Tutorial 5

1. Consider a multiple linear regression (MLR) model with one response (Y) and two predictors (X_1 and X_2), $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. Suppose that there are n observations, i.e. we want to analyze the following table of numbers,

Y	X_1	X_2
y_1	x_{11}	x_{12}
y_2	x_{21}	x_{22}
$\vdots$	$\vdots$	$\vdots$
y_n	x_{n1}	x_{n2}

   (i) Write out the matrix form of the above MLR model, including write out the matrices, $\mathbf{Y}$, $\mathbf{X}$ and ϵ, and the corresponding assumptions.

   Solution Refer to the solutions to Question 1 in Tutorial 4.

   (ii) Derive the distribution of $\mathbf{c}'\widehat{\beta}$, where $\mathbf{c} = (c_0, c_1, c_2)'$ and $\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is the least squares estimator.

   Solution Since $\widehat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$ we have

$$\mathbf{c}'\widehat{\beta} \sim N(\mathbf{c}'\beta, \sigma^2\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}),$$

   where we use the property of multivariate normal distribution, given in Chapter 3.

   If we are further given that $n = 25$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 2.779 & -0.0112 & -0.106 \\ -0.0112 & 0.146 \times 10^{-3} & 0.175 \times 10^{-3} \\ -0.106 & 0.175 \times 10^{-3} & 0.479 \times 10^{-2} \end{pmatrix}, \quad \mathbf{X}'\mathbf{Y} = \begin{pmatrix} 235.6 \\ 11821.432 \\ 4831.86 \end{pmatrix}$$

   and $s^2 = 0.4377$.

(iv) Calculate the least squares estimator $\widehat{\beta}$.

Solution It is straightforward to check that

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (10.1552, -0.0672, 0.2398)'.$$

(v) Test for the hypothesis that $\beta_1 + \beta_2 = 0$ at the significant level of 0.05. (Some quantiles of t-distribution: $t_{25}^{0.05/2} = 2.060$, $t_{24}^{0.05/2} = 2.064$ and $t_{22}^{0.05/2} = 2.074$)

Solution The hypothesis that $\beta_1 + \beta_2 = 0$ can be rewritten as $\mathbf{c}'\beta$ with $\mathbf{c}' = (0, 1, 1)$ and $\beta' = (\beta_0, \beta_1, \beta_2)$. Therefore we may use $\mathbf{c}'\widehat{\beta}$ to estimate $\mathbf{c}'\beta$. From (ii) we know the distribution of $\mathbf{c}'\widehat{\beta}$. However since σ^2 is unknown we use s^2 to estimate σ^2. As a result,

$$\frac{\mathbf{c}'\widehat{\beta} - \mathbf{c}'\beta}{s\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}} \sim t_{n-p-1}.$$

Given the values of s^2 and $\mathbf{c}$ one may obtain

$$s\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}} = \sqrt{0.4377 \times 0.0053} = 0.0482, \quad \mathbf{c}'\widehat{\beta} = 0.1725.$$

This shows that

$$\frac{\mathbf{c}'\widehat{\beta}}{s\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}} = 3.5798 > t_{22}^{0.05/2} = 2.074.$$

We therefore reject the null hypothesis.

2. Many people believe that the length of one's life is linearly related to the length of lifeline on one's left hand. Dr. L.E. Mather and Dr. M.E. Wilson conducted a experiment to study this belief in 1974. In their experiment, two variables were involved,

Y = Age of person at death (to nearest year) and
X = Length of lifeline on left hand in centimeters (to nearest 0.15cm),

and 50 pairs of observations were collected. The first four observations are listed as below.

Case	Y = Age (Year)	X = Length (cm)
1	19	9.75
2	40	9.00
3	42	9.60
4	42	9.75
⋮	⋮	⋮
50	94	9.00

The summary of the data is as follows.

$$\sum y = 3333, \quad \sum y^2 = 231933, \quad \sum xy = 30549.75,$$

$$\sum x = 459.9, \ \sum x^2 = 4308.57.$$

Consider a simple linear regression (SLR) model to fit the above data.

(i) Write out the fitted SLR model, including the estimation of β_0, β_1 and σ^2.

Solution From the summary of the data we have

$$S_{yy} = \sum y^2 - n(\bar{y})^2 = 9755.2, \quad S_{xy} = \sum xy - n\bar{x}\bar{y} = -107.184.$$

$$S_{xx} = \sum x^2 - n(\bar{x})^2 = 78.4098, \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = -1.366972. \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 79.23340833,$$

$$SSR = \hat{\beta}_1^2 S_{xx} = 146.5175, \quad SSE = S_{yy} - SSR = 9608.683,$$

and

$$s^2 = \frac{SSE}{n-2} = \frac{9608.683}{48} = 200.1809.$$

It follows that the fitted SLR model is

$$\hat{y}_i = 79.23340833 - 1.366972 x_i.$$

(ii) Use t test to check whether or not the length of lifeline can affect the lifetime at a size of 5%.

Solution The t statistic is

$$|\frac{\hat{\beta}_1}{\sqrt{s^2/S_{xx}}}| = |\frac{-1.366972}{\sqrt{200.1809/78.4098}}| = |-0.8555264| < t_{48}^{0.025} = 2.0106.$$

So these data do not confirm the idea that the length of life is related to the length of lifeline.

(iii) Construct the ANOVA table, and check whether the fitted SLR model is significant.

Solution

Source	df	SS	MS	F	p-value
Regression	1	$SSR = 146.5175$	146.5175	$146.5175/200.1809 = 0.7319255$	
Residual	48	$SSE = 9608.683$	200.1809		
Total	49	$S_{yy} = 9755.2$			

Note that $0.731 < F_{1.48}^{0.05} = 4.05$. So these data do not confirm the idea that the length of life is related to the length of lifeline, which is consistent with the conclusion made by the t statistic. the fitted SLR model is not significant.

(iv) Calculate the value of R^2. Give some comments.

$$R^2 = \frac{SSR}{S_{yy}} = \frac{146.5175}{9755.2} = 1.5\%,$$

which shows that it is not a good fit.

(v) The length of lifeline on my left hand is 9.32 cm. Tell me the mean lifetime of persons with this length of lifeline. Given me an interval with 95% confidence.

Solution The predicted value of the mean lifetime of persons with such a length is

$$\hat{y}_0 = 79.23340833 - 1.366972 x_0 = 79.23340833 - 1.366972 \times 9.32 = 66.49323.$$

The standard error of $\hat{y}_0$ is

$$s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} t_{n-2}^{\frac{\alpha}{2}} = \sqrt{200.1809(\frac{1}{50} + \frac{(9.32 - 9.198)^2}{78.4098})} \times 2.0106 = 4.042065.$$

It follows that the prediction interval is

$$[66.49323 - 4.042065, 66.49323 + 4.042065] = [62.45116, 70.53529].$$