M241 Coursework

up930537

1) a) $H_0: \eta = \eta_0 = 0$
$H_1: \eta \neq 0$

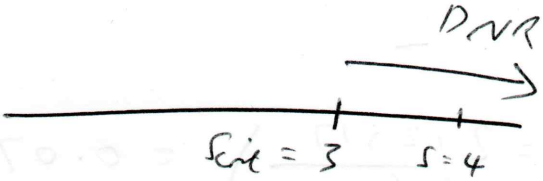
Null hypothesis: There is no significant difference between the median reaction times of drivers under 'Test' and 'Control' conditions.

Alternative hypothesis: There is a significant difference between the median reaction times of drivers under 'Test' and 'Control' conditions.

Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Test ①	324	286	260	242	179	264	233	190	276	251	279	257	339	256	274	196
Control ②	336	323	315	372	301	300	242	254	243	220	309	259	295	265	273	254
① − ②	-12	-37	-55	-130	-122	-36	-9	-64	+33	+31	-30	-2	+44	-9	+1	-58

Test statistic $S_+ = 4$, $S_- = 12$ $\Rightarrow S = 4$

Critical value $n = 16$, $S_{crit\,2.5\%}(n) = S_{crit\,2.5\%}(16) = 3$



Conclusion

$4 > S_{crit\,2.5\%}(16)$, so there is insufficient evidence to reject the null hypothesis that there is no ~~median~~ difference between median reaction times of the drivers at the 5% level of significance. So we can conclude

that there is no significant difference in reaction times between the drivers under 'Test' and 'Control' 'conditions' at the 5% level of significance.

b) $n = 16, p = 0.5$ $\quad P(x) = {}^{n}C_{x}(0.5)^{x}(1-0.5)^{n-x}$

Assuming H_0 is true, sampling distribution of S_+ will be binomial

$S_+ \sim B(16, 0.5)$

$p\text{-value} = P(S_+ \leq 4)$ because $S = 4$

$p\text{-value} = \left[P(0) + P(1) + P(2) + P(3) + P(4)\right] \times 2$ (two-tailed)

$= \Big[{}^{16}C_0 (0.5)^0 (0.5)^{16} + {}^{16}C_1 (0.5)^1 (0.5)^{15} + {}^{16}C_2 (0.5)^2 (0.5)^{14}$

$+ {}^{16}C_3 (0.5)^3 (0.5)^{13} + {}^{16}C_4 (0.5)^4 (0.5)^{12} \Big] \times 2$

$= 2\left[\frac{1}{65536} + \frac{1}{4096} + \frac{15}{8192} + \frac{35}{4096} + \frac{455}{16384}\right]$

$= 2\left(\frac{2517}{65536}\right) = 0.076812\,74414 \simeq 0.0768$

So $p\text{-value} \simeq 0.0768$ $\quad (p\text{-value} > 0.025 \Rightarrow \text{DNR } H_0)$

If $\alpha = 10\%$

Critical value: $S_{crit\,\frac{\alpha}{2}}(\wedge) = S_{crit\,5\%}(16) = 4$

$S = 4$, $4 \leq 4$, $\Rightarrow$ Reject H_0

$S \leq S_{crit}$

p-value

$P(x \leq 4) \times 1 = p\text{-value}$

$p\text{-value} = P(x \leq 4) = \frac{2517}{65536} = 0.038406372070$

$p\text{-value} \simeq 0.0384$, $p\text{-value} \leq 0.05$ $\Rightarrow$ Reject H_0

If the significance level were 10%, the conclusion would change. The p-value is less than the level of significance, so we would reject the null hypothesis that there is no difference in reaction time at the 10% level of significance.

c) Wilcoxon Signed Rank Test → Samples are paired

"Test" − "Control" 1 − 2	|1 − 2|	Rank of |1 − 2|	+'ve Ranks	−'ve Ranks
−12	12 (5)	5		5
−37	37 (10)	10		10
−55	55 (12)	12		12
−130	130 (16)	16		16
−122	122 (15)	15		15
−36	36 (9)	9		9
−9	9 (3+4/2)	3.5		3.5
−64	64 (14)	14		14
+33	33 (8)	8	8	
+31	31 (7)	7	7	
−30	30 (6)	6		6
−2	2 (2)	2		2
+44	44 (11)	11	11	
−9	9 (3+4/2)	3.5		3.5
+1	1 (1)	1	1	
−58	58 (13)	13		13
		SUM =	8+7+... = 27	5+10+12+... = 109

$Sum = W_+ + W_- = 27 + 109 = 136$ ✓

Check: $Sum = \frac{n(n+1)}{2}$, $n = 16 \Rightarrow \frac{16(16+1)}{2} = \frac{16(17)}{2} = \frac{272}{2} = 136$

$H_0: m_D = 0$

$H_1: m_D \neq 0$ where $D = 1-2$

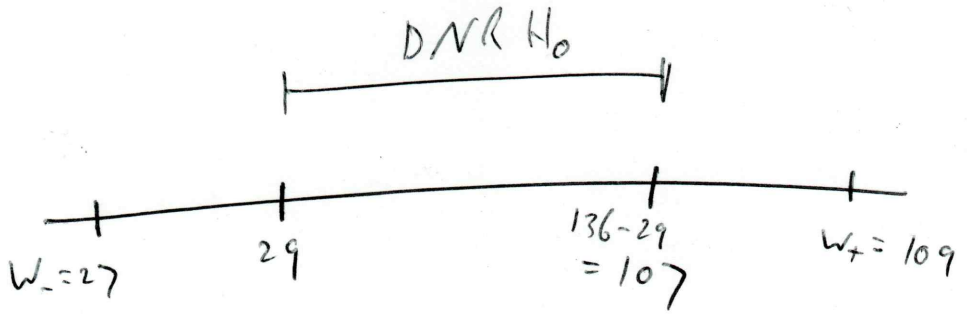
Null hypothesis: There is no significant difference between the median reaction times of drivers under 'Test' and 'Control' conditions

Alternative hypothesis: There is a significant difference.

Test statistic $W = 27$

Critical values $\alpha = 5\%$, $\frac{\alpha}{2} = 2.5\%$, $n = 16$

$W_{crit\,2.5\%}(16) = W_{crit\,2.5\%}(16) = 29$



Conclusion ~~If~~ $W < 29$, so there is sufficient evidence to reject the null hypothesis that there is no median difference in reaction times between 'Test' and Control 'conditions' at the 5% level of significance.

So we can conclude that there is a significant difference in reaction times between 'Test' and 'Control' conditions of the drivers at the

5% level of significance. ~~There is a significant increase in reaction times under 'Test' condition, as the overall test score under 'Test' condition, as~~

There is a significant decrease in overall ~~test~~ test score under 'Test' condition as $W_- > W_+$.

Therefore, the reaction times of the drivers are significantly greater under 'Test' condition compared to 'Control' condition at the 5% level of significance.

