STAT 2509 A
Assignment #2

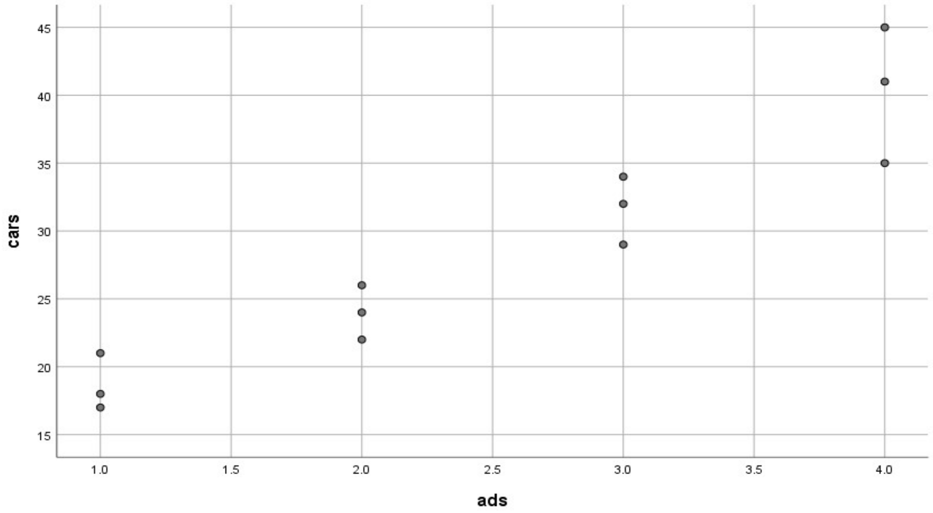
SOLUTION // 60

1. [42 marks]

[1] a) Identify independent (*x*) and dependent (*y*) variables.

***x* = # of ads per day [1/2]**
***y* = # of cars sold [1/2]**

[2] b) **[1]**



Scatter plot indicates [1/2] approximately straight line (or linear relationship) with [1/2] positive slope.

[3] c)

Model: $y = \beta_0 + \beta_1 x + \varepsilon$ **[1/2], *n* = 12**

Assumptions: (i) *x*'s are observed without error [1/2]
(ii) *y*'s (or ε's) are independently [1/2] distributed with mean $E(y) = \beta_0 + \beta_1 x$
(or $E(\varepsilon) = 0$**) [1/2]**
(iii) variance of *y*'s (or ε's) is constant [1/2], σ^2 **for all *x*'s**
(iv) *y* ~ $N\left(E(y), \sigma^2\right)$ **[1/2] for any value of *x* (or** $\varepsilon \sim N\left(0, \sigma^2\right)$ **for any value of *x*)**

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

[4] d)

$$\text{[1/2]}\quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{969 - \frac{(30)(344)}{12}\ \text{[1/2]}}{90 - \frac{(30)^2}{12}\ \text{[1/2]}} =$$

$$= \frac{109}{15} = \underline{\underline{7.266666667 \doteq 7.2667}} \quad \text{[1/2]}$$

