1 The Multiple Regression Model

- The multiple regression model with k explanatory variables:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + .. + \beta_k X_{ki} + \varepsilon_i,$$

where i subscripts to denote individual observations and $i = 1, .., N$.

- This chapter discusses how statistical derivations for simple regression extend to multiple regression.

- Also derives some results (e.g. about omitted variables bias and multicollinearity) that were intuitively motivated in Chapter 2.

1.1 Basic Theoretical Results

In Chapter 3 derived theoretical results using simple regression model with classical assumptions

$$Y_i = \beta X_i + \varepsilon_i$$

1. $E(\varepsilon_i) = 0$ – mean zero errors.

2. $var(\varepsilon_i) = E\left(\varepsilon_i^2\right) = \sigma^2$ – constant variance errors (homoskedasticity).

3. $E\left(\varepsilon_i \varepsilon_j\right) = 0$ for $i \neq j$ — ε_i and ε_j are independent of one another.

4. ε_i is Normally distributed

5. X_i is fixed. It is not a random variable.

Statistical results using this model are basically the same as for the simple regression model. For instance, OLS is still unbiased and confidence intervals and hypothesis tests are derived in the same way. Gauss Markov theorem still says that, under the classical assumptions, OLS is BLUE.

Formulae do get messier (which is why, in more advanced courses, matrix algebra is used with the multiple regression model).

- For instance, with an intercept and two explanatory variables, we have

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i.$$

- The OLS estimator is the one which minimizes the sum of squared residuals and turns out to be:

$$\widehat{\beta}_1 = \frac{\left(\sum x_{1i} y_i\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{2i} y_i\right)\left(\sum x_{1i} x_{2i}\right)}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i} x_{2i}\right)^2}$$

$$\widehat{\beta}_2 = \frac{\left(\sum x_{2i} y_i\right)\left(\sum x_{1i}^2\right) - \left(\sum x_{1i} y_i\right)\left(\sum x_{1i} x_{2i}\right)}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i} x_{2i}\right)^2}$$

$$\widehat{\alpha} = \overline{Y} - \widehat{\beta}_1 \overline{X}_1 - \widehat{\beta}_2 \overline{X}_2.$$

where variables with bars over them are means (e.g. $\overline{X}_2 = \frac{\sum X_{2i}}{N}$)

- The previous formulae used small letters to indicate deviations from means. That is,

$$\begin{aligned} y_i &= Y_i - \overline{Y} \\ x_{1i} &= X_{1i} - \overline{X}_1 \\ x_{2i} &= X_{2i} - \overline{X}_2 \end{aligned} .$$

- It can be shown that OLS is unbiased (e.g. $E\left(\widehat{\beta}_j\right) = \beta_j$ for $j = 1, 2$) and an unbiased estimator for σ^2 is:

$$s^2 = \frac{\sum \widehat{\varepsilon}_i^2}{N - k - 1}$$

where

$$\widehat{\varepsilon}_i = y_i - \widehat{\alpha} - \widehat{\beta}_1 X_{1i} + \widehat{\beta}_2 X_{2i}$$

- An estimate of the variance of the OLS estimator can also be calculated:

$$var\left(\widehat{\beta}_1\right) = \frac{s^2}{\left(1 - r^2\right) \sum x_{1i}^2}$$

$$var\left(\widehat{\beta}_2\right) = \frac{s^2}{\left(1-r^2\right)\sum x_{2i}^2}$$

where r is the correlation between x_1 and x_2. ($var\left(\widehat{\alpha}\right)$ formula not provided here).

- Do not memorize these formulae. I provide them here mostly to show you how messy things the formulae get (and to motivate why we have been deriving key theoretical results in the simple regression case).

- Interpretation:

"β_j is the marginal effect of the j^{th} explanatory variable on y, holding all the other explanatory variables constant"

1.2 Measures of Fit

The most popular measure of model fit, R^2, has the same definition as before:

$$R^2 = 1 - \frac{SSR}{TSS} = 1 - \frac{\sum \widehat{\varepsilon}_i^2}{\sum \left(Y_i - \overline{Y}\right)^2}.$$

Interpretation: R^2 is the proportion of the variability in the dependent variable which can be explained by the explanatory variables.

Note: When new explanatory variables are added to a model, the R^2 will always rise (even if new variables are insignificant). Why? By adding a new variable, $\sum \widehat{\varepsilon}_i^2$ will always get at least a little bit smaller.