2) a) Ordered Female data

i	Female	
1	6	
2	9	
3	10	
4	16	← Q_1
5	17	
6	18	
7	21	
8	28	← median
9	32	
10	35	
11	35	
12	39	← Q_3
13	52	
14	65	
15	71	

$Q_1 = \frac{1}{4}(n+1)^{th}$ value

$= \frac{1}{4}(15+1)$

$= \frac{1}{4}(16) = 4^{th}$ value

$Q_3 = \frac{3}{4}(n+1)^{th}$ value

$= \frac{3}{4}(16) = 12^{th}$ value

Estimated median $= \left(\frac{n+1}{2}\right)^{th}$ value $= \left(\frac{15+1}{2}\right)^{th}$ value $= \left(\frac{16}{2}\right)^{th}$ value $= 8^{th}$ value

$= 28$

95% CI:

$$np \pm 1.96\sqrt{np(1-p)}$$

$n = 15$
$p = 0.5$
$q = 1 - 0.5 = 0.5$

$$\Rightarrow 15(0.5) \pm 1.96\sqrt{15(0.5)(0.5)}$$

$$= 7.5 \pm 1.96\sqrt{3.75} = \left(3.704476321^{th}\text{ value}, 11.29552368^{th}\text{ value}\right)$$

$$\simeq \left(4^{th}\text{ value}, 12^{th}\text{ value}\right)$$

So 95% CI: $(16, 39)$ Same as $(Q1, Q3)$

$$\text{Exact level of confidence} = \sum_{r=i}^{j-1} {}^{n}C_r \times 0.5^{n} \times 100\%$$

$i = 4$
$j = 12$
$n = 15$

$$\Rightarrow \left[{}^{15}C_4 + {}^{15}C_5 + {}^{15}C_6 + {}^{15}C_7 + {}^{15}C_8 + {}^{15}C_9 + {}^{15}C_{10} + {}^{15}C_{11}\right] \times 0.5^{15} \times 100\%$$

$$= 96.484375\% \simeq 96.48\%$$

The exact level of confidence achieved by the 95% Confidence Interval is approximately 96.48%.

2) b) The Data samples are not paired, so use a Mann Whitney U Test.

$H_0: n_1 = n_2$

$H_1: n_1 \neq n_2$ (two-tailed)

where n_1 = median male time (smaller sample)

n_2 = median female time

Null hypothesis: There is no difference in median time spent in days on psychiatric ward after being sectioned between 'Males' and 'Females.'

Alternative hypothesis: There is a difference in median time spent in days on psychiatric ward after being sectioned between 'Males' and 'Females'.

Male (1)	Rank Male	Female (2)	Rank Female
96 (25)	25	39 (18)	18
113 (26)	26	71 (23)	23
195 (28)	28	17 (5)	5
22 (8)	8	9 (2)	2
56 (20)	20	35 ($\frac{5+6}{2}$)	15.5
155 (27)	27	6 (1)	1
30 (12)	12	57 (21)	21
50 (19)	19	10 (3)	3
29 (11)	11	18 (6)	6
38 (17)	17	65 (22)	22
25 (9)	9	35 ($\frac{15+16}{2}$)	15.5
79 (24)	24	21 (7)	7
34 (14)	14	32 (13)	13
		28 (10)	10
		16 (4)	4
SUM =	~~800~~ 240		166

Check: $n_1 = 13$ (males)

$n_2 = 15$ (Females)

$n = 13 + 15 = 28$

$\frac{n(n+1)}{2} = \frac{28(28+1)}{2} = \frac{28(29)}{2} = 406$ ✓

$R_1 + R_2 = 240 + 166 = 406$

Test Statistic $R_1 = 166 = W_1$

$\Rightarrow W_1 = 166$, ~~$n_1 = 13, n_2 = 15$~~ $n_1 = 15, n_2 = 13$

$n_1, n_2 > 10$, so moderately large sample, so compare test statistic to standard normal distribution.

$$U_1 = n_1 n_2 + 0.5 n_1 (n_1 + 1) - W_1$$

~~$U_1 = 13(15) + 0.5(13)(13+1) - 166$~~

$$U_1 = 15(13) + 0.5(15)(15+1) - 166$$

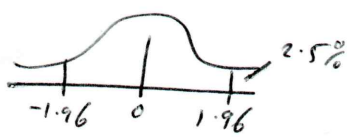
$$U_1 = 149$$

$$Z_0 = \frac{\left| U_1 - \frac{n_1 n_2}{2} \right| - \frac{1}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} = \frac{\left| 149 - \frac{15(13)}{2} \right| - \frac{1}{2}}{\sqrt{\frac{15(13)(15+13+1)}{12}}}$$

$$= \frac{|51.5| - \frac{1}{2}}{\sqrt{471.25}} = \frac{51.5 - \frac{1}{2}}{\sqrt{471.25}} = \frac{51}{\sqrt{471.25}} = 2.34\,933\,254\,2$$

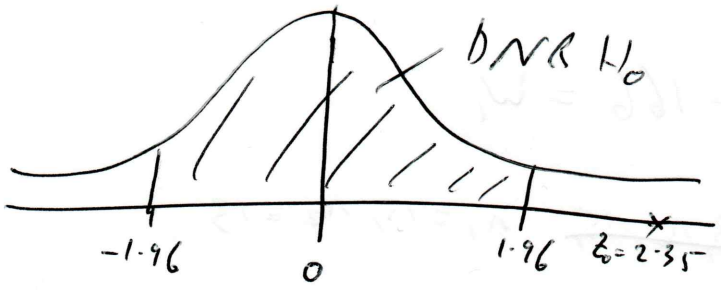
$$\simeq 2.35$$

Critical values 2-tailed $\alpha = 5\%$, $\frac{\alpha}{2} = 2.5\%$



~~$Z = \ldots$~~ $Z = \pm 1.96$

(value closest to 0.025 in table)



Conclusion

$|Z_0| > 1.96$, The test statistic is greater than the critical value, so there is sufficient evidence to reject the null hypothesis that there is no significant median difference at the 5% level of significance.

So we can conclude there is a significant difference in median time spent (in days) on psychiatric ward after being sectioned between 'Males' and 'Females' at the 5% level of significance.
There is evidence that the 'Male' times are significantly higher than the 'Female' times.

2) c) probability that randomly chosen male patient tire greater than randomly chosen female patient tire.

Male	No of female tires less than each male tire
96	15
113	15
195	15
22	7
56	12
155	15
30	8
50	12
29	8
38	11
25	7
79	15
34	9
Sum	15+15+15+7+... +9 = 149

← Since 96 is greater than any of the female tires.

Let Male = Y
Female = X

All possible ways of randomly choosing male and female patient = 15×13
= 195

So $P(Y > X) = \frac{149}{195} = 0.764102564 \simeq 0.7641$

So the probability that a randomly chosen male patient will spend longer on the psychiatric ward than a randomly chosen female patient is 0.7641.

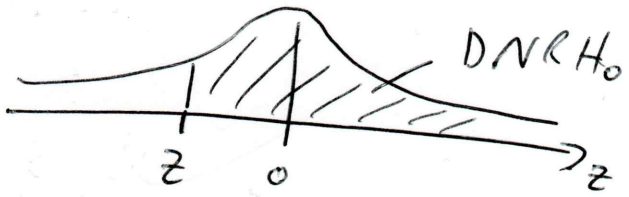
3) a) $X \sim N(200, 8^2)$

$$X \approx N\left(200, \frac{8^2}{4}\right)$$

$\mu = 200 = \mu_0$

$\sigma^2 = 8^2$

$n = 4$



$$z = \frac{X - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{195 - 200}{\frac{\sqrt{8^2}}{\sqrt{4}}}$$

$$= \frac{195 - 200}{\frac{\sqrt{64}}{\sqrt{4}}} = -1.25$$

$P(z < -1.25) = P(z > 1.25)$ (because distribution is symmetric)

$P(z > 1.25) = 0.1056$ Using z table

$\Rightarrow \alpha = 0.1056 = 10.56\% \simeq 10\%$ (1-tailed)

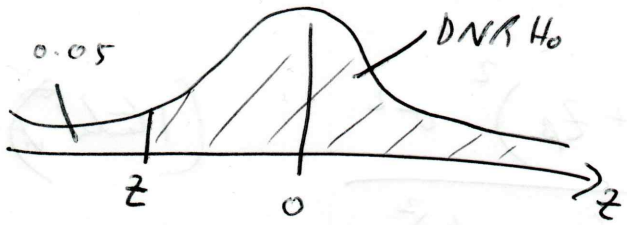
The significance level of the test is approximately 10%.

