14.32 Econometric Data Science
Professor Anna Mikusheva
September, 2019

Lectures 7-9

Multivariate regression

Omitted variable bias

The goal of many econometric studies is to estimate a causal relation. For example, in the case of the effect of a class size on student performance, the goal is to predict what would happen to the students' test scores if one were able to reduce the class size by a specified number of students. That is, we wish to compare the same school and the same students but in two different states: when kids are educated in small classes and when they are educated in large classes. One of the way of doing this is to run an experiment when school/students are randomly assigned to large/small classrooms, educate kids and compare their performances. We often cannot run an experiment like this, but we have observational data. In Stock and Watson, the following regression was run on observational data:

$$Test\ score_i = \alpha + \beta Class\ size_i + e_i,$$

the estimated coefficients are $\widehat{\alpha} = 698, \widehat{\beta} = -2.28$. Can we plausibly say that these estimates are close to the true coefficients? No, the class size is not randomly assigned, but rather the schools with small and large class sizes are quite different.

Omitted variable bias appears if there exists a variable Z that satisfies the following two conditions:

(1) Z is a determinant of Y;

(2) Z is correlated with X

In the class size example 'family income' (average family income in the school district) is one such variable. Indeed, richer families devote more resources to their kids' education outside of school- thus a determinant of tests scores. School districts in higher-income neighborhoods have more resources to hire additional teachers and tend to have smaller classes (correlated with class size).

Intuition: family income has a direct effect on Y, thus, enters e in our regression equation. Given condition (2) it produces the correlation between X and e, which contradicts assumption 1 (exogeneity) of the regression analysis.

Here we derive the asymptotic bias due to omitted variable:

$$\widehat{\beta} - \beta = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})e_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} \to^p \frac{cov(X,e)}{Var(X)}.$$

This formula allows us to determine the sign of the bias by thinking about the sign of the covariance between X and e. For example, higher family income has a positive effect on test scores, and higher family income is negatively correlated with class size. Thus, positive times negative, we would expect a negative bias. We *a priori* expect a negative coefficient in our regression (smaller class size leads to better test scores), thus, with bias we will get a negative effect, but one that is larger in size than we should (we overstate the effect due to bias).

Motivation of multivariate regression

So, the problem is that the regression implicitly compares quite different schools, not just of different class sizes but also of different incomes. If one wants to compare schools with different class sizes but the same

income level, then s/he should consider the multivariate regression:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + e_i,$$

here, for example, $X_{1,i}$ may be the class size, while $X_{2,i}$ is family income. In this regression the coefficients have somewhat different interpretations: β_0 is still an intercept. β_1 is the expected effect on Y of a one-unit change in X_1 *holding* X_2 *constant*. Indeed, compare two schools with the same X_2 but slightly changed X_1:

$$\begin{aligned} EY &= \beta_0 + \beta_1 X_1 + \beta_2 X_2, \\ E(Y + \Delta Y) &= \beta_0 + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2, \\ E\Delta Y &= \beta_1 \Delta X_1. \end{aligned}$$

Ordinary Least Squares (OLS)

The method of estimation of a multivariate regression is OLS. Consider the following model:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + ... + \beta_k X_{k,i} + e_i,$$

and we have data $\{Y_i, X_{1,i}, ..., X_{k,i}, i = 1, ..., n\}$. The OLS estimator is the solution to the following optimization problem:

$$(\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_k) = \arg \min_{(b_0, b_1, ..., b_k)} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1,i} - ... - b_k X_{k,i})^2 = \arg \min_{(b_0, b_1, ..., b_k)} S(b_0, b_1, ..., b_k).$$

Let us write the first-order condition for this optimization:

$$\frac{\partial S(b)}{\partial b_0}|_{b=\widehat{\beta}} = -2 \sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_{1,i} - ... - \widehat{\beta}_k X_{k,i}) = 0,$$

$$\frac{\partial S(b)}{\partial b_j}|_{b=\widehat{\beta}} = -2 \sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_{1,i} - ... - \widehat{\beta}_k X_{k,i}) X_{j,i} = 0,$$

or if we introduce the residuals $\widehat{e}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_{1,i} - ... - \widehat{\beta}_k X_{k,i}$, then:

$$\sum_{i=1}^{n} \widehat{e}_i = 0, \quad \text{and } \sum_{i=1}^{n} \widehat{e}_i X_{j,i} = 0 \text{ for all } j = 1, ..., k.$$

