/90

STAT 2509C
Test#1
SOLUTION

1. A computer-equipment outlet sells an imported personal computer (PC) on a franchise basis and performs preventive maintenance and repair service on this PC. The following data have been collected from 16 recent calls on users to perform routine preventive maintenance service. Suppose they are interested in knowing how the number of machines serviced influences the total number of minutes spent by the service person.

# of serviced machines	Total # of minutes spent
6	86
5	95
1	18
5	69
4	62
7	101
4	39
4	53
2	33
8	102
5	65
2	25
7	105
1	17
4	55
5	68

$\sum x_i = 70$, $\sum x_i^2 = 372$, $\sum y_i = 993$, $\sum y_i^2 = 75\,187$, $\sum x_i y_i = 5\,246$

[1] (a) The response variable, y, is: **Total number of minutes spent** (1)

[1] (b) The explanatory variable, x, is: **# of serviced machines** (1)

[6] (c) State a SLR model making sure you give all assumptions necessary for statistical inference.

Model: $y = \beta_0 + \beta_1 x + \varepsilon$, **$n$ = 16** (1)

Assumptions: (i) *x*'s are observed without error (1)

(ii) *y*'s (or ε's) are independently distributed with mean $E(y) = \beta_0 + \beta_1 x$ (1)

(or $E(\varepsilon) = 0$) (1)

(iii) variance of *y*'s (or ε's) is constant, σ^2 for all *x*'s

(iv) $y \sim N(E(y), \sigma^2)$ for any value of *x* (or $\varepsilon \sim N(0, \sigma^2)$ for any value of *x*) (1)

[5] (d) Find the least squares estimates of β_0 and β_1. Find the least squares fitted regression line.

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{5246 - \frac{(70)(993)}{16}}{372 - \frac{(70)^2}{16}} = \frac{901.625}{65.75} = 13.71293 \cong \underline{13.713}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{993}{16} - (13.71293)\left(\frac{70}{16}\right) = 62.0625 - 59.99406 =$$

$$= 2.068441 \cong \underline{2.068}$$

∴ **the least squares fitted regression line is given by:** $\hat{y} = \underline{2.068 + 13.713}\, x$

Assuming no violations of the assumptions, answer the following questions:

[6] (e) Find s^2, an estimate of σ^2.

$$s^2 = \frac{SSE}{n-2} = \frac{S_{yy} - \frac{S_{xy}^2}{S_{xx}}}{n-2} = \frac{\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right] - \frac{\left[\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}\right]^2}{\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]}}{n-2} =$$

$$= \frac{\left[75187 - \frac{(993)^2}{16}\right] - \frac{(901.625)^2}{65.75}}{14} = \frac{13558.94 - 12363.92}{14} = \frac{1195.019}{14} = \underline{85.3585}$$

∴ $s = \sqrt{s^2}$ = **9.238966**

[6] (f) Use the t-test to test whether there is a significant linear relationship between the number of machines serviced and the total number of minutes spent. Use α = 0.10.

$H_0 : \beta_1 = 0$ $\quad \alpha = 0.10 \Rightarrow \alpha/2 = 0.05$

$H_a : \beta_1 \neq 0$

test-statistics: $t = \dfrac{\hat{\beta}_1}{s/\sqrt{S_{xx}}} = \dfrac{13.71293}{9.238966/\sqrt{65.75}} = \underline{\mathbf{12.03524}}$

R.R: **we reject** H_0 **if** $t < -t_{\alpha/2;n-2} = -t_{0.05;14} = -1.761$

or $t > t_{\alpha/2;n-2} = t_{0.05;14} = 1.761$

Since *t* = 12.03524 > 1.761, we reject H_0 and conclude that at 10% level of significance there is an evidence to say that the number of machines serviced and the total number of minutes spent are linearly related.

[4] (g) Find a 90% confidence interval for the true population slope, β_1.

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

$$\beta_1 \in \left(\hat{\beta}_1 \pm t_{\alpha/2;n-2} \, s/\sqrt{S_{xx}} \right) = \left(13.71293 \pm t_{0.05;14} \, 9.238966/\sqrt{65.75} \right) = (13.71293 \pm 1.761(1.139398)) =$$

$$= (13.71293 \pm 2.00648) = (11.70645, 15.71941) \cong (11.706, 15.719)$$

i.e. We are 90% confident that in repeated sampling the true value of the population slope would lie in the interval (11.706 , 15.719).

