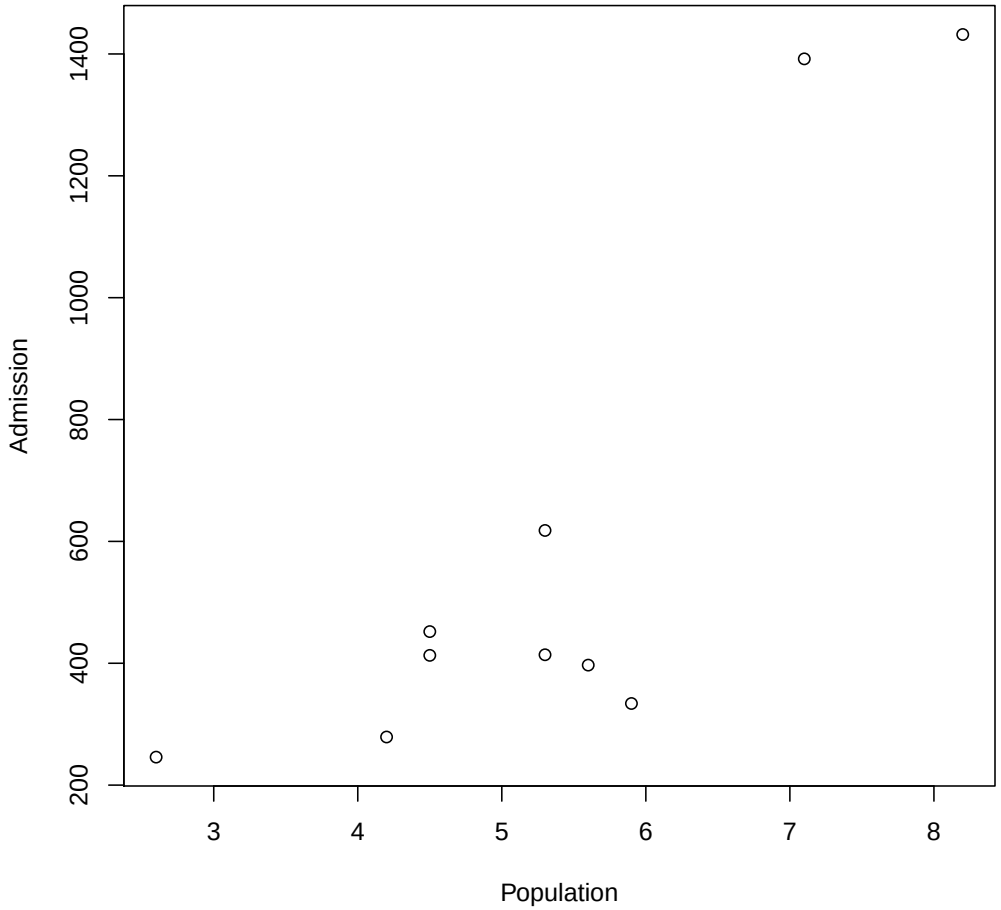
Assessed Coursework 2 Solutions



1. (a)

   The scatterplot implies there is a non-linear relationship between population and admissions so a non-parametric test is appropriate.

   (b) We rank each variable and then calculate the sum of squares of the difference in ranks.

Region	A	P	Rank(A)	Rank(P)	d	d^2
North East	246	2.6	1	1.0	0	0.00
North West	1392	7.1	9	9.0	0	0.00
Yorkshire	618	5.3	8	5.5	2.5	6.25
East Midlands	452	4.5	7	3.5	3.5	12.25
West Midlands	397	5.6	4	7.0	3	9.00
East of England	334	5.9	3	8.0	5	25.00
London	1432	8.2	10	10.0	0	0.00
South East	413	4.5	5	3.5	1.5	2.25
South Central	279	4.2	2	2.0	0	0.00
South West	414	5.3	6	5.5	0.5	0.25

Therefore $\sum d_i^2 = 55$. The test statistic is

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 55}{10 \times 99} = \frac{2}{3}.$$

From tables, the critical value is 0.65. Therefore we reject the null hypothesis.

2.

