Lecture 6

The Multiple regression Model

Costas Meghir

The Model and its interpretation

- The Multiple regression model takes the form

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_k X_{ik} + u_i$$

- There are *k* regressors (explanatory Variables) and a constant
- Hence there will be *k*+1 parameters to estimate

Assumption M.1

- We will keep the **basic least squares assumption** - We will assume that the **error term is mean independent of all regressors** (loosely speaking - all *X*s are uncorrelated with the error term, i.e.

$$E(u_i \mid X_1, X_2, ..., X_k) = E(u_i \mid X) = 0$$

Interpretation of the coefficients

- Since the error term is mean independent of the *Xs* (M.1) varrying the X's does not have an impact on the error term.

- Thus under Assumption M.1 the coefficients in the regression model have the following simple interpretation:

$$b_j = \frac{\partial Y_i}{\partial X_{ij}}$$

- Thus each coefficient measures the impact of the corresponding *X* on *Y* **keeping all other factors (*Xs* and u) constant.** A *ceteris paribus* effect.

Example:
Male wages at 33 and the Student Teacher ratio in Secondary School
National Child Development Survey

Data on all people born in the second week of March 1958

```
. regress lhw_5 strat_3 lowabil payrsed mayrsed

                                     Number of obs =    1523
                                      F(  4,  1518) =   32.59
                                      Prob > F      =  0.0000
                                      R-squared     =  0.0761
                                      Root MSE      =  .40571

------------------------------------------------------------------------------

Log Wage rate at 33 (Men)         lhw_5 |      Coef.      Std. Err.       t       P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 Student Teacher Ratio (sec sch) strat_3 |  -.0231986   .005419      -4.28   0.000    -.0338281    -.012569
 Below median ability           lowabil |  -.1870087   .0216397    -8.64   0.000    -.2294554     -.1445619
 Father's Years of education    payrsed |   .012372    .0047924     2.58   0.010     .0029716      .0217723
 Mother's Years of education    mayrsed|  -.0058386   .0050984    -1.15   0.252    -.0158393      .0041621
                                  _cons | 2.476687     .0952986    25.99   0.000    2.289756      2.663618
------------------------------------------------------------------------------
```

Least Squares in the Multiple Regression Model

- We maintain the same set of assumptions as in the two variable regression model.
- We modify assumption 1 to assumption M1 to take into account the existence of many regressors.
- The OLS estimator is chosen to minimise the residual sum of squares exactly as before.
- Thus $b_0, b_1, b_2, ..., b_k$ are chosen to minimise

$$S = \sum_{i=1}^{N} u_i^2 = \sum_{i=1}^{N} (Y_i - b_0 - b_1 X_{i1} - b_2 X_{i2} - ... - b_k X_{ik})^2$$

The Normal Equations

- Differentiating *S* with respect to each coefficient in turn we obtain a set of *k*+1 equations constituting the **first order conditions** for minimising the residual sum of squares *S*. These equations are called the **Normal Equations.**

The Normal Equations

$$\frac{\partial S}{\partial b_0} = -2\sum_{i=1}^{N} (Y_i - b_0 - b_1 X_{i1} - \ldots - b_k X_{ik}) = 0$$

$$\frac{\partial S}{\partial b_1} = -2\sum_{i=1}^{N} X_{i1}(Y_i - b_0 - b_1 X_{i1} - \ldots - b_k X_{ik}) = 0$$

.

.

$$\frac{\partial S}{\partial b_k} = -2\sum_{i=1}^{N} X_{ik}(Y_i - b_0 - b_1 X_{i1} - \ldots - b_k X_{ik}) = 0$$

- Solving the normal equations for $b_0, b_1, b_2, ..., b_k$ provides the OLS estimator.
- We have dealt with the special case of k=1.
- From the first equation corresponding to the constant we get that

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}_1 - ... - \bar{b}_k \bar{X}_k$$

- In the above the *bar* denotes sample mean and the *hat* denotes the solution to the normal equations.
- This is a direct generalisation of the result for the constant term that we had in the two variable regression model.
- We substitute this expression in the remaining equations and obtain

$$\sum_{i=1}^{N} X_{i1}((Y_i - \overline{Y}) - b_1(X_{i1} - \overline{X}_1) - \ldots - b_k(X_{ik} - \overline{X}_k)) = 0$$

$$\cdot$$

$$\cdot$$

$$\sum_{i=1}^{N} X_{ik}((Y_i - \overline{Y}) - b_1(X_1 - \overline{X}_1) - \ldots - b_k(X_k - \overline{X}_k)) = 0$$

A solution for two regressors

- With two regressors this represents a two equation system with two unknowns, i.e. b_1, b_2
- We have already solved for the constant term
- The solution for b_1 is

$$b_1 = \frac{\sum_{i=1}^{N} (X_{i2} - \overline{X}_2) X_{i2} \sum_{i=1}^{N} (Y_i - \overline{Y}) X_{1i} - \sum_{i=1}^{N} (X_{i2} - \overline{X}_2) X_{i1} \sum_{i=1}^{N} (Y_i - \overline{Y}) X_{2i}}{\sum_{i=1}^{N} (X_{i2} - \overline{X}_2) X_{i2} \sum_{i=1}^{N} (X_{i1} - \overline{X}_1) X_{1i} - \sum_{i=1}^{N} (X_{i2} - \overline{X}_2) X_{i1} \sum_{i=1}^{N} (X_{i1} - \overline{X}_1) X_{2i}}$$

- This formula can also be written as

$$b_1 = \frac{\operatorname{cov}(Y, X_1)Var(X_2) - \operatorname{cov}(X_1 X_2)\operatorname{cov}(Y, X_2)}{Var(X_1)Var(X_2) - \operatorname{cov}(X_1 X_2)^2}$$

- Similarly we can derive the formula for the other coefficient (b_2)
- Note that the formula for b_1 is now different from the formula we had in the two variable regression model. This now takes into account the presence of the other regressor(s)
- The extent to which the two formulae differ depends on the covariance of X_1 and X_2.
- When this covariance is zero we are back to the formula for the one variable regression model.
- This result is of significance and will be discussed in the context of ommitted variable bias in a later lecture

Assumption M.4

- The Gauss Markov Theorem is valid for the multiple regression model. We need however to modify assumption A.4.
- Define the covariance matrix of the regressors *X* to be

$$\operatorname{cov}(X) = \begin{bmatrix} \operatorname{var}(X_1) & \operatorname{cov}(X_1, X_2) & \ldots & \operatorname{cov}(X_1, X_k) \\ \operatorname{cov}(X_1, X_2) & \operatorname{var}(X_2) & \ldots & \operatorname{cov}(X_2, X_k) \\ . & . & . & . \\ \operatorname{cov}(X_1, X_k) & \operatorname{cov}(X_2, X_k) & & \operatorname{var}(X_k) \end{bmatrix}$$

- We assume that *cov(X)* **positive definite** and hence can be inverted.

The Gauss Markov Theorem

- **Theorem**: Under Assumptions M.1 A.2 and A3 and M.4 the Ordinary Least Squares Estimator (OLS) is Best in the class of Linear Unbiased estimators (BLUE).
- As before this means that OLS provides estimates that are least sensitive to changes in the data - given the stated assumptions.

The Coefficient of Determination R^2

- The R^2 is defined in exactly the same way as in the two variable regression model and measures the goodness of fit of the model.
- By goodness of fit we mean the proportion of the variance of the dependent variable that is explained by the model.

Omitted Variable Bias

- Suppose the true regression relationship has the form

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + u_i$$

- Instead we decide to estimate

$$Y_i = b_0 + b_1 X_{i1} + v_i$$

- We will show that *in general* this omission will lead to a biased estimate of X_1

- Suppose we use OLS on the second equation. As we know we will obtain:

$$\tilde{b}_1 = b_1 + \frac{\sum_{i=1}^{N} (X_{1i} - \bar{X}_1) v_i}{\sum_{i=1}^{N} (X_{1i} - \bar{X}_1)^2}$$

- The question is : What is the expected value of the last expression on the right hand side. For an unbiased estimator this will be zero. Here we will show that it is not zero.
- First note that according to the **true model** (i.e. the model on the top of the previous slide) we have that

$$v_i = b_2 X_{i2} + u_i$$

- We can substitute this into the expression for the OLS estimator to obtain

$$\tilde{b}_1 = b_1 + \frac{1}{\sum_{i=1}^{N} (X_{1i} - \overline{X}_1)^2} \left(\sum_{i=1}^{N} (X_{1i} - \overline{X}_1) b_2 X_{2i} + \sum_{i=1}^{N} (X_{1i} - \overline{X}_1) u_i \right)$$

- Now we can take expectations of this expression.

$$E(\tilde{b}_1 \mid X) = b_1 + \frac{1}{\sum_{i=1}^{N} (X_{1i} - \overline{X}_1)^2} \left(\sum_{i=1}^{N} E\left[(X_{1i} - \overline{X}_1) b_2 X_{2i} \mid X \right] + \sum_{i=1}^{N} E\left[(X_{1i} - \overline{X}_1) u_i \mid X \right] \right)$$

- The last expression is zero under the assumption that *u* is mean independent of *X* [Assumption M.1]

The omitted variable bias expression

- Thus we are left with an expression for the OLS. The last term on the right hand side is now a bias term due to the omission of a regressor.

$$E(\tilde{b}_1 \mid X) = b_1 + \frac{1}{\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)^2} b_2 \left(\sum_{i=1}^{N} E\left[(X_{1i} - \bar{X}_1) X_{2i} \mid X \right] \right)$$

- This expression can be written more compactly as

$$E(\tilde{b} \mid X) = b_1 + b_2 \frac{\operatorname{cov}(X_1, X_2)}{Var(X_1)}$$

- The bias will be zero in two cases:

  - When the coefficient b_2 is zero. In this case the regressor X_2 obviously does not belong to the regression.

  - When the covariance between the two regressors X_1 and X_2 is zero

- Thus in general omitting regressors which have an impact on Y (b_2 non-zero) will bias the OLS estimator of the coefficients on the included regressors **unless the omitted regressors are uncorrelated with the included ones**.