3) b) $X \sim N(\mu, \sigma^2)$, $X \approx N\left(\mu, \frac{\sigma^2}{n}\right)$

$X \sim N(200, 8^2)$, $X \approx N\left(200, \frac{8^2}{9}\right)$

$\mu_0 = 200$
$\sigma = 8$
$n = 9$

1-tailed



Find the z value which corresponds to C.

Using z table backwards, value closest to 0.05:

$z = -1.645$ (halfway between 1.64 and 1.65)

$$z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

$$-1.645 = \frac{C - 200}{\frac{8}{\sqrt{9}}} \Rightarrow -1.645\left(\frac{8}{\sqrt{9}}\right) + 200 = C$$

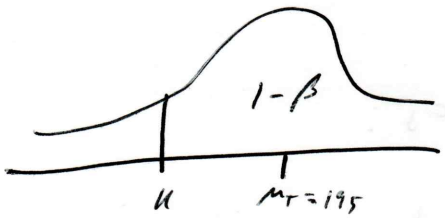
$$\Rightarrow C = \frac{14671}{75} = 195.61333\ldots$$

$C = 195.61333\ldots$

$\boxed{C = 196}$

The value of C such that the probability of a type I error is 0.05 ($\alpha = 0.05$) is $C = 196$. The resultant decision rule is to reject the null hypothesis if the sample mean of the four measurements falls below 196.

3) c) $n \geq \dfrac{(z_\alpha + z_\beta)^2 \sigma^2}{(\mu_T - \mu_0)^2}$ (1 tailed)



Power $= 1 - \beta$

$z_\alpha = 1.645$ (always positive)

$z_\beta = ?$

$\sigma^2 = 8^2$

$\mu_T = 200$

$\mu_0 = 195$

$n = 9$

$$9 \geq \frac{(1.645 + z_\beta)^2 \times 8^2}{(195 - 200)^2}$$

The sample size is $n = 9$, so change the sign to equals.

$$9 = \frac{(1.645 + Z_\beta)^2 \times 8^2}{(195-200)^2}$$

$$9 = \frac{(1.645 + Z_\beta)^2 \times 64}{(-5)^2}$$

$$\frac{9(-5)^2}{64} = (1.645 + Z_\beta)^2$$

$$1.645 + Z_\beta = \sqrt{\frac{9(-5)^2}{64}}$$

$$Z_\beta = \sqrt{\frac{9(-5)^2}{64}} - 1.645$$

$$Z_\beta = \frac{23}{100} = 0.23$$

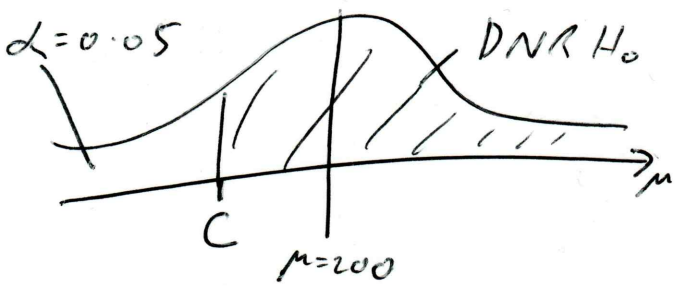
Using z distribution table:

$\beta = 0.4090$

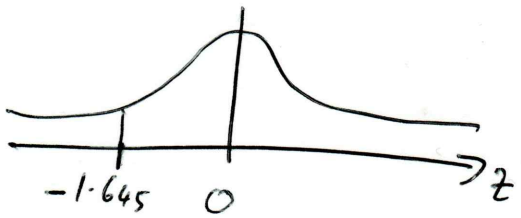
Power $= 1 - \beta = 1 - 0.4090 = 0.591 \simeq 59.1\%$

The power of the test is 59.1% if the true mean is 195.

3) d) $\alpha = 0.05$, $\beta = 0.10$ when $\mu = 195$



when assuming H_0 is true, so $\mu = 200$



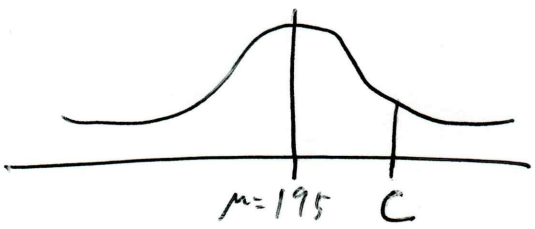
$z_\alpha = z_{0.05} = -1.645$ (value closest to 0.05 using z table, basically, halfway between 1.64 and 1.65)

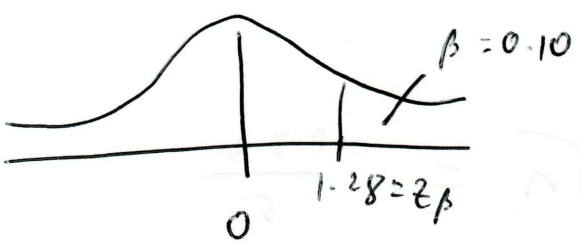
$\bar{x} = C$

$\mu = 200$

$\sigma = 8$

$$z_\alpha = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \Rightarrow \boxed{-1.645 = \frac{C - 200}{\frac{8}{\sqrt{n}}}} \quad -(1)$$





$z_\beta = z_{0.1} = 1.28$

$$\Rightarrow \boxed{1.28 = \frac{C - 195}{\frac{8}{\sqrt{n}}}} \quad -(2)$$

Solving (1) for C:

$$-1.645 = \frac{C-200}{\frac{8}{\sqrt{n}}}$$

$$-1.645\left(\frac{8}{\sqrt{n}}\right) + 200 = C$$

$$C = \frac{-13.16}{\sqrt{n}} + 200$$

Substitute C into eqn (2):

$$1.28 = \frac{C-195}{\frac{8}{\sqrt{n}}} \Rightarrow 1.28\left(\frac{8}{\sqrt{n}}\right) = C-195$$

~~$C = 1.28\left(\frac{8}{\sqrt{n}}\right)$~~ $\Rightarrow 1.28\left(\frac{8}{\sqrt{n}}\right) + 195 = C$

$$\frac{10.24}{\sqrt{n}} + 195 = C$$

$$\frac{10.24}{\sqrt{n}} + 195 = \frac{-13.16}{\sqrt{n}} + 200$$

$$\frac{10.24}{\sqrt{n}} + \frac{13.16}{\sqrt{n}} = 200 - 195$$

$$\frac{23.4}{\sqrt{n}} = 5 \Rightarrow \sqrt{n} = \frac{23.4}{5}$$

So $n = \left(\frac{23.4}{5}\right)^2 = \frac{13689}{625} = 21.9024$

$$\boxed{n = 22}$$

Substitute n back into eqn for C:

$$C = 1.28\left(\frac{8}{\sqrt{21.9024}}\right) + 195 = \frac{23071}{117} = 197.188034\overline{2}$$

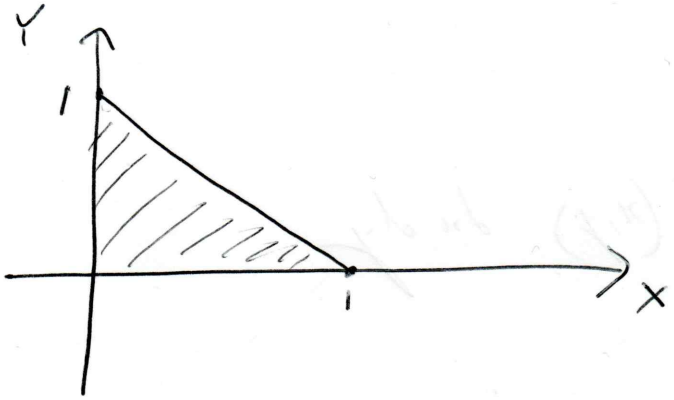
$$\boxed{C \simeq 197}$$

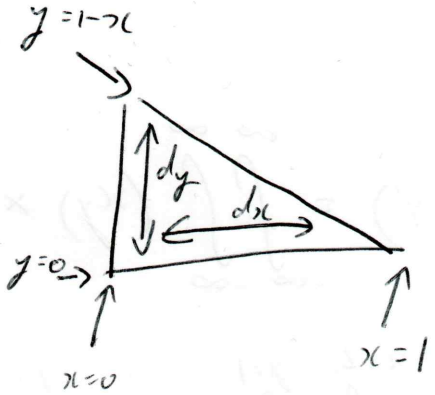
The suggested value for n with $\alpha = 0.05$ and $\beta = 0.10$ when $\mu = 195$ is 22. Therefore suggested sample size is 22.

