Lecture 7B

Goodness of Fit

→ consider the regression model:

$$y = \alpha + \beta x + \varepsilon$$

→ we want to know how well the model explains the variance of y

→ the predictions

$$\hat{y} = \hat{\alpha} + \hat{\beta} x \quad \text{where: } \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \quad \hat{\beta} = \frac{cov(y,x)}{var(x)}$$

must be compared ~~with~~ with the observed values of y

→ so compute the observed residuals

$$\hat{\varepsilon} = y - \hat{y}$$

→ note that

$$(y - \bar{y}) = (\hat{y} - \bar{y}) + \hat{\varepsilon}$$

→ also note that we use $(y-\bar{y})$ to compute the variance of y

→ so we're essentially going to look at variance today

$$\Sigma(y-\bar{y})^2 = \Sigma\left[(\hat{y}-\bar{y}) + \hat{\varepsilon}\right]^2$$

$$= \Sigma\left[(\hat{y}-\bar{y})^2 + 2\cdot(\hat{y}-\bar{y})\hat{\varepsilon} + \hat{\varepsilon}^2\right]$$

$$= \Sigma(\hat{y}-\bar{y})^2 + \underbrace{2\cdot\Sigma(\hat{y}-\bar{y})\hat{\varepsilon}}_{\text{FOCUS on this term}} + \Sigma\hat{\varepsilon}^2$$

$$\Sigma(\hat{y}-\bar{y})\hat{\varepsilon} = \Sigma(\hat{\alpha} + \hat{\beta}x - \bar{y})\hat{\varepsilon}$$

$$= \hat{\alpha}\underbrace{\Sigma\hat{\varepsilon}}_{\text{ZERO}} + \hat{\beta}\underbrace{\Sigma x\hat{\varepsilon}}_{\text{ASSUMED ZERO}} - \bar{y}\underbrace{\Sigma\hat{\varepsilon}}_{\text{ZERO}}$$

(by Gauss-Markov assumptions)

→ so if Gauss Markov assumption of $cor(\varepsilon, x) = 0$ holds then:

$$\sum (y - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum \hat{\varepsilon}^2$$

TOTAL SUM of SQUARES (TSS)	=	EXPLAINED SUM of SQUARES (ESS)	+	RESIDUAL SUM of SQUARES (RSS)

→ if we have a good regression model, then RSS will be low and ESS will be high relative to TSS

$$TSS = ESS + RSS$$

$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS}$$

$$\boxed{R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}}$$

→ In theory, we could use either ESS/TSS or $1 - \frac{RSS}{TSS}$ to compute R^2, but in practice we use the latter

$$R^2 = 1 - \frac{\sum \hat{\varepsilon}^2}{\sum (y - \bar{y})^2}$$

→ Notice that when residuals are smaller R^2 is higher

→ This measure essentially compares the variance of the residuals (remember that expected value of residual is zero) to the variance of y

→ A higher R^2 implies a better model fit

→ So R^2 is great, but if you just keep adding variables to the model, R^2 will rise

→ "Adjusted R^2" penalizes you for adding terms, but rewards you for better model fit

$$\text{Adj } R^2 = 1 - \frac{\sum \hat{\varepsilon}^2 / (n-k)}{\sum (y-\bar{y})^2 / (n-1)}$$

where k is number of estimated parameters

→ degrees of freedom

- $(n-1)$ is df when computing the variance of y
- $(n-k)$ is df of regression model

→ degrees of freedom

- suppose you have 100 observations
- to compute variance of y you first must compute $\bar{y}$, so subtract one ~~[illegible]~~ to get df

$$df = 100 - 1 = 99$$

◆ ◆ ◆

- in the case of a regression model you have estimated k parameters

$$y = \alpha + \beta x$$

here $k = 2$ (one for α, one for β)

so $df = 100 - 2 = 98$

◆ ◆ ◆

- the reverse occurs when working with explained sum of squares

↗

- when working w/ ESS

$$\hat{y} = \hat{\alpha} + \hat{\beta} x$$

so only <u>one</u> <u>variable</u> determines $\hat{y}$

$df = 1$

- Note however that if:

$$\hat{y} = \hat{\alpha} + \hat{\beta} x_1 + \hat{\gamma} x_2$$

then there would be $df = 2$

F-test

→ in previous lectures we focused on estimating coefficients & comparing their values to the std error (testing for statistical significance of coefficient)

→ But we may also want to test the overall model or compare one model to another

→ For that purpose, we use F-test

→ Simple example: test $H_0: \beta = 0$

- regression model: $y = \alpha + \beta x + \varepsilon$
- if $\beta = 0$, then $\bar{y} = \hat{\alpha}$

  because $\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}$

- so we want to test

  $y = \alpha + \varepsilon$ against $y = \alpha + \beta x + \varepsilon$

- the regression model has $(n-2)$ degrees of freedom
- the test imposes ONE Restriction

  i.e. $H_0: \beta = 0$, so $m = 1$

F-test

$$F = \frac{(RSS_R - RSS_{UR})/m}{RSS_{UR}/(n-k)} = \frac{(R^2_{UR} - R^2_R)/m}{(1-R^2_{UR})/(n-k)}$$

RSS_R is restricted sum of squares in the "restricted case" (ie when $\beta = 0$)

RSS_{UR} is unrestricted RSS

m is "numerator df" or in this case its the number of restrictions (in this case: one)

$(n-k)$ is "denominator df" or in this case its df of regression model

Note that when $\beta = 0$ the regression model becomes:

$$\boxed{\begin{aligned} y &= \alpha + \varepsilon \\ \hat{\alpha} &= \bar{y} \end{aligned}}$$

so $\sum \hat{\varepsilon}^2 = \sum (y - \bar{y})^2$

$RSS = TSS$ and $R_R^2 = 0$