Jane Downer
A20452471
September 11, 2021

MATH 484: HW 1

Problem 1 The purity of oxygen (Y) produced by a fractional distillation process is thought to be related to the percentage of hydrocarbons (X) in the main condensor of the processing unit. Twenty samples were measured and shown in the attached data sheet.

1) Fit a simple linear regression model to the data, provide the liner regression line equation $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$.

$\bar{x} = 91.818$

$\bar{y} = 1.1825$

$$\hat{\beta}_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \quad \begin{matrix} = 12.5678 \\ = 1.064975 \end{matrix}$$

$= 11.80103$

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

$\boxed{\hat{Y} = 77.863 + 11.801X}$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$= 91.818 - 11.80103 \times 1.1825$

$= 77.86328$

2) Test the hypothesis $H_0 : \beta_1 = 0$ against the alternative $H_A : \beta_1 \neq 0$, and conclude if there is significant linear relationship between the purity of oxygen and the percentage of hydrocarbons.

$SSE = \Sigma(Y_i - \hat{Y})^2 = 232.8344$

$\hat{\sigma}^2 = \frac{SSE}{n-2} = 232.8344/(20-2) = 12.93524$

$s.e.(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\Sigma(X_i - \bar{X})^2}} = \sqrt{12.93524/1.064975} = 3.4851$

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_a : \beta_1 \neq 0 \end{cases}$$

t-test: $T = \frac{\hat{\beta}_1 - \beta_1^0}{s.e.(\hat{\beta}_1)} = \frac{11.801 - 0}{3.4851} = 3.386$

$t_{\alpha/2, n-2} = t_{0.025, 18} = 2.10$

$T > t_{\alpha/2, n-2} \rightarrow \boxed{\text{reject } H_0}$

3) Calculate the coefficient of determination, r^2.

$$r^2 = 1 - \frac{SSE}{SST} \qquad SSE = 232.834 \qquad SST = \Sigma(y_i - \bar{y})^2 = 381.147$$

$$\boxed{r^2 = 0.389}$$

4) Find a 95% confidence interval on the slope.

$$95\% \text{ C.I.}: \hat{\beta}_1 \pm t_{0.05/2,\, n-2} \times S.E.(\hat{\beta}_1)$$

$$11.801 \pm 2.10 \times 3.485 \longrightarrow \boxed{[4.4791, 19.123]}$$

5) Find a 95% confidence interval on the mean purity when the hydrocarbon percentage is 1.05.

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 \times 1.05$$
$$= 77.863 + 11.801 \times 1.05$$
$$= 90.2541$$

95% C.I. for $E(\hat{Y}_0)$:

$$\hat{Y}_0 \pm t_{\alpha/2,\, n-2} \times S.E.(\hat{Y}_0)$$

$$90.2541 \qquad 2.1 \times 0.8984$$

$$S.E.(\hat{Y}_0) = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\Sigma(x_i - x_0)^2}\right)}$$

$$= \sqrt{\frac{SSE}{n-2} \times \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\Sigma(x_i - x_0)^2}\right)}$$

$$= \sqrt{\frac{232.834}{20-2}\left(\frac{1}{20} + \frac{(1.05 - 1.1825)^2}{\underbrace{[\Sigma(x_i - 1.05)^2]}_{= 1.416}}\right)}$$

$$= \sqrt{0.807}$$

$$= 0.8984$$

$$\boxed{[88.367, 92.142]}$$

6) What is the correlation coefficient between Y and X?

$$r = \sqrt{r^2} = \sqrt{0.389} = \boxed{0.6238}$$

7) Test the hypothesis: $H_0 : \rho = 0$ against $H_A : \rho \neq 0$ using a t-test based on the correlation coefficient computed from the previous step.

$$\begin{cases} H_0: \rho = 0 \\ H_a: \rho \neq 0 \end{cases}$$

identical to test for $\hat{\beta}_1 = 0$

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$= \frac{0.6238\sqrt{18}}{\sqrt{1-0.389}}$$

$$= 3.386$$

$$T = 3.386$$

$$t_{\alpha/2, n-2} = 2.1$$

$$T > t_{\alpha/2, n-2} \rightarrow \boxed{\text{reject } H_0}$$

Problem 2 Consider the simple linear regression model $Y = \beta_0 + \beta_1 X + \epsilon$. Assume that $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$, and ϵ_i are independent of each other. Show that $Cov(\bar{Y}, \hat{\beta}_1) = 0$.