[23] (h) Complete the following ANOVA table and hence test whether there is a significant linear relationship between the number of machines serviced and the total number of minutes spent. Use α = 0.10.

$TSS = S_{yy} =$ **13 558.94** **(given; also calculated in part (e))**

$SSE =$ **1 195.019** **(calculated in part (e))**

$SSR = TSS - SSE = \dfrac{S_{xy}^2}{S_{xx}} =$ **12 363.92** **(also calculated in part (e))**

$MSR = \dfrac{SSR}{1} =$ **12 363.92**

$MSE = \dfrac{SSE}{n-2} = \dfrac{1195.019}{14} =$ **85.3585** **(= s^2, calculated in part (e))**

$$F = \frac{MSR}{MSE} = \underline{\mathbf{144.8469}}$$

Source	d.f.	SS	MS	F
Regression	**1**	**12 363.92**	**12 363.92**	**144.8469**
Error	**14**	**1 195.019**	**85.3585**	
Total	**15**	**13 558.94**		

$H_0 : \beta_1 = 0 \qquad \alpha = 0.10$

$H_a : \beta_1 \neq 0$

test-statistics: $F = \dfrac{MSR}{MSE} = \underline{\mathbf{144.8469}}$

R.R: **we reject** H_0 **if** $F > F_{\alpha(1,n-2)} = F_{0.10(1,14)} = \mathbf{3.10}$

Since *F* = 144.8469 > 3.10, we reject H_0 and conclude that at 10% level of significance there is an evidence to say that the number of machines serviced and the total number of minutes spent are linearly related.

[5] (i) Find the values of the coefficient of correlation, *r*, and coefficient of determination, r^2, and interpret their meanings in this problem. What is your conclusion about the model?

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{901.625}{\sqrt{(65.75)(13558.94)}} = \mathbf{0.954916} \cong \underline{\mathbf{95.49\%}}$$

i.e. the number of machines serviced and the total number of minutes spent are strongly positively correlated (related) with the strength of their relationship of 95.49%.

$$r^2 = \frac{SSR}{TSS} = \mathbf{0.911865} \cong \underline{\mathbf{91.19\%}}$$

i.e. approximately 91.19% of the total variation in the data is explained by the regression line (and only 8.81% is due to error). I.e. model is a good fit.

[5] (j) If the number of machines serviced is 3, find a 95% Confidence Interval of the average time in minutes spent servicing the computers.

95% C.I. for *E(y)* when x_p = 3:

$\hat{y}$ **= 2.068 + 13.713(3) = 43.207 and** $1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$

$$\therefore E(y) \in \left(\hat{y} \pm t_{\alpha/2;n-2} s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}} \right) = \left(43.207 \pm t_{0.025;14} (9.238966) \sqrt{\frac{1}{16} + \frac{(3-4.375)^2}{65.75}} \right) =$$

$$= (43.207 \pm 2.145(2.790944)) = (43.207 \pm 5.986575) = (37.22042, 49.19358) \cong$$

$$\cong (37.22, 49.19)$$

i.e. We are 95% confident that in repeated sampling the average time spent servicing 3 computers would be between 37.22 and 49.19 minutes.

[5] (k) Find a 95% Prediction Interval of the time in minutes spent servicing the computers if the service person services 3 machines.

95% P.I. for y when $x_p = 3$:

$\hat{y} = 2.068 + 13.713(3) = 43.207$ and $1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$

$$\therefore y \in \left(\hat{y} \pm t_{\alpha/2;n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}} \right) = \left(43.207 \pm t_{0.025;14} (9.238966) \sqrt{1 + \frac{1}{16} + \frac{(3-4.375)^2}{65.75}} \right) =$$

$$= (43.207 \pm 2.145(9.651314)) = (43.207 \pm 20.70207) = (22.50493, 63.90907) \cong$$

$$\cong (22.51, 63.91)$$

i.e. We are 95% confident that in repeated sampling the time spent servicing 3 computers would be between 22.51 and 63.91 minutes.

2. Refers to question 1.

# of machines serviced (x_i)	**Total # of minutes spent (y_{ij})**	n_i	$\bar{y}_i$	$\sum_j (y_{ij} - \bar{y}_i)^2$
1	18, 17	2	17.5	0.5
2	33, 25	2	29	32
4	62, 39, 53, 55	4	52.25	278.75
5	95, 69, 65, 68	4	74.25	582.75
6	86	1	86	0
7	101, 105	2	103	8
8	102	1	102	0

[5] (a) Decompose SSE into the sum of squares due to the pure error, SSPE, and sum of squares due to the lack of fit, SSLF.

