Econ 272

Midterm II Study Guide

The Classical Assumptions

I The regression model is linear, is correctly specified, and has an additive error term.

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_k X_{i,k} + \epsilon_i$$

II The error term has a zero population mean.

$$E[e_i] = 0 \text{ or } E[e_i \mid X_{i,j}] = 0.$$

III All explanatory variables are uncorrelated with the error term. (Strict Exogeneity)

$$\text{Cov}(X_{i,j}, \epsilon_i) = 0 \ \forall j.$$

IV Observations of the error term are uncorrelated with each other (no serial correlation).

$$\text{Cov}(\epsilon_j, \epsilon_i) = 0 \ \forall i \neq j.$$

V The error term has a constant variance. (Homoskedasticity).

$$\text{Var}(\epsilon_i) = \sigma^2 \ \forall i.$$

VI No explanatory variable is a perfect linear function of any other explanatory variable(s).

Omitted Variable Bias

Definition. An *omitted variable* is an important explanatory variable that has been left out of a regression equation.

Definition. An *omitted variable bias* is the bias caused by leaving an omitted variable out of an OLS estimation.

Example. Suppose a true regression model is given by:

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

If we omit X_2 from the equation, we get:

$$Y_i = \beta_0^* + \beta_1^* X_{i,1} + \epsilon_i^*.$$

Recall that the *stochastic error term* is a term which is added to a regression equation to introduce all the variation of Y that cannot be explained by the included X's. This means the stochastic error term includes the effects of any omitted variables, giving:

$$\epsilon_i^* = \epsilon_i + \beta_2 X_{i,2}.$$

There is a reason our omitted variable equation includes β_0^* and β_1^*. Note that β_1 is the impact of a one-unit increase in X_1 on Y *while holding* X_2 *constant*. But since X_2 isn't in our omitted variable equation, the OLS can't hold it constant, and as a result β_1^* is the impact of a one-unit increase in X_1 on Y *not holding* X_2 *constant*.

If we leave an important variable out of an equation, we violate Classical Assumption III. Most pairs of variables are correlated to some degree, so X_1 and X_2 from our true regression model are almost surely correlated. When X_2 is omitted from the equation, the impact of X_2 goes into ϵ^*, so ϵ^* and X_2 are correlated, violating strict exogeneity.

These paragraphs use words to describe why Classical Assumption III fails, which I think is dumb. From the above equations, note that:

$$\begin{aligned} E[\epsilon_i^*] &= E[\epsilon_i + \beta_2 X_{i,2}] \\ &= E[\epsilon_i] + E[\beta_2 X_{i,2}] \\ &= \beta_2 E[X_{i,2}]. \end{aligned}$$

If β_2 and $E[X_{i,2}]$ are nonzero, then

Example. We want to quantify the amount and direction of bias. Suppose

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

is our true regression equation. Since we know most pairs of variables are correlated, we can infer that:

$$X_{i,2} = \alpha_0 + \alpha_1 X_{i,1} + u_i.$$

Subbing this into our true regression equation, we get:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i,1} + \beta_2(\alpha_0 + \alpha_1 X_{i,1} + u_i) + \epsilon_i \\ &= (\beta_0 + \beta_2\alpha_0) + (\beta_1 + \beta_2\alpha_1) X_{i,1} + (\epsilon_i + \beta_2 u_i). \end{aligned}$$

The *bias* is given by $\beta_2\alpha_1$. Furthermore, note that:

$$\begin{aligned} E[\widehat{\beta_1}] &= E[\beta_1 + \beta_2\alpha_1] \\ &= \beta_1 + \beta_2\alpha_1. \end{aligned}$$

In general, this bias exists unless:

1. the true coefficient of our omitted variable equals zero;
2. the included and omitted variables are uncorrelated in the sample.

Question. Consider the following population regression function:

$$\text{Unemp}_{cs} = \beta_0 + \beta_1 \text{MinWage}_s + \delta \mathbf{X}_{cs} + \epsilon_{cs}.$$

Regression using OLS provides $\widehat{\beta_1} = -0.024$. An omitted variable is the history of labour movements in the state. Assume that states with a strong labour movement have higher minimum wages, or $\text{Cov}(\text{Unions}_s, \text{MinWage}_s) > 0$. Is the estimated β_1 an over-estimate or an under-estimate of the true impact of minimum wage policies on unemployment? Show your steps and clearly state any assumptions you make.

Solution. Note that our "true" regression function would be:

$$\text{Unemp}_{cs} = \beta_0 + \beta_1 \text{MinWage}_s + \beta_2 \text{Union}_s + \delta \mathbf{X}_{cs} + \epsilon_{cs}.$$

Typical economic theory presumes higher labour movements results in higher unemployment, whence $\beta_2 > 0$. Since most pairs of variables are correlated to some degree, we can assume:

$$\text{MinWage}_s = \alpha_0 + \alpha_1 \text{Unions}_s + \epsilon_s.$$

By the problem statement, we can assume $\alpha_1 > 0$. Thus our bias, which is denoted by $\alpha_1\beta_2$, must be positive. Whence:

$$\begin{aligned} -0.024 &= \beta_1 + \beta_2\alpha_1 \\ &\iff \\ -0.024 - \beta_2\alpha_1 &= \beta_1. \end{aligned}$$

Thus our estimator is []-biased.

Interaction Terms

Definition. An *interaction term* is an independent variable in a regression equation that is a multiple of two or more other independent variables.