So R^2 should not be used to decide whether to add a new variable to a regression.

What you should do is a hypothesis test (test whether new variable significant and, if not, drop it).

Alternatively, use $\overline{R}^2$ which is similar to R^2 but does not always rise when new variables are added. If you add a new variable and $\overline{R}^2$ increases, you can be confident this new variable should be included.

If you have two regressions (with the same dependent variable but different explanatory variables), then the one with the higher $\overline{R}^2$ is the better one.

Definition:

$$\overline{R}^2 = 1 - \frac{var(\varepsilon)}{var(Y)} = 1 - \frac{s^2}{\frac{1}{N-1}\sum\left(Y_i - \overline{Y}\right)^2}$$

The only problem with $\overline{R}^2$ is that it CANNOT be interpreted simply as reflecting the proportion of the variability

in the dependent variable which can be explained by the explanatory variables.

Summary: R^2 has a nice simple interpretation as a measure of fit, but should not be used for choosing between models.

$\overline{R}^2$ does not have a nice simple interpretation as a measure of fit, but can be used for choosing between models.

2 Hypothesis Testing in the Multiple Regression Model

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + .. + \beta_k X_{ki} + \varepsilon_i.$$

- We know how to do t-tests of $H_0 : \beta_j = 0$
- In Chapter 2 presented a test for whether $R^2 = 0$
- This is equivalent to a test of the hypothesis:

$$H_0 : \beta_1 = .. = \beta_k = 0.$$

- Remember: testing the hypothesis $H_0 : \beta_1 = .. = \beta_k = 0$ is not the same as testing the k individual hypotheses $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$ through $H_0 : \beta_k = 0$

- But there are other hypotheses that you may want to test. E.g.

$$H_0 : \beta_1 = \beta_3 = 0.$$

or

$$H_0 : \beta_1 - \beta_3 = 0, \beta_2 = 5$$

etc. etc.

- F-tests, are suitable for testing hypotheses involving any number of linear combinations of regression coefficients.

- Likelihood ratio tests, can do the same, but can also be used for nonlinear restrictions and can be used with models other than the regression model.

- In this course, we only have time to do F-tests

2.1 F-tests

- We will not provide proofs relating to F-tests.
- Distinguish between unrestricted and restricted model
- Unrestricted model is the multiple regression model.
- Restricted model is the multiple regression model with the restrictions in H_0 imposed
- Examples using $k = 3$:
- The unrestricted model:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

- Consider testing

$$H_0 : \beta_1 = \beta_2 = 0.$$

- Then restricted model is:

$$Y_i = \alpha + \beta_3 X_{3i} + \varepsilon_i$$

- Now consider testing

$$H_0 : \beta_1 = 0, \beta_2 + \beta_3 = 0.$$

- Imposing this restriction yields a restricted model:

$$Y_i = \alpha + \beta_2 (X_{2i} - X_{3i}) + \varepsilon_i.$$

- This restricted model is just a regression with $(X_2 - X_3)$ being the explanatory variable.

- Now consider testing

$$H_0 : \beta_1 = 1, \beta_2 = 0.$$

- Imposing this restriction yields a restricted model:

$$Y_i - X_{1i} = \alpha + \beta_3 X_{3i} + \varepsilon_i.$$

- This restricted model is just a regression with $(Y_i - X_{1i})$ as dependent variable and X_3 being the explanatory variable.

- In general: You can show that for any set of linear restrictions on the unrestricted model you can write out a new restricted model (which is still a linear regression model but with dependent and explanatory variables which may be different).

- For testing such hypotheses (involving more than one linear restriction on the regression coefficients), the following test statistic is used:

$$F = \frac{(SSR_R - SSR_{UR})/q}{SSR_{UR}/(N-k-1)},$$

where SSR is the familiar sum of squared residuals and the subscripts UR and R distinguish between the SSR from the "unrestricted" and "restricted" regression models.

- Since $SSR_R > SSR_{UR}$ (i.e. the model with fewer restrictions can always achieve the lower SSR), it can be seen that F is positive.

- The number of restrictions being tested is q (Note: $q = 2$ in the examples above).

- Note: F is positive and large values of F indicate the null hypothesis is incorrect.

- As with any hypothesis test, you calculate the test statistic (here F) and compare it to a critical value. If F is greater than the critical value you reject H_0 (else you accept H_0).

- To get the critical value you need to specify a level of significance (usually .05) and, using this, obtain a critical value from statistical tables.

