

5. Specification Analysis

Analysis so far assumes that model $y = X\beta + u$ is correct

Omission of Relevant Variables

True: $y = X_1\beta_1 + X_2\beta_2 + e$

Estimate: $y = X_1\beta_1 + u$

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'y$$

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'X_1\beta_1 + (X_1'X_1)^{-1}(X_1'X_2)\beta_2 + (X_1'X_1)^{-1}X_1'e$$

$$\hat{\beta}_1 = \beta_1 + (X_1'X_1)^{-1}(X_1'X_2)\beta_2 + (X_1'X_1)^{-1}X_1'e$$

so

$$E(\hat{\beta}_1) = \beta_1 + (X_1'X_1)^{-1}(X_1'X_2)\beta_2 \neq \beta_1$$

estimates of the coefficients on the set of X_1 variables are biased in the presence of omitted variables

{and sign of bias depends on a) the effect of the omitted variables on y , β_2 ,
b) the covariance of X_1 and X_2

Not only is mean biased so is OLS estimates of parameter variances

Since partitioned matrix

$$\begin{bmatrix} x_1'x_1 & x_1'X_2 \\ X_2'x_1 & X_2'X_2 \end{bmatrix} = X'X$$

Then try variance/covariance matrix of parameter estimates is given by upper left hand block of $\sigma^2(X'X)^{-1}$

rules on inverse of partitioned matrices imply that inverse of top left hand element is

$$\text{Var}(\hat{\beta}_{1,2}) = \sigma^2[X_1'X_1 - X_1'X_2(X_2'X_2)^{-1}X_2'X_1]^{-1}$$

If X_2 are omitted then

$$\text{Var}(\hat{\beta}_1) = \sigma^2[X_1'X_1]^{-1}$$

$$\text{So } \text{Var}(\hat{\beta}_1) < \text{Var}(\hat{\beta}_{1,2})$$

ie estimated parameters will have a smaller variance in omitted variable case (and so t statistics will be biased upward)

unless X_1 and X_2 are orthogonal

The more highly correlated X_1 and X_2 the larger the difference between the OLS estimates and the larger the variance of the true estimates compared to the omitted variable case.

However

Since σ^2 is unobserved it is always estimated by $s^2 = \frac{\hat{u}'\hat{u}}{N-k}$

However can show (see Greene) that , like the coefficient estimates, the estimate of s^2 is biased in the presence of omitted variables

$$E(\hat{u}_1'\hat{u}_1) = \beta_2' X_2' M_1 X_2 \beta_2 + \sigma^2 (N - k_1)$$

ie biased up by an amount equal to the increase in ESS when X_2 added to the regression (and to take account of the bias need information on (unknown) effects of X_2)

this goes in the opposite direction to the bias on $\text{Var}(\hat{\beta})$ when σ^2 is known

So can only conclude that in the presence of omitted variables both the estimates of β and their variances are biased.

Hence any inference based on these values (eg t or F tests) will also be biased.

Inclusion of Irrelevant Variables

True: $y = X_1\beta_1 + u$

Estimate: $y = X_1\beta_1 + X_2\beta_2 + e$

Think of true model as failing to impose the (correct) restriction that $\beta_2=0$

But we know that OLS will give unbiased estimates of the true values of all the coefficients and so

$$E(\hat{\beta}) = \begin{bmatrix} E(\hat{\beta}_1) \\ E(\hat{\beta}_2) \end{bmatrix} = \begin{bmatrix} \beta_1 \\ 0 \end{bmatrix}$$

ie if variables are irrelevant then OLS estimates of effects should be zero

By similar reasoning the OLS estimate of σ^2 is also unbiased

The only cost is the loss in precision of parameter estimates caused by a rise in the number of right hand side variables

$$\text{Var}(\hat{\beta}_1) < \text{Var}(\hat{\beta}_{1,2})$$

so standard errors on estimates in presence of irrelevant variables will be higher than should be (and the more correlated X_1 and X_2 are the more the standard errors are inflated)

As a result any t and F statistics will be lower than should be (leading to Type II error)

Trade off bias with gain in efficiency

How to choose a “good” model then?

Look for a range of well behaved diagnostics

- 1) the value of the \bar{R}^2
- 2) well behaved residuals (net of leverage)
- 3) parameter stability across different time periods/sub-groups
- 4) do individual coefficient estimates agree with economic priors?

5) Ramsey RESET Test

- if model is good fit then addition of extra variables should not be statistically significant

rather than add higher order terms of original variables a more parsimonious alternative is to use fact that

$$\hat{y} = X \hat{\beta}$$

so predicted values are linear function of all the X variables (weighted by their estimated coefficients)

and hence $(\hat{y})^j = (X \hat{\beta})^j$

are linear functions of higher powers of all the X variables

$$y = X\beta + \delta_2 \hat{y}^2 + \delta_3 \hat{y}^3 + \dots + \delta_j \hat{y}^j + u$$

and test null $H_0: \delta_2 = \delta_3 = \dots \delta_j = 0$

LM Test of Omitted Variables

1. Run restricted regression (no higher order terms)
2. save residuals
3. Regress residuals on unrestricted model (containing higher order values of X (or the \hat{y}_j) - the auxiliary regression

Can show

$$NR^2_{aux} \stackrel{a}{\sim} \chi^2_{(No.ofrestrictions)}$$