For this section our interaction terms will strictly be the product of two independent variables. Interaction terms can involve two quantitative variables, or two dummy variables, but the most frequent application if interaction terms involves one quantitative variable and one dummy variable.

Example. We will often be asked to interpret the coefficients of an OLS estimation containing interaction terms. We start with a very general example. Let:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 X_i D_i + \epsilon_i$$

be a given linear model with X_i a quantitative variable and D_i a dummy variable. There are only two cases to consider.

Case 1: $D_i = 1$. Then:

$$Y_i = (\beta_0 + \beta_1) + (\beta_1 + \beta_3)X_i + \epsilon_i.$$

From this, we can see that β_3 represents a "change of change"; it measures how much Y_i changes per unit-change of X_i *assuming* D_i is true. Another interpretation would be it measures how much the effect of X_i for D_i differs from the effect of X_i without D_i. Similarly, our new constant term, $\beta_0 + \beta_3$, represent the mean value of the dependent variable when D_i is true.

Case 2: $D_i = 0$. Then:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

Notice that β_1 isn't merely the change in Y_i per unit change in X_i, rather it is the *change in* Y_i *per unit change in* X_i *when* D_i *is false.* The constant term β_0 is the mean value of our dependent variable assuming D_i to be false.

Example.

Linear Probability Models

Definition. A *linear probability model* is a OLS equation used to explain a dummy dependent variable:

$$D_i = \beta_0 + \beta_1 X_{i,1} + ... + \beta_k X_{i,k} + \epsilon_i.$$

The term *probability model* comes from the fact that taking the expected value of a dummy variable measures the probability that $D_i = 1$. This means:

$$\begin{aligned} E[D_i] = P(\widehat{D_i = 1}) &= E[\beta_0 + \beta_1 X_{i,1} + ... + \beta_k X_{i,k} + \epsilon_i] \\ &= \widehat{\beta_0} + \widehat{\beta_1} X_{i,1} + ... + \widehat{\beta_k} X_{i,k}, \end{aligned}$$

where $P(D_i = 1)$ indicates the probability that $D_i = 1$ for the i^{th} observation. So a unit change of X_i results in a β change in probability for $D_i = 1$ to occur.

Using OLS to estimate the coefficients of an equation with a dummy dependent variable faces at least three problems:

(1) $\overline{R}^2$ is not an accurate measure of overall fit.

(2) $\widehat{D_i} = P(\widehat{D_i = 1})$ is not bounded by 0 and 1. Any prediction that a probability equals something less than zero or greater than one is meaningless.

(3) The error term is neither homoskedastic nor normally distributed. In practice these problems on OLS estimation is minor, so it is typically ignored.

Fixed Effects

Definition. *Panel data*, or *longitudinal data* combines time-series and cross-sectional, by including observations on the same variables from the same cross-sectional sample from two or more different time periods.

Example. Suppose we surveyed 200 students when they graduated from college and then administered the same questionnaire to each student five years later. This would be a panel data set.

We are interested in estimating panel data equations.

Definition. The *fixed effects model* estimates panel data equations by including enough dummy variables to allow each cross-sectional entity (like a state or country) and each time period to have a different intercept:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \alpha_2 E_2 + ... + \alpha_n E_n + \rho_2 T_2 + ... + \rho_m T_m + \epsilon_{it},$$

where each E_i is an entity fixed dummy variable (equal to 1 for the ith entity and 0 otherwise) and T_i is a time fixed dummy variable (equal to 1 for the ith period and 0 otherwise).

There's a reason this equation looks so complicated (even though it's just $n + m$ added dummy variables). If we estimated our model without accounting for the fact that our observations are from a panel data set. Then our equation looks like:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + V_{it},$$

where V_{it} represents the error term. To understand V, it's best to look at an example which is less general. Suppose our cross-sectional data is on the 50 U.S. states, and suppose our time-series data is from the years 2000-2010. Clearly no two states are alike —they have different cultures, histories, institutions, governments, etc. Likewise, a state's history and culture are relatively constant from year to year. Even if they are impossible to measure, we know that they don't change a lot, and in particular they distinguish each state from all the others. It is very likely that the unchanging and unmeasured differences between states are correlated with X_{it}, giving us omitted variable bias. Furthermore, it's also likely that the time-series data would add even more omitted variables (see example in text).

The solution lies in examining what V_{it} represents. We can break it into three components:

$$V_{it} + \epsilon_{it} + a_i + z_t,$$

where ϵ_{it} is the classical error term, a_i refers to the entity characteristics omitted from the equation, and z_t refers to the time characteristics omitted from the equation. If a_i and z_t are correlated with X_{it}, we will violate Classical Assumption III, and our estimate of β_1 will be biased. But simply including dummy variables for every entity (but one) and every time period (but one), we can control for the unchanging entity-effects and the time-fixed effects. Including each of the n dummy variables for every entity and each of the m dummy variables for each time-period results in the entity and time fixed effects not being omitted variables (because they are represented by dummy variables). We arrive at our original equation with the $m + n$ dummy variables.

The major advantages of the fixed effects model is that it avoids bias due to omitted variables that don't change over time, or that change over time equally for all entities. The beauty lies in the fact we don't have to know exactly what things go into the entity and time fixed effects, the dummy variables include them all.

There are drawbacks, however. No substantive explanatory variable that varies across entities, but not over time within each entity, can be used. They would create perfect multicollinearity.