Carleton University – School of Mathematics and Statistics
STAT 2509 – Test 1 – **SOLUTION**

**45**

**1. [36 marks]**

[0.5] (a) The response variable, $y$, is: **# of errors** **[1/2]**

[0.5] (b) The explanatory variable, $x$, is: **# of hours without sleep** **[1/2]**

[1] (c) **We have approximately positive [1/2] linear [1/2] relationship between the # of hours without sleep and the # of errors made.**

[3] (d)

**Model:** $y = \beta_0 + \beta_1 x + \varepsilon$ **[1/2]**, ***n* = 10**

**Assumptions:**

**(i) *x*'s are observed without error [1/2]**

**(ii) *y*'s (or $\varepsilon$'s) are independently [1/2] distributed with mean** $E(y) = \beta_0 + \beta_1 x$ **[1/2]**

**(or** $E(\varepsilon) = 0$ **[1/2])**

(Students might write i.i.d. This will be accepted in place of "independent". )

**(iii) variance of *y*'s (or $\varepsilon$'s) is constant [1/2],** $\sigma^2$ **for all *x*'s**

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

NOTE: Assumptions (ii) – (iv) can be summarized also as $y \overset{i.i.d.}{\sim} N(E(y), \sigma^2)$ **(or** $\varepsilon \overset{i.i.d.}{\sim} N(0, \sigma^2)$**)**

[5] (e)

**1st plot:** **(** $\hat{y}_i$'s vs $e_i$'s **)** Since there is a random scatter of points above and below zero (i.e. no obvious pattern) **[1/2]**, the **plot of predicted values vs residuals** suggests that the assumption of the independence **[1/2]** (and of linearity) is not violated. **[1/2]**

**2nd plot:** **(** $x_i$'s vs $e_i$'s **)** Since there is a random scatter of points above and below zero (i.e. no obvious pattern) **[1/2]**, the **plot of x's vs residuals** suggests that the assumption of equality of variance **[1/2]** is not violated. **[1/2]**

**3rd plot:** **(**histogram of errors**)** Since it looks bi-modal **[1/2]** and not symmetric **[1/2]**. Therefore, there might be (since sample size is small, i.e. *n* = 10) a violation **[1/2]** of the assumption of errors (or *y*'s) being normally **[1/2]** distributed.

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

[2.5] (f)

[1/2] $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{1848 - \frac{(160)(106)}{10}}{2880 - \frac{(160)^2}{10}} = \frac{152}{320} = \underline{\underline{0.475}}$ [1/2]

[1/2] $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{106}{10} - (0.475)\left(\frac{160}{10}\right)$ = 10.6 – 7.6 = $\underline{\underline{3}}$ [1/2]

**Fitted regression line:** $\hat{y} = 3 + 0.475\,x$ [1/2]

[0.5] (g) $\hat{y} = 3 + 0.475(10)$ = 7.75 [1/2]

[8] (h)

| Source | *d.f.* | *SS* | MS | *F* |
|---|---|---|---|---|
| Regression | 1 | **72.2** | 72.2 | 14.368 |
| Error | **8 [1/2]** | **40.2** | **5.025** | |
| Total | 9 | 112.4 | | |

[1 mark for entering the calculated values into ANOVA table]

[1/2] $SSR = \frac{S_{xy}^2}{S_{xx}} = \frac{(152)^2}{320}$ $\underline{\underline{72.2}}$ [1/2] **or MSR=SSR/1, hence SSR=MSR**

[1/2] $SSE = TSS - SSR = \underline{\underline{40.2}}$ [1/2]

[1/2] $MSE = \frac{SSE}{n-2} = \frac{40.2}{8} = \underline{\underline{5.025}}$ [1/2] **or F=MSR/MSE, hence MSE=MSR/F**

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

$H_a : \beta_1 \neq 0$ [1]

**test-statistics**: $F = \frac{MSR}{MSE} = \underline{14.368}$

**R.R:** we reject $H_0$ if $F > F_{\alpha(1,n-2)} = F_{0.05(1,8)} =$ **5.32** [1]

Since $F$ = 14.368 > 5.32 [1/2], we reject $H_0$ [1/2] and conclude that at 5% level of significance there is an evidence to say that a linear relationship between the # of hours without sleep and # of errors made exists. [1/2]

[1.5] (i)

