11/90

STAT 2509B
Test#1
SOLUTION

1. Suppose we were interested in knowing how the amount of fertilizer influences the corn yield. The yield per plot in bushels of corn was observed on 10 plots that had been fertilized in varying degrees. The data are given in the following table:

Yield (in bushels)	Fertilizer (in pounds)
12	2
13	2
13	3
14	3
15	4
15	4
14	5
16	5
17	6
18	6

$\sum y_i = 147$, $\sum x_i = 40$

$\sum y_i^2 = 2\,193$, $\sum x_i^2 = 180$, $\sum x_i y_i = 611$

[1] (a) The response variable, y, is: **corn yield** (1)

[1] (b) The explanatory variable, x, is: **fertilizer amount** (1)

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

Model: $y = \beta_0 + \beta_1 x + \varepsilon$, **$n = 10$** (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.

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{611 - \frac{(40)(147)}{10}}{180 - \frac{(40)^2}{10}} = \frac{23}{20} = \underline{1.15}$$

(1) (1)

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \left(\frac{\sum_{i=1}^{n} x_i}{n} \right) = \frac{147}{10} - (1.15)\left(\frac{40}{10}\right) = 14.7 - 1.15(4) = \underline{10.1}$$

$\therefore$ **the least squares fitted regression line is given by:** $\hat{y} = \underline{10.1 + 1.15x}$

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

[6] (e) Find s^2, an estimate of σ^2.

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

$$= \frac{\left[2193 - \frac{(147)^2}{10}\right] - \frac{(23)^2}{20}}{8} = \frac{32.1 - 26.45}{8} = \frac{5.65}{8} = \underline{0.70625}$$

$\therefore s = \sqrt{s^2}$ **= 0.840387**

[6] (f) Use the t-test to test whether there is a significant linear relationship between the amount of fertilizer applied and the corn yield. Use α = 0.10.

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

test-statistics: $t = \frac{\hat{\beta}_1}{s/\sqrt{S_{xx}}} = \frac{1.15}{0.840387/\sqrt{20}} = \underline{6.119749}$

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**

Since $t = 6.119 > 1.860$, we reject H_0 and conclude that at 10% level of significance there is an evidence to say that the amount of fertilizer applied and the corn yield are linearly related.

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

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

$$\beta_1 \in \left(\hat{\beta}_1 \pm t_{\alpha/2;n-2}\, s/\sqrt{S_{xx}}\right) = \left(1.15 \pm t_{0.05;8}\, 0.840387/\sqrt{20}\right) = (1.15 \pm 1.860(0.187916)) =$$

$$= (1.15 \pm 0.349524) = (0.800476, 1.499524) \cong (0.8005, 1.4996)$$

i.e. We are 90% confident that in repeated sampling the true value of the population slope would lie in the interval (0.8005 , 1.4996).

[23] (h) Complete the following ANOVA table and hence test whether there is a significant linear relationship between the amount of fertilizer applied and the corn yield. Use α = 0.10.

$TSS = S_{yy} =$ **32.10** **(given) (also calculated in part (e))**

$SSE =$ **5.65** **(calculated in part (e))**

$SSR = TSS - SSE = \dfrac{S_{xy}^2}{S_{xx}} =$ **26.45** **(also calculated in part (e))**

$MSR = \dfrac{SSR}{1} =$ **26.45**

$MSE = \dfrac{SSE}{n-2} = \dfrac{5.65}{8} =$ **0.70625** **(= s^2) (also calculated in part (e))**

$F = \dfrac{MSR}{MSE} =$ **37.45133**

Source	*d.f.*	*SS*	*MS*	*F*
Regression	1	26.45	26.45	37.45133
Error	8	5.65	0.70625	
Total	9	32.10		

$H_0 : \beta_1 = 0$ $\quad \alpha = 0.10$
$H_a : \beta_1 \neq 0$

Test-statistics: $F = \dfrac{MSR}{MSE} =$ **37.45133**

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

Since *F* = 37.45 > 3.46, we reject H_0 **and conclude that at 10% level of significance there is an evidence to say that the amount of fertilizer applied and the corn yield 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?

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{23}{\sqrt{(20)(32.10)}} = 0.907738 \cong 0.908$$

i.e. the amount of fertilizer applied and the corn yield are positively correlated (related) with the strength of their relationship approx. 90.8%.

$$r^2 = \frac{SSR}{TSS} = 0.823988 \cong 0.8240$$

i.e. approximately 82.40% of the total variation in the data is explained by the regr. line (and approx. 17.60% is due to error). I.e. model is a good fit.

[5] (j) Find a 95% Confidence Interval of the average corn yield per plot if the amount of fertilizer added is 4.75 lbs.

95% C.I. for *E(y)* when x_p = 4.75:
$\hat{y}$ **= 10.1 + 1.15(4.75) = 15.5625 and** $1-\alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$

$$\therefore E(y) \in \left(\hat{y} \pm t_{\alpha/2;n-2} s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}} \right) = \left(15.5625 \pm t_{0.025;8} (0.840387) \sqrt{\frac{1}{10} + \frac{(4.75-4)^2}{20}} \right) =$$

$$= (15.5625 \pm 2.306(0.300812767)) = (15.5625 \pm 0.693674) = (14.86883, 16.25617) \cong$$

$$\cong (14.8688, 16.2562)$$

i.e. We are 95% confident that in repeated sampling the average corn yield per plot when the amount of fertilizer applied is 4.75 lbs will fall in the interval (14.8688 , 16.2562).

[5] (k) Find a 95% Prediction Interval of the corn yield per plot if the amount of fertilizer added is 4.75 lbs.

