Lecture 5

Chapters 1.5-1.6

1.5 Results from OLS

Review from yesterday:

1. Simple linear regression model – $Y_i = \beta_1 + \beta_2 X_i + u_i$
2. Fitted regression model – $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$
3. Normal equations for $\widehat{\beta_1}$ and $\widehat{\beta_2}$ – $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$ $\qquad \hat{\beta}_2 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

Exercise 1: Show that the mean value of the residuals is zero.

$$e_i = Y_i - \hat{Y}_i$$

$$\frac{1}{n}\sum e_i = \frac{1}{n}\sum Y_i - \frac{1}{n}\sum \hat{Y}_i$$

$$\frac{1}{n}\sum e_i = \frac{1}{n}\sum Y_i - \frac{1}{n}\sum(\hat{\beta}_1 + \hat{\beta}_2 X_i)$$

$$\bar{e} = \boxed{\frac{1}{n}\sum Y_i} - \frac{1}{n}\sum\hat{\beta}_1 - \boxed{\frac{1}{n}}\boxed{\sum}\hat{\beta}_2\boxed{X_i}$$

$$\frac{1}{n}\sum Y_i = \bar{Y}$$

$$\frac{1}{n}\sum X_i = \bar{X}$$

$$\bar{e} = \bar{Y} - \frac{1}{\cancel{n}}\cdot\cancel{n}\hat{\beta}_1 - \hat{\beta}_2\bar{X}$$

$$\bar{e} = \bar{Y} - (\bar{Y} - \hat{\beta}_2\bar{X}) - \hat{\beta}_2\bar{X}$$

$$\bar{e} = \bar{Y} - \bar{Y} + \hat{\beta}_2\bar{X} - \hat{\beta}_2\bar{X}$$

$$\boxed{\bar{e} = 0}$$

Note: This means $E(Y) = E(\hat{Y})$

$$E(e) = E(Y - \hat{Y})$$

$$0 = E(Y) - E(\hat{Y})$$

$$E(\hat{Y}) = E(Y)$$

Exercise 2: Show that the sample correlation coefficient between X and e is zero.

$$r_{X,e} = \frac{\sum(x_i-\bar{x})(e_i-\bar{e})}{\sqrt{\sum(x_i-\bar{x})^2\sum(e_i-\bar{e})^2}}$$

$\bar{e}=0$

For division to be $=0$, only the numerator needs to equal 0.

$$\sum(x_i-\bar{x})e_i$$

$$=\sum(x_ie_i-\bar{x}e_i)$$

$$=\sum(x_ie_i)-\sum(\bar{x}e_i)$$

$$=\sum x_ie_i-\bar{x}\sum e_i$$

$\bar{e}=\frac{1}{n}\sum e_i$

$$=\sum x_ie_i-\bar{X}n\bar{e}$$

$n\bar{e}=\sum e_i$

$\bar{e}=0$

$$=\sum x_ie_i$$

$e_i=Y_i-\hat{\beta}_1-\hat{\beta}_2X_i$

$$=\sum x_i(Y_i-\hat{\beta}_1-\hat{\beta}_2X_i)$$

$$=\sum X_iY_i-\hat{\beta}_1X_i-\hat{\beta}_2X_i^2$$

Yesterday when solving for the normal equations we found this as $\frac{\partial RSS}{\partial b_2}$ which $=0$. So

$$r_{X,e}=\frac{0}{\sqrt{\sum(x_i-\bar{x})^2\sum(e_i-\bar{e})^2}}$$

$$\boxed{r_{X,e}=0}$$

Note that we can also show that $\rho_{e,\hat{Y}}=0$.

1.6 Goodness of Fit: R^2

Often, we want to know how good a job the OLS estimates do at fitting Y.