- F is distributed as $F_{q,N-k-1}$ (in words, critical values should be obtained from the F-distribution with q degrees of freedom in the numerator and $N-k-1$ degrees of freedom in the denominator).

- Statistical tables for the F-distribution are available in most places (including the textbook).

- In practice, the more sophisticated econometrics package will provide you with a P-value for any test you do.

- Remember: Useful (but not quite correct) intuition: "P-value is the probability that H_0 is true"

- A correct interpretation: "P-value equals the smallest level of significance at which you can reject H_0"

2.2 Multicollinearity

A fairly common practical problem in empirical work. Intuition: If the explanatory variables are very highly correlated with one another you run into problems.

Informally speaking, if two variables are highly correlated they contain roughly the same information. The OLS estimator has trouble estimating two separate marginal effects for two such highly correlated variables.

2.2.1 Perfect Multicollinearity

An exact linear relationship exists between the explanatory variables.

The correlation between two explanatory variables equals 1.

Example: A regression relating to the effect of studying on student performance.

$y =$ student grade on test

$X_1 =$ family income

$X_2 =$ hours studies per day

$X_3 =$ hours studied per week.

But $X_3 = 7X_2$ – an exact linear relationship between two explanatory variables (they are perfectly correlated).

This is a care of perfect multicollinearity.

OLS estimates cannot be calculated (i.e. Excel will not be able to find a solution and will give an error message).

Intuition: β_2 will measure the marginal effect of hours studied per day on student grade, *holding all other explanatory variables constant*.

With perfect multicollinearity there is no way of "*holding all other explanatory variables constant*" – when X_3 changes, then X_4 will change as well (it cannot be held constant).

2.2.2 Regular Multicollinearity

In practice you will never get perfect multicollinearity, unless you do something that does not make sense (like put in two explanatory variables which measure the exact same thing).

And if you ever do try and estimate a model with perfect multicollinearity you will find out quickly — Excel will not run properly and will give you an error message.

But you may get very highly correlated explanatory variables.

Example: Macroeconomic regression involving the interest rate.

$X_1 =$ interest rate set by Bank of England

$X_2 =$ interest rate charged by banks on mortgages.

X_1 and X_2 will not be exactly the same, but will be very highly correlated (e.g. $r = .99$).

Multicollinearity of this form can cause problems too.

Basic idea of technical proofs is based on variances of OLS estimators. Previously, we wrote:

$$var\left(\widehat{\beta}_1\right) = \frac{s^2}{\left(1 - r^2\right)\sum x_{1i}^2}$$

$$var\left(\widehat{\beta}_2\right) = \frac{s^2}{\left(1 - r^2\right) \sum x_{2i}^2}$$

If r is near 1 (or near -1), then $\left(1 - r^2\right)$ will be near zero. Variances of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ become very large. This feeds through into very wide confidence intervals (i.e. inaccurate estimates) and very small t-statistics (i.e. hypothesis tests indicate β_1 and β_2 are insignificant).

Common symptom of multicollinearity problem:

Some or all explanatory variables appear insignificant, even though the model is fitting well (has a high R^2).

Common way to investigate if multicollinearity is a problem:

Calculate a correlation matrix for your explanatory variables. See if any correlations are very high.

Note: What does we mean by a correlation being "high"? There is no hard and fast rule. As a rough guideline, if you find correlations between your explanatory variables $|r| >$ .9 then you probably have a multicollinearity problem.

Solutions to multicollinearity problem:

1. Get more data (often not possible).

2. Drop out one of the highly correlated variables.

Example: Macroeconomic regression involving the interest rate (continued)

If you include both X_1 and X_2 you will run into a multicollinearity problem. So include one or the other (not both).

2.3 Omitted Variables Bias

- Discussed intuitively in Chapter 2.

- Assume true model is:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i,$$

and the classical assumptions hold.

- Correct OLS estimate is:

$$\widehat{\beta}_1 = \frac{\left(\sum x_{1i} y_i\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{2i} y_i\right)\left(\sum x_{1i} x_{2i}\right)}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i} x_{2i}\right)^2}$$

- What if you mistakenly omit X_2:

$$Y_i = \alpha + \beta_1 X_{1i} + \varepsilon_i.$$

- A simple extension of the derivations of Chapter 3 show OLS estimator for β_1 is:

$$\widetilde{\beta}_1 = \frac{\sum x_{1i} y_i}{\sum x_{1i}^2}$$

- Note: use notation $\widetilde{\beta}_1$ to distinguish it from the correct OLS estimator

- $\widetilde{\beta}_1$ is biased.