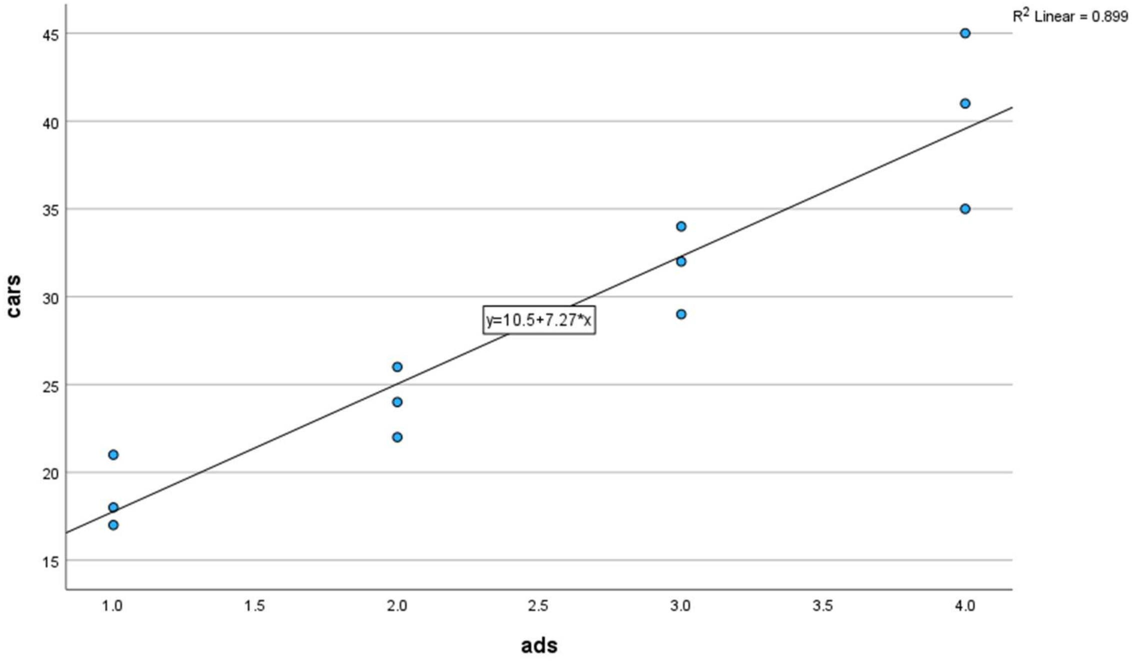
$$\text{[1/2]}\quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \left(\frac{\sum_{i=1}^{n} x_i}{n} \right) = \frac{344}{12} - (7.266666667)\left(\frac{30}{12}\right) =$$

= 28.66666667 – 18.16666667 = $\underline{10.5}$ [1/2]

∴ **the least squares fitted regression line is given by:** $\hat{y}$ $\underline{\underline{\textbf{= 10.5 + 7.2667}}}$ x [1]

[1] e)

[1]



[4] f)

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

$$= \frac{\left[10\,742 - \frac{(344)^2}{12}\right] - \frac{(109)^2}{15}}{10} = \frac{\overset{[1/2]}{880.6666667} - \overset{[1/2]}{792.0666667}}{10\,[1/2]} = \frac{88.6}{10} = \underline{\mathbf{8.86}} \quad [1/2]$$

$\therefore$ [1/2] $s = \sqrt{s^2} = \underline{\underline{2.976575213}} \cong \underline{\underline{2.9766}}$ [1/2]

[4.5] g)

$H_0 : \beta_1 = 0$ [1/2] $\quad \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$
$H_a : \beta_1 \neq 0$ [1/2]

test-statistics: [1/2] $t = \frac{\hat{\beta}_1}{s/\sqrt{S_{xx}}} = \frac{7.266666667}{2.976575213/\sqrt{15}} = \underline{9.45505} \cong \underline{9.455}$ [1/2]

$$\textbf{R.R:}\ \text{we reject } H_0 \text{ if } \left.\begin{array}{l} t < -t_{\alpha/2;n-2} = -t_{0.025;10} = -\,2.228 \\ \text{or } t > t_{\alpha/2;n-2} = t_{0.025;10} = 2.228 \end{array}\right\} \quad [1]$$

Since t = $\underline{9.455}\overset{[1/2]}{\boxed{>}}$ 2.228, we reject H_0 [1/2] and conclude that at 5% level of significance there is an evidence to say that the No. of ads per day and the No. of cars sold are linearly related. [1/2]

[2.5] h) $1-\alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$

[1/2]
$$\beta_1 \in \left(\hat{\beta}_1 \pm t_{\alpha/2;n-2}\, {}^{s}\!\big/\!\sqrt{S_{xx}}\right) = \left(7.2667 \pm t_{0.025;10}\, {}^{2.976575213}\!\big/\!\sqrt{15}\right) = \left(7.2667 \pm \overset{[1/2]\,\checkmark}{\underline{2.228}}(0.768548415)\right) =$$

$$= (7.2667 \pm 1.712325869) = (5.554374131\ ,\ 8.979025869) \cong \underline{\underline{(5.5544\ ,\ 8.979)}} \quad [1]$$

(1/2 mark for each correct interval value)

i.e. We are 95% confident that in repeated sampling the true value of the population slope would lie in the interval (5.5544 , 8.979). [1/2]

[12] i)

[1/2] $TSS = S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = \underline{\underline{880.6666667}}$ [1/2] **(as calculated in part (f))**