4) $f(x,y) = \frac{1}{3}(x+y)$ on $0 < x < 2$, $0 < y < 1$

$f(x,y) = 0$ otherwise

a) $x + y < 1 \Rightarrow y < 1 - x$





$$P(x + Y < 1) = \int_{x=0}^{x=1} \int_{y=0}^{y=1-x} \frac{1}{3}(x+y)\, dy\, dx$$

$$= \int_{x=0}^{x=1} \int_{y=0}^{y=1-x} \frac{1}{3}x + \frac{1}{3}y\, dy\, dx = \int_{x=0}^{x=1} \left[\frac{1}{3}xy + \frac{1}{6}y^2\right]_{y=0}^{y=1-x} dx$$

$$= \int_{x=0}^{x=1} \left(\frac{1}{3}x(1-x) + \frac{1}{6}(1-x)^2\right) - 0\, dx$$

$$= \int_{x=0}^{x=1} \frac{1}{3}x - \frac{1}{3}x^2 + \frac{1}{6}(1 - 2x + x^2)\, dx = \int_{x=0}^{x=1} \frac{1}{3}x - \frac{1}{3}x^2 + \frac{1}{6} - \frac{1}{3}x + \frac{1}{6}x^2\, dx$$

$$= \int_{x=0}^{x=1} -\frac{1}{6}x^2 + \frac{1}{6}\, dx = \left[-\frac{1}{18}x^3 + \frac{1}{6}x\right]_{x=0}^{x=1}$$

$$= \left(-\frac{1}{18}(1)^3 + \frac{1}{6}(1)\right) - 0 = \frac{1}{9}$$

$\therefore P(X+Y<1) = \frac{1}{9}$

b) $E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (xy) \times f(x,y)\, dx\, dy$

$$\Rightarrow E(XY) = \int_{x=0}^{x=2}\int_{y=0}^{y=1} (xy)\left(\frac{1}{3}x + \frac{1}{3}y\right) dy\, dx$$

$$= \int_{x=0}^{x=2}\int_{y=0}^{y=1} \frac{1}{3}x^2y + \frac{1}{3}xy^2\, dy\, dx = \int_{x=0}^{x=2}\left[\frac{1}{6}x^2y^2 + \frac{1}{9}xy^3\right]_{y=0}^{y=1} dx$$

$$= \int_{x=0}^{x=2}\left[\left(\frac{1}{6}x^2(1)^2 + \frac{1}{9}x(1)^3\right) - 0\right] dx = \int_{x=0}^{x=2} \frac{1}{6}x^2 + \frac{1}{9}x\, dx$$

$$= \left[\frac{1}{18}x^3 + \frac{1}{18}x^2\right]_{x=0}^{x=2} = \left(\frac{1}{18}(2)^3 + \frac{1}{18}(2)^2\right) - 0$$

$$= \frac{4}{9} + \frac{2}{9} = \frac{2}{3}$$

So $E(XY) = \frac{2}{3}$

c) ~~$Cov(X,Y)$~~ $\quad Cov(X,Y) = E(XY) - E(X)E(Y)$

$$Cov(X,Y) = \left(\frac{2}{3}\right) - \left(\frac{11}{9}\right)\left(\frac{5}{9}\right) = -\frac{1}{81}$$

$$\Rightarrow Cov(X,Y) = -\frac{1}{81}$$

d) $T = 2X - 3Y - 1$

$E(aX+bY+c) = aE(X) + bE(Y) + c$
$V(aX+bY+c) = a^2V(X) + b^2V(Y) + 2ab\,Cov(X,Y)$

mean: $E(T) = E(2X-3Y-1) = 2E(X) - 3E(Y) - 1$

$$= 2\left(\frac{11}{9}\right) - 3\left(\frac{5}{9}\right) - 1 = -\frac{2}{9}$$

Variance: $V(T) = V(2X-3Y-1)$ ~~$= 2^2V(X) + (-3)^2V(Y)$~~ ~~$+ 2(2)(-3)$~~

$$= 2^2V(X) + (-3)^2V(Y) + 2(2)(-3)\left(-\frac{1}{81}\right)$$

$$= 2^2\left(\frac{23}{81}\right) + (-3)^2\left(\frac{13}{162}\right) + 2(2)(-3)\left(-\frac{1}{81}\right)$$

$$= \frac{325}{162}$$

e) $f_x(x) = \int_{y=0}^{y=1} \frac{1}{3}x + \frac{1}{3}y \, dy = \left[\frac{1}{3}xy + \frac{1}{6}y^2\right]_{y=0}^{y=1}$

$= \left(\frac{1}{3}x(1) + \frac{1}{6}(1)^2\right) - 0$

$= \frac{1}{3}x + \frac{1}{6}$ on $0 < x < 2$

f) $f(y|x) = \frac{f(x,y)}{f(x)}$ because $f(x) > 0$, and the variables are continuous.

$f(y|x) = \frac{\frac{1}{3}(x+y)}{\frac{1}{3}x + \frac{1}{6}} = \frac{\frac{1}{3}(x+y)}{\frac{1}{3}\left(x + \frac{1}{2}\right)} = \frac{x+y}{x+\frac{1}{2}}$

g) Conditional Expectation: $E(Y|x) = \int_{\text{all } Y} y \times f(y|x)\, dy$

because the variables are continuous

$E(Y|x) = \int_{y=0}^{y=1} y\left(\frac{x+y}{x+\frac{1}{2}}\right) dy = \int_{y=0}^{y=1} \frac{(x+y)y}{x+\frac{1}{2}} \, dy$

$= \int_{y=0}^{y=1} \frac{xy+y^2}{x+\frac{1}{2}} \, dy = \int_{y=0}^{y=1} \frac{xy}{x+\frac{1}{2}} + \frac{y^2}{x+\frac{1}{2}} \, dy$

$$= \left[\frac{1}{2} \cdot \frac{xy^2}{\left(x+\frac{1}{2}\right)} + \frac{1}{3} \cdot \frac{y^3}{\left(x+\frac{1}{2}\right)} \right]_{y=0}^{y=1}$$

$$= \left[\frac{xy^2}{2x+1} + \frac{y^3}{3x+\frac{3}{2}} \right]_{y=0}^{y=1} = \frac{x}{2x+1} + \frac{1}{3x+\frac{3}{2}}$$

$$= \frac{x\left(3x+\frac{3}{2}\right)}{(2x+1)\left(3x+\frac{3}{2}\right)} + \frac{1(2x+1)}{(2x+1)\left(3x+\frac{3}{2}\right)} = \frac{3x^2+\frac{3}{2}x+2x+1}{6x^2+3x+3x+\frac{3}{2}}$$

$$= \frac{3x^2+\frac{7}{2}x+1}{6x^2+6x+\frac{3}{2}} = \frac{(3x+2)\cancel{\left(x+\frac{1}{2}\right)}}{(6x+3)\cancel{\left(x+\frac{1}{2}\right)}} = \frac{3x+2}{6x+3} = \frac{\cancel{3}\left(x+\frac{2}{3}\right)}{\cancel{3}(2x+1)}$$

