/90

STAT 2509B
Test#1
SOLUTION

1. A diabetic is interested in determining how the amount of aerobic exercise impacts his/her blood sugar. The data are given in the following table:

Distance (miles)	2	2	2.5	2.5	3	3	3.5	3.5	4	4	4.5	4.5
Blood Sugar (mg/dL)	136	146	131	125	120	116	104	95	85	94	83	75

$\sum y_i = 1\,310$ $\qquad$ $\sum x_i = 39$

$\sum y_i^2 = 148\,870$ $\qquad$ $\sum x_i^2 = 135.5$

$\sum x_i y_i = 4\,035.5$

[1] (a) The response variable, y, is: **Blood Sugar level** (1)

[1] (b) The explanatory variable, x, is: **Distance run (aerobic exercise)** (1)

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

Model: (1) $y = \beta_0 + \beta_1 x + \varepsilon$, **$n = 12$**

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

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

(1) $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}} = \dfrac{\sum_{i=1}^{n} x_i y_i - \dfrac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \dfrac{4035.5 - \dfrac{(39)(1310)}{12}}{135.5 - \dfrac{(39)^2}{12}} = \dfrac{-222}{8.75} = -25.3714$ (1)

(1) $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \dfrac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \left(\dfrac{\sum_{i=1}^{n} x_i}{n}\right) = \dfrac{1310}{12} - (-25.3714)\left(\dfrac{39}{12}\right)$ **= 109.1667 – (–82.4571) =**

= 191.6238 (1)

(1) **∴ the least squares fitted regression line is given by:** $\hat{y} = 191.6238 - 25.3714\,x$

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

[6] (e) Find s^2, an estimate of σ^2.

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

$$= \frac{\left[148870 - \frac{(1310)^2}{12}\right] - \frac{(-222)^2}{8.75}}{10} = \frac{5861.667 - 5632.457}{10} = \frac{229.2095}{10} = \underline{\mathbf{22.92095}}$$

$\therefore s = \sqrt{s^2}$ **= 4.787583**

[6] (f) Use the t-test to test whether there is a significant linear relationship between the distance run and the levels of blood sugar. Use α = 0.05.

$H_0 : \beta_1 = 0$
$H_a : \beta_1 \neq 0$
$\alpha = 0.05 \Rightarrow \alpha/2 = 0.025$

test-statistics: $t = \frac{\hat{\beta}_1}{s/\sqrt{S_{xx}}} = \frac{-25.3714}{4.787583/\sqrt{8.75}} = \underline{\mathbf{-15.6759}}$

R.R: **we reject** H_0 **if** $t < -t_{\alpha/2;n-2} = -t_{0.025;10} = \mathbf{-2.228}$
or $t > t_{\alpha/2;n-2} = t_{0.025;10} = \mathbf{2.228}$

Since t = –15.6759 < –2.228, we reject H_0 and conclude that at 5% level of significance there is an evidence to say that a linear relationship between the distance run and blood sugar level exists.

[4] (g) Find a 95% confidence interval for the true population slope, β_1.

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

$$\beta_1 \in \left(\hat{\beta}_1 \pm t_{\alpha/2;n-2}\, s/\sqrt{S_{xx}}\right) = \left(-25.3714 \pm t_{0.025;10}\, 4.787583/\sqrt{8.75}\right) = (-25.3714 \pm 2.228(1.618499)) =$$

$$= (-25.3714 \pm 3.606015) = (-28.9774, -21.7654) \cong (-28.98, -21.76)$$

i.e. We are 95% confident that in repeated sampling the true value of the population slope would lie in the interval (–28.98 , –21.76).

[23] (h) Complete the following ANOVA table and hence test whether there is a significant linear relationship between the distance run and the levels of blood sugar. Use $\alpha = 0.05$.

$TSS = S_{yy} =$ **5 861.66667** **(given; also calculated in part (e))**

$SSE =$ **229.2095** **(calculated in part (e))**

$SSR = TSS - SSE = \dfrac{S_{xy}^2}{S_{xx}} =$ **5 632.457** **(also calculated in part (e))**

$MSR = \dfrac{SSR}{1} =$ **5 632.457**

$MSE = \dfrac{SSE}{n-2} = \dfrac{229.2095}{10} =$ **22.92095** **(= s^2, calculated in part (e))**

$F = \dfrac{MSR}{MSE} =$ **245.734**

Source	d.f.	SS	MS	F
Regression	1	5 632.457	5 632.457	245.734
Error	10	229.2095	22.92095	
Total	11	5 861.66667		

$H_0 : \beta_1 = 0$ $\alpha = 0.05$

$H_a : \beta_1 \neq 0$

test-statistics: $F = \dfrac{MSR}{MSE} =$ **245.734**

R.R: **we reject** H_0 **if** $F > F_{\alpha(1,n-2)} = F_{0.05(1,10)} =$ **4.96**

Since $F = 245.734 > 4.96$, we reject H_0 and conclude that at 5% level of significance there is an evidence to say that a linear relationship between the distance run and blood sugar level exists.

[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?

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{-222}{\sqrt{(8.75)(5861.667)}} = \underline{\underline{-0.98025}}$$

i.e. the distance run and the blood sugar level are strongly negatively correlated (related) with the strength of their relationship of 98.025%.

$$r^2 = \frac{SSR}{TSS} = 0.960897 \cong \underline{\underline{96.09\%}}$$

i.e. approximately 96.09% of the total variation in the data is explained by the regression line (and 3.91% is due to error). I.e. model is a very good fit.

[5] (j) Find a 95% confidence Interval of the average blood sugar level of people who run the distance of 3.1 miles.