[1/2] $SSR = \frac{S_{xy}^2}{S_{xx}} = \underline{\underline{792.0666667}}$ [1/2] **(as calculated in part (f))**

[1/2] $SSE = TSS - SSR = \underline{\underline{88.6}}$ [1/2] **(calculated in part (f))**

[1/2] $MSR = \frac{SSR}{1} = \underline{\underline{792.0666667}}$ [1/2]

[1/2] $MSE = \frac{SSE}{n-2} = \frac{88.6}{10} = \underline{\underline{8.86}}$ [1/2] **(= s^2) (as calculated in part (f))**

[1/2] $F = \frac{MSR}{MSE} = \underline{\underline{89.39804}}$ [1/2]

Source	d.f.	SS	MS	F
Regression	1	792.0666667	792.0666667	89.398
Error	10	88.6	8.86	
Total	11	880.6666667		
	[1/2]	[1/2]	[1/2]	[1/2]

one mark for each column, if values are entered correctly

$$\left.\begin{aligned} H_0 &: \beta_1 = 0 \\ H_a &: \beta_1 \neq 0 \end{aligned}\right\}$$ $\alpha = 0.05$ [1]

test-statistics: $F = \frac{MSR}{MSE} = \underline{\mathbf{89.398}}$ **[1/2]**

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

[1/2]

Since *F* = 89.398 > 4.96, 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 No. of ads per day and the No. of cars sold exists. [1/2]**

[5] j)

[1/2] $r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{109}{\sqrt{(15)(880.6666667)}} = \underline{\mathbf{0.94836}} \cong \underline{\mathbf{0.95}}$ [1/2]

i.e. the No. of ads per day and the No. of cars sold are positively [1/2]correlated (related) with the strength of their relationship approx. 95%. [1/2]

[1/2] $r^2 = \frac{SSR}{TSS} = \underline{\underline{0.89939}} \cong \underline{\underline{0.90}}$ [1/2]

i.e. approximately 90% of the total variation in the data is explained by the regression line (and approx. 10% is due to error). [1]

The model is a very good fit (or it is a very good model). [1]

[3] k)

Model Summary[b] [1]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.948[a]	.899	.889	2.977

← S

a. Predictors: (Constant), ads

b. Dependent Variable: cars

[1] **ANOVA[a]**

p-value < 0.05 ⇒ reject Ho

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	792.067	1	792.067	89.398	.000[b]
	Residual	88.600	10	8.860		
	Total	880.667	11			

a. Dependent Variable: cars

b. Predictors: (Constant), ads

[1] **Coefficients[a]**

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	10.500	2.105		4.989	.001	5.810	15.190
	ads	7.267	.769	.948	9.455	.000	5.554	8.979

a. Dependent Variable: cars

$\hat{\beta}_0$

$\hat{\beta}_1$

t-test statistics

$\beta_1 \in (5.554, 8.979)$

p-value < 0.05 ⇒ reject Ho

2. [9 marks]

[6] a) **95% C.I. for *E(y)* when x_p = 0:**

$\hat{y}$ **= 10.5 + 7.2667 (0) = 10.5 [1/2] and** $1-\alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$

[1/2] $\therefore E(y) \in \left(\hat{y} \pm t_{\alpha/2;n-2}\, s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}} \right) = \left(10.5 \pm t_{0.025;10}(2.976575213)\sqrt{\frac{1}{12} + \frac{(0-2.5)^2}{15}} \right) =$

[1/2] $= (10.5 \pm 2.228(2.104756518)) = (10.5 \pm 4.689397522) = (5.810602478\ ,\ 15.18939752) \cong$

$\cong \underline{\underline{(5.8106\ ,\ 15.1894)}}$ **[1] (1/2 mark for each correct interval value)**

i.e. We are 95% confident that in repeated sampling the <u>average value</u> of the No. of cars sold when the 0 ads were run, will fall in the interval (5.8106 , 15.1894). [1/2]

and

95% P.I. for *y* when x_p = 0:

$\hat{y}$ **= 10.5 + 7.2667 (0) = 10.5 and** $1-\alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$

[1/2] $\therefore y \in \left(\hat{y} \pm t_{\alpha/2;n-2}\, s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}} \right) = \left(10.5 \pm t_{0.025;10}(2.976575213)\sqrt{1 + \frac{1}{12} + \frac{(0-2.5)^2}{15}} \right) =$