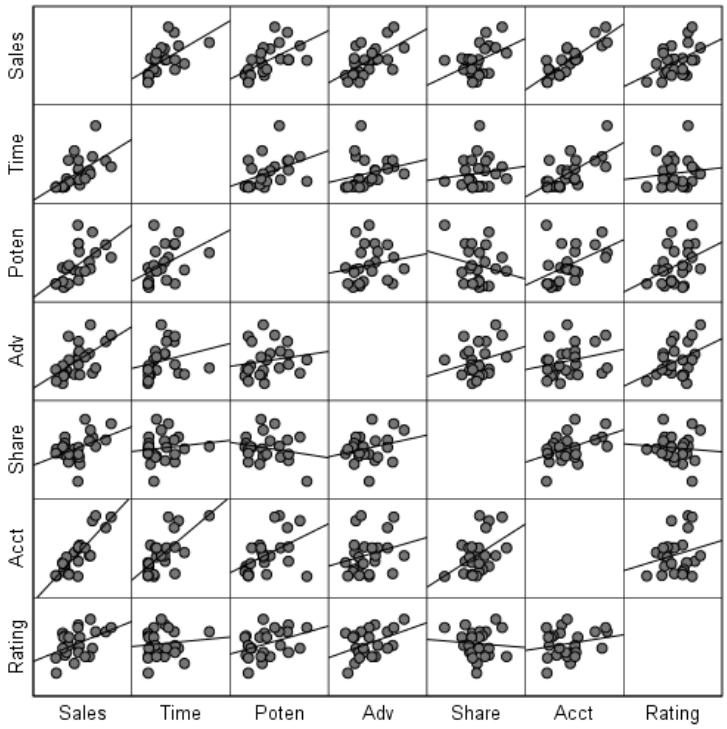
$$\Rightarrow E(Y|X) = \frac{x+\frac{2}{3}}{2x+1}$$

$$E(Y|X=1.5) = \frac{(1.5)+\frac{2}{3}}{2(1.5)+1} = \frac{13}{24}$$

5)

a)

Scatterplot matrix including the single dependent variable and the six continuous variables:



Pearson correlation matrix of the six continuous variables:

Correlations

		Time	Poten	Adv	Share	Acct	Rating
Time	Pearson Correlation	1	.434*	.243	.127	.684**	.107
	Sig. (2-tailed)		.034	.252	.554	.000	.617
	N	24	24	24	24	24	24
Poten	Pearson Correlation	.434*	1	.173	-.210	.479*	.380
	Sig. (2-tailed)	.034		.419	.324	.018	.067
	N	24	24	24	24	24	24
Adv	Pearson Correlation	.243	.173	1	.258	.244	.415*
	Sig. (2-tailed)	.252	.419		.223	.250	.044
	N	24	24	24	24	24	24
Share	Pearson Correlation	.127	-.210	.258	1	.459*	-.082
	Sig. (2-tailed)	.554	.324	.223		.024	.705
	N	24	24	24	24	24	24
Acct	Pearson Correlation	.684**	.479*	.244	.459*	1	.221
	Sig. (2-tailed)	.000	.018	.250	.024		.299

	N	24	24	24	24	24	24
Rating	Pearson Correlation	.107	.380	.415*	-.082	.221	1
	Sig. (2-tailed)	.617	.067	.044	.705	.299	
	N	24	24	24	24	24	24

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

Second correlation matrix displaying just single dependent variable against six continuous predictor variables:

Correlations

		Sales
Time	Pearson Correlation	.604**
	Sig. (2-tailed)	.002
	N	24
Poten	Pearson Correlation	.601**
	Sig. (2-tailed)	.002
	N	24
Adv	Pearson Correlation	.584**
	Sig. (2-tailed)	.003
	N	24
Share	Pearson Correlation	.431*
	Sig. (2-tailed)	.036
	N	24
Acct	Pearson Correlation	.848**
	Sig. (2-tailed)	.000
	N	24
Rating	Pearson Correlation	.443*
	Sig. (2-tailed)	.030
	N	24

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Syntax used on SPSS:

CORRELATIONS

```
 /VARIABLES=Time Poten Adv Share Acct Rating WITH Sales
 /PRINT=TWOTAIL NOSIG
 /MISSING=PAIRWISE.
```

Preliminary examination of the relationships between "sales" and the six continuous predictor variables:

Comments on the scatterplot matrix and the second correlation matrix:

The scatterplots of "Sales" against "Time", "Potent" and "Adv" appear to have a moderately positive correlation. This is confirmed by looking at the output from the second correlation matrix. The Pearson correlation for each is approximately +0.6 (to 1 significant figure), and the p-values are approximately 0.002, 0.002 and 0.003 respectively, which are less than 0.05 therefore showing that the correlation is statistically significant.

Both the scatterplots of "Sales" against "Share" and "Rating" appear to have a moderately weak positive correlation. This can also be confirmed by looking at the output from the second correlation matrix. The correlations are approximately +0.4 for both, with p-values of approximately 0.036 and 0.03 respectively. Since the p-values are less than 0.05, they show that the correlations are still statistically significant, despite being slightly weaker.

Finally, the scatterplot of "Sales" against "Acct" shows a very strong positive correlation. From looking at the output from the second correlation matrix, we can see the Pearson correlation is +0.8. The p-value is approximately zero, which is less than 0.05, so the correlation is statistically significant.

Preliminary examination of the relationships between the continuous predictor variables:

Comments on the scatterplot matrix and the first correlation matrix:

The scatterplots of "Time" against "Poten", "Poten" against "Acct", "Adv" against "Rating" and "Share" against "Acct" all show a moderately positive correlation. This can also be seen by looking at the first correlation matrix where the Pearson values are +0.4, +0.5, +0.4, and +0.5 (to 1 significant figure) respectively. They have p-values of approximately 0.034, 0.018, 0.044, and 0.024, which are all less than 0.05, so they show statistical significance.

The scatterplots of "Time" against "Adv", "Share" and "Rating", "Poten" against "Adv" and "Rating", "Adv" against "Share", "Acct" and "Acct" against "Rating" appear to all have a very weak positive correlation with Pearson values of +0.2, +0.1, +0.1, +0.2, +0.4, +0.3, +0.2, +0.2 (to 1 significant figure) respectively. They have p-values of approximately 0.252, 0.554, 0.617, 0.419, 0.067, 0.223, 0.250, and 0.299. These are all greater than 0.05, so the correlation is not statistically significant.

The scatterplot of "Time" against "Acct" appears to have a moderately strong positive correlation, with a Pearson value of +0.7 (to 1 significant figure) and a p-value of approximately zero, which shows statistical significance.

Finally, the scatterplots of "Poten" against "Share", and "Share" against "Rating" both appear to have a weak negative correlation. The Pearson values are -0.2 and 0.1 (to 1 significant figure)

respectively. The p-values are approximately 0.324 and 0.705, which are both greater than 0.05, which shows it is not statistically significant.

b)

The best single predictor of "Sales" is "Acct", the number of accounts assigned to the salesperson. The Pearson correlation is +0.848 which shows a very strong positive correlation. Increasing "Acct" by 1 is expected to increase "Sales" by 0.848 on average. The p-value is approximately zero, which is less than 0.05, which shows statistical significance. Therefore "Acct" is the best predictor of "Sales"

From the Model summary table below, the R Square value is 0.719, which implies that 71.9% of the variance in the "Sales" is explained by "Acct".