[1/2] $s^2 = MSE = \frac{SSE}{n-2} = 5.025 \quad \Rightarrow$ [1/2] $s = \sqrt{5.025}$ = **2.241651** [1/2]

[4.5] (j)

$\left.\begin{array}{l} H_0 : \beta_1 \geq 0 \\ H_a : \beta_1 < 0 \end{array}\right\}$ [1] $\alpha = 0.05$

**test-statistics:** [1/2] $t = \frac{\hat{\beta}_1}{s/\sqrt{S_{xx}}} = \frac{0.475}{2.241651/\sqrt{320}} =$ **3.790535** $\approx$ **3.79** [1/2]

**R.R:** **we reject** $H_0$ **if** $t < -t_{\alpha;n-2} = -t_{0.05;8} = -$ **1.860** [1]

**Since** $t = 3.79 \not< -1.860$ **[1/2], we do not reject** $H_0$ **[1/2] and conclude that at 5% level of significance there is not evidence to say that the # of hours without sleep and the # of errors made are negatively linearly related.[1/2]**

[3] (k) [1/2] $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{152}{\sqrt{(320)(112.4)}} =$ **0.801467** $\cong$ **0.80** [1/2]

**i.e. the # of hours without sleep and the # of errors made are strongly positively correlated (related) with the strength of their relationship approx. 80%.** [1/2]

[1/2] $r^2 = \frac{SSR}{TSS} =$ **0.642349** $\cong$ **64.24%** [1/2]

**i.e. approximately 64.24% of the total variation in the data is explained by the regression line (and 35.76% is due to error).** [1/2]

[2.5] (l)

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

[1/2] $\beta_1 \in \left(\hat{\beta}_1 \pm t_{\alpha/2;n-2} \, s/\sqrt{S_{xx}}\right) = \left(0.475 \pm t_{0.025;8} \, 2.241651/\sqrt{320}\right) = \left(0.475 \pm 2.306(0.125312)\right) =$

**[1/2] for correct t-value**

$= (0.475 \pm 0.28897) = (0.18603,\ 0.76397) \cong (0.186,\ 0.764)$ [1]

**(1/2 mark for each confidence limit)**

**i.e. We are 95% confident that in repeated sampling the true value of the population slope would lie in the interval (0.186, 0.764).** [1/2]

[3.5] (m)

**95% P.I. for $y$ when $x_p$ = 10:**
$\hat{y}$ **= 3 + 0.475(10) = 7.75 and** $1-\alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$

[1] $$\therefore y \in \left( \hat{y} \pm t_{\alpha/2;n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}} \right) = \left( 7.75 \pm t_{0.025;8} (2.241651) \sqrt{1 + \frac{1}{10} + \frac{(10-16)^2}{320}} \right) =$$

$$= (7.75 \pm 2.306(2.468362)) = (7.75 \pm 5.692043) = (2.057957,\ 13.44204) \cong (2.06,\ 13.44)$$ [1]

[1/2] [1/2] (1/2 mark for each confidence limit)

**i.e. We are 95% confident that (in repeated sampling) the # of errors made by a person who was without sleep for 10 hours would be between 2.06 and 13.44. [1/2]**

2. **[9 marks]** Refers to question 1.

[1/2] $SSE = SSPE + SSLF$, where **SSE =** 40.2 [1/2] (calculated in part h))

and **SSPE =** $SSPE = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2$ = **38** [1/2]

$\therefore SSLF = SSE - SSPE =$ **2.2** [1/2]

$H_0$ : *model is appropriate* $\alpha = 0.05$
$H_a$ : *model is not appropriate* [1]

[1/2] for d.f.

**test-statistics:** [1/2] $$F_{LF} = \frac{MSLF}{MSPE} = \frac{SSLF \Big/ \left[ (n-2) - \sum_i (n_i - 1) \right]}{SSPE \Big/ \sum_i (n_i - 1)} = \frac{2.2/(8-5)}{38/5} =$$

[1/2]

$$= \frac{0.73333}{7.6} = 0.09649$$ [1/2]

([1/2] 0.73333, [1/2] 7.6)

(Note: ½ mark for correct values of each: SSLF df, SSPE df, MSLF and MSPE)

**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 [1]**

**Since $F$ = 0.965 $\not>$ 5.41 [1/2], we do not reject $H_0$ [1/2] and conclude that at 5% level of significance there is not enough evidence to say that a linear model is not appropriate. [1/2]**

Model is a good fit. [1/2]