$$\text{cov}(\bar{Y}, \hat{\beta}_1) = ?$$

$$\hat{\beta}_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2} \xrightarrow{\text{simplify}}$$

$$\Sigma(X_i - \bar{X})(Y_i - \bar{Y}) = \Sigma Y_i * (X_i - \bar{X}) - \Sigma \bar{Y} * (X_i - \bar{X})$$

$$= \Sigma Y_i * (X_i - \bar{X}) - (\bar{Y}\,\Sigma X_i - \bar{Y}\,\Sigma \bar{X})$$

$$= \Sigma Y_i * (X_i - \bar{X}) - (\bar{Y} * \bar{X} n - \bar{Y} * \bar{X} n)$$

$$= \Sigma Y_i * (X_i - \bar{X})$$

$$= \frac{\Sigma Y_i * (X_i - \bar{X})}{\Sigma(X_i - \bar{X})^2}$$

let $k_i = \dfrac{(X_i - \bar{X})}{\Sigma(X_i - \bar{X})^2}$

$$= \Sigma k_i Y_i$$

$$\bar{Y} = \frac{\Sigma Y_i}{n}$$

$$\text{cov}(\bar{Y}, \hat{\beta}_1) = \text{cov}\left(\frac{\Sigma Y_i}{n}, \Sigma k_i Y_i\right)$$

$$= \underbrace{\sum_{i=1}^{n} \frac{1}{n} \cdot k_i \cdot \text{var}(Y_i)}_{i = j} + \underbrace{\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{1}{n} \cdot k_i \cdot \cancelto{0}{\text{cov}(Y_i, Y_j)}}_{i \neq j}$$

$$= \sum_{i=1}^{n} \frac{1}{n} \cdot k_i \cdot \text{var}(Y_i)$$

$$= \frac{\sigma^2}{n} \sum_{i=1}^{n} k_i$$

$$= \frac{\sigma^2}{n} \frac{\cancelto{0}{\Sigma(X_i - \bar{X})}}{\Sigma(X_i - \bar{X})^2}$$

$$= 0$$

Problem 3 Show that the sample correlation coefficient r between X and Y is a value between -1 and 1, i.e., $-1 \leq r \leq 1$. Use two different methods to show it, with one using only the data, and the other one in the context of simple linear regression.

method 1:

$$r^2 = 1 - \frac{SSE}{SST}$$

$$= 1 - \frac{\Sigma(Y_i - \hat{Y})^2}{\Sigma(Y_i - \bar{Y})^2}$$

$$= 1 - \frac{232.834}{381.147}$$

$$= 0.389$$

$r = \pm 0.624$

It will be the same sign as $\hat{\beta}_1$

$\hat{\beta}_1$ is positive

$\therefore r = 0.624$

↳ satisfies condition $-1 \leq r \leq 1$ ✓✓

method 2:

$$SSE = \Sigma(y_i - \hat{y}_i)^2$$

$$SSR = \Sigma(\hat{y}_i - \bar{y})^2$$

$$SST = SSE + SSR = \Sigma(y_i - \bar{y})^2$$

$$r^2 = 1 - \frac{SSE}{SST}$$

$$= 1 - \frac{SSE}{SSE + SSR}$$

SSE will be 0 if $y_i = \hat{y}_i$ for all i
Otherwise, SSE will be greater than or equal to zero, since it is a squared value.

SSR is also a squared value — greater than or equal to zero

i.e., $SSE \leq SSE + SSR$

$\frac{SSE}{SSE + SSR}$ is a fraction between 0 and 1, inclusive

$$\min\left(1 - \frac{SSE}{SSE + SSR}\right) = 1 - 1 = 0$$

$$\max\left(1 - \frac{SSE}{SSE + SSR}\right) = 1 - 0 = 1$$

$\therefore r^2 \leq 1$ and $-1 \leq r \leq 1$ ✓✓

Problem 4 Consider the simple linear regression model $Y = \beta_0 + \beta_1 X + \epsilon$. Assume that $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$. Find the variance of $\hat{Y}_h$, where $\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h$.

$$\hat{\beta}_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2} = \frac{\Sigma Y_i * (X_i - \bar{X})}{\Sigma(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\begin{aligned} var(\hat{Y}_h) &= var(\hat{\beta}_0 + \hat{\beta}_1 X_h) \\ &= var(\bar{Y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 X_h) \\ &= var\left(\frac{\Sigma Y_i}{n} + \hat{\beta}_1(\bar{X} - X_h)\right) \\ &= \left(\frac{1}{n}\right)^2 \cdot n \cdot var(Y_i) + (\bar{X} - X_h)^2 \cdot var(\hat{\beta}_1) \\ &= \frac{1}{n}\sigma^2 + (\bar{X} - X_h)^2 \cdot \frac{1}{\Sigma(X_i - \bar{X})^2}\sigma^2 \\ &= \boxed{\sigma^2\left[\frac{1}{n} + \frac{(\bar{X} - X_h)^2}{\Sigma(X_i - \bar{X})^2}\right]} \end{aligned}$$

$$\begin{aligned} var(\hat{\beta}_1) &= var\left(\frac{\Sigma Y_i * (X_i - \bar{X})}{\Sigma(X_i - \bar{X})^2}\right) \\ &= \frac{\Sigma(X_i - \bar{X})^2}{\left[\Sigma(X_i - \bar{X})^2\right]^2} var(Y_i) \\ &= \frac{1}{\Sigma(X_i - \bar{X})^2}\sigma^2 \end{aligned}$$

Problem 5 Exercise 2.9 from the TEXT.

2.9 Let Y and X denote the labor force participation rate of women in 1972 and 1968, respectively, in each of 19 cities in the United States. The regression output for this data set is shown in Table 2.10. It was also found that SSR = 0.0358 and SSE = 0.0544. Suppose that the model $Y = \beta_0 + \beta_1 X + \varepsilon$ satisfies the usual regression assumptions.

Table 2.10 Regression Output When Y is Regressed on X for Labor Force Participation Rate of Women

Variable	Coefficient	s.e.	t-Test	p-value
Constant	β_0 0.203311	0.0976	2.08	0.0526
X	β_1 0.656040	0.1961	3.35	< 0.0038
$n = 19$	$R^2 = 0.397$	$R_a^2 = 0.362$	$\hat{\sigma} = 0.0566$	df = 17

(a) Compute Var(Y) and Cor(Y, X).

$$\mathrm{var}(Y) = \hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{0.0544}{19-2} = 0.0032 = 0.0566^2 \checkmark$$

$$\mathrm{cor}(Y,X) = \sqrt{\frac{SSR}{SST}} = \sqrt{\frac{SSR}{SSR + SSE}} = \sqrt{\frac{0.0358}{0.0358 + 0.0544}} = 0.63$$

$$\boxed{\begin{array}{l} \mathrm{var}(Y) = 0.0032 \\ \mathrm{cor}(Y,X) = 0.63 \end{array}}$$

(b) Suppose that the participation rate of women in 1968 in a given city is 45%. What is the estimated participation rate of women in 1972 for the same city?

$$Y = 0.203311 + 0.65604 * 0.45 = 0.4985$$

$\boxed{\text{Estimated participation: } 49.85\%}$

(c) Suppose further that the mean and variance of the participation rate of women in 1968 are 0.5 and 0.005, respectively. Construct the 95% confidence interval for the estimate in (b).

$$t_{0.05/2,\, 19-2} = 2.11$$

$$\text{C.I.} = 0.4985 \pm t_{\alpha/2,\, n-2} * \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(x_i - \bar{x})^2}}$$

$$\mathrm{var}(x) = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$$

$$\hookrightarrow \Sigma(x_i - \bar{x})^2 = 18 \times 0.005 = 0.09$$

$$= 0.4985 \pm 2.11 * 0.0566\sqrt{\frac{1}{19} + \frac{(0.45 - 0.5)^2}{0.09}}$$

$$= \boxed{[0.465, 0.532]}$$

(d) Construct the 95% confidence interval for the slope of the true regression line, β_1.

Given: $\hat{\beta}_1 = 0.203311$

$s.e.(\hat{\beta}_1) = 0.1961$ (given)

$$C.I. = \hat{\beta}_1 \pm t_{\alpha/2,\, n-2} * s.e.(\hat{\beta}_1)$$

$$t_{0.05/2,\, 19-2} = 2.11$$

$$= 0.65604 \pm 2.11 * 0.1961$$

$$= \boxed{[0.2423,\ 1.06981]}$$

(e) Test the hypothesis: $H_0 : \beta_1 = 1$ versus $H_1 : \beta_1 > 1$ at the 5% significance level.

$$T = \frac{\hat{\beta}_1 - \beta_1^0}{s.e.(\hat{\beta}_1)}$$

Reject if $T > t_{\alpha,\, n-2}$

$$= \frac{0.65604 - 1}{0.1961}$$

$$= -1.754$$

$$t_{\alpha,\, n-2} = t_{0.05,\, 17}$$

$$= 1.74$$

T is not greater than t, so we fail to reject H_0.

(f) If Y and X were reversed in the above regression, what would you expect R^2 to be?

$$r^2 = (cor(Y, X))^2$$

$$= \frac{[\Sigma(X_i - \bar{x})(y_i - \bar{y})]^2}{\Sigma(X_i - \bar{x})^2 \cdot \Sigma(y_i - \bar{y})^2}$$

This formula will be the same regardless

i.e., $[cor(Y, X)]^2 = [cor(X, Y)]^2$

$\therefore$ $\boxed{r^2 \text{ will not change}}$