95% P.I. for y when $x_p = 4.75$:

$\hat{y} = 10.1 + 1.15(4.75) = 15.5625$ **and** $1-\alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$

$$\therefore y \in \left(\hat{y} \pm t_{\alpha/2;n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}} \right) = \left(15.5625 \pm t_{0.025;8} (0.840387) \sqrt{1 + \frac{1}{10} + \frac{(4.75-4)^2}{20}} \right) =$$

$$= (15.5625 \pm 2.306(0.892602168)) = (15.5625 \pm 2.05834) = (13.50416, 17.62084) \cong$$

$$\cong (13.5042, 17.62084)$$

i.e. We are 95% confident that the corn yield per plot when the amount of fertilizer applied is 4.75 lbs will fall in the interval (13.5042 , 17.6208).

2. Refers to question 1.

Fertilizer x_i	**Yield** y_{ij}	n_i	$\bar{y}_i$	$\sum_j (y_{ij} - \bar{y}_i)^2$
2	12, 13	2	12.5	0.5
3	13, 14	2	13.5	0.5
4	15, 15	2	15	0
5	14, 16	2	15	2
6	17, 18	2	17.5	0.5

[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: $SSPE = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2$ = 3.5

$\sum x_i = 40$, $\sum x_i^2 = 180$, $\sum y_i = 147$, $\sum y_i^2 = 2\,193$, $\sum x_i y_i = 611$

Solution:

$SSE = SSPE + SSLF$

$SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$ = **32.10 – 26.45 =** 5.65 **(calculated in 1e))**

$SSPE =$ **3.5 (given)**

$\therefore SSLF = SSE - SSPE =$ **2.15**

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

$y = \beta_0 + \beta_1 x + \varepsilon, \quad \alpha = 0.05$

H_0 : *model is appropriate*
H_a : *model is not appropriate*

test-statistics: $F = \dfrac{MSLF}{MSPE} = \dfrac{SSLF \Big/ \left[(n-2) - \sum_i (n_i - 1)\right]}{SSPE \Big/ \sum_i (n_i - 1)} = \dfrac{2.15/(8-5)}{3.5/5} =$

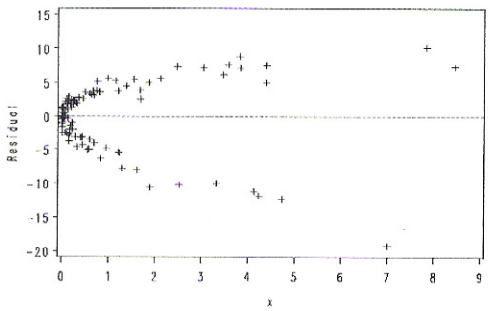
$$= \frac{0.716667}{0.7} = \underline{\mathbf{1.02381}}$$

R.R: **we reject** H_0 **if** $F > F_{\alpha(n-2-\sum_i (n_i-1), \sum_i (n_i-1))} = F_{0.05(3,5)} =$ **5.41**

Since *F* = 1.02381 < 5.41, 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 (i.e. model is fine).

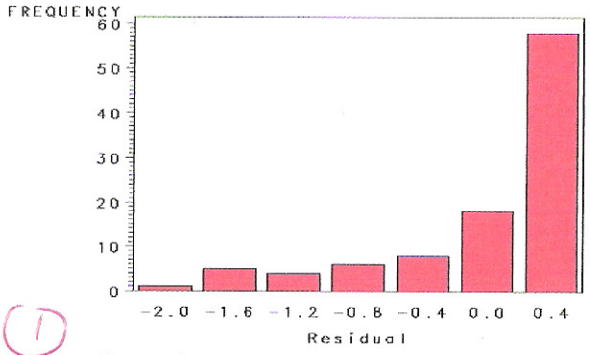
3. State which violations of the SLR model (if any) are indicated by each of the following residual plots. Give reasons for your answer.

[3] (a)



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

[3] (b)

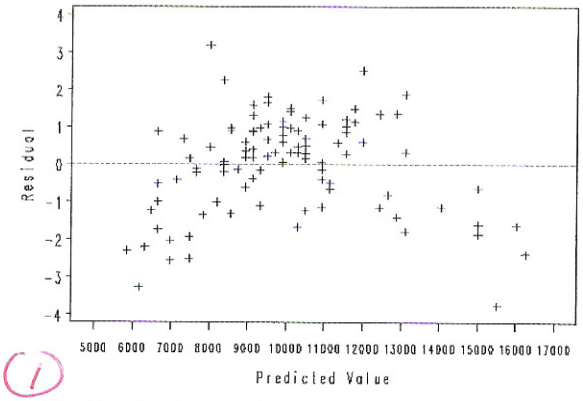


(1)

- **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)

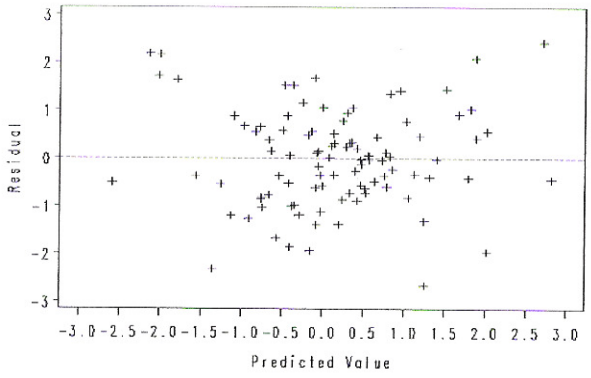
[3] (c)



(1)

- **Violation of linearity (or independence), since we have a curve-linear pattern** (2)

[3] (d)



(1)

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

(2)