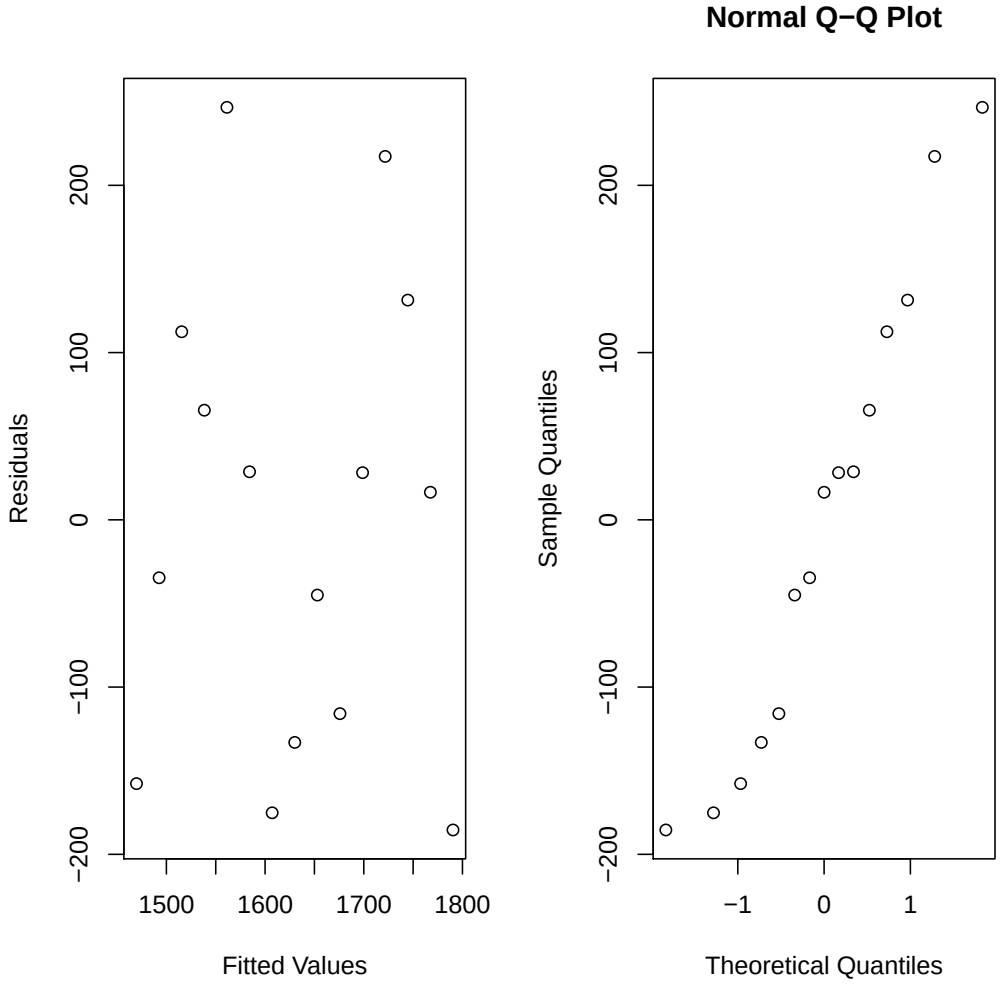
$$\begin{aligned}
t_2 &= \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{(n-1)s_x^2}} = \hat{\beta}_1\sqrt{\frac{(n-1)s_x^2}{\hat{\sigma}^2}} \\
&= \hat{\beta}_1\sqrt{\frac{(n-1)s_x^2(n-2)}{(n-1)(s_y^2 - \hat{\beta}_1^2 s_x^2)}} && \text{since } \hat{\sigma}^2 = \frac{n-1}{n-2}(s_y^2 - \hat{\beta}_1^2 s_x^2) \\
&= \frac{s_{xy}}{s_x^2}\sqrt{\frac{s_x^2(n-2)}{(s_y^2 - s_{xy}^2/s_x^2)}} && \text{since } \hat{\beta}_1 = \frac{s_{xy}}{s_x^2} \\
&= \frac{s_{xy}}{s_x^2}\sqrt{\frac{s_x^2(n-2)}{s_y^2(1 - s_{xy}^2/s_x^2 s_y^2)}} = \frac{s_{xy}}{s_x s_y}\sqrt{\frac{n-2}{1 - s_{xy}^2/s_x^2 s_y^2}} \\
&= r\sqrt{\frac{n-2}{1-r^2}} && \text{since } r = \frac{s_{xy}}{s_x s_y} \\
&= t_1
\end{aligned}$$

3. (a) From `R` or otherwise, we find $\hat{\beta}_0 = -44275.848$ and $\hat{\beta}_1 = 22.907$.

   (b) From `R` or otherwise, we find $\hat{\sigma} = 143.7$ and $s_x^2 = 20$. Therefore a 95% confidence interval for β_1 is

$$\begin{aligned}
\hat{\beta}_1 \pm t_{n-2,0.025}\frac{\hat{\sigma}}{\sqrt{(n-1)s_x^2}} &= 22.907 \pm 2.16\frac{143.7}{\sqrt{14 \times 20}} \\
&= 22.907 \pm 18.549 = (4.358, 41.456)
\end{aligned}$$

   The confidence interval does not span 0 so there is evidence for the existence of regression.