95% C.I. for $E(y)$ when $x_p = 3.1$:

$\hat{y} = 191.6238 - 25.3714(3.1) = 112.9725$ and $1-\alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$

$$\therefore E(y) \in \left(\hat{y} \pm t_{\alpha/2;n-2} s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}} \right) = \left(112.9725 \pm t_{0.025;10} (4.787583) \sqrt{\frac{1}{12} + \frac{(3.1-3.25)^2}{8.75}} \right) =$$

$$= (112.9725 \pm 2.228(1.403217)) = (112.9725 \pm 3.126368) = (109.8461, 116.0988) \cong$$

$$\cong (109.85, 116.10)$$

i.e. We are 95% confident that in repeated sampling the average blood sugar level of people who run the distance of 3.1 miles will lie in (109.85 , 116.10) mg/dl.

[5] (k) Find a 95% Prediction Interval of the blood sugar level of an individual who run the distance of 3.1 miles.

95% P.I. for y when $x_p = 3.1$:

$\hat{y} = 191.6238 - 25.3714(3.1) = 112.9725$ and $1-\alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$

$$\therefore y \in \left(\hat{y} \pm t_{\alpha/2;n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}} \right) = \left(112.9725 \pm t_{0.025;10} (4.787583) \sqrt{1 + \frac{1}{12} + \frac{(3.1 - 3.25)^2}{8.75}} \right) =$$

$$= (112.9725 \pm 2.228(4.988985)) = (112.9725 \pm 11.11546) = (101.857, 124.0879) \cong$$

$$\cong (101.86, 124.09)$$

i.e. We are 95% confident that in repeated sampling the blood sugar level of a person who run the distance of 3.1 miles would lie in (101.86 , 124.09) mg/dl.

2. Refers to question 1.

Distance x_i	**Blood sugar** y_{ij}	n_i	$\bar{y}_i$	$\sum_j (y_{ij} - \bar{y}_i)^2$
2	136, 146	2	141	50
2.5	131, 125	2	128	18
3	120, 116	2	118	8
3.5	104, 95	2	99.5	40.5
4	85, 94	2	89.5	40.5
4.5	83, 75	2	79	32

[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: $SSPE = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2$ = 189

$\sum x_i$ = 39 $\sum x_i^2$ = 135.5 $\sum y_i$ =1 310 $\sum y_i^2$ = 148 870 $\sum x_i y_i$ = 4 035.5

Solution:

$SSE = SSPE + SSLF$

$SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}} =$ **229.2095** **(calculated in Q.1(e))**

$SSPE =$ **189** **(given)**

$\therefore SSLF = SSE - SSPE =$ **40.2095**

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

test-statistics: $F = \dfrac{MSLF}{MSPE} = \dfrac{SSLF \Big/ \left[(n-2) - \sum_i (n_i - 1)\right]}{SSPE \Big/ \sum_i (n_i - 1)} = \dfrac{40.2095/(10-6)}{189/6} =$

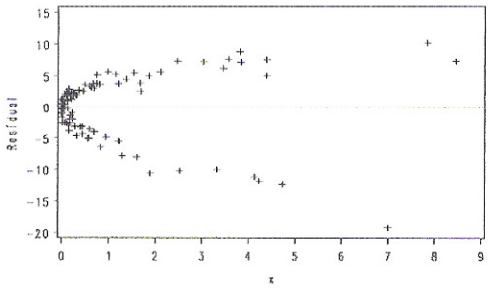
$$= \frac{10.05238}{31.5} = \underline{\mathbf{0.319123}}$$

R.R: we reject H_0 if $F > F_{\alpha(n-2-\sum_i (n_i-1), \sum_i (n_i-1))} = F_{0.05(4,6)} =$ **4.53**

Since F = 0.3191232 < 4.53, 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.

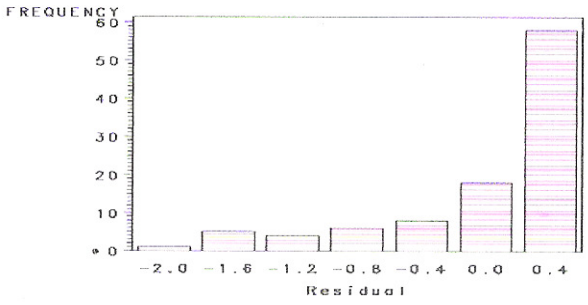
3. State which violations of the SLR model (if any) are indicated by each of the following residual plots. Give reasons for your answer.

[3] (a)



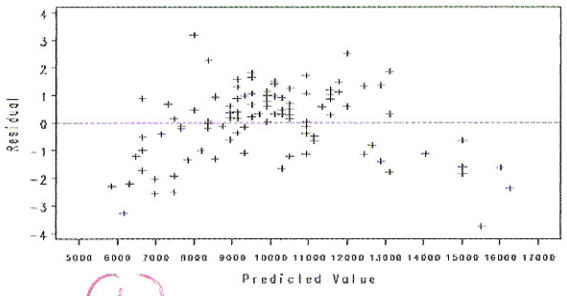
- **Violation of the assumption of constant variance, since the residuals are increasing with x's**

[3] (b)



- **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)**

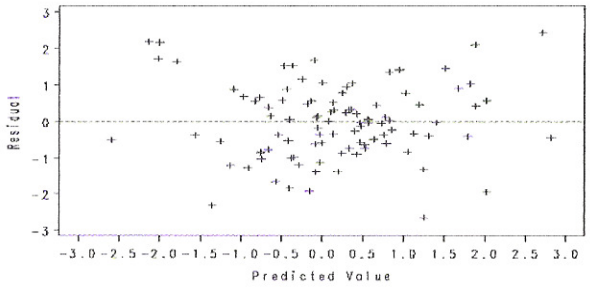
[3] (c)



(1) (2)

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

[3] (d)



(1)

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

(2)