling distribution of the biased OLS estimator
hold on
ksdensity(B_hat_1_sim(:,1))
ksdensity(B_hat_1_sim(:,10))
line([mean(B_hat_1_sim(:,1)) mean(B_hat_1_sim(:,1))],ylim, ...
    'Color','black')
line([mean(B_hat_1_sim(:,10)) mean(B_hat_1_sim(:,10))],ylim, ...
    'Color','black')
title(['The Effect of Omitting a Variable on the Distribution ' ...
    'of the OLS estimator'])
legend('Correlation between x\_1 and x\_2 is 0',['Correlation ' ...
    'between x\_1 and x\_2 is almost perfect'], ...
    'B\_hat\_1\_sim\_mean')
ylabel('Density')
xlabel('B\_hat\_1')
hold off
```

Fig. 1. The effect of omitting a variable on the distribution of the OLS estimator



5. Other experiments

We can conduct additional experiments. For example, we can repeat the simulation increasing the sample size for each draw in the simulation: `N_obs`. Does the simulated sampling distribution of the OLS estimate suggest that it is consistent when the zero conditional mean assumption is violated due to omitting a relevant variable? It does not. When we increase the sample size, the OLS estimator does not come closer to the true coefficient value. This means that the OLS estimator is inconsistent when the zero conditional mean assumption is violated, as well as that it is biased as shown above already. This kind of exploration helps to see the real nature of the DGP and how the chosen statistical estimator performs.

6. Final notes

This file is prepared and copyrighted by Tunga Kantarcı. Parts of the simulation exercise are based on Carsey, T. M., and Harden, J. J., 2014. Monte Carlo simulation and resampling methods for social science. SAGE Publications. This file and the accompanying MATLAB files are available on GitHub and can be accessed via this link.