(c)

The residual plot shows a band of random scatter about zero, indicating that the model assumptions are satisfied. The QQ plot shows that the residuals are not normally distributed and therefore the assumptions of the test for the existence of regression are not satisfied.

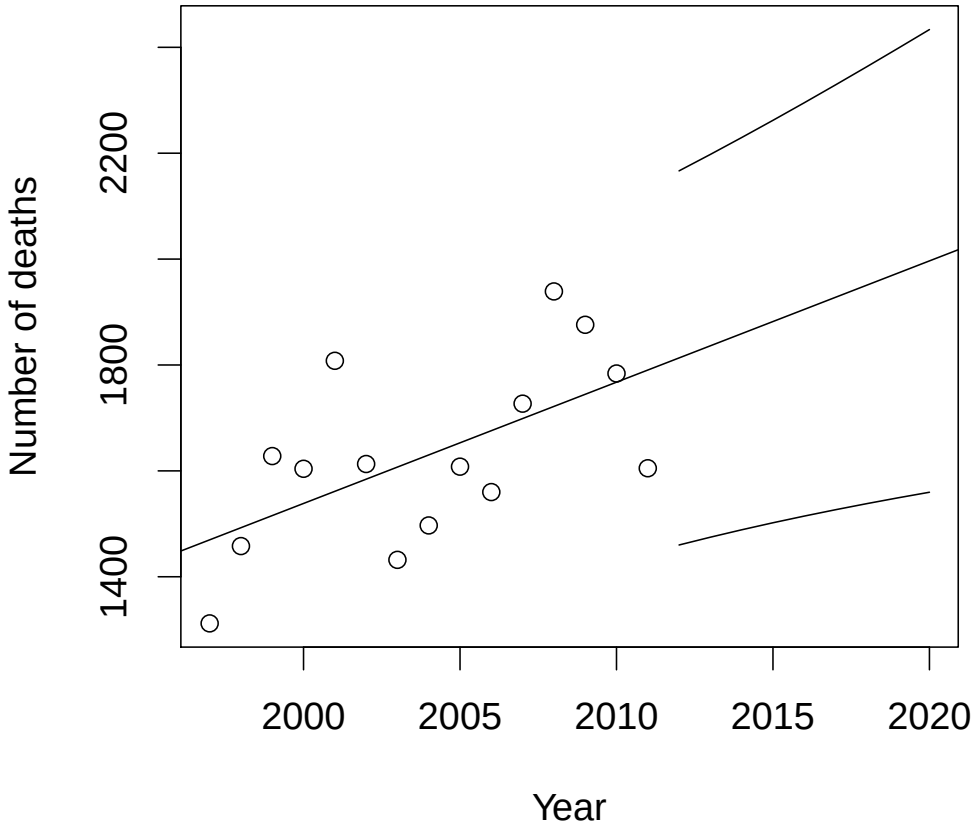
(d) The 95% prediction intervals are given by

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{n-2,0.025}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}}$$

$$= -44275.848 + 22.907x_0 \pm 2.16 \times 143.7\sqrt{1 + \frac{1}{15} + \frac{(x_0 - 2004)^2}{14 \times 20}}$$

for $x_0 = 2012, 2013, \ldots, 2020$. Using `R` or otherwise the prediction intervals are

Year	Interval
2012	(1460.0, 2166.6)
2013	(1474.7, 2197.7)
2014	(1488.7, 2229.6)
2015	(1502.0, 2262.1)
2016	(1514.6, 2295.3)
2017	(1526.6, 2329.1)
2018	(1538.1, 2363.4)
2019	(1549.1, 2398.2)
2020	(1559.6, 2433.5)



(e)

4. (a) The sample proportion is $p = 30/40 = 0.75$. The test statistic is

$$z = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} = \frac{0.75 - 0.5}{\sqrt{0.5(1-0.5)/40}} = \sqrt{10} = 3.162.$$

We reject H_0 if $z > z_\alpha = z_{0.05} = 1.645$. Therefore we reject the null hypothesis and conclude the proportion of deaths where the subject is male is greater than 0.5.

(b) We reject H_0 in (a) if $\dfrac{p-\pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} > z_{0.05}$, i.e.

$$p > \pi_0 + z_{0.05}\sqrt{\frac{\pi_0(1-\pi_0)}{n}} = 0.5 + 1.645\sqrt{\frac{0.5(1-0.5)}{40}} = 0.63.$$

The probability of a type II error for $\pi = \pi_1$ is

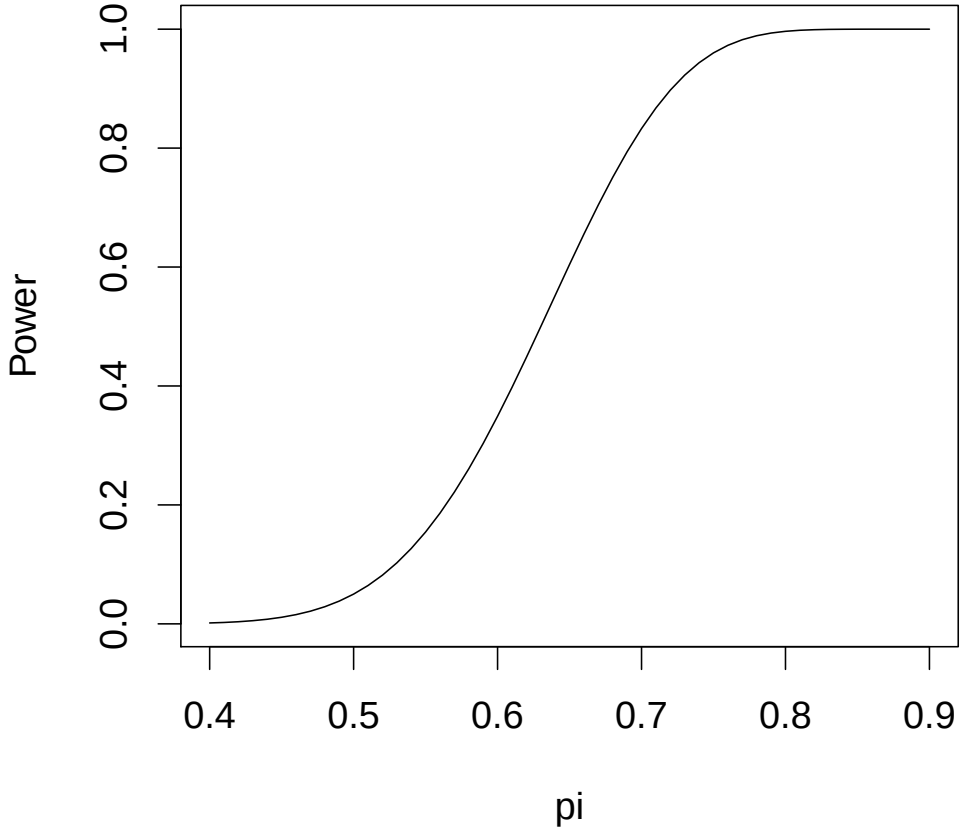
$$\begin{aligned}\beta &= P(\text{accepting } H_0 | H_0 \text{ false}) = P(p < 0.63 | p \sim N(\pi_1, \pi_1(1-\pi_1)/n)) \\ &= P\left(Z = \frac{p-\pi_1}{\sqrt{\pi_1(1-\pi_1)/n}} < \frac{0.63-\pi_1}{\sqrt{\pi_1(1-\pi_1)/n}} | Z \sim N(0,1)\right)\end{aligned}$$

For $\pi_1 = 0.6$, $\beta = P\left(Z < \dfrac{0.63-0.6}{\sqrt{0.6(1-0.6)/40}}\right) = P(Z < 0.39) = 0.65$.

For $\pi_1 = 0.7$, $\beta = P\left(Z < \dfrac{0.63-0.7}{\sqrt{0.7(1-0.7)/40}}\right) = P(Z < -0.97) = 1 - P(Z < 0.97) = 1 - 0.83 = 0.17$.

For $\pi_1 = 0.8$, $\beta = P\left(Z < \dfrac{0.63-0.8}{\sqrt{0.8(1-0.8)/40}}\right) = P(Z < -2.69) = 1 - P(Z < 2.69) = 1 - 0.9964 = 0.0036$.

(c) The power is given by $1 - \beta = 1 - P\left(Z < \dfrac{0.63-\pi_1}{\sqrt{\pi_1(1-\pi_1)/n}}\right)$. The power curve for $0.4 \leq \pi_1 \leq 0.9$ is



(d) Let π_1 and π_2 represent the proportion for under 40 year olds and over 40 year olds respectively. The respective sample proportions are $p_1 = 16/20 = 0.8$ and $p_2 = 14/20 = 0.7$. The overall proportion is $\hat{\pi} = \dfrac{16+14}{20+20} = \dfrac{30}{40} = 0.75$.
We wish to test $H_0 : \pi_1 = \pi_2$ vs. $H_1 : \pi_1 \neq \pi_2$. The test statistic is

$$z = \frac{p_1 - p_2}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.8 - 0.7}{\sqrt{0.75(1-0.75)\frac{2}{20}}} = 0.73.$$

We reject H_0 if $|z| > z_{0.025} = 1.96$. Therefore we do not reject the null hypothesis.