THE UNIVERSITY OF SYDNEY

STAT2012 STATISTICAL TESTS

Semester 2 **Solution to Tutorial Week 11** 2015

Tutorial questions

1. Regression analysis: From last week, we have

$$\begin{aligned} S_{yy} &= 130.4, \quad S_{xy} = 3.34, \quad S_{xx} = 0.0958, \quad \overline{y} = 7.6, \quad \overline{x} = 0.15 \\ \hat{\beta} &= 34.8643, \quad \hat{\alpha} = 2.370355, \quad SSR = 13.95324, \quad s^2 = 1.744154, \quad s = 1.3207 \end{aligned}$$

(a) The standard error estimate for the residuals is

$$\begin{aligned} SSR &= SST_o - SST = S_{yy} - \hat{\beta}S_{xy} = 130.4 - 34.8643 \cdot 3.34 = 130.4 - 116.4468 = 13.95324, \\ s^2 &= \frac{SSR}{n-2} = \frac{13.95324}{10-2} = 1.744154, \\ s &= \sqrt{1.744154} = 1.320664. \end{aligned}$$

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

(b) The 95% confidence interval for β is

$$\begin{aligned} \text{CI for } \beta &= \left(\hat{\beta} - t_{n-2,0.975}\frac{s}{\sqrt{S_{xx}}},\ \hat{\beta} + t_{n-2,0.975}\frac{s}{\sqrt{S_{xx}}}\right) \\ &= \left(34.8643 - 2.306 \cdot \frac{1.321}{\sqrt{0.0958}},\ 34.8643 + 2.306 \cdot \frac{1.321}{\sqrt{0.0958}}\right) \\ &= (25.0249,\ 44.7037) \end{aligned}$$

As the CI for β does not contain 0, β is significantly greater than 0.

(c) The test for the significance of the regression model is

1. **Hypothesis:** H_0: $\beta = 0$ vs H_1: $\beta \neq 0$

2. **Test statistic:** $t_0 = \dfrac{\hat{\beta}}{s/\sqrt{S_{xx}}} = \dfrac{34.8643}{1.3207/\sqrt{0.0958}} = 8.1709$

   or $f_0 = t_0^2 = 8.1709^2 \overset{\text{or}}{=} \dfrac{SST}{s^2} = \dfrac{116.4468}{1.74415} = 66.7636$

3. **Assumption:** $Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$ and Y_i are independent.

4. **P-value:** p-value $= 2\Pr(t_8 > 8.1709)) < 0.002$
   ($t_{8,0.999} = 4.501$; 0.0000 from R)

   or p-value $= \Pr(F_{1,8} > 66.7636)) < 0.002$ ($F_{1,8,0.999} = 25.4$).

5. **Conclusion:** Reject H_0 and conclude that the regression model is significant at $\alpha = 0.05$.

(d) We have

$$\begin{aligned} SST &= \hat{\beta} S_{xy} = 34.8643 \times 3.34 = 116.4468, \\ r^2 &= \frac{SST}{SST_o} = \frac{116.4468}{130.4} = 0.892997, \\ r &= \sqrt{0.892997} = 0.944985 \end{aligned}$$

As 89.3% of variation in Y is explained by the model, the model fit is good.

(e) The estimate of the average number of accidents when the amount of precipitation in inches is $x = 0.11$:

$$\begin{aligned} \widehat{y}|\, x_0 = 0.11 &= \hat{\alpha} + \hat{\beta}\, x_0 = 2.370355 + 34.8643 \cdot 0.11 = 6.2054. \\ \text{s.e.}(\widehat{y}|x = 0.11) &= s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} = 1.320664 \sqrt{\frac{1}{10} + \frac{(0.11 - 0.15)^2}{0.0958}} = 0.45116. \end{aligned}$$

The 95% Estimation Interval for the mean number of accident when the amount of precipitation in inches is $x = 0.11$:

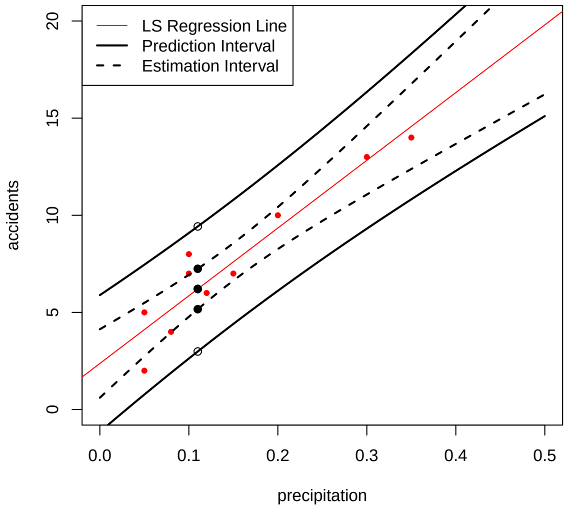
$$\begin{aligned} &\left[(\hat{\alpha} + \hat{\beta} x_0) - t_{\alpha/2,n-2}\; s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}, (\hat{\alpha} + \hat{\beta} x_0) + t_{\alpha/2,n-2}\; s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right] \\ = \; &(6.2054 \; - \; 2.3060 \times 0.45116, \; 6.2054 \; + \; 2.3060 \times 0.45116) = (5.16505, \; 7.245806). \end{aligned}$$

(f) The predicted number of accidents on a day when the amount of precipitation in inches is $x_0 = 0.11$ is the same as (c). The s.e. for the predicted number is:

$$\text{s.e.}(\widehat{y}|x = 16) = s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} = 1.320664 \sqrt{1 + \frac{1}{10} + \frac{(0.11 - 0.15)^2}{0.0958}} = 1.3956.$$

The 95% Prediction Interval for the number of accidend when the amount of precipitation in inches is $x_0 = 0.11$:

$$\begin{aligned} &\left[(\hat{\alpha} + \hat{\beta} x_0) - t_{\alpha/2,n-2}\; s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}, (\hat{\alpha} + \hat{\beta} x_0) + t_{\alpha/2,n-2}\; s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right] \\ = \; &(6.2054 \; - \; 2.3060 \times 1.3956, \; 6.2054 \; + \; 2.3060 \times 1.3956) = (2.987168, \; 9.423688). \end{aligned}$$



(g) No since the regression model applies well only within the range of x, that is from 0.05 to 0.20 inches. Since $x_0 = 2$ inches for the predicted point lies well outside the range of x, the regression model may be invalid. There may be other functional form of x or other factors x_k which will affect y, the number of automobile accidents.

(h) No since the conditions for the relationship between x and y change in other cities. Note that the regression model is quite 'local' in the sense that it can only be applied to situations which are similar to the situations under which the study is conducted.

Extra problems

1. We have

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

$$\overline{y} = 144.3429, \ \overline{x} = 2148.0953, \ S_{yy} = 75,716.09, \ S_{xy} = 828,778.72, \ S_{xx} = 13,056,523.82,$$

$$\widehat{\beta} = 0.063476, \ \widehat{\alpha} = 7.9899.$$

(a) The estimate s^2 of the variance for residual, σ^2, is

$$\begin{aligned} SSR &= S_{yy} - \widehat{\beta}_1 S_{xy} = 75,716.09169 - 0.063476215 \times 828,778.7165 = 23,108.36. \\ s^2 &= \frac{SSR}{n-2} = \frac{23,108.36}{21-2} = 1216.229. \\ s &= \sqrt{1216.229} = 34.87448. \end{aligned}$$

The test for the significance of $\widehat{\beta}$ is

1. **Hypothesis:** H_0: $\beta = 0$ vs H_1: $\beta \neq 0$
2. **Test statistic:** $t_0 = \dfrac{\hat{\beta} - \beta}{s/\sqrt{S_{xx}}} = \dfrac{0.063476 - 0}{34.8745/\sqrt{13,056,523.82}} = 6.5768$
3. **Assumption:** $Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$ and Y_i are independent.
4. **P-value:** p-value $= 2\Pr(t_{19} \geq 6.5768) < 0.01$ ($t_{19,0.999} = 3.579$, 0.000 from R)
5. **Conclusion:** Reject H_0 and conclude that β 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.