Total sum of squares – sum of the squared deviations about the sample mean of Y. (TSS)

Formula – TSS

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Exercise 3: Deconstruct the TSS formula into its explained and unexplained parts.

$$
\begin{aligned}
TSS &= \sum (Y_i - \bar{Y})^2 \\
&= \sum (\hat{Y}_i + e_i - \bar{Y})^2 \\
&= \sum [(\hat{Y}_i - \bar{Y}) + e_i]^2 \\
&= \sum (\hat{Y}_i - \bar{Y})^2 + 2\sum (Y_i - \bar{Y}) e_i + \sum e_i^2 \\
&= \sum (\hat{Y}_i - \bar{Y})^2 + 2\sum Y_i e_i - 2\bar{Y}\sum e_i + \sum e_i^2 \\
&= \underbrace{\sum (\hat{Y}_i - \bar{Y})^2}_{ESS} + \underbrace{\sum e_i^2}_{RSS}
\end{aligned}
$$

$\sum e_i = n\bar{e} = 0$

$\sum Y_i e_i = 0$

(not shown)

$$TSS = ESS + RSS$$

Explained sum of squares – sum of squared deviations of fitted Y about its sample mean. (ESS)

Coefficient of determination – the proportion of the total sum of squares that is explained by the regression line, R^2.

Formula – R^2

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

Properties of R^2

1. It always lies between 0 and 1.
2. When it is 1, RSS is 0.
3. When it is 0, ESS is 0, and TSS=RSS.

Exercise 4: Rewrite the R^2 formula in terms of the RSS.

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

Exercise 5: Find the R^2 given the following information.

Observation	X	Y	$\hat{Y}$
1	1	4	2.9
2	2	3	4.3
3	3	5	5.7
4	4	8	7.1
Mean	2.5	5	5

e_i

$4-2.9=1.1$

$3-4.3=-1.3$

$5-5.7=-0.7$

$8-7.1=0.9$

$$R^2=\frac{ESS}{TSS}$$

$$ESS=(2.9-5)^2+(4.3-5)^2+(5.7-5)^2+(7.1-5)^2$$

$$=(-2.1)^2+(-0.7)^2+(0.7)^2+(2.1)^2$$

$$=4.41+0.49+0.49+4.41=\boxed{9.8}$$

$$TSS=(4-5)^2+(3-5)^2+(5-5)^2+(8-5)^2$$

$$=(-1)^2+(-2)^2+(0)^2+(3)^2$$

$$=1+4+0+9=\boxed{14}$$

$$R^2=\frac{9.8}{14}=\boxed{0.7}$$

Check: $R^2=1-\frac{RSS}{TSS}$

$$RSS=\sum e_i^2=(1.1)^2+(-1.3)^2+(-0.7)^2+(0.9)^2$$

$$=1.21+1.69+0.49+0.81$$

$$=\boxed{4.2}$$

$$R^2=1-\frac{4.2}{14}=1-0.3=\boxed{0.7}$$

Exercise 6: If the fitted regression line does a good job, then the fitted values of Y should be highly correlated with the true values of Y. Show that $r_{Y,\hat{Y}} = \sqrt{R^2}$

$$r_{Y,\hat{Y}} = \frac{\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sqrt{\sum (Y_i - \bar{Y})^2 \sum (\hat{Y}_i - \bar{Y})^2}} = \frac{\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sqrt{TSS \cdot ESS}}$$

simplify numerator with $Y_i = \hat{Y}_i + e_i$

$$\sum (\hat{Y}_i + e_i - \bar{Y})(\hat{Y}_i - \bar{Y})$$

$$= \sum ([\hat{Y}_i - \bar{Y}] + e_i)(\hat{Y}_i - \bar{Y})$$

$$= \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i (\hat{Y}_i - \bar{Y})$$

$$= ESS + \sum e_i \hat{Y}_i - \sum e_i \bar{Y}$$

$$= ESS + 0 - \bar{Y} n \underbrace{\bar{e}}_{\to 0}$$

$$= ESS$$

multiply by 1

$$r_{Y,\hat{Y}} = \frac{ESS}{\sqrt{TSS}\sqrt{ESS}} \cdot \frac{\sqrt{ESS}}{\sqrt{ESS}}$$

$$= \frac{\cancel{ESS}\sqrt{ESS}}{\sqrt{TSS}\ \cancel{ESS}}$$

$$= \frac{\sqrt{ESS}}{\sqrt{TSS}}$$

$$R^2 = \frac{ESS}{TSS}$$

$$\boxed{r_{Y,\hat{Y}} = \sqrt{\frac{ESS}{TSS}} = \sqrt{R^2}}$$