THE UNIVERSITY OF SYDNEY

STAT2012 STATISTICAL TESTS

| Semester 2 | **Solution to Tutorial Week 10** | 2015 |
|---|---|---|

## Tutorial questions

1. We have $n = 8$,

$$\sum_{i=1}^{8} x_i = 429.5;\ \sum_{i=1}^{8} y_i = 149.4;\ \sum_{i=1}^{8} x_i^2 = 26,533.95;\ \sum_{i=1}^{8} y_i^2 = 5501.1;\ \sum_{i=1}^{8} x_i y_i = 11,008.59.$$

(a) We have $\overline{y} = \dfrac{149.4}{8} = 18.675$, $\overline{x} = \dfrac{429.5}{8} = 53.6875$,

$$\begin{aligned} S_{xy} &= \sum_i x_i y_i - n\overline{xy} = 11,008.59 - 8 \times 53.6875 \times 18.675 = 2987.6775, \\ S_{xx} &= \sum_i x_i^2 - n\overline{x}^2 = 26,533.95 - 8 \times 53.6875^2 = 3475.16875. \end{aligned}$$

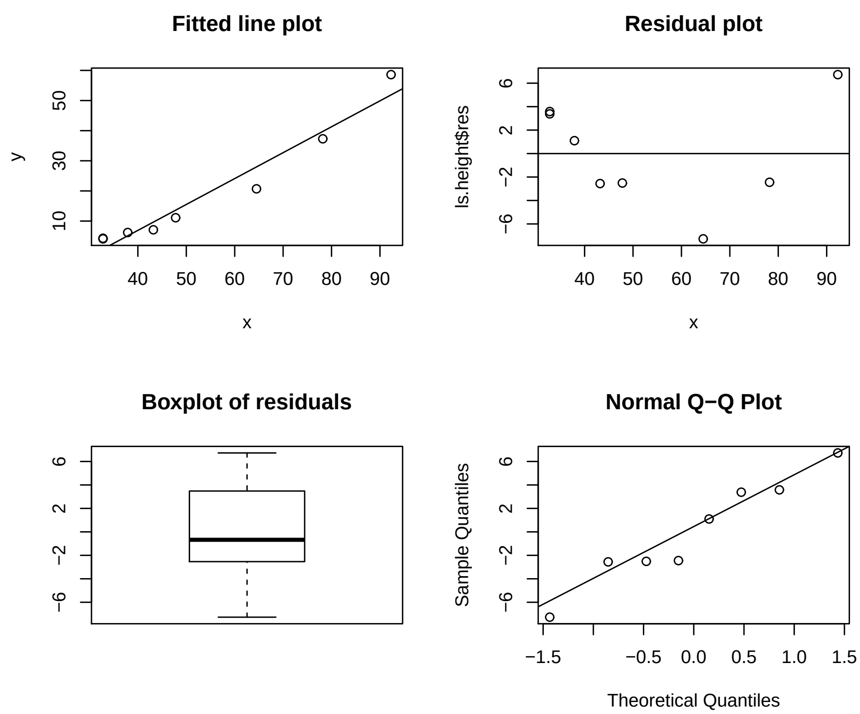
Therefore,

$$\begin{aligned} \widehat{\beta} &= \frac{S_{xy}}{S_{xx}} = \frac{2987.6775}{3475.16875} = 0.85972156, \\ \widehat{\alpha} &= \bar{y} - \widehat{\beta}\bar{x} = 18.675 - 0.85972156 \times 53.6875 = -27.4813013. \end{aligned}$$

The estimated regression line is

Stopping dist. = -27.4813013 + 0.85972156 $\times$ Velocity.

From the plot, the relationship between the stopping distance $Y$ and the velocity $X$ is not linear.



(b) Now the revised summary statistics are

$$\sum_{i=1}^{8} y_i = 30.89688; \quad \sum_{i=1}^{8} y_i^2 = 149.4; \quad \sum_{i=1}^{8} x_i y_i = 1980.780511.$$

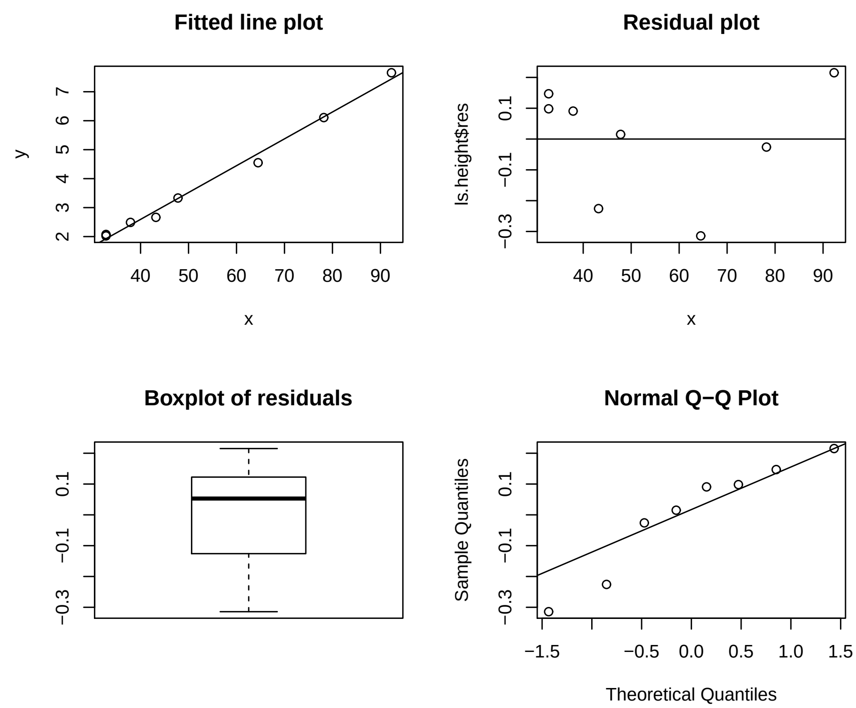
Hence $\overline{y} = \dfrac{30.89688}{8} = 3.86211$, and

$$S_{xy} = \sum_i x_i y_i - n\overline{x}\overline{y} = 1980.780511 - 8 \times 53.6875 \times 3.86211 = 322.004266.$$

Therefore,

$$\begin{aligned} \widehat{\beta} &= \frac{S_{xy}}{S_{xx}} = \frac{322.004266}{3475.16875} = 0.092658598. \\ \widehat{\alpha} &= \bar{y} - \widehat{\beta}\bar{x} = 3.86211 - 0.092658598 \times 53.6875 = -1.112498465. \end{aligned}$$

The estimated regression line is
Square root of stopping dist. = -1.112498465 + 0.092658598 $\times$ Velocity.
From the plot, the model fit has improved and the relationship between the square root of stopping distance $\sqrt{Y}$ and the velocity $X$ is closer to linear.




(c) When the velocity at the time of braking is 50km/hr, the stopping distance in m is

$$y_0 = (\hat{\alpha} + \hat{\beta}x_0)^2 = (-1.112498465 + 0.092658598 \times 50)^2 = 12.39343749.$$

2. We have

$$\sum_{i=1}^{10} x_i = 1.5, \ \sum_{i=1}^{10} y_i = 76, \ \sum_{i=1}^{10} x_i^2 = 0.328, \ \sum_{i=1}^{10} y_i^2 = 708, \ \sum_{i=1}^{10} x_i y_i = 14.74.$$

(a) We have $\overline{y} = \dfrac{\sum_i y_i}{n} = \dfrac{76}{10} = 7.6$ and $\overline{x} = \dfrac{\sum_i x_i}{n} = \dfrac{1.5}{10} = 0.15$. Hence

$$\begin{aligned} S_{yy} &= \sum_i y_i^2 - n\overline{y}^2 = 708 - 10 \cdot 7.6^2 = 130.4. \\ S_{xy} &= \sum_i x_i y_i - n\overline{xy} = 14.74 - 10 \cdot 0.15 \cdot 7.6 = 3.34. \\ S_{xx} &= \sum_i x_i^2 - n\overline{x}^2 = 0.328 - 10 \cdot 0.15^2 = 0.0958. \end{aligned}$$

Then

$$\begin{aligned} \hat{\beta} &= \frac{S_{xy}}{S_{xx}} = \frac{3.34}{0.0958} = 34.8643, \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} = 7.6 - 13.95324 \cdot 0.15 = 2.370355 \end{aligned}$$

The regression line is

$$\text{Number of accident} = 2.370355 + 34.8643 \times \text{precipitation in inches.}$$

The slope is positive indicating that increase in precipitation (in inches) increases the number of accidents.

(b) The standard error estimate for the residuals is

$$\begin{aligned} SSR &= S_{yy} - \hat{\beta} S_{xy} = 130.4 - 34.8643 \cdot 3.34 = 13.95324, \\ s^2 &= \frac{SSR}{n-2} = \frac{13.95324}{10-2} = 1.744154, \\ s &= \sqrt{1.744154} = 1.320664. \end{aligned}$$

The s.e. estimate for the residuals is 1.320664.

(c) The 95% confidence interval for $\alpha$ and $\beta$ are respectively

$$\begin{aligned} \text{CI for } \beta &= \left(\hat{\beta} - t_{n-2,0.975}\frac{s}{\sqrt{S_{xx}}},\ \hat{\beta} + t_{n-2,0.975}\frac{s}{\sqrt{S_{xx}}}\right) \\ &= \left(34.8643 - 2.306 \cdot \frac{1.321}{\sqrt{0.0958}},\ 34.8643 + 2.306 \cdot \frac{1.321}{\sqrt{0.0958}}\right) \\ &= (25.0249,\ 44.7037) \\ \text{CI for } \alpha &= \left(\hat{\alpha} - t_{n-2,0.975}\ s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}},\ \hat{\alpha} + t_{n-2,0.975}\ s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\right) \\ &= \left(2.3704 - 2.306 \cdot 1.321\sqrt{\frac{1}{10} + \frac{0.15^2}{0.0958}},\ 2.3704 + 2.306 \cdot 1.321\sqrt{\frac{1}{10} + \frac{0.15^2}{0.0958}}\right) \\ &= (0.6080,\ 4.1327) \end{aligned}$$

As the CI for $\beta$ does not contain 0, $\beta$ is significantly greater than 0.