(b) The predicted mortality rate for a country with the annual cigarette consumption $x_0 = 3000$ per adult per year is

$$\widehat{Y}|X = 3,000 \ = \widehat{\beta}_0 + \widehat{\beta}_1 \times 3,000 \ = \ 7.9899 + 0.0635 \ \times 3,000 \ = 198.4185.$$

The 95% Prediction Interval for the mortality rate if the annual cigarette consumption is $x_0 = 3000$ per adult per year is

$$\begin{aligned} & \left[(\hat{\beta}_0 + \hat{\beta}_1 x_0) - t_{\alpha/2,n-2} \ s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}, \ (\hat{\beta}_0 + \hat{\beta}_1 x_0) + t_{\alpha/2,n-2} \ s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right] \\ = & \left[(7.9899 + 0.0635 \cdot 3000) - 2.093 \times 34.8745\sqrt{1 + \frac{1}{21} + \frac{(3000 - 2148.095)^2}{13056523.82}}, \right. \\ & \left. (7.9899 + 0.0635 \cdot 3000) + 2.093 \times 34.8745\sqrt{1 + \frac{1}{21} + \frac{(3000 - 2148.095)^2}{13056523.82}} \right] \\ = & \ (198.4185 - 76.6664, \ 198.4185 + 76.6664) = (121.7513, \ 275.0858). \end{aligned}$$

(c) No. As the reference periods are not identical, it is not fair to assume that the mortality rates would follow the same relationship. The mortality rate in 1999 may have substantially reduced due to the advanced technology and the advancement in medical treatment.

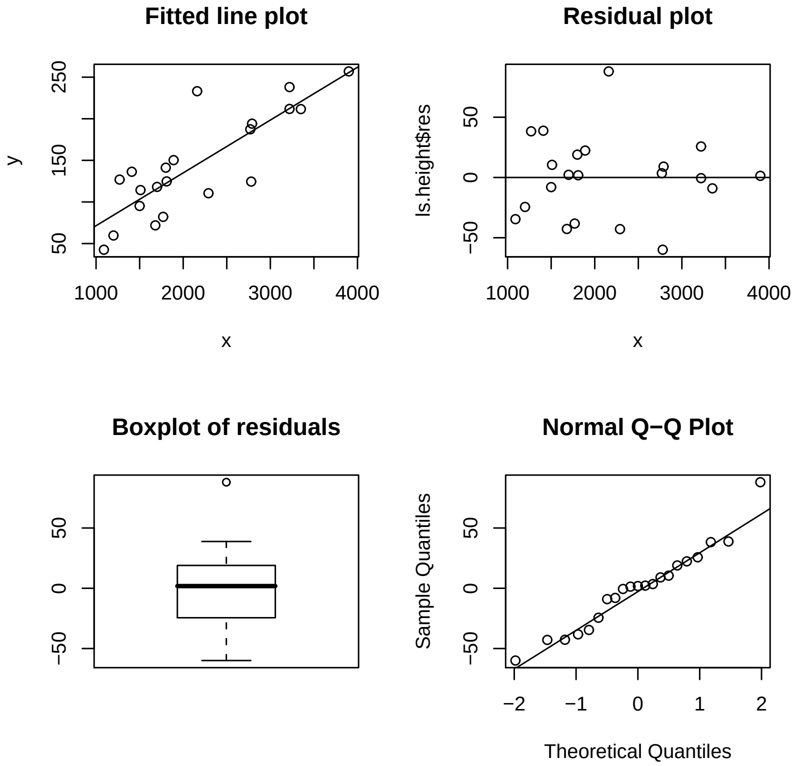
(d) The correlation coefficient is

$$\begin{aligned} r^2 &= \frac{SST}{s^2} = \frac{\beta S_{xy}}{s^2} = \frac{0.0635(828,778.72)}{1216.229} = 0.6948026 \\ r &= \sqrt{0.6948026} = 0.8335482 \end{aligned}$$

The test for the significance of the correlation ρ between the annual cigarette consumption x and its mortality rate due to coronary heart disease y (per adult per year) is

1. **Hypothesis:** $H_0 : \rho = 0 \quad \text{vs} \quad H_1 : \rho \neq 0.$
2. **Test statistic:** $t_0 = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \dfrac{0.8335482\sqrt{21-2}}{\sqrt{1-0.8335482^2}} = 6.576837.$
3. **Assumption:** The data are taken from a bivariate normal population.
4. ***P*-value:** p-value $= 2\ \Pr(t_{19} \geq 6.576837) < 0.001$ ($t_{19,0.999} = 4.501$; 0.0000 from R)
5. **Decision:** Since the p-value < 0.05, we reject H_0. There are strong evidence in the data that the mortality rate and the annual cigarette consumptionthe are dependent.

(e) Yes since $r^2 = 0.6948$ is reasonably high and the slope coefficient β is tested to be significant. Besides, there is no particular pattern in the residual plot and the normality plot resembles a straight line.


Fitted line plot
y
x
Residual plot
ls.height$res
x
Boxplot of residuals
Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles