STAT 2509 A - TEST #1 //90

SOLUTION

Q.1. [1] a) the response variable, y, is lung damage (1)

[1] b) the explanatory variable, x, is yrs of smoking (1)

[6] c) SLR model: $y = \beta_0 + \beta_1 x + \varepsilon$ (1), $n = 10$

where (i) x_i's are observed without error (1)
(ii) y_i's are independently (1) distributed with mean $E(y) = \beta_0 + \beta_1 x$ (1)
(iii) variance of y's is constant (1) for all x's
(iv) $y \sim$ (1) $N(E(y), \sigma^2)$ for any value of x

or, equivalently

(i) x's observed without error
(ii) ε_i's are indep. distributed with mean $E(\varepsilon) = 0$
(iii) variance of ε's is constant for all x's
(iv) $\varepsilon \sim N(0, \sigma^2)$ for any value of x

[5] d) $\hat{\beta}_1$ (1) $= \frac{S_{xy}}{S_{xx}} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} =$

$= \frac{18\,055 - \frac{(319)(530)}{10}}{11\,053 - \frac{(319)^2}{10}} = \frac{1148}{876.9} = 1.309157 \doteq$

$\doteq \underline{1.309}$ (1)

$\hat{\beta}_0$ (1) $= \bar{y} - \hat{\beta}_1 \bar{x} = \frac{\sum y_i}{n} - \hat{\beta}_1 \frac{\sum x_i}{n} =$

$= \frac{530}{10} - (1.309157)\left(\frac{319}{10}\right) = 11.237788$ (1) $\doteq \underline{11.238}$

∴ the least-squares fitted regr. line is given by

$$\hat{y} = 11.238 + 1,309x \quad (1)$$

e) [6] $s^2 = \frac{SSE}{n-2}$ (1), where $SSE = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$ (1) =

$$= \left[\sum_i y_i^2 - \frac{(\sum_i y_i)^2}{n}\right] - \frac{\left[\sum_i x_i y_i - \frac{(\sum_i x_i)(\sum_i y_i)}{n}\right]^2}{\sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n}} =$$

$$= \left[30\,600 - \frac{(530)^2}{10}\right] - \frac{(1148)^2}{876,9} = 2510 - 1502.913 \quad (1)(1)$$

$$= 1007.087 \quad (1)$$

$$\therefore s^2 = \frac{SSE}{n-2} = \frac{1007,087}{8} = 125,8859 \quad (1)$$

$$\therefore s = \sqrt{s^2} = 11,21989$$

f) [6] $H_0: \beta_1 = 0$ } (1); $\alpha = 0,10 \Rightarrow \alpha/2 = 0,05$
$H_a: \beta_1 \neq 0$ }

- test-stat: $t = \frac{\hat{\beta}_1}{s/\sqrt{S_{xx}}}$ (1) $= \frac{1,309157}{11,21989/\sqrt{876,9}} = 3,45524$ (1)

R.R. - we reject H_0 if $t < -t_{\alpha/2;n-2} = -t_{0,05;8} = -1,860$
or $t > t_{\alpha/2;n-2} = t_{0.05;8} = 1,860$ } (1)

- since $t = 3,45524$ (✓1) $> 1,860$, we reject (1/2) H_0 and conclude that at 10% level of signif. there is an evidence that a linear relationship between yrs of smoking and lung damage exists. (1/2)

g) 90% C.I. for β_1 : [4]

$1-\alpha = 0.90$
$\alpha = 0.10$
$\alpha/2 = 0.05$

$$\beta_1 \in \left(\hat{\beta}_1 \pm t_{\alpha/2;\,n-2} \times S/\sqrt{S_{xx}}\right) =$$

$$= \left(1.309 \pm \underbrace{t_{0.05;\,8}}_{\text{"}1.860} \times \frac{11.21989}{\sqrt{876.9}}\right) =$$

$$= (1.309 \pm 0.704736) = (0.604264,\ 2.013736)$$

ie. we are 90% confident that in repeated sampling the true value of the popue^n slope β_1 will lie in the interval $(0.604,\ 2.014)$.

h) [23]

TSS $= S_{yy} = 2510$ (given)

SSE $= 1007.087$ (calculated in part (e))

SSR $=$ TSS $-$ SSE $= 1502.913$

$$MSR = \frac{SSR}{1} = 1502.913$$

$$MSE = \frac{SSE}{n-2} = \frac{1007.087}{8} = 125.8859$$

$$F = \frac{MSR}{MSE} = 11.93869$$

ANOVA table:

Source of Variation	d.f.	SS	MS	F
Regression	1	1502.913	1502.913	11.93869
Error	8	1007.087	125.8859	
Total	9	2510		

$H_0: \beta_1 = 0$
$H_a: \beta_1 \neq 0$ } (1) $\alpha = 0.10$

- test-stat. $F = \frac{MSR}{MSE} = 11.93869$ (1)

- R.R. - we reject H_0 if $F > F_\alpha(1, n-2) = F_{0.10}(1, 8) = 3.46$ (1)

- since $F = 11.93869$ (1) > 3.46, we reject (1) H_0 and conclude that at 10% level of signif. there is an evidence that a linear relationship between yrs of smoking and lung damage exists. (1)

i) [5] $r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = \frac{1148}{\sqrt{(276.9)(2510)}} = 0.773802 \doteq$ (1)

$\doteq 0.774$

ie. the lung damage and yrs of smoking are positively correlated (related) with the strength of the relationship approx. 77.4%. (1/2)

$r^2 = \frac{SSR}{TSS} = 0.59877 \doteq 0.599$ (1)

ie. approx. 60% of the total variation in the data is explained by the regression line (and 40% is due to error). (1/2)

ie. Model is not good (1) as 60% is quite low.

j) [5] 95% C.I. for $E(y)$ when $x_p = 40$

$\Rightarrow \hat{y} = 11.238 + 1.309(40) = 63.598$ (1)

and $1 - \alpha = 0.95$
$\alpha = 0.05$
$\alpha/2 = 0.025$

$\Rightarrow E(y) \in \left(\hat{y} \pm t_{\alpha/2;n-2} \times s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}}\right) =$ (1)

$= \left(63.598 \pm \underbrace{t_{0.025;8}}_{=2.306} \times (11.21989)\sqrt{\frac{1}{10} + \frac{(40-31.9)^2}{876.9}}\right) =$ (1)

$= (63.598 \pm 10.81792) = (52.78008, 74.41592) \doteq$

$\doteq (52.78, 74.42)$ (1)

ie. We are 95% confident that an average lung damage of people who smoked for 40 years falls in interval $(52.78, 74.42)$. (1)

k) 95% P.I. for y when $x_p = 40$

[5] $\Rightarrow \hat{y} = 63.598$ (1)

$y \in \left(\hat{y} \pm t_{\alpha/2;n-2} \times s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}}\right) =$ (1)

$= \left(63.598 \pm \underbrace{t_{0.025;8}}_{=2.306} \times (11.21989)\sqrt{1 + \frac{1}{10} + \frac{(40-31.9)^2}{876.9}}\right) =$ (1)

$= (63.598 \pm 28.04359) = (35.55441, 91.64159) \doteq$

$\doteq (35.55, 91.64)$ (1)

ie. We are 95% confident that the lung damage for an individual who smoked for 40 years falls in interval $(35.55, 91.64)$. (1)

Q.2. a) $SSE = S_{yy} - \frac{(S_{xy})^2}{S_{xx}} =$ (1)

[5] $= \left[\sum_i y_i^2 - \frac{(\sum_i y_i)^2}{n}\right] - \frac{\left[\sum_i x_i y_i - \frac{(\sum_i x_i)(\sum_i y_i)}{n}\right]^2}{\sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n}} =$

$n = 12$

$= \left[2854 - \frac{(178)^2}{12}\right] - \frac{\left[373 - \frac{(27)(178)}{12}\right]^2}{\left(64.5 - \frac{(27)^2}{12}\right)} =$ (1)

$= 12.000004 \cong 12$

and $SSE = SSPE + SSLF$ (1), where $SSPE = 7.3333333$ given

So $SSLF = SSE - SSPE =$ (1)

$= 4.6666707$ (1)

[6] b) $y = \beta_0 + \beta_1 x + \varepsilon$, $\alpha = 0.05$

H_0: model is appropriate
H_a: model is not appropriate (1)

-test-stat: $F = \frac{MSLF}{MSPE} = \frac{SSLF / [n-2-\sum_i (n_i - 1)]}{SSPE / \sum_i (n_i - 1)} =$ (1)

$= \frac{4.6666707/(10-8)}{7.3333333/8} = \frac{2.3333354}{0.9166666} = 2.5454568$ (1)

-R.R: - we reject H_0 if $F > F_{\alpha, [n-2-\sum_i (n_i-1),\ \sum_i (n_i-1)]} =$

$= F_{0.05, [2, 8]} = 4.46$ (1)

-since $F = 2.5454 \not> 4.46$ (1), we do not reject H_0 and conclude that at 5% level of signif. there is not enough evidence to say that the lin. model is not appropriate. (1/2) (1/2)

Q.3:

a) [3] • violation of the assumption of constant variance, since the residual values are increasing with x (1) (2)

b) [3] • violation of the assumption of errors being normally distributed, since the histogram of residuals is not bell-shaped nor is it symmetric (1) (1) (It is in fact negatively skewed)

c) [3] • violation of linearity (or independence), since curve-linear pattern (1) (2)

d) [3] • no violations, since residuals are randomly scattered around their mean (i.e. no pattern) (1) (2)