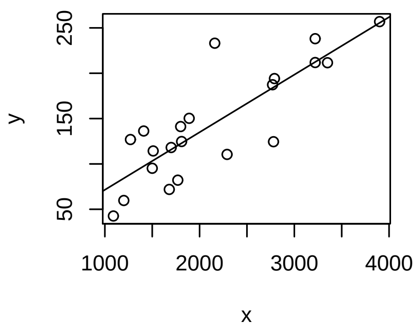
3. $SST = \sum_i n_i \bar{y}_i^2 - N\bar{y}^2 = 4(11.25^2 + 17^2 + 23.75^2) - 3605.33 = 313.17$,
$SSR = \sum_i (n_i - 1)s_i^2 = 3(20.2 + 46 + 20.25) = 259.35$,
$F = \frac{SST/(g-1)}{SSR/(N-g)} = \frac{313.17/2}{259.35/9} = 5.4338$,
$0.025 < \Pr(F_{2,9} \geq 5.4338) < 0.05$

## Extra problems

1. We have

$$\sum_{i=1}^{21} x_i = 45,110, \ \sum_{i=1}^{21} y_i = 3031.2, \ \sum_{i=1}^{21} x_i^2 = 109,957,100, \ \sum_{i=1}^{21} y_i^2 = 513,248.16, \ \sum_{i=1}^{21} x_i y_i = 7,340,085.$$

(a) The plot is




The plot shows that a straight line relationship between $X$ and $Y$ is appropriate.

(b) Hence $\overline{y} = \frac{3,031.2}{21} = 144.3428571$ and $\overline{x} = \frac{45,110}{21} = 2148.095238$.

$$\begin{aligned} S_{yy} &= \sum_i y_i^2 - n\overline{y}^2 = 513,248.16 - 21 \times 144.3428571^2 = 75,716.09169 \\ S_{xy} &= \sum_i x_i Y_i - n\overline{x}\overline{Y} = 7,340,085 - 21 \times 2,148.095238 \times 144.3428571 = 828,778.7165 \\ S_{xx} &= \sum_i x_i^2 - n\overline{x}^2 = 109,957,100 - 21 \times 2,148.095238^2 = 13,056,523.82. \end{aligned}$$

$$\begin{aligned} \widehat{\beta} &= \frac{S_{xy}}{S_{xx}} = \frac{828,778.7165}{13,056,523.82} = 0.063476215 \\ \widehat{\alpha} &= \bar{y} - \widehat{\beta}_1 \bar{x} = 144.3428571 - 0.063476215 \times 2,148.095238 = 7.989901933. \end{aligned}$$

(c)

$$\begin{aligned} SSR &= S_{yy} - \widehat{\beta}_1 S_{xy} = 75,716.09169 - 0.063476215 \times 828,778.7165 = 23,108.35569. \\ s^2 &= \frac{SSR}{n-2} = \frac{23,108.35569}{21-2} = 1216.229247. \\ s &= \sqrt{1,216.229247} = 34.87447845. \end{aligned}$$

The $\alpha = 0.05$ level of significance test for $\widehat{\beta}$ is

1. **Hypothesis:** $H_0$: $\beta = 0$ vs $H_1$: $\beta \neq 0$
2. **Test statistic:** $t_0 = \dfrac{\hat{\beta} - \beta}{s/\sqrt{S_{xx}}} = \dfrac{0.063476 - 0}{34.8745/\sqrt{13,056,523.82}} = 6.577$
3. **Assumption:** $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$ and $Y_i$ are independent.
4. **P-value:** $p$-value $= 2\Pr(t_{19} \geq 6.577) < 0.01$ (0.000)
5. **Conclusion:** Reject $H_0$ and conclude that $\beta$ is greater than 0. The mortality rate due to coronary heart disease depends linearly on the annual cigarette consumption in a way that higher cigarette consumption leads to higher mortality rate.

(d) The 95% CI for $\alpha$:

$$\begin{aligned}
&\left(\hat{\alpha} - t_{\alpha/2,n-2}\ s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}},\ \hat{\alpha} + t_{\alpha/2,n-2}\ s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\right) \\
= &\left(7.9899 - 2.093 \times 34.8745\sqrt{\frac{1}{21} + \frac{2,148.095^2}{13,056,523.82}},\ 7.9899 + 2.093 \times 34.8745\sqrt{\frac{1}{21} + \frac{2,148.095^2}{13,056,523.82}}\right) \\
= &(7.9899 - 46.2237,\ 7.9899 + 46.2237) = (-38.2344, 54.2142). \\
&\text{since } SE(\hat{\alpha}) = 34.8745\sqrt{\frac{1}{21} + \frac{2148.095^2}{13,056,523.82}} = 22.08492506
\end{aligned}$$