It is easier to solve this system of equations in a matrix form. Let us stack data and parameters in the following matrix/vectors:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ ... \\ Y_n \end{pmatrix}, X = \begin{pmatrix} 1 & X_{1,1} & X_{2,1} & ... & X_{k,1} \\ 1 & X_{1,2} & X_{2,2} & ... & X_{k,2} \\ & & ... & & \\ 1 & X_{1,n} & X_{2,n} & ... & X_{k,n} \end{pmatrix}, e = \begin{pmatrix} e_1 \\ e_2 \\ ... \\ e_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ ... \\ \beta_k \end{pmatrix}.$$

Sizes are: Y is $n \times 1$, X is $n \times k$, e is $n \times 1$, β is $k \times 1$. Then the model is

$$Y = X\beta + e.$$

The first-order conditions we wrote above are:

$$(1, 1, ...)(Y - X\widehat{\beta}) = 0,$$

and

$$(X_{j,1}, X_{j,2}, ..., X_{j,n})(Y - X\widehat{\beta}) = 0.$$

If all conditions are stuck together, we get:

$$X'(Y - X\widehat{\beta}) = 0,$$

or when solved for $\widehat{\beta}$:

$$\widehat{\beta} = (X'X)^{-1}X'Y.$$

This works only if $X'X$ is an insertable matrix; apparently, this is the second-order condition for the optimization as well. This condition is called 'no multicollinearity' and is formulated as 'no regressor is a linear combination of the other regressors including constant'. Also notice,

$$\widehat{\beta} - \beta = (X'X)^{-1}X'Y - \beta = (X'X)^{-1}X'(X\beta + e) - \beta = (X'X)^{-1}X'e.$$

Assumptions for OLS

Model

$$Y = X\beta + e.$$

Assumption 1 Exogeneity: $E(e|X) = 0$.

Assumption 2 Data $(Y_i, X_{1,i}, ..., X_{k,i})$ are i.i.d. draws from $k+1$ dimensional distribution.

Assumption 3 Large outliers are unlikely: $EX_{j,i}^4 < \infty$, $Ee_i^4 < \infty$.

Assumption 4 No perfect multicollinearity: no regressor is a perfect linear combination of the others.

Homoskedasticity is a historic assumption we will not impose or use. Here it is formulated for your awareness only:

$$\text{Homoskedasticity: } E(ee'|X) = \sigma^2 I_n,$$

here I_n is an $n \times n$ identity matrix. All comments from bivariate regression about homoskedasticity versus heteroskedasticity apply here as well. Please, remember that STATA uses the wrong default(!!!!!) One can prove that if Assumptions 1-4 hold and, in addition, homoskedasticity is also satisfied, then OLS is the best linear unbiased estimator (BLUE).

Statement 1. If Assumptions 1-4 hold then the OLS estimator is unbiased, consistent and asymptotically gaussian.

Proof.

$$E\widehat{\beta} - \beta = E\left[(X'X)^{-1}X'e\right] = E\left(E\left[(X'X)^{-1}X'e|X\right]\right) = E\left((X'X)^{-1}X'\left[e|X\right]\right) = 0.$$

Let us also introduce the notation for one draw: $X_i = (1, X_{1,i}, ..., X_{k,i})$ is $1 \times (k+1)$ vector.

$$\widehat{\beta} - \beta = \left(\frac{1}{n}X'X\right)^{-1}\frac{1}{n}X'e = \left(\frac{1}{n}\sum_{i=1}^n X_i'X_i\right)^{-1}\frac{1}{n}\sum_{i=1}^n X_i'e_i.$$

We use the Law of Large Numbers:

$$\frac{1}{n}\sum_{i=1}^n X_i'X_i \to^p E[X_i'X_i] = \Sigma_X; \quad \frac{1}{n}\sum_{i=1}^n X_i'e_i \to^p E[X_i'e_i] = (0, ..., 0)'.$$

The last equality is due to Assumption 1. Thus

$$\widehat{\beta} \to^p \beta.$$

Finally,

$$\sqrt{n}(\widehat{\beta}-\beta)=\left(\frac{1}{n}X'X\right)^{-1}\frac{1}{\sqrt{n}}X'e=\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_i'e_i.$$

We consider the last sum and apply the Central Limit Theorem:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_i'e_i=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i.$$

Notice that ξ_i are i.i.d. draws of a random vector with $E\xi_i=0$ (due to Assumption 1) and $Var(\xi_i)=E[e_i^2X_i'X_i]=Q$. Thus,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i\Rightarrow N(0,Q).$$

Thus, we get

$$\sqrt{n}(\widehat{\beta}-\beta)\Rightarrow N(0,\Sigma_X^{-1}Q\Sigma_X^{-1}).$$

We also can get a consistent estimator of the covariance matrix:

$$\widehat{\Sigma}_\beta=\left(\frac{1}{n}\sum_{i=1}^{n}X_i'X_i\right)^{-1}\frac{1}{n-k-1}\sum_{i=1}^{n}\widehat{e}_i^2X_i'X_i\left(\frac{1}{n}\sum_{i=1}^{n}X_i'X_i\right)^{-1}.$$

This is a heteroskedasticity-robust formula. □

Measure of fit

The predicted values and residuals are defined similarly to the bivariate case:

$$\widehat{Y}_i=X_i\widehat{\beta}=\widehat{\beta}_0+\widehat{\beta}_1X_{1,i}+...+\widehat{\beta}_kX_{k,i};\quad \widehat{e}_i=Y_i-\widehat{Y}_i.$$

The following quantities are defined in the same way as before:

$$TSS=\sum_{i=1}^{n}(Y_i-\bar{Y})^2;\qquad ESS=\sum_{i=1}^{n}(\widehat{Y}_i-\bar{Y})^2;\qquad SSR=\sum_{i=1}^{n}\widehat{e}_i^2.$$

Also as before we can prove that

$$TSS=ESS+SSR.$$

We define

$$R^2=\frac{ESS}{TSS}=1-\frac{SSR}{TSS}.$$

However, if one adds an additional regressor to a regression, then R^2 will always increase even if the regressor is not really related to the dependent variable. It happens, because the smaller model is nested in the large one, and the unconstrained minimization leads to smaller minimum values than does the constrained version. Define $b=(\bar{b},b_k)$, where $\bar{b}=(b_0,b_1,...,b_{k-1})$. Consider the OLS optimization problem

$$SSR_{k+1}=\min_b S(b)\le\min_{\bar{b}}S(\bar{b},0)=SSR_k,$$

where SSR_{k+1} is the sum of squared residuals in a regression with $k+1$ regressors, while SSR_k is from the regression without the last regressor.

This gives us that R^2 always increases with the size of the model, not just with a fit. In order to account for the size of the model, adjusted R^2 is introduced:

$$R^2_{adj}=1-\frac{n-1}{n-k-1}\cdot\frac{SSR}{TSS}=1-\frac{s_e^2}{s_Y^2}<R^2.$$

Adjusted R^2 may take negative values.

Remember, that 'maximizing R^2_{adj}' or 'maximizing R^2' is rarely the answer to any meaningful economic question.

- Increase in R^2 or R^2_{adj} does not mean that the added variable is significant.
- High $R^2(R^2_{adj})$ does not mean that the regressors cause the depenent variable.
- High R^2 or R^2_{adj} does not mean that there is no omitted variable bias
- We add regressors in order to fix omitted variable bias, not to increase R^2.

Testing

Test of a hypothesis about a single coefficient

If the null hypothesis is about one coefficient, we do our testing as before using t-statistics. For example, if

$$H_0 : \beta_j = \beta_{j,0} \quad vs \quad H_1 : \beta_j \neq \beta_{j,0}.$$

Consider the (heteroskedasticity-robust) variance estimate $\widehat{\Sigma}_\beta$, it is a $(k+1) \times (k+1)$ dimensional matrix which contains on its diagonal estimates of the asymptotic variance of individual coefficient estimates.

$$s.e.(\widehat{\beta}_j) = \frac{1}{\sqrt{n}}(\widehat{\Sigma}_\beta)_{(j+1,j+1)}.$$

Then we form the t- statistics $t = \frac{\widehat{\beta}_j - \beta_{j,0}}{s.e.(\widehat{\beta}_j)}$. We reject the null at the 95% significance level iff $|t| > 1.96$.

A 95% confidence set for one coefficient is $[\widehat{\beta}_j - 1.96 s.e.(\widehat{\beta}_j), \widehat{\beta}_j + 1.96 s.e.(\widehat{\beta}_j)]$.

Testing a hypothesis about a one-dimensional linear combination of coefficients

Imagine that one wants to test the following one-dimensional hypothesis:

$$H_0 : a_0\beta_0 + a_1\beta_1 + ... + a_k\beta_k = C \quad vs \quad H_1 : a_0\beta_0 + a_1\beta_1 + ... + a_k\beta_k \neq C;$$

another way of writing the null is $a'\beta = C$.

Example. Imagine one is interested in racial wage gap and runs a regression:

$$Wage_i = \beta_0 + \beta_1 Black_i + \beta_2 Hispanic_i + e_i,$$

where both regressors are dummy variables. Then

$$E(Wage|not\ Black \ \ or \ \ Hispanic) = \beta_0; \quad E(Wage|Black) = \beta_0 + \beta_1; \quad E(Wage|Hispanic) = \beta_0 + \beta_2.$$

If one wants to test whether there is a wage gap between black and hispanic populations, then the hypothesis of interest is $H_0 : \beta_1 = \beta_2$, which is equivalent to $H_0 : \beta_2 - \beta_1 = 0$. □

In order to test a linear one-dimensional hypothesis let us define a new parameter $\gamma = a'\beta$, then estimator $\widehat{\gamma} = a'\widehat{\beta}$ is unbiased, consistent, and asymptotically gaussian with asymptotic variance

$$AVar(\widehat{\gamma}) = a' AVar(\widehat{\beta}) a.$$

Thus

$$s.e.(\widehat{\gamma}) = \frac{1}{\sqrt{n}}\sqrt{a'\widehat{\Sigma}_\beta a}.$$

As before the t-statistics of interest is $t = \frac{\widehat{\gamma} - C}{s.e.(\widehat{\gamma})}$.

Notice that in our example above $\gamma = \beta_2 - \beta_1$, and $Var(\widehat{\gamma}) = Var(\widehat{\beta}_2) + Var(\widehat{\beta}_1) + 2cov(\widehat{\beta}_1, \widehat{\beta}_2)$. Also, notice that matrix $\widehat{\Sigma}_\beta$ contains estimated covariances as off-diagonal elements.

Testing multiple restrictions

Imagine that you wish to test several restrictions simultaneously. For example,

$$H_0 : \beta_1 = \beta_{1,0}, \beta_2 = \beta_{2,0}, ..., \beta_q = \beta_{q,0};$$

the dimensionality of the null hypothesis is equal to the number of linearly independent restrictions. In the example above it is q. What is a proper alternative, and can we do multiple sequential testing (testing all restrictions one-by-one)?

Let's consider a simple case of

$$H_0 : \beta_1 = 0 \text{ and } \beta_2 = 0.$$

The proper alternative in this case is the negation of the null statement, that is,

$$H_1 : \text{ either } \beta_1 \neq 0, \text{ or } \beta_2 \neq 0, \text{ or both.}$$

The answer to the second question is 'no, we cannot do sequential testing'. Let us discuss why.

We form two t-statistics: $t_1 = \frac{\widehat{\beta}_1}{s.e.(\widehat{\beta}_1)}$ and $t_2 = \frac{\widehat{\beta}_2}{s.e.(\widehat{\beta}_2)}$. Imagine that we perform sequential testing, that is we reject H_0 when either $|t_1| > 1.96$ or $|t_2| > 1.96$ or both. This test is not of the correct size. Indeed,

$$P\{|t_1| > 1.96 \text{ or } |t_2| > 1.96\} > P\{|t_1| > 1.96\} = 0.05.$$

In fact, if t_1 and t_2 are independent random variables

$$P\{|t_1| > 1.96 \text{ or } |t_2| > 1.96\} = 1 - P\{|t_1| \leq 1.96 \text{ and } |t_2| \leq 1.96\} = 1 - P\{|t_1| \leq 1.96\} \cdot P\{|t_2| \leq 1.96\} = 1 - 0.95 \cdot 0.95 = 0.0975.$$

If t_1 and t_2 are dependent, then it will be more complicated.

The correct approach to test a multidimensional hypothesis is an F-test. Specifically, for the two-dimensional hypothesis above

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho} t_1 t_2}{1 - \hat{\rho}^2}\right),$$

here $\hat{\rho}$ is an estimate of the correlation between t_1 and t_2. For example, if t_1 and t_2 are independent then

$$F = \frac{1}{2}(t_1^2 + t_2^2) \Rightarrow \frac{\chi_2^2}{2}.$$

In a more general multi-dimensional case when the hypothesis is $H_0 : A\beta = C$, where A is $q \times (k+1)$ matrix of rank q, the estimate of tested parameters is $A\widehat{\beta}$, its estimate of asymptotic variance matrix is $A\widehat{\Sigma}_\beta A'$- which is $q \times q$. The proper F-statistics is :

$$F = \frac{1}{q}(\widehat{\beta}' A' - C')(A\widehat{\Sigma}_\beta A')^{-1}(A\widehat{\beta} - C).$$

Under the null the distribution of this statistics is $F_{q,\infty} = \frac{\chi_q^2}{q}$. Both the F and χ^2 distributions are tabulated. The students are responsible for knowing how to use tables for testing. The null is rejected if the value of the F statistic exceeds the 95% quantile of the $F_{q,\infty}$ distribution.

