Use this page for extra answer or working space

Answer to Q1:

(a) We assume that

$$\mathbb{E}_{Y_i|X_i}[Y_i|\mathbf{x}_i] = \beta_0 + \beta_1 x_{i1} \qquad \text{Var}_{Y_i|X_i}[Y_i|\mathbf{x}_i] = \sigma^2$$

with $Y_1, \ldots, Y_n$ presumed either uncorrelated or independent. Thus in vector form

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}|\mathbf{X}] = \mathbf{X}\beta \qquad \text{Var}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}|\mathbf{X}] = \sigma^2\mathbf{I}_n$$

where $\mathbf{X}$ is the $(n \times 2)$ design matrix, with first column $\mathbf{1}_n$, and second column $\underset{\sim}{x}_1$, $\beta = (\beta_0, \beta_1)^\top$, and $\mathbf{I}_n$ is the $(n \times n)$ identity matrix. 4 Marks

(b) The hat matrix, $\mathbf{H}$, is the matrix that linearly transforms the observed data into the fitted values. We have $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\beta} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \mathbf{H}\mathbf{y}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$.

2 Marks

Then

$$\sum_{i=1}^n (y_i - \widehat{y}_i)^2 = (\mathbf{y} - \widehat{\mathbf{y}})^\top(\mathbf{y} - \widehat{\mathbf{y}}) = \{(\mathbf{I}_n - \mathbf{H})\mathbf{y}\}^\top(\mathbf{I}_n - \mathbf{H})\mathbf{y} = \mathbf{y}^\top(\mathbf{I}_n - \mathbf{H})^\top(\mathbf{I}_n - \mathbf{H})\mathbf{y}$$

However, we verify by direct calculation that $\mathbf{H}^\top\mathbf{H} = \mathbf{H}$, and so

$$(\mathbf{I}_n - \mathbf{H})^\top(\mathbf{I}_n - \mathbf{H}) = (\mathbf{I}_n - \mathbf{H})$$

and the result follows.

4 Marks

(c) We have $\text{SS}_\text{T} = \text{SS}_\text{Res} + \text{SS}_\text{R}$, as

$$\begin{aligned}
\sum_{i=1}^n (y_i - \overline{y})^2 &= \sum_{i=1}^n (y_i - \widehat{y}_i + \widehat{y}_i - \overline{y})^2 \\
&= \sum_{i=1}^n (y_i - \widehat{y}_i)^2 + \sum_{i=1}^n (\widehat{y}_i - \overline{y})^2 + 2\sum_{i=1}^n (y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}) \\
&= \sum_{i=1}^n (y_i - \widehat{y}_i)^2 + \sum_{i=1}^n (\widehat{y}_i - \overline{y})^2
\end{aligned}$$

as the third term is zero. Now, we use the result that the left hand side can be written

$$\sum_{i=1}^n (y_i - \overline{y})^2 = (\mathbf{y} - \mathbf{1}_n\overline{y})^\top(\mathbf{y} - \mathbf{1}_n\overline{y})$$

and we may write $n = \mathbf{1}_n^\top\mathbf{1}_n$, so in fact

$$\overline{y} = \frac{1}{n}\sum_{i=1}^n y_i = (\mathbf{1}_n^\top\mathbf{1}_n)^{-1}\mathbf{1}_n^\top\mathbf{y} \quad \Longrightarrow \quad \mathbf{1}_n\overline{y} = \mathbf{1}_n(\mathbf{1}_n^\top\mathbf{1}_n)^{-1}\mathbf{1}_n^\top\mathbf{y} = \mathbf{H}_1\mathbf{y}$$

say. Then, as we may write

$$(\mathbf{I}_n - \mathbf{H}_1) = (\mathbf{I}_n - \mathbf{H}) + (\mathbf{H} - \mathbf{H}_1)$$

and the result follows.

4 Marks

(d) Here is the full output:

```
 1 > summary(lm(y~x))
 2 Coefficients:
 3             Estimate Std. Error t value Pr(>|t|)
 4 (Intercept)   1.3824     1.5325   0.902   0.3693
 5 x             0.3382     0.1394   2.425   0.0172 *
 6 ---
 7 Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1
 8
 9 Residual standard error: 6.057 on 94 degrees of freedom
10 Multiple R-squared:  0.05889,   Adjusted R-squared:  0.04887
11 F-statistic: 5.882 on 1 and 94 DF,  p-value: 0.01721
```

From the output identify

(i) the sample size n; $n = 96$ (from line 9). 1 Mark

(ii) the conclusion of the test of the null hypothesis $\mathrm{H}_0 : \beta_0 = 0$ omitted from line 4: test statistic is $1.3824/1.5325 = 0.902$, which does not exceed the critical value, so the hypothesis is not rejected. 1 Mark

(iii) the `Estimate` omitted from line 5; estimate is $0.3382 = 0.1394 \times 2.425$. 1 Mark

(iv) whether `x` is a useful predictor of y; the p-value from line 5 indicates that there is mild evidence of a significant contribution, however, the R^2 statistic is very low. 1 Mark

(v) the three terms in the sums of squares decomposition

$$\mathrm{SS_T} = \mathrm{SS_{Res}} + \mathrm{SS_R} :$$

We can compute

$$\mathrm{SS_{Res}} = (n-2)\hat{\sigma}^2 = 94 \times 6.057^2 = 3448.601.$$

Then,

$$\mathrm{SS_T} = \frac{\mathrm{SS_{Res}}}{1 - R^2} = \frac{3448.601}{1 - 0.0589} = 3664.436$$

so therefore

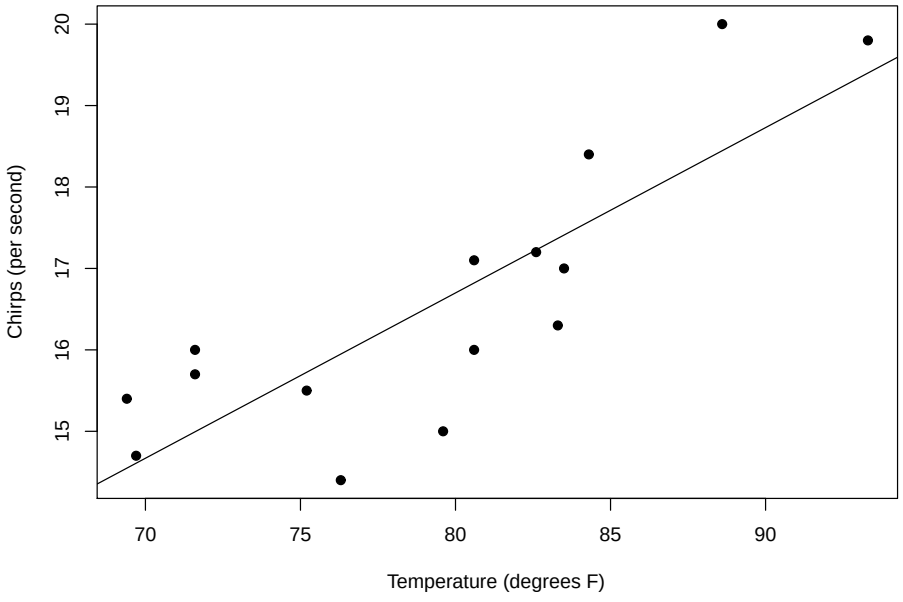
$$\mathrm{SS_R} = \mathrm{SS_T} - \mathrm{SS_{Res}} = 3664.436 - 3448.601 = 215.835.$$

2 Marks

Q2 The following data, originally published by G. W. Pierce in 1948, show the relationship between chirps per second (y) of a striped ground cricket and the ground temperature measured in degrees Fahrenheit (x). Fifteen independent measurements were made, and analysis of these data in `R` is presented below (some of the output has been omitted):

```
 1 > fit.Chirp<-lm(y~x)
 2 > abline(fit.Chirp)
 3 > summary(fit.Chirp)
 4 Coefficients:
 5             Estimate Std. Error t value Pr(>|t|)
 6 (Intercept)  0.45931    2.98920   0.154 0.880239
 7 x            0.20300    0.03754   XXXXX XXXXXXXX
 8 ---
 9 Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1
10
11 Residual standard error: 0.986 on 13 degrees of freedom
12 Multiple R-squared:  XXXX,    Adjusted R-squared:  XXXX
13 F-statistic: XXXXX on 1 and 13 DF,  p-value: XXXXX
14
15 > SST<-sum((y-mean(y))^2)
16 > SST
17 [1] 41.07333
18 > mean(x)
19 [1] 79.34667
20 > sum((x-mean(x))^2)
21 [1] 690.0173
```

The straight line fit is depicted below



(a) Is temperature a useful predictor of the number of chirps per second ? Justify your answer

4 Marks

(b) Compute the R^2 statistic for these data.

4 Marks

Question continued on the next page.