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.848[a]	.719	.706	139.43166	.719	56.227	1	22	.000

a. Predictors: (Constant), Acct

c)

Fitting a multiple regression model predicting sales:

Descriptive Statistics

	Mean	Std. Deviation	N
Sales	705.4583	257.14367	24
Time	89.8750	87.66652	24
Poten	39.0000	15.70932	24
Adv	43.3750	27.40646	24
Share	7.3500	2.72429	24
Acct	121.5417	40.70357	24
Rating	3.7375	.97927	24

Correlations

		Sales	Time	Poten	Adv	Share	Acct	Rating
Pearson Correlation	Sales	1.000	.604	.601	.584	.431	.848	.443
	Time	.604	1.000	.434	.243	.127	.684	.107
	Poten	.601	.434	1.000	.173	-.210	.479	.380
	Adv	.584	.243	.173	1.000	.258	.244	.415
	Share	.431	.127	-.210	.258	1.000	.459	-.082
	Acct	.848	.684	.479	.244	.459	1.000	.221
	Rating	.443	.107	.380	.415	-.082	.221	1.000
Sig. (1-tailed)	Sales	.	.001	.001	.001	.018	.000	.015
	Time	.001	.	.017	.126	.277	.000	.309
	Poten	.001	.017	.	.210	.162	.009	.034
	Adv	.001	.126	.210	.	.111	.125	.022
	Share	.018	.277	.162	.111	.	.012	.352
	Acct	.000	.000	.009	.125	.012	.	.149
	Rating	.015	.309	.034	.022	.352	.149	.
N	Sales	24	24	24	24	24	24	24
	Time	24	24	24	24	24	24	24
	Poten	24	24	24	24	24	24	24
	Adv	24	24	24	24	24	24	24
	Share	24	24	24	24	24	24	24
	Acct	24	24	24	24	24	24	24
	Rating	24	24	24	24	24	24	24

From the correlations table above, we can see that the most significant non-zero correlation with regards to the dependent variable is between “Sales” and “Acct”, as the p-value is approximately equal to zero.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.964[a]	.928	.903	79.99970

a. Predictors: (Constant), Rating, Share, Time, Adv, Poten, Acct

b. Dependent Variable: Sales

The R-square value of 0.928 shows that 92.8% of variance in the dependent variable is explained by the model, therefore 92.8% of the variance in “Sales” is explained by the model, which is very high.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1412026.784	6	235337.797	36.772	.000[b]
	Residual	108799.174	17	6399.951		
	Total	1520825.958	23			

a. Dependent Variable: Sales

b. Predictors: (Constant), Rating, Share, Time, Adv, Poten, Acct

The p-value for the ANOVA is approximately zero, which is less than 0.05, and 0.005, so this suggests one or more of the six continuous predictor variables have a significant effect on the "Sales". This also suggests that the model is useful.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-233.120	90.323		-2.581	.019	-423.684	-42.555
	Time	.040	.284	.014	.140	.890	-.559	.638
	Poten	4.790	1.506	.293	3.181	.005	1.613	7.968
	Adv	2.969	.733	.316	4.052	.001	1.423	4.515
	Share	16.987	9.077	.180	1.871	.079	-2.165	36.139
	Acct	3.262	.791	.516	4.122	.001	1.592	4.931
	Rating	26.252	20.731	.100	1.266	.222	-17.486	69.990

a. Dependent Variable: Sales

The hypotheses for the parameters in the model are:

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$

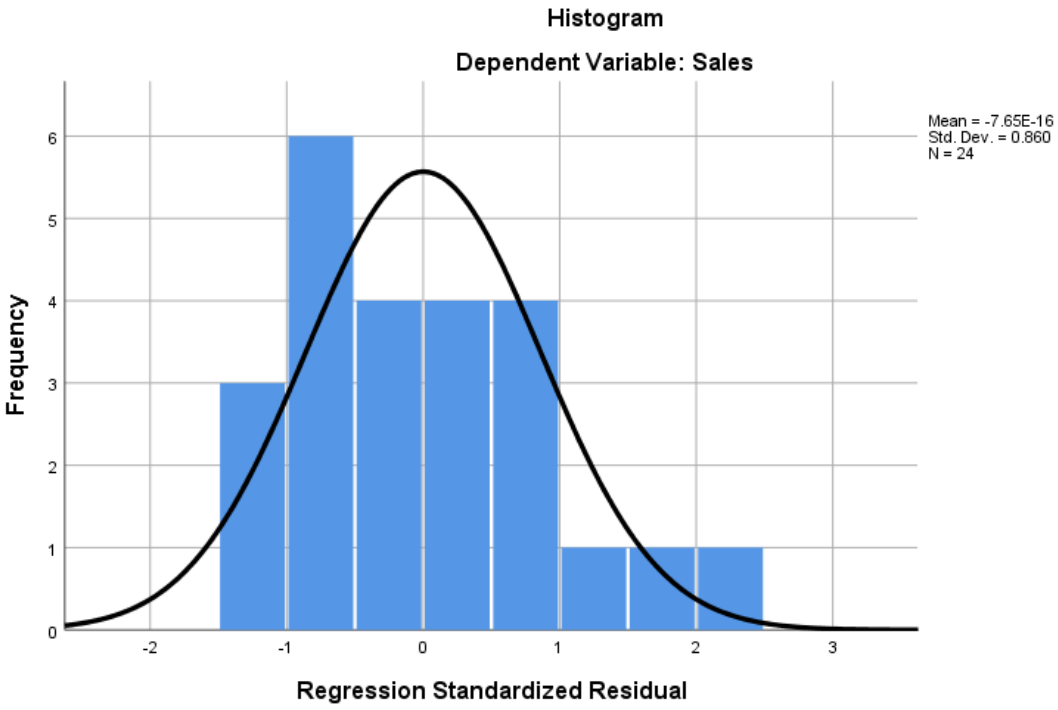
H_1: $\beta_1 \neq \beta_2 \neq \beta_3 \neq \beta_4 \neq \beta_5 \neq \beta_6 \neq 0$

Null hypothesis: The predictor variables have no effect in predicting the "Sales" of a salesperson.

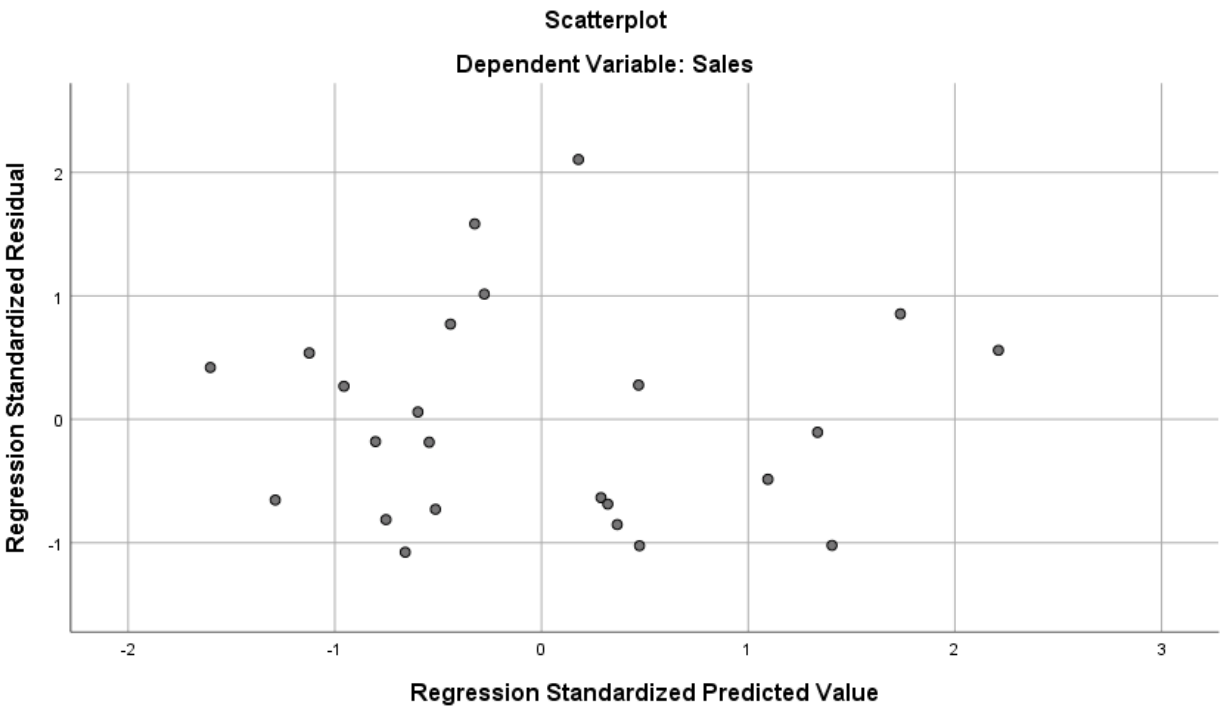
Alternative hypothesis: One or more of the predictor variables do have an effect in predicting the "Sales" of a salesperson.