[1/2] $= (10.5 \pm 2.228(3.645545226)) = (10.5 \pm 8.122274764) = (2.377725236\ ,\ 18.62227476) \cong$

$\cong \underline{\underline{(2.377\ ,\ 18.622)}}$ **[1] (1/2 mark for each correct interval value)**

i.e. We are 95% confident that in repeated sampling the No. of cars sold when the 0 ads were run, will lie in the interval (2.377 , 18.622). [1/2]

<u>Conclusion</u>:

- **The P.I. is <u>wider</u> [1/2] than C.I. (as expected), since the variability in the error for predicting a single value of y is always greater than the variability of the error for the estimation of the mean/average value of y.**

[3] b) **95% C.I. for *E(y)* when x_p = 0 : (5.81031 , 15.18969) [1]**

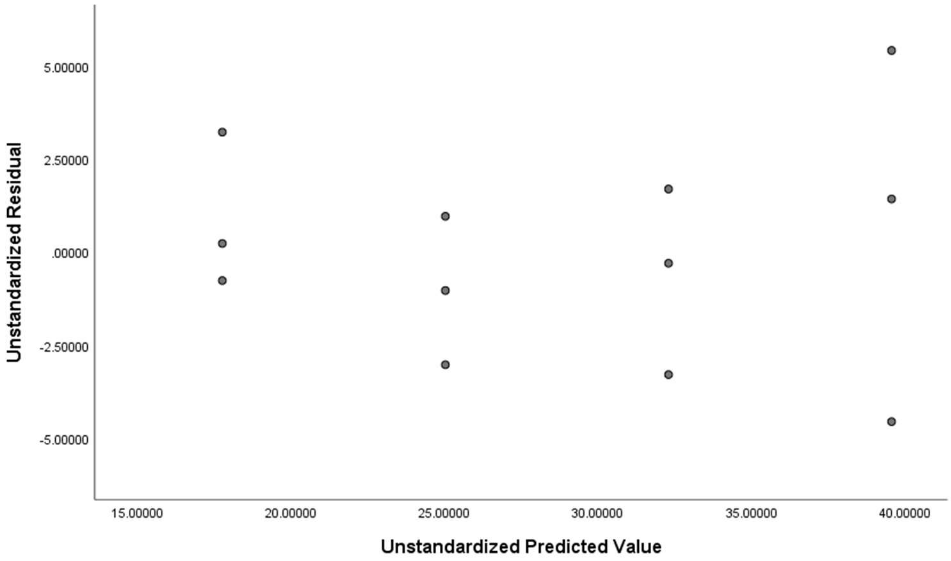
95% P.I. for *y* when x_p = 0 : (2.37722 , 18.62278) [1]

$\hat{y}$ **= 10.50000 when x_p = 0**

[1]

3. [9 marks]

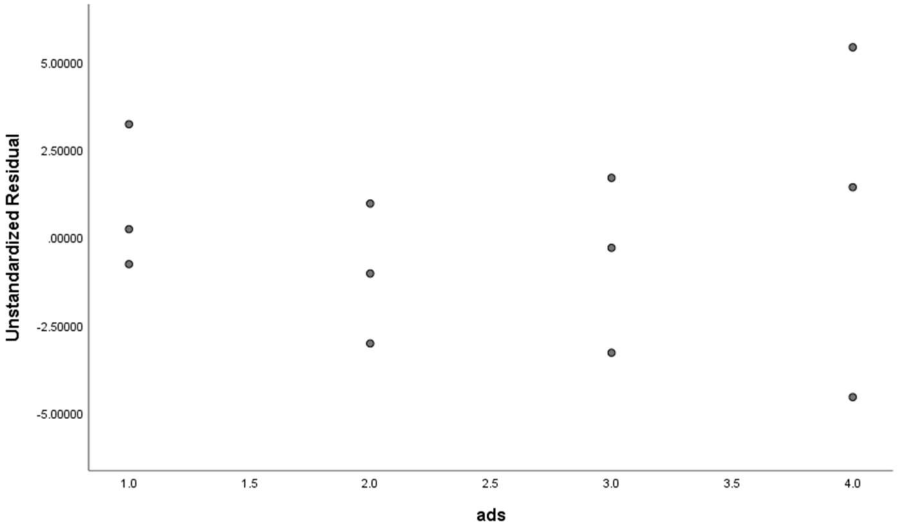
[1]



- Residuals seem to be randomly scattered around zero (i.e. no pattern) ⇒ [1/2] no violations of independence (and linearity) [1/2]

NOTE: if students saw and indicated a pattern, then there is a **[1/2]** violation of the assumption of the independence of the errors (and/or linearity) **[1/2]**.

