T6

1. The relationship between the concentration of ozone in New York and other environmental factors is under investigation. The concentration Y, the solar radiation X_1 and the temperature in New York X_2 are collected for 30 consecutive days.

   Consider a multiple linear regression model. Let $\mathbf{Y}$ be the vector of the responses and $\mathbf{X}$ be the design matrix. The following matrices are given for the regression analysis.

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 30 & 5733 & 2085 \\ 5733 & 1463179 & 411076 \\ 2085 & 411076 & 147243 \end{pmatrix} \quad \mathbf{X}'\mathbf{Y} = \begin{pmatrix} 83.84 \\ 17097.38 \\ 5953.47 \end{pmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 2.2157 & 6.176 \times 10^{-4} & -3.310 \times 10^{-2} \\ 6.176 \times 10^{-4} & 3.341 \times 10^{-6} & -1.807 \times 10^{-5} \\ -3.310 \times 10^{-2} & -1.807 \times 10^{-5} & 5.259 \times 10^{-4} \end{pmatrix}$$

   $\mathbf{Y}'\mathbf{Y} = 251.0472$

   (i) Find the least squares estimators, $\widehat{\beta}_0, \widehat{\beta}_1$ and $\widehat{\beta}_2$, for the regression coefficients.

   (ii) Is the overall regression significant at a 5% level of significance ?

   Solution (i)

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (-0.736, 0.0013, 0.0468).$$

   (ii) Note that $\sum_i y_i$ is the first component of $\mathbf{X}'\mathbf{Y}$. It follows that $\bar{y} = 83.84/30 = 2.795$ and

$$S_{yy} = \mathbf{Y}'\mathbf{Y} - n(\bar{y})^2 = 251.0472 - 234.304 = 16.7423.$$

   Also we have

$$SSE = \mathbf{Y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} = 251.0472 - 239.966 = 11.1. \quad s = \sqrt{\frac{11.1}{30-2-1}} = 0.64.$$

   This implies that

$$SSR = S_{yy} - SSE = 16.7423 - 11.1 = 5.66.$$

It follows that the F statistic

$$F = \frac{SSR/p}{SSE/(n-p-1)} = 6.899 > F_{2,27}^{0.95} = 3.35,$$

which suggests that the model is significant.

Consider a reduced regression model without the predictor X_1, $y_i = \beta_0 + \beta_1 x_{2i} + \varepsilon_i$. Denote the design matrix in the reduced model by

$$(\mathbf{X}_r'\mathbf{X}_r)^{-1} = \begin{pmatrix} 2.1015 & -0.029756 \\ -0.029756 & 0.000428 \end{pmatrix}$$

(iii) Find the least squares estimators $\widehat{\beta}_0, \widehat{\beta}_1$ in the reduced model.

(iv) Find the standard error for each estimator. Is the overall regression significant at a 5% level of significance ?

(v) Use an F test to check the significance of the reduced model at a 5% level of significance ?

Solution (iii) Note that

$$\mathbf{X}_r'\mathbf{Y} = \begin{pmatrix} 83.84 \\ 5953.47 \end{pmatrix}.$$

We obtain

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'\mathbf{Y} = \begin{pmatrix} -0.96 \\ 0.053 \end{pmatrix}$$

(iv)

$$SSE = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}_r(\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'\mathbf{Y} = 251.0472 - 236.9423 = 14.1049.$$

This implies

$$s^2 = \frac{SSE}{n-p-1} = \frac{236.9423}{28} = 0.5037.$$

In view of this we have

$$s.e.(\hat{\beta}_0) = \sqrt{s^2 \times 2.1015} = 1.0288, \quad s.e.(\hat{\beta}_1) = \sqrt{s^2 \times 0.000428} = 1.468 \times 10^{-2}.$$

(v) One should note that S_{yy} in the reduced model and full model are the same. ANOVA table:

Source	df	SS	MS	F	p-value
Regression fitting ω	1	16.74-14.1049=2.6351	2.6351		
Extra	1	14.1049-11.1=3.0049	3.0049	3.0049/0.41=7.319	
Residual	27	11.1	0.41		
Total	29	16.7423			

Note that $F = 7.319 > F_{1,27}^{0.05} = 4.21$. This suggests that using the reduced model is not enough.

2. Consider a multiple regression model with two predictors $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$ and its reduced model $y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$,. Let SSE_ω be the residual sum of square of the reduced model and SSE_Ω the residual sum of square of the full model. Show that $SSE_\Omega \leq SSE_\omega$.

   Solution : Denote the least square estimators in the full model by $\hat{\beta}_0^\Omega, \hat{\beta}_1^\Omega, \hat{\beta}_2^\Omega$ and in the reduced model by $\hat{\beta}_0^\omega, \hat{\beta}_1^\omega$. Note that $\hat{\beta}_0^\Omega, \hat{\beta}_1^\Omega, \hat{\beta}_2^\Omega$ are obtained by minimizing $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2$. Therefore

$$SSE_\Omega = \sum_{i=1}^{n}(y_i - \hat{\beta}_0^\Omega - \hat{\beta}_1^\Omega x_{i1} - \hat{\beta}_2^\Omega x_{i2})^2 \leq \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2$$

   for any $\beta_0, \beta_1, \beta_2$. Now choose $\beta_0 = \hat{\beta}_0^\omega$, $\beta_1 = \hat{\beta}_1^\omega$, $\beta_2 = 0$ in the last sum. It follows that

$$SSE_\Omega = \sum_{i=1}^{n}(y_i - \hat{\beta}_0^\Omega - \hat{\beta}_1^\Omega x_{i1} - \hat{\beta}_2^\Omega x_{i2})^2 \leq \sum_{i=1}^{n}(y_i - \hat{\beta}_0^\omega - \hat{\beta}_1^\omega x_{i1} - 0x_{i2})^2 = SSE_\omega.$$