

Problem Set VIII

Content: Simple Linear Regression III

Question 1

How do we make sure that our model's predictive performance is good?

Solution:

$$u_p = Y_p - E(Y_p) \quad \text{AND} \quad e_p = Y_p - \hat{Y}_p$$

this eqn show us prediction error:

$$e_p = Y_p - Y_p + E(Y_p) - E(Y_p) = [Y_p - E(Y_p)] + [E(Y_p) - Y_p]$$

where $[Y_p - E(Y_p)] = u_p$ and $E(\hat{Y}_p) = E(Y_p)$

For a good prediction we should test:

$$H_0 : e_p = 0 \text{ and } H_A : e_p \neq 0$$

$$t = \frac{e_p}{\hat{\sigma}_{e_p}} \longrightarrow \hat{\sigma}_{e_p}^2 = \hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{x_p^2}{\sum x_i^2} \right) \quad \text{where } x_p^2 = (X_p - \bar{X})^2$$

Confidence interval for Y_p : $\hat{Y}_p \mp t_{\alpha/2} \hat{\sigma}(e_p)$

$\sigma(e_p)$ is different from the std deviation of the Y_p

$$\sigma_{\hat{Y}_p}^2 = \hat{\sigma}^2 \left(\frac{1}{T} + \frac{x_p^2}{\sum x_i^2} \right)$$

Question 2

We have the regression line for the retail sales per households which depends disposable income. $\hat{y}_{n+1} =$

$1.922 + 0.3815x_{n+1}$. Also we got the information that

$$n = 22, \bar{x} = 10799, \quad \sum (x_i - \bar{x})^2 = 34110178, \quad s_e^2 = 21806.$$

Find 95% confidence interval for next year's retail sales when disposable income is 12000\$.

Solution:

$$\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} s_e = \sqrt{\left[1 + \frac{1}{22} + \frac{(12000 - 10799)^2}{34110178} \right]} \sqrt{21806} = 154.01$$
$$t_{n-2, \alpha/2} = t_{20, 0.025} = 2.086$$

When we put the given disposable income to our regression line we found that

$$y_{n+1} = 1.922 + (0.3815)(12000) = 6501$$

So our confidence interval is: $6501 \mp (2.086)(154.01) = 6501 \mp 321$

Question 3

A sample of 25 blue-collar employees at a production plant was taken. Each employee was asked to assess his or her own job satisfaction(x) on a scale from 1 to 10. In addition, the numbers of days absent(y) from work during the last year were found for these employees. The sample regression line $\hat{y} = 12.6 - 1.2x$ was estimated by least squares for this data. Also found were

$$\bar{x} = 6, \quad \sum_{i=1}^{25} (x_i - \bar{x})^2 = 130, \quad s_e^2 = 80.6$$

A particular employee has a job satisfaction level 4 . Find a %90 interval for the number of days this employee would absent from work in a year.

Solution:

For the prediction of the conditional expectation $E[Y_{n+1} | X_{n+1}]$ the confidence interval is:

$$\hat{y}_{n+1} \pm t_{n-2, \alpha/2} \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{25} (x_i - \bar{x})^2}} s_e \quad \longrightarrow \quad \sqrt{1 + \frac{1}{25} + \frac{(6-4)^2}{130}} (\sqrt{80.6}) \cong 9.29$$

$$y_{n+1} = 12.6 - 1.2X4 = 7.8 \text{ and } t_{23, 0.1/2} = t_{23, 0.05} = 1.714$$

So our confidence interval is

$$7.8 \pm 1.714X9.29$$

Question 4

Suppose you estimate the consumption function:

$$Y_i = \alpha X_i + u_{1i}$$

And the saving function

$$Z_i = \beta X_i + u_{2i}$$

Where Y = consumption, Z = savings, X =income, and $X = Y + Z$, that is income is equal to consumption plus savings.

- What is the relationship, if any, between $\hat{\alpha}$ and $\hat{\beta}$?
- Will the SSR be the same for the two models?
- Can you compare the R^2 terms of the two models?

Solution:

- $X = Y + Z$

$$\begin{aligned} \hat{\alpha} &= \frac{\sum X_t Y_t}{\sum X_t^2} & \hat{\beta} &= \frac{\sum X_t Z_t}{\sum X_t^2} \\ \hat{\alpha} &= \frac{\sum X_t (X_t - Z_t)}{\sum X_t^2} = 1 - \frac{\sum X_t Z_t}{\sum X_t^2} = 1 - \hat{\beta} \\ \hat{\alpha} + \hat{\beta} &= 1 \quad (MPC + MPS = 1) \end{aligned}$$

- b. $\hat{U}_1 = Y_t - \hat{Y}_t \quad \hat{U}_2 = Z_t - \hat{Z}_t$
 $\hat{U}_1 = Y_t - \hat{\alpha}X_t \quad \hat{U}_2 = Z_t - \hat{\beta}X_t$
 $\sum \hat{U}_1^2 = \sum (Y_t - \hat{\alpha}X_t)^2 \quad \sum \hat{U}_2^2 = \sum (Z_t - \hat{\beta}X_t)^2$
 $\hat{\alpha} = 1 - \hat{\beta}$
 $\sum \hat{U}_{t1}^2 = \sum (Y_t - X_t + \beta X_t)^2$
 $\sum \hat{U}_{t1}^2 = \sum (Z_t - \hat{\beta}X_t)^2$
 $\sum \hat{U}_{t2}^2 = \sum (Z_t - \hat{\beta}X_t)^2$
- c. Their R^2 's cannot be compared because their dependent variables are different.

Question 5

A researcher wishes to estimate the parameters of the following model:

Model 1: $Y_t = \beta_0 + \beta_1 X_t + u_t$

However, her assistant confuses the dependent and independent variables and estimates the following model:

Model 2: $X_t = \alpha_0 + \alpha_1 Y_t + v_t$, and reports

$X_t = -1.44 + 0.709Y_t$

$(-1.55) \quad (30.68)$

$R_2^2 = 0.986, \text{SSE}_2 = 35.76, T = 15$

- Using the results for Model 2, estimate β_1 and its standard error.
- Find the coefficient of determination for the first model.
- Test if β_1 is not significantly different from zero.

Solution:

$$\begin{aligned} \text{i. } R_1^2 &= \frac{\sum \hat{y}_t^2}{\sum y_t^2} = 1 - \frac{\sum \hat{u}_t^2}{\sum y_t^2} = \hat{\beta}_1 \frac{\sum x_t y_t}{\sum y_t^2} = \hat{\beta}_1^2 \frac{\sum x_t^2}{\sum y_t^2} = \frac{(\sum x_t y_t)^2}{\sum y_t^2 \sum x_t^2} \\ \hat{\beta}_1 &= \frac{\sum x_t y_t}{\sum x_t^2} \quad \hat{\alpha}_1 = 0.709 = \frac{\sum x_t y_t}{\sum y_t^2} \\ R_2^2 &= \alpha_1 \frac{\sum x_t y_t}{\sum x_t^2} = 0.986 \quad 0.986 = (0.709) \frac{\sum x_t y_t}{\sum x_t^2} \Rightarrow \beta_1 = \frac{0.986}{0.709} = 1.39 \\ \text{Var}(\hat{\beta}_0) &= \frac{\sum u_t^2}{(T-k-1) \sum x_t^2} = \frac{\sum \hat{u}_t^2}{(T-2) \sum x_t^2} \quad R_2^2 = 1 - \frac{\sum \hat{v}_t^2}{\sum x_t^2} = 0.986 = 1 - \frac{\text{SSE}}{\sum x_t^2} \quad \text{SSE} = 35.76 \\ &\Rightarrow \sum x_t^2 = 2554 \\ R_2^2 &= \hat{\alpha}_1^2 \frac{\sum y_t^2}{\sum x_t^2} = \frac{(0.709)^2 \sum y_t^2}{2554} = 0.986 \\ \sum y_t^2 &= 5009 \quad R_1^2 = R_2^2 = 1 - \frac{\sum \hat{u}_t^2}{\sum y_t^2} = 0.986 \Rightarrow \sum \hat{u}_t^2 = 70.126 \\ \text{Var}(\hat{\beta}_1) &= \frac{70.126}{13(2554)} = 0.0021 \quad \sigma_{\hat{\beta}_1} = \sqrt{0.0021} = 0.049 \end{aligned}$$

ii. $R^2 = r_{xy}^2$
 $r^2 = 0.986$

iii. $H_0 = \beta_1 = 0$

$H_a = \beta_1 \neq 0$
Reject H_0 !

Question 6

Let x denote the biochemical oxygen demand (BOD) mass loading and y denote BOD mass removal. A linear regression model is formed

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

And the following information is given about the model:

$$n = 14, \sum x_i = 517, \sum y_i = 346, \sum x_i^2 = 39,095, \sum x_i y_i = 25,825$$

- Obtain the OLS estimators.
- Predict the value of *BOD* mass removal for a single observation made when BOD mass loading is 35.
- Calculate the SSE and then a point estimate of σ .
- What portion of observed variation in removal can be explained by the approximate linear relationship between two variables?

Solution:

a.

$$S_{xx} = 39,095 - \frac{(517)^2}{14} = 20002.929 \quad S_{xy} = 25,825 - \frac{(517)(346)}{14} = 13047.714$$

From these we can conclude $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = 0.652$ and $\hat{\beta}_0 = 0.626$

b. From part a our fitted equation is

$$\hat{y} = 0.626 + 0.652x$$

When we substitute $x = 35$ we are left with $\hat{y} = 0.626 + 0.652(35) = 23.456$

c. First we compute S_{yy}

$$S_{yy} = 17,454 - \frac{(346)^2}{14} = 8902.857 \quad SSE = 8902.857 - (0.652)(13047.714) = 395.747$$

Then we can estimate the variance by $\hat{\sigma} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{395.747}{12}} = 5.743$

d. $SST = S_{yy} = 8902.857$ then we can compute R^2 by

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{395.747}{8902.857} = 0.956$$

So the model accounts for about 95 % of the variation