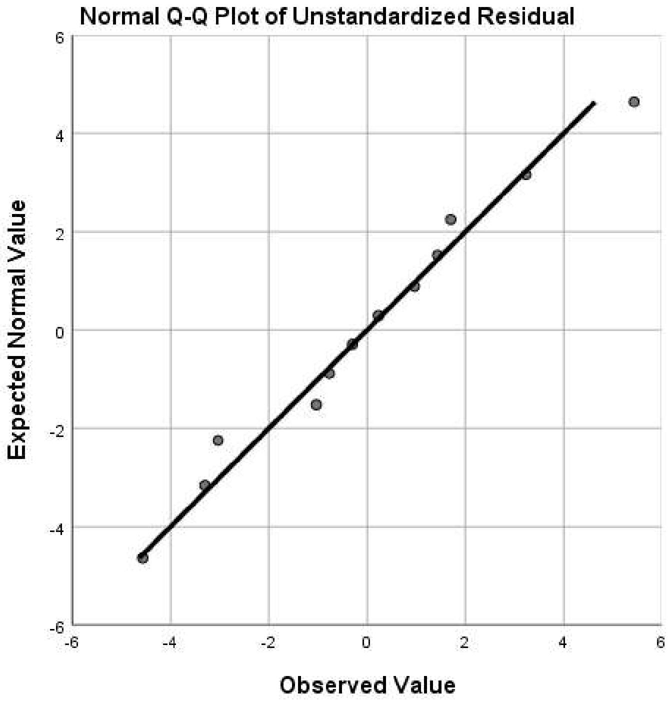
[1]



- **Residuals seem to be randomly scattered around zero (i.e. no pattern) ⇒ [1/2] no violations of constant variance [1/2]**

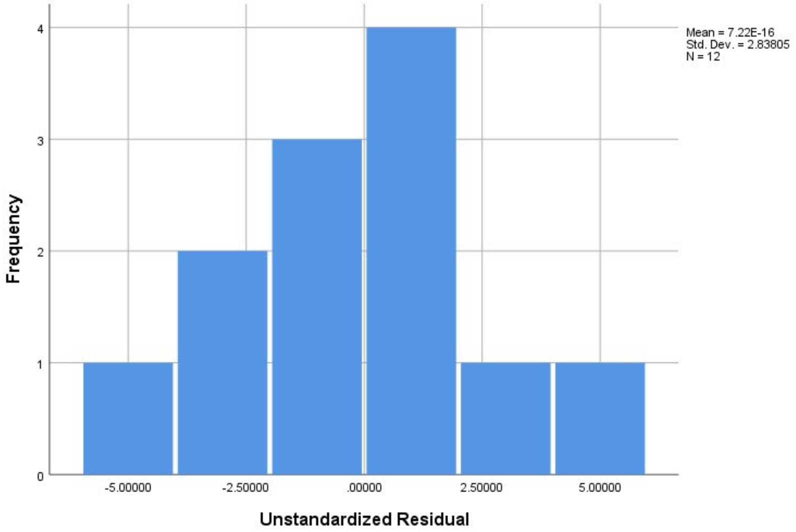
NOTE: if students saw and indicated a pattern, then there is a **[1/2]** violation of the assumption of the constant variance **[1/2]**.

[1]



- **Q-Q Plot of residuals shows approximately the straight line ⇒ [1/2] no violations of normality of the errors [1/2]**

[1]



Histogram of the errors looks approx. bell-shaped [1/2]. It is not really symmetric [1/2] (but it is most likely due to small sample size, as *n* = 12). Therefore, since Q-Q plot did not show any violations, errors are normally distributed. [1/2]

All the plots suggest that the model assumptions are (reasonably) satisfied [1/2]