We can see that the "Time", "Share" and "Rating" variables all have p-values that are greater than 0.05. They also include zero in the 95% confidence intervals. So, these three predictor variables are not useful in predicting the "Sales".

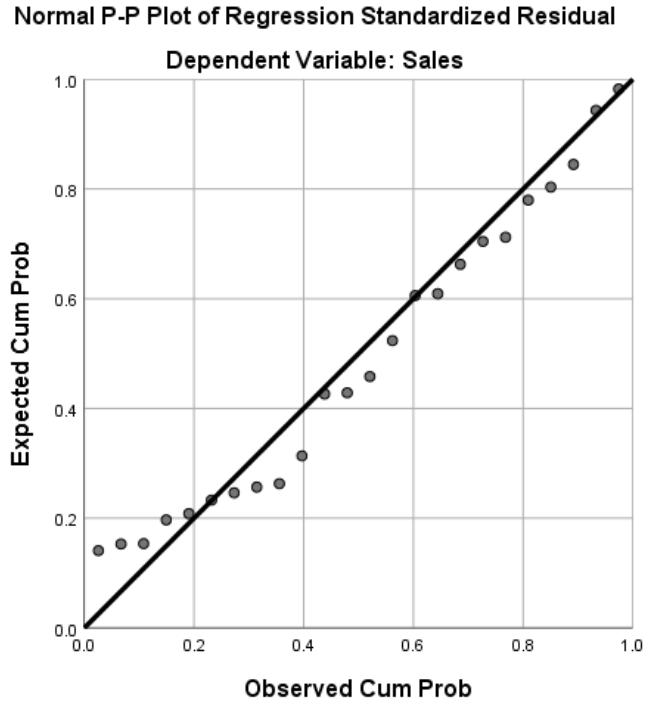
We can also see that the "Poten", "Adv" and "Acct" variables have p-values that are less than 0.05. Zero is not included in their confidence intervals, so we can conclude that these three predictor variables are useful in predicting the "Sales".



From looking at the histogram above we see that the standardised residuals appear to be normally distributed about zero.



The regression standardised residuals and regression standardised predicted values appear to be roughly constant, so we can conclude homoscedasticity.



The expected values appear very close to the observed values, so this shows the model also has a good fit, as well as having a high R-square value. There is evidence of redundancy however in three of the predictor variables, “Time”, “Share” and “Rating” since the p-values are less than 0.05.

Conclusion:

This multiple regression model for predicting “Sales” is suitable, since three of the predictor variables are significant in predicting the “Sales”, which is the same as one or more. The model also explains 92.8% of the variance in “Sales” which is very high. And the residuals appear to be normally distributed about zero with constant variance.

d)

For each additional account given to a salesperson, we would expect to see in increase in sales of 3.262 units, which means an increase in £3262 worth of sales.

95% Confidence Interval about estimate:

Let Acct=x_5

Degrees of freedom = 17

$(b_{x_5} - t_{2.5\%}\,(17)(SE(b_{x_5}), b_{x_5} + t_{2.5\%}\,(17)(SE(b_{x_5}))$

= (3.262 – 2.125(0.791), 3.262+2.125(0.791))

= (1.581125,4.942875)

So, the confidence interval will be:

(3262-1.581125, 3262+4.942875)

= (3260.418875, 3266.942875)

≈ (£3260, £3267)

e)

Using backwards elimination to produce a reduced model:

Step 1: Remove the predictor variable "Time" since it has the highest p-value of approximately 0.890. The coefficients table is below.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-233.857	87.679		-2.667	.016
	Poten	4.794	1.464	.293	3.275	.004
	Adv	2.995	.691	.319	4.337	.000
	Share	16.595	8.396	.176	1.976	.064
	Acct	3.331	.601	.527	5.540	.000
	Rating	25.591	19.629	.097	1.304	.209

a. Dependent Variable: Sales

Step 2: Now remove the variable "Rating" since that now has the highest p-value of approximately 0.209.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-163.084	70.111		-2.326	.031
	Poten	5.151	1.464	.315	3.518	.002
	Adv	3.359	.643	.358	5.226	.000
	Share	14.902	8.447	.158	1.764	.094
	Acct	3.393	.610	.537	5.560	.000

a. Dependent Variable: Sales

Step 3: Remove the variable “Share”, since it has the highest p-value of approximately 0.094.

This is the reduced model:

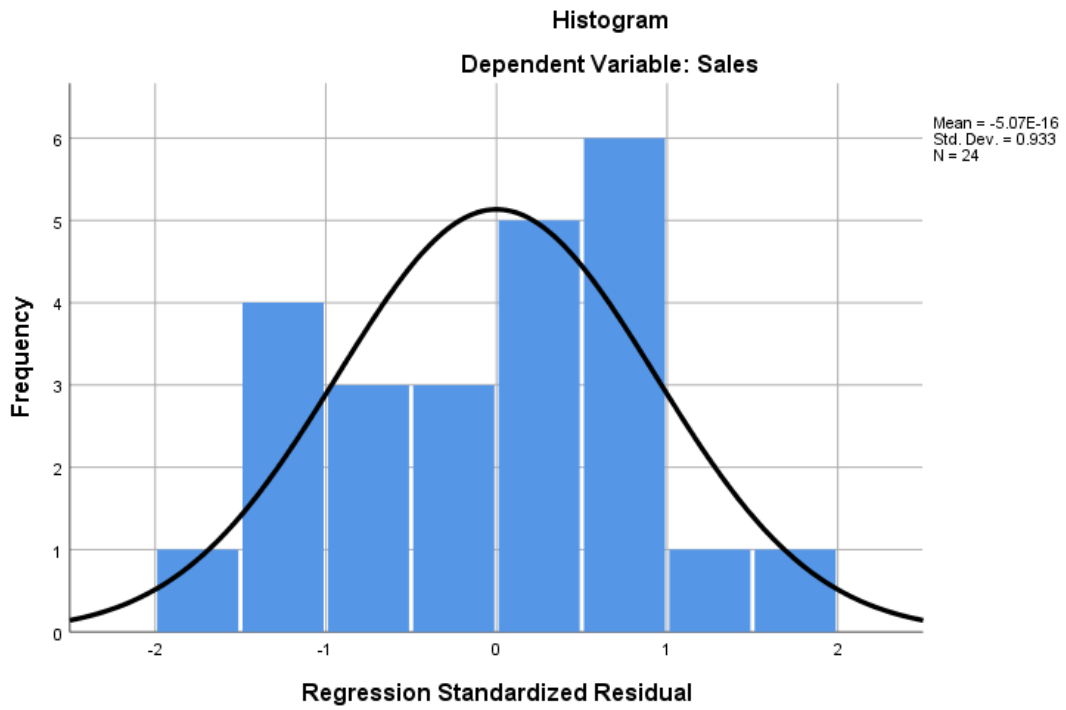
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-91.329	60.049		-1.521	.144
	Poten	3.672	1.262	.224	2.909	.009
	Adv	3.640	.655	.388	5.559	.000
	Acct	4.078	.495	.646	8.242	.000