Discussion of F test in homoskedastic case

The F statistic has a different interpretation in the homoskedastic case, which is useful for forming an intuition about what it is doing. Imagine for concreteness that we have regression

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + e_i, \tag{1}$$

and the hypothesis tested is $H_0 : \beta_1 = 2$, and $\beta_2 = 0$. Then regression (1) is called unrestricted, and the sum of the squared residuals from this regression is denoted as SSR_U. The restricted regression is one in which the null hypothesis is imposed, that is:

$$Y_i = \beta_0 + 2X_{1,i} + \beta_3 X_{i,3} + e_i.$$

The OLS in such a case minimizes the same sum of squares but subject to restriction imposed by the null:

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_{1,i} - b_2 X_{2,i} - b_3 X_{3,i})^2 \to \min_{b:b_1=2,b_2=0}.$$

The minimum achieved is SSR_R and is always bigger than SSR_U.

$$F = \frac{(SSR_R - SSR_U)/q}{SSR_U/(n-k-1)} = \frac{(R_U^2 - R_R^2)/q}{(1-R_U^2)/(n-k-1)},$$

here $q = 2$ is the number of restrictions tested and $k = 3$ the number of regressors. Thus the F statistic characterizes how much harder it is to fit the regression with the restriction than without. The F-statistics characterizes how binding the restrictions are.

A comment about F vs t test

The F-test, if applied to a one-dimensional restriction, is equivalent to a t-test with a two sided alternative. For a multi-dimensional restriction, F- test and several t-tests often give incompatible results, which may go either way: for a given result of an F test you may obtain any combination of acceptances/rejections of individual t tests. You should rely on the F test.

Controls vs variables of interest

Imagine that one is interested in the causal effect of X on Y and decides to run a regression

$$Y = \alpha + \beta X + e.$$

Unfortunately, as she recognizes there is an omitted variable bias due to the variable W, to fix it, she runs:

$$Y = \alpha + \beta X + \gamma W + e. \tag{2}$$

But then she realizes that another variable, Z, is a determinant of Y and is correlated with W (though not with X). The researcher thinks that now Z would lead to omitted variable bias, should she add Z to the regression? When does this process stop?

Apparently, we can make a distinction between X (variable of interest) and W (a control), as we are interested in estimating β well, while we do not care if γ is consistent or estimates the causal effect (or if there is any causal effect from W)- we include W to the regression only to fix the omitted variable bias.

We can use an alternative assumption in place of Assumption 1.

Assumption 1' : $E[e|X, W] = E[e|W]$.

It means that if we consider the sub-population with the same value of $W = w$ (this is conditioning) then for this sub-population e is uncorrelated with any function of X. This assumption is weaker than Assumption 1, as we do not force $E[e|W] = 0$.

If Assumption 1' holds instead of Assumption 1, then we can prove that $\widehat{\beta}$ is unbiased, consistent, asymptotically gaussian, but $\widehat{\gamma}$ may be not consistent for the causal coefficient on W. Thus, if one wants to argue for an ommitted variable bias to β, s/he should find a variable Z which is a determinant of Y and correlated with X even when one consider sub-populations with fixed values of the controls.

Slides with empirical example

Slides available on the stellar web-site. Questions to discuss:

- What is the causal effect we are trying to estimate? What is a mental experiment to evaluates it? Can you run this experiment?
- Bivariate regression: interpretation of coefficients, are they of correct sign? Is the size expected?
- Do you expect omitted variable bias? Can smoking cause omitted variable bias? What direction of the bias you expect? Same question for 'attention to your health'
- Longer regressions: does inclusion of additional variables change the coefficient of interest in any reasonable way? Do you think there was OVB that you fixed? What regression would you prefer?
- Inclusion of a group of dummy variables: one can test whether they are significant as a group by using an F test. Compare the F test to multiple t tests. Which one you prefer? Do you think it makes much sense to include some but not all dummies?
- Is the coefficient on *unmarried* causal or not? Do we care?
- Notice how to report your results in a table. Check that you understand all entries.