Lecture 3. Goodness of Fit

Useful to have summary measure of how well the OLS regression line fits the data
Given

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

$$y'y = (X\hat{\beta} + \hat{u})'(X\hat{\beta} + \hat{u})$$

$$y'y = \hat{\beta}'X'X\hat{\beta} + \hat{u}'\hat{u}$$

Since $y'y \neq \sum_i (y_i - \bar{y})^2$

Need to subtract $N\bar{y}^2$ from both sides

$(y'y - N\bar{y}^2)$	$= (\hat{\beta}'X'X\hat{\beta} - N\bar{y}^2)$	$+ \quad \hat{u}'\hat{u}$
Total sum of Squares	= Explained sum of squares	+ residual sum of squares

Hence measure of how well regression fits the data is given by

R^2 (the coefficient of determination) = ESS/TSS = 1 –(RSS/TSS)

- the proportion of the total variation in y accounted for by the variation in the regressors

$0 <= R^2 <= 1$

Major problem with using r^2 is that will never fall when add new explanatory variables to the model

Use instead the Adjusted R^2 , $\bar{R}^2 = 1 - \frac{(N-1)}{(N-k)}(1 - R^2)$

Whether this rises or falls depends on whether variable added to model has a t ratio greater than one

There are other goodness of fit criteria that impose different weights to additional numbers of regressors

1. Schwarz $\quad Log_e(RSS/N) + k/N \, log_e(N)$

2. Akaike $\quad Log_e(RSS/N) + 2k/N$

these penalties are larger than those imposed in the adjusted R^2 and so will favour simpler models

Assumptions Underlying OLS

In order to assess the accuracy or the precision of OLS estimates need to make assumptions about the statistical process generating the observations in the dataset.

1. regression model is linear in parameters

2. X values are fixed in repeated sampling

3. Mean value of (true, unobserved) residuals is zero

4. No covariance between residuals and independent variables

5. Equal variance of individual residual terms (homoskedasticity)

6. No correlation between individual residual terms (no autocorrelation)

When moving from 2 to k variable model need to make 1 additional assumption

7. No exact colinearity between X variables

Statistical Properties of OLS Estimators

Given $\hat{\beta} = (X'X)^{-1}X'y$ = $(X'X)^{-1}X'(X\beta + u)$ $= \beta + (X'X)^{-1}X'u$

So

$$E(\hat{\beta}) = \beta + (X'X)^{-1}X'E(u)$$

OLS estimators are unbiased

$$Var(\hat{\beta}) = E\left[(\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))'\right]$$

$$Var(\hat{\beta}) = E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right]$$ = E[$(X'X)^{-1}X'uu'X(X'X)^{-1}$]

= $(X'X)^{-1}X'E(uu')\,X(X'X)^{-1}$

= $(X'X)^{-1}X'\sigma^2 I\, X(X'X)^{-1}$

$\sigma^2(X'X)^{-1}$

which is a $k_x k$ matrix with variances on the main diagonal and covariances on the off diagonal

Gauss-Markov Theorem

Consider another linear unbiased estimator $\tilde{\beta} = Cy$

If $\tilde{\beta}$ is unbiased then $E(\tilde{\beta}) = E(Cy) = E[CX\beta + Cu] = \beta$

Hence CX=I and $\tilde{\beta}$ = β + Cu

So $$Var(\tilde{\beta}) = E\left[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)'\right]$$

$= E[Cuu'C']$

$= \sigma^2 CC'$

Let D be the difference between the OLS and alternative estimated explanatory component ie

$$D = C-(X'X)^{-1}X'$$

So

$$Var(\tilde{\beta}) = \sigma^2\left[(D + (X'X)^{-1}X')(D + +(X'X)^{-1}X')'\right]$$

Since $CX= I = DX + (X'X)^{-1}X'X$ the $DX = 0$

Cross product terms vanish and

$$Var(\tilde{\beta}) = \sigma^2 DD' + \sigma^2(X'X)^{-1} = \sigma^2 DD' + Var(\hat{\beta})_{OLS}$$

ie variance of alternative estimator equals that of OLS plus a non-negative definite matrix (see problem set 0)

Hence OLS estimate has minimum variance property (BLUE – Best Linear Unbiased Estimator). Main reason for widespread use of OLS, will always provide estimators with smaller standard errors

Inference

Minimum variance and unbiased properties are main reasons for widespread use of OLS. Properties exist without saying anything about the distribution of the disturbance term.

However if want to test hypotheses about values of particular variable(s) need to make certain distributional assumptions

Specifically assume the (unobserved) true residual is normally distributed

$$u \sim N(0, \ \sigma^2 I)$$

Follows that $\hat{\beta} \sim N(\ \beta, \sigma^2(X'X)^{-1}\ )$

Since σ^2 never observed use instead an unbiased estimate based on observed OLS residuals

$$s^2 = \frac{\hat{u}'\hat{u}}{N-k} = \frac{RSS}{N-k}$$

using fact that $(u'u) = (N-k)\,\sigma^2$ (see Exercise 2 for proof)

Hence an unbiased estimate of $Var(\hat{\beta}\ ) = \sigma^2(X'X)^{-1}$ is

$Var(\hat{\beta}\ ) = s^2(X'X)^{-1}$