Proof that $\widetilde{\beta}_1$ is biased

- Step 1

$$\begin{aligned}\overline{Y} &= \frac{\sum Y_i}{N} = \frac{\sum (\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i)}{N} \\ &= \alpha + \beta_1 \overline{X}_1 + \beta_2 \overline{X}_2 + \overline{\varepsilon}.\end{aligned}$$

- Step 2

$$\begin{aligned}y_i &= Y_i - \overline{Y} \\ &= (\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i) - \\ &\quad \left(\alpha + \beta_1 \overline{X}_1 + \beta_2 \overline{X}_2 + \overline{\varepsilon}\right) \\ &= \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i - \overline{\varepsilon}.\end{aligned}$$

- Step 3: replace y_i in formula for $\widetilde{\beta}_1$:

$$\begin{aligned}
\widetilde{\beta}_1 &= \frac{\sum x_{1i}\left(\beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i - \overline{\varepsilon}\right)}{\sum x_{1i}^2} \\
&= \frac{\beta_1 \sum x_{1i}^2}{\sum x_{1i}^2} + \frac{\beta_2 \sum x_{1i} x_{2i}}{\sum x_{1i}^2} + \frac{\sum x_{1i}\left(\varepsilon_i - \overline{\varepsilon}\right)}{\sum x_{1i}^2} \\
&= \beta_1 + \frac{\beta_2 \sum x_{1i} x_{2i}}{\sum x_{1i}^2} + \frac{\sum x_{1i}\left(\varepsilon_i - \overline{\varepsilon}\right)}{\sum x_{1i}^2}.
\end{aligned}$$

- Step 4: take expected value of both sides of this equation:

$$\begin{aligned}
E\left(\widetilde{\beta}_1\right) &= E\left(\beta_1 + \frac{\beta_2 \sum x_{1i} x_{2i}}{\sum x_{1i}^2} + \frac{\sum x_{1i}\left(\varepsilon_i - \overline{\varepsilon}\right)}{\sum x_{1i}^2}\right) \\
&= \beta_1 + \frac{\beta_2 \sum x_{1i} x_{2i}}{\sum x_{1i}^2},
\end{aligned}$$

- Thus, $E\left(\widetilde{\beta}_1\right) \neq \beta_1$ and if we omit an explanatory variable in the regression which should be included, we obtain a biased estimate of the coefficient on the included explanatory variable.

- This is omitted variables bias.

- Note that omitted variables bias does not exist if $\beta_2 = 0$

- But if $\beta_2 = 0$ then X_2 does not belong in the regression so it is okay to omit it.

- Note that omitted variables bias does not exist if $\frac{\sum x_{1i}x_{2i}}{\sum x_{1i}^2} = 0$.

- Can show (see Chapter 1) that this implies correlation between X_1 and X_2 is zero

2.4 Inclusion of Irrelevant Explanatory Variables

- Now reverse role of the two models discussed in omitted variables bias section

- True model:

$$Y_i = \alpha + \beta_1 X_{1i} + \varepsilon_i$$

and classical assumptions hold

- Incorrect model adds an irrelevant variable:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i.$$

- Using incorrect model get OLS estimate:

$$\widetilde{\beta}_1 = \frac{\left(\sum x_{1i} y_i\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{2i} y_i\right)\left(\sum x_{1i} x_{2i}\right)}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i} x_{2i}\right)^2}$$

- But correct OLS estimate is:

$$\widehat{\beta}_1 = \frac{\sum x_{1i} y_i}{\sum x_{1i}^2}.$$

- Gauss-Markov theorem tells us that $\widehat{\beta}_1$ is the best linear unbiased estimator.

- Thus, $\widehat{\beta}_1$ has a smaller variance than any other unbiased estimator.

- If we can show that $\widetilde{\beta}_1$ is unbiased, then Gauss-Markov theorem tell us $var\left(\widetilde{\beta}_1\right) > var\left(\widehat{\beta}_1\right)$

- This proves that including irrelevant explanatory variables will lead to less precise estimates.

- Proof that $\widetilde{\beta}_1$ is unbiased is given in the textbook

Important Message for Empirical Practice

- Omitted variables bias says:

"you should always try to include all those explanatory variables that could affect the dependent variable"

- Inclusion of irrelevant explanatory variables section says:

"you should always try not to include irrelevant variables, since this will decrease the accuracy of the estimation of all the coefficients (even the ones that are not irrelevant)"

- How do you play off these considerations?