Hint: $SSPE = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 = 902$

$\sum x_i = 70$, $\sum x_i^2 = 372$, $\sum y_i = 993$, $\sum y_i^2 = 75\,187$, $\sum x_i y_i = 5\,246$

Solution:

$SSE = SSPE + SSLF$

$SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}} =$ **1 195.019** **(calculated in Q.1(e))**

$SSPE =$ **902** **(given)**

$\therefore SSLF = SSE - SSPE =$ **293.019**

[6] (b) Test whether the linear model $y = \beta_0 + \beta_1 x + \varepsilon$ is appropriate.
Use $\alpha = 0.05$.

H_0 : *model is appropriate* $\alpha = 0.05$
H_a : *model is not appropriate*

test-statistics: $F = \frac{MSLF}{MSPE} = \frac{SSLF \Big/ \left[(n-2) - \sum_i (n_i - 1)\right]}{SSPE \Big/ \sum_i (n_i - 1)} = \frac{293.019/(14-9)}{902/9} =$

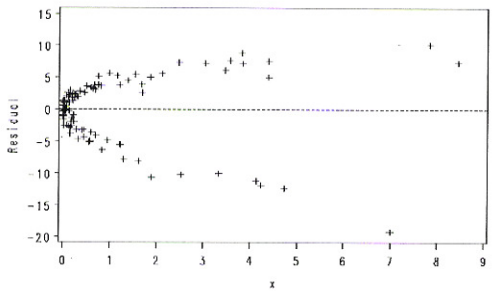
$= \frac{58.6038}{100.2222} =$ **0.584739**

R.R: **we reject** H_0 **if** $F > F_{\alpha(n-2-\sum_i (n_i-1), \sum_i (n_i-1))} = F_{0.05(5,9)} =$ **3.48**

Since *F* = 0.584739 < 3.48, we do not reject H_0 and conclude that at 5% level of significance there is not enough evidence to say that a linear model is not appropriate.

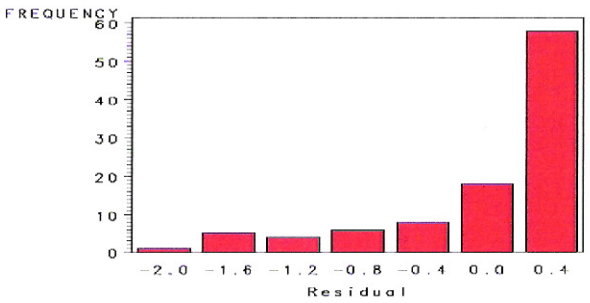
3. State which violations of the SLR model (if any) are indicated by each of the following residual plots. Give reasons for your answer.

[3] (a)



- **Violation of the assumption of constant variance, since the residuals are increasing with x's**

[3] (b)

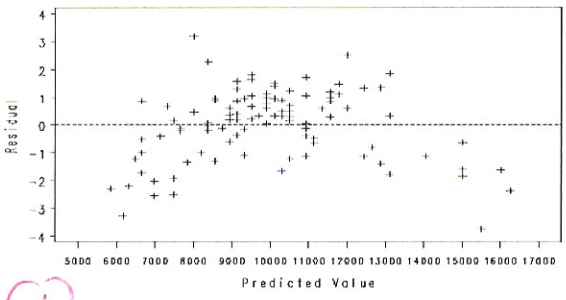


1

- **Violation of the assumption of errors being normally distributed, since the histogram of errors is not bell-shaped, nor is it symmetric (it is negatively skewed)**

1 1

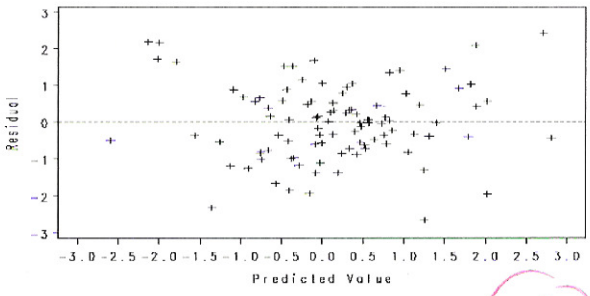
[3] (c)



1 2

- **Violation of independence (or linearity), since we have a curve-linear pattern**

[3] (d)



1 2

- **No violations, since residuals are randomly scattered around their mean (i.e. no pattern)**