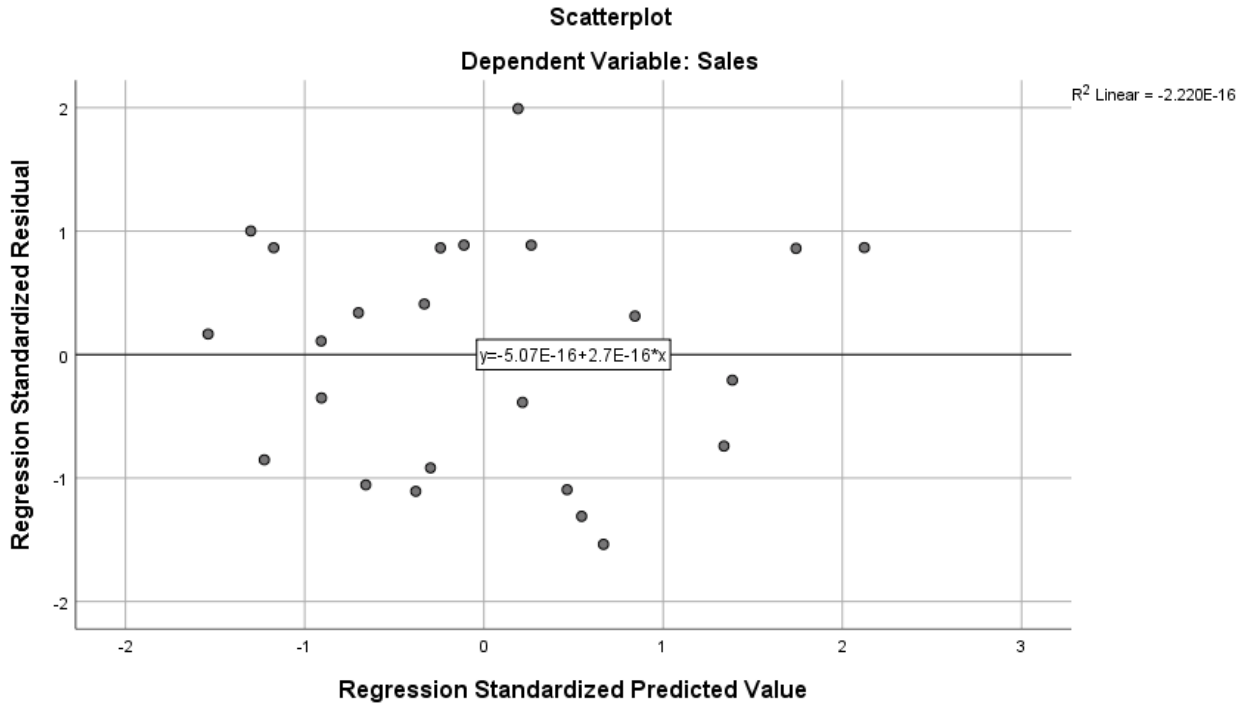
a. Dependent Variable: Sales

The three continuous variables left in the reduced model are all significant at the 5% level of significance, since their p-values are less than 0.05. The constant has a p-value greater than 0.05, however it is left in the model otherwise it could introduce bias.

f)



Looking at the plot above, the regression standardised residuals do not appear to be normally distributed. They do not fit the symmetric bell curve.



From this scatterplot above of the regression standardised residuals and regression standardised predicted values, it appears to be roughly constant, so we can conclude from this homoscedasticity.

Therefore, the usual assumptions are not reasonable for regression analysis.

g)
Using the three significant predictors, the unstandardized predicted value obtained for the annual income ("Sales") generated by this salesperson is 701.8841549471924 = £701884.1549 ≈ £701884.15. This makes sense, because calculating by hand:
-91.329 + 3.762(Poten) +3.64(Adv) + 4.078(Acct)
= -91.329 + 3.762(39) +3.64(43) + 4.078(121)
= 701.837
= £701837, which is similar to the value from SPSS.

The 95% prediction interval is (524.5634211507937, 879.2048887435911) = (£524563.4212, £879204.8887) ≈ (£524563.42, 879204.89). Therefore, the range is £879204.89 - £524563.42 = £354641.47. The range is quite high, which implies the prediction interval is not very reliable.

h)
Creating an "indicator" for the categorical variable "training" and using the Mann Whitney U test, since the two samples are independent. And using the reduced model:
H_0: $\eta_1 = \eta_2$
H_1: $\eta_1 > \eta_2$
Where:
1 = "Yes", 2 = "No" (If the salesperson participated in company training in the last twelve months)

Null hypothesis: There is no significant difference between the median sales of salesperson's that participated in company training in the last twelve months.
Alternative hypothesis: The median sales of salespersons that participated in company training in the last twelve months was higher than those that didn't.

Ranks

	Training	N	Mean Rank	Sum of Ranks
Unstandardized Predicted Value	Yes	6	11.00	66.00
	No	18	13.00	234.00
	Total	24		

Test Statistics[a]

	Unstandardized Predicted Value
Mann-Whitney U	45.000
Wilcoxon W	66.000
Z	-.600
Asymp. Sig. (2-tailed)	.549
Exact Sig. [2*(1-tailed Sig.)]	.581[b]

a. Grouping Variable: Training

b. Not corrected for ties.

Conclusion:
The one tailed p-value is approximately 0.581, which is greater than 0.05, so we cannot reject the null hypothesis that there is no median difference between salesperson's with training and those without at the 5% level of significance. Therefore, we conclude that attending annual training is not beneficial to the company at the 5% level of significance, using the reduced model.

i)
Two Sample T-test because the two samples are independent:
$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$
This becomes:
$H_0: \mu_D = 0$
$H_1: \mu_D \neq 0$
Where D = 1 – 2, (1 = "Yes, 2 = "No" to having participated in company training in the last twelve months)

Null hypothesis: There is no significant mean difference between the duration of employment (months) of salespersons and whether or not they have participated in company training in the last twelve months.
Alternative hypothesis: There is a significant mean difference between the duration of employment (months) of salespersons and whether or not they have participated in company training in the last twelve months.

Group Statistics

	Training	N	Mean	Std. Deviation	Std. Error Mean
Time	Yes	6	101.5000	84.84044	34.63596
	No	18	86.0000	90.64929	21.36624

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Time	Equal variances assumed	.004	.951	.368	22	.716	15.50000	42.12577	-71.86351	102.86351
	Equal variances not assumed			.381	9.140	.712	15.50000	40.69602	-76.34603	107.34603

Conclusion:
The p-value is approximately 0.716, which is greater than 0.05, so we cannot reject the null hypothesis that there is no significant mean difference between the duration of employment (months) of salespersons and whether they have participated in company training in the last twelve months at the 5% level of significance.

We can conclude that there is not a statistically significant difference between the duration of time employed and whether or not the salespersons have participated in company training at the 5% level of significance.

Furthermore the 95% confidence interval (-71.86351, 102.86351) includes zero, which suggests no difference.

j)

Assumptions for the t-test in part (i):

1) The group that participated in company training and the group that didn't are assumed to have equal variances.
2) The data for each group is assumed to be normally distributed.
3) The two groups are independent of each other.

k)

Linear Regression Model with "Time" as the dependent variable and "Training" as the independent variable:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	117.000	75.943		1.541	.138
	Training	-15.500	42.126	-.078	-.368	.716

a. Dependent Variable: Time

From the SPSS output, the t-statistic for “Training” is approximately -0.368, which is the same as 0.368 in part (i), since the distribution is symmetric. The p-value is 0.716, which is also the same as part (i). So, this linear regression model produces the same t-statistic and p-value as part (i).