Lecture 3: Multiple Regression

R.G. Pierse

1 The General Linear Model

Suppose that we have k explanatory variables

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i \quad , \quad i = 1, \cdots, n \tag{1.1}$$

or

$$Y_i = \sum_{j=1}^{k} \beta_j X_{ji} + u_i \quad , \quad i = 1, \cdots, n$$

where $X_{1i} = 1 \quad , \quad i = 1, \cdots, n$ is an intercept.

1.1 Assumptions of the Model

$$E(u_i) = 0 \quad , \quad i = 1, \cdots, n \tag{A1}$$

$$E(u_i^2) = \sigma^2 \quad , \quad i = 1, \cdots, n \tag{A2}$$

$$E(u_i u_j) = 0 \quad , \quad i, j = 1, \cdots, n \quad j \neq i \tag{A3}$$

$$\text{X values are } \textit{fixed in repeated sampling} \tag{A4}$$

We now need to make an additional assumption about the explanatory variables:

$$\text{The variables } X_j \text{ are } \textit{not perfectly collinear.} \tag{A4$'$}$$

Perfect collinearity occurs when there is one or more variables X_m such that

$$X_{mi} = \sum_{j \neq m} c_j X_{ji} \quad , \quad i = 1, \cdots, n$$

where c_j are fixed constants. In this case, one of the variables is redundant and should be dropped from the regression.When a set of variables is perfectly collinear, dropping any one of them will solve the problem. An example of a perfectly collinear set of variables would be GDP plus all the components of the GDP identity.

Y_i is a random variable with the following properties:

$$E(Y_i) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} \quad , \quad i = 1, \cdots, n$$

$$\begin{aligned} Var(Y_i) &= E(Y_i - E(Y_i))^2 \\ &= E(u_i^2) = \sigma^2 \quad , \quad i = 1, \cdots, n \end{aligned}$$

2 Interpretation of the OLS Coefficients

β_j represents the effect of changing the variable X_j on the dependent variable Y *while holding all other variables constant*. In mathematical terms this is the partial derivative of Y with respect to X_j

$$\beta_j = \frac{\partial Y}{\partial X_j}$$

3 The Ordinary Least Squares Estimator

$$Y_i = \sum_{j=1}^{k} \widehat{\beta}_j X_{ji} + e_i \quad , \quad i = 1, \cdots, n \tag{3.1}$$

The OLS estimator is the estimator that minimises the sum of squared residuals $s = \sum_{i=1}^{n} e_i^2$.

$$\min_{\widehat{\beta}_1, \cdots, \widehat{\beta}_k} \; s = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{k} \widehat{\beta}_j X_{ji} \right)^2$$

Differentiating this expression with respect to the k parameters $\widehat{\beta}_j$ gives k first order conditions:

$$\frac{\partial s}{\partial \widehat{\beta}_1} = -2 \sum_{i=1}^{n} \left(Y_i - \widehat{\beta}_1 - \widehat{\beta}_2 X_{2i} - \cdots - \widehat{\beta}_k X_{ki} \right) = 0 \tag{3.2}$$

$$\frac{\partial s}{\partial \widehat{\beta}_2} = -2 \sum_{i=1}^{n} \left(Y_i - \widehat{\beta}_1 - \widehat{\beta}_2 X_{2i} - \cdots - \widehat{\beta}_k X_{ki} \right) X_{2i} = 0 \tag{3.3}$$

and so on up to

$$\frac{\partial s}{\partial \widehat{\beta}_k} = -2\sum_{i=1}^{n}\left(Y_i - \widehat{\beta}_1 - \widehat{\beta}_2 X_{2i} - \cdots - \widehat{\beta}_k X_{ki}\right) X_{ki} = 0 \tag{3.4}$$

Note that the first order conditions can be rewritten as

$$\sum_{i=1}^{n} e_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} X_{ji} e_i = 0,\ j = 2, \cdots, k \tag{3.5}$$

Rearranging the equations gives the k normal equations in the k unknowns $\widehat{\beta}_j$

$$\sum_{i=1}^{n} Y_i = n\widehat{\beta}_1 + \widehat{\beta}_2 \sum_{i=1}^{n} X_{2i} + \cdots + \widehat{\beta}_k \sum_{i=1}^{n} X_{ki}$$

$$\sum_{i=1}^{n} Y_i X_{2i} = \widehat{\beta}_1 \sum_{i=1}^{n} X_{2i} + \widehat{\beta}_2 \sum_{i=1}^{n} X_{2i}^2 + \cdots + \widehat{\beta}_k \sum_{i=1}^{n} X_{2i} X_{ki}$$

and so on up to

$$\sum_{i=1}^{n} Y_i X_{ki} = \widehat{\beta}_1 \sum_{i=1}^{n} X_{ki} + \widehat{\beta}_2 \sum_{i=1}^{n} X_{2i} X_{ki} + \cdots + \widehat{\beta}_k \sum_{i=1}^{n} X_{ki}^2.$$

Solving these k equations defines the OLS estimators. Note that from the first normal equation

$$\widehat{\beta}_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i - \sum_{j=2}^{k} \widehat{\beta}_j \frac{1}{n}\sum_{i=1}^{n} X_{ji} = \overline{Y} - \sum_{j=2}^{k} \widehat{\beta}_j\, \overline{X}_j \tag{3.6}$$

where $\overline{Y} = \frac{1}{n}\sum_i^n Y_i$ and $\overline{X}_j = \frac{1}{n}\sum_i^n X_{ji}$ are the sample means of Y and X_j respectively.

3.1 Special case ($k = 3$)

In the general case, the formulae for the estimators $\widehat{\beta}_j$ are messy in summation notation. Elegant expressions require matrix notation. However, for the case $k = 3$ we have

$$\widehat{\beta}_2 = \frac{\sum y_i x_{2i} \sum x_{3i}^2 - \sum y_i x_{3i} \sum x_{2i} x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - \left(\sum x_{2i} x_{3i}\right)^2}$$

and

$$\widehat{\beta}_3 = \frac{\sum y_i x_{3i} \sum x_{2i}^2 - \sum y_i x_{2i} \sum x_{2i} x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - \left(\sum x_{2i} x_{3i}\right)^2}$$

where, as before, lower case letters denotes deviations from sample means.

4 Properties of the OLS Estimator

The OLS estimator in the general model has the same properties as in the simple regression model. Proofs will not be given here.

4.1 The OLS Estimator $\widehat{\beta}_j$ is Unbiased

It can be shown that

$$E(\widehat{\beta}_j) = \beta_j$$

4.2 OLS is BLUE

It can be shown that the OLS estimators $\widehat{\beta}_j$ are Best Linear Unbiased Estimators (BLUE) in the general regression model.

4.3 The Variance of OLS Estimators

For the general case, again the expressions are messy in summation notation, and elegant expressions require matrix notation. However, for the $k = 3$ case we have

$$Var(\widehat{\beta}_1) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{X}_2^2 \sum x_{3i}^2 + \overline{X}_3^2 \sum x_{2i}^2 - 2\overline{X}_2 \overline{X}_3 \sum x_{2i} x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - \left(\sum x_{2i} x_{3i}\right)^2} \right]$$

$$Var(\widehat{\beta}_2) = \frac{\sigma^2 \sum x_{3i}^2}{\sum x_{2i}^2 \sum x_{3i}^2 - \left(\sum x_{2i} x_{3i}\right)^2}$$

and

$$Var(\widehat{\beta}_3) = \frac{\sigma^2 \sum x_{2i}^2}{\sum x_{2i}^2 \sum x_{3i}^2 - \left(\sum x_{2i} x_{3i}\right)^2}.$$

We now also have an expression for the covariance between $\widehat{\beta}_2$ and $\widehat{\beta}_3$

$$Cov(\widehat{\beta}_2, \widehat{\beta}_3) = \frac{-\sigma^2 \sum x_{2i} x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - \left(\sum x_{2i} x_{3i}\right)^2}.$$

4.4 The OLS estimator of σ^2

It can be shown that the estimator

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - k} \tag{4.1}$$

is an unbiased estimator of σ^2.

5 Matrix Representation

Writing out all n observations in the model as a set of equations

$$\begin{aligned}
Y_1 &= \beta_1 X_{11} + \beta_2 X_{21} + \beta_3 X_{31} + \cdots + \beta_k X_{k1} + u_1 \\
Y_2 &= \beta_1 X_{12} + \beta_2 X_{22} + \beta_3 X_{32} + \cdots + \beta_k X_{k2} + u_2 \\
&\vdots \\
Y_n &= \beta_1 X_{1n} + \beta_2 X_{2n} + \beta_3 X_{3n} + \cdots + \beta_k X_{kn} + u_n
\end{aligned}$$

this can be rewritten in matrix notation as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{11} & X_{21} & \cdots & X_{k1} \\ X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

or

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$$

where $\mathbf{y}$ and $\mathbf{u}$ are an $n \times 1$ column vectors, β is a $k \times 1$ column vector and $\mathbf{X}$ is an $n \times k$ matrix.

In matrix notation the OLS estimator $\widehat{\beta}$ is given by the expression

$$\widehat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{y}$$

6 Goodness of Fit and R^2

Recall that

$$Y_i = \sum_{j=1}^{k} \widehat{\beta}_j X_{ji} + e_i = \widehat{Y}_i + e_i$$

or, taking mean deviations,

$$y_i = \sum_{j=2}^{k} \widehat{\beta}_j x_{ji} + e_i = \widehat{y}_i + e_i.$$

Squaring and summing over all observations gives

$$\begin{aligned}
\sum_{i=1}^{n} y_i^2 &= \sum_{i=1}^{n} \widehat{y}_i^2 + \sum_{i=1}^{n} e_i^2 + 2\sum_{i=1}^{n} \widehat{y}_i e_i \\
&= \sum_{i=1}^{n} \widehat{y}_i^2 + \sum_{i=1}^{n} e_i^2
\end{aligned}$$

since

$$\sum_{i=1}^{n} \widehat{y}_i e_i = \sum_{j=2}^{k} \widehat{\beta}_j \sum_{i=1}^{n} x_{ji} e_i$$

and

$$\sum_{i=1}^{n} x_{ji} e_i = 0 \quad , \quad \forall j$$

from (3.5).

Defining $\sum_{i=1}^{n} y_i^2$ = Total Sum of Squares (TSS), $\sum_{i=1}^{n} \widehat{y}_i^2$ = Explained Sum of Squares (ESS), and $\sum_{i=1}^{n} e_i^2$ = Residual Sum of Squares (RSS), we have the identity

$$TSS = ESS + RSS$$

and define the *Multiple Coefficient of Determination*, R^2 as

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n} \widehat{y}_i^2}{\sum_{i=1}^{n} y_i^2}$$

or, alternatively,

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n} \widehat{e}_i^2}{\sum_{i=1}^{n} y_i^2}.$$

R^2 measures the proportion of the total variation of Y_i about its mean that is explained by all the regressors jointly. This is the goodness-of-fit of the regression.

6.1 Properties of R^2

As long as an intercept is included in the set of regressors then

$$0 \leq R^2 \leq 1.$$

(If an intercept is not included then it is possible to get a negative value for R^2).

R^2 will increase as k increases i.e. more regressors are added. In the limit when $k = n$ then $R^2 = 1$. This suggests that choosing the equation with the highest R^2 will not always be a sensible strategy.

The value of R^2 depends on the scale of the dependent variable. R^2 measures cannot be compared between equations with different dependent variables.

Example: Consider the equation

$$Y_t = a + bY_{t-1} + u_t.$$

Subtracting Y_{t-1} from both sides gives

$$\Delta Y_t \equiv Y_t - Y_{t-1} = a + (b-1)Y_{t-1} + u_t$$

but does not affect the error term or the coefficients. This is merely a renormalisation of the equation and the RSS of the two will be the same. However, the R^2 statistic will give very different values, purely because of the different scales of the dependent variables.

6.2 Correcting for degrees of freedom: $\overline{R}^2$

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k}$$

$\overline{R}^2$ makes a correction to R^2 based on the number of parameters in the model k. Note that $\overline{R}^2 < R^2$ in general, and that $\overline{R}^2$ can be negative.

7 Hypothesis Testing

In order to make statistical inferences on the parameter estimates $\widehat{\beta}$ we must make a further assumption:

$$u_i \sim iid\, N(0, \sigma^2) \quad , \quad i = 1, \cdots, n \tag{A5}$$

that the errors u_i are distributed as independent normal variables.

For hypotheses involving a single parameter $\widehat{\beta}_j$,

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{\widehat{Var}(\widehat{\beta}_j)}} \sim t_{n-k}$$

where t_{n-k} is the Student t distribution with $n - k$ degrees of freedom.

Note that $\sqrt{\widehat{Var}(\widehat{\beta}_j)} \equiv \widehat{\sigma}_{\widehat{\beta}_j}$ is the *standard error* of the estimator $\widehat{\beta}_j$.

This is just as in the simple regression case. However, there is now the additional possibility of tests involving more than one parameter, or more than one restriction.

7.1 Testing equality of two coefficients

Without loss of generality, let the null hypothesis be

$$H_0 : \beta_2 = \beta_3$$

or, alternatively,

$$H_0 : \beta_2 - \beta_3 = 0.$$

This is a single restriction although it involves two parameters. It can be shown that

$$\frac{\widehat{\beta}_2 - \widehat{\beta}_3}{\sqrt{\widehat{Var}(\widehat{\beta}_2 - \widehat{\beta}_3)}} \sim t_{n-k}$$

where

$$\widehat{Var}(\widehat{\beta}_2 - \widehat{\beta}_3) = \widehat{Var}(\widehat{\beta}_2) + \widehat{Var}(\widehat{\beta}_3) - 2\widehat{Cov}(\widehat{\beta}_2, \widehat{\beta}_3)$$

and the estimated variances and covariances needed to compute this test are available in any computer regression package.

7.2 Jointly testing all the coefficients

One interesting hypothesis is that

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_k = 0$$

or, alternatively,

$$H_0 : \beta_j = 0 \quad , \quad j = 2, \cdots, k$$

so that the model collapses to the intercept term

$$Y_i = \beta_1 + u_i \quad , \quad i = 1, \cdots, n$$

Note that the hypothesis H_0 involves *jointly* testing $k-1$ restrictions.

It can be shown that a test of this hypothesis is given by

$$\frac{ESS/(k-1)}{RSS/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \sim F_{k-1,n-k}$$

where $F_{k-1,n-k}$ is the F distribution with $k-1$ and $n-k$ degrees of freedom.

The Student t distribution and the F distribution are related. If x is distributed as t_{n-k} then x^2 is distributed as $F_{1,n-k}$.

7.3 General F Testing

More generally, we may often be interested in jointly testing several linear restrictions. Suppose that there are r restrictions. Then the null hypothesis can be written

$$H_0 : \sum_{j=1}^{k} a_{ij}\beta_j - a_{i0} = 0, \quad i = 1, \cdots, r$$

where a_{ij} are fixed constants.

To compute a test of this hypothesis we need to estimate the model both under the null hypothesis (imposing the restrictions) and on the alternative hypothesis (unrestricted). Let RSS_U be the sum of squared residuals for the unrestricted model, and RSS_R be the sum of squared residuals of the restricted model. Then it can be shown that a test of H_0 is given by

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

7.4 Testing σ^2

Finally, as with the simple regression model

$$(n-k)\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-k} \tag{7.1}$$

where χ^2_{n-k} is the Chi-squared distribution with $n-k$ degrees of freedom.

8 Functional Form

We require that the OLS model is *linear in parameters.* It is not necessary that it is *linear in variables.* It is often possible to tranform variables in order to make an equation linear in parameters.

8.1 Example: Cobb-Douglas Production Function

$$Q = AK^{\gamma}L^{\delta}$$

where $\gamma + \delta = 1$ implies constant returns to scale. Taking logarithms we get

$$\ln Q = \ln A + \gamma \ln K + \delta \ln L$$

or

$$Y = b_0 + b_1X_1 + b_2X_2 + u$$

where $Y = \ln Q$, $X_1 = \ln K$, $X_2 = \ln L$, $b_0 = \ln A$, $b_1 = \gamma$, and $b_2 = \delta$, which is linear in parameters. Constant return to scale can be tested in this model by the linear hypothesis that $b_1 + b_2 = 1$.

Note that an error term u has been added to the final equation. If we assume that $u \sim N(0, \sigma^2)$ then this implies that the error term on the original equation is e^u and is distributed with a *log-normal* distribution.

Specifications cannot always be transformed to make them linear in parameters. Consider the CES production function

$$Q = A[\gamma K^{-\rho} + \delta L^{-\rho}]^{-\mu/\rho}.$$

This is intrinsically nonlinear in parameters. Similarly, the equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_0 \beta_1 X_2 + u$$

is linear in variables but nonlinear in parameters.

8.2 Log-linear Specifications

The Cobb-Douglas example is one in which all the variables in the econometric equation are in logarithms. Such specifications are called *log-linear* or *log-log* and are very common in economics. Compare a linear and a log-linear specification

$$Y = a + bX + u$$

and

$$\ln Y = c + d \ln X + v.$$

In the linear specification $b = \partial Y / \partial X$ is the marginal response of Y to a change in X. Note that the effect of doubling X on Y depends on the level of X, because the elasticity is not constant. In the log-linear specification $d = \partial \ln Y / \partial \ln X$ which is the elasticity of Y with respect to X. Log-linear models impose the assumption that elasticities are constant.

In log-linear models the slope parameters are independent of the scale of measurement of the variables. Suppose we rescale X by multiplying by a factor k, denoting the new variable kX as X^*. In the linear model the effect is that the parameter b becomes $b^* = b/k$. In the log-linear model we have $\ln X^* = \ln k + \ln X$ so that the rescaling factor is constant and changes the intercept c but not the elasticity d.

Many economic variables represent money flows or stocks or prices that logically cannot ever be negative. In principle, a linear specification can give a negative value for the dependent variable if u is large and negative enough. In the log-linear specification, the dependent variable can never become negative since $Y = e^{\ln Y} \geq 0$.

8.3 Linearity as an Approximation

Consider a general function of a single variable

$$Y = g(X).$$

Weierstrass's theorem states that, under certain very general conditions,

$$Y \simeq a_0 + a_1 X + a_2 X^2 + a_3 X^3 + \cdots$$

so that the function $g(X)$ can be approximated, to any desired degree of precision, by a polynomial function in X. This polynomial function is linear in parameters so it can be estimated by OLS, treating X, X^2, X^3 etc. as separate regressors. Note that X, X^2, X^3, although they are exact functions of each other, are not *linearly* dependent so that there is no perfect collinearity here.