- Begin with as many explanatory variables as possible, then use hypothesis testing procedures to discard those that are irrelevant (and then re-run the regression with the new set of explanatory variables).

2.5 Nonlinear Regression

The linear regression model is:

$$Y_i = \alpha + \beta_1 X_{1i} + .. + \beta_k X_{ki} + \varepsilon_i$$

Sometimes you may think the relationship between your explanatory and dependent variables is nonlinear.

$$Y_i = f\left(X_{1i}, .., X_{ki}\right) + \varepsilon_i$$

where $f\left(\right)$ is some nonlinear function.

In some cases, nonlinear regressions is a bit more complicated (using techniques beyond those covered in this textbook). However, in many cases a nonlinear function

can be transformed into a linear one – and then linear regression techniques can be used.

Example:

$$Y_i = \beta_0 X_{1i}^{\beta_1} X_{2i}^{\beta_2} .. X_{ki}^{\beta_k}$$

becomes:

$$\ln(Y_i) = \alpha + \beta_1 \ln(X_{1i}) + .. + \beta_k \ln(X_{ki})$$

where $\alpha = \ln(\beta_0)$. So you can run a regression of the logged dependent variable on logs of the explanatory variables.

2.5.1 How to decide which nonlinear form?

It can be hard to decide which nonlinear form is appropriate. Here are a few pieces of advice.

- Sometimes economic theory suggests a particular functional form. For instance, the previous example arises when one is using a Cobb-Douglas production function.

- Experiment with different functional forms and use hypothesis testing procedures or $\overline{R}^2$ to decide between them.

Example: Is there a quadratic pattern?

Run OLS regressions on two models:

$$Y_i = \alpha + \beta_1 X_{1i} + \varepsilon_i$$

and

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i$$

and choose the quadratic model if its $\overline{R}^2$ is higher than the linear model.

Alternatively, run OLS on quadratic model and test whether $\beta_2 = 0$.

Warning: you can only use $\overline{R}^2$ to compare models involving nonlinear transformations of the explanatory variables. You cannot use it to compare models which tranform the dependent variable in different ways. Remember,

$$\overline{R}^2 = 1 - \frac{var(\varepsilon)}{var(Y)}.$$

In order to use it for choosing a model, all models must have the same Y.

Example: Comparing two models:

$$Y_i = \alpha + \beta_1 X_{1i} + \varepsilon_i$$

and

$$\ln(Y_i) = \alpha + \beta_1 X_{1i} + \varepsilon_i$$

You CANNOT use $\overline{R}^2$ to decide which of these models to use.

Textbook describes a test for linear versus log-linear regression.

2.5.2 Interpretation of Coefficients when Variables are Logged

- Consider log-linear regression

$$\ln(Y_i) = \alpha + \beta_1 \ln(X_{1i}) + .. + \beta_k \ln(X_{ki}) + \varepsilon_i.$$

- Interpretation of coefficients is as an elasticity: if X_j increases by one percent, Y tends to increase by β_j percent (ceteris paribus)

- Now consider a regression where some variables are not logged such as:

$$\ln(Y_i) = \alpha + \beta_1 \ln(X_{1i}) + \beta_2 X_{2i} + \varepsilon_i.$$

- β_1 has elasticity interpretation, but β_2 has interpretation: if X_2 increases by one unit, Y tends to increase by β_2 percent (ceteris paribus)

2.6 Chapter Summary

- Much of this chapter builds on Chapter 2 (Non-technical introduction to regression) and Chapter 3 (on deriving statistical results for simple regression)

- Issues in Chapter 2 (i.e. omitted variables bias, the impact of including irrelevant variables and multi-collinearity) treated more formally in a statistical sense.

- Most derivations (e.g. of confidence intervals, t-tests, etc.) in Chapter 3 extend in conceptually straight-forward manner to multiple regression

- This chapter introduced a new framework for hypothesis testing: F-tests. These are useful for testing multiple hypotheses about regression coefficients.

- This chapter discussed selection of the appropriate functional form of a regression.

- Many nonlinear relationships can be made linear through an appropriate transformation (nothing is new, simply run a regression with transformed variables).

- Care must be taken with interpretation of coefficients in nonlinear regression.

- Care has to be taken when choosing between models which have different nonlinear transformations of the dependent variable.