STATISTICAL MODELLING

PRACTICAL I
SOLUTIONS

I.1 Suppose that Y is a random variable that represents the actual contents of a 1-lb can of coffee. The model proposed for the distribution of Y is the uniform distribution over the interval $\left[15.5, 17.0\right]$

$$f\left(y\right)=\frac{1}{1.5},\quad 15.5\le y\le 17.0$$

a) What is the probability that a can will contain less than 16oz?

The required probability is $\int_{15.5}^{16} f\left(y\right)dy=\int_{15.5}^{16}\frac{1}{1.5}dy=\left.\frac{y}{1.5}\right|_{15.5}^{16}=\frac{0.5}{1.5}=0.3333$.

b) Find the population mean and standard deviation for these cans.

$$E\left[Y\right]=\int_{-\infty}^{\infty} yf\left(y\right)dy=\int_{15.5}^{17} y\frac{1}{1.5}dy=\int_{15.5}^{17}\frac{y}{1.5}dy=\left.\frac{y^2}{3}\right|_{15.5}^{17}=\frac{17^2-15.5^2}{3}=16.25$$

$$\begin{aligned}\mathrm{var}\left[Y\right]&=\int_{-\infty}^{\infty}\left(y-\psi_Y\right)^2 f\left(y\right)dy=\int_{15.5}^{17}\left(y-16.25\right)^2\frac{1}{15}dy=\frac{1}{15}\int_{15.5}^{17}\left(y-16.25\right)^2 dy\\&=\frac{1}{15}\left.\frac{\left(y-16.25\right)^3}{3}\right|_{15.5}^{17}\\&=\frac{1}{15}\left(\frac{\left(17-16.25\right)^3}{3}-\frac{\left(15.5-16.25\right)^3}{3}\right)\\&=\frac{0.421875-\left(-0.421875\right)}{45}\\&=0.01875\end{aligned}$$

I.2 Verify that $\mathbf{V} = E\left[\left(\mathbf{Y} - E[\mathbf{Y}]\right)\left(\mathbf{Y} - E[\mathbf{Y}]\right)'\right]$ is equivalent to the following expression for $\mathbf{V}$ by obtaining an expression for the *ij*th element of $E\left[\left(\mathbf{Y} - E[\mathbf{Y}]\right)\left(\mathbf{Y} - E[\mathbf{Y}]\right)'\right]$.

$$\mathbf{V} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1i} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2i} & \cdots & \sigma_{2n} \\ \vdots & \vdots & & \vdots & & \vdots \\ \sigma_{1i} & \sigma_{2i} & \cdots & \sigma_i^2 & \cdots & \sigma_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_{in} & \cdots & \sigma_n^2 \end{bmatrix}$$

The ijth element of $\mathbf{V} = E\left[\left(\mathbf{Y} - E[\mathbf{Y}]\right)\left(\mathbf{Y} - E[\mathbf{Y}]\right)'\right]$ *is the expectation of the product of the ith and jth elements of* $\left(\mathbf{Y} - E[\mathbf{Y}]\right)$. *The ith element of* $\left(\mathbf{Y} - E[\mathbf{Y}]\right)$ *is* $\left(\mathbf{Y}_i - E[\mathbf{Y}_i]\right)$ *so that the ijth element of* $\mathbf{V} = E\left[\left(\mathbf{Y} - E[\mathbf{Y}]\right)\left(\mathbf{Y} - E[\mathbf{Y}]\right)'\right]$ *is* $E\left[\left(\mathbf{Y}_i - E[\mathbf{Y}_i]\right)\left(\mathbf{Y}_j - E[\mathbf{Y}_j]\right)\right]$. *By definition this is* σ_{ij} *which for* $i = j$ *is* σ_i^2. *The two expressions are equivalent.*

I.3 Prove that $\frac{1}{3}\mathbf{J}_3$ is idempotent, where $\mathbf{J}_3$ is the 3×3 matrix all of whose elements are equal to 1.

On noting that $\mathbf{J}_3\mathbf{J}_3 = 3\mathbf{J}_3$ *we have that* $\frac{1}{3}\mathbf{J}_3\frac{1}{3}\mathbf{J}_3 = \frac{1}{9}\mathbf{J}_3\mathbf{J}_3 = \frac{1}{9} \times 3\mathbf{J}_3 = \frac{1}{3}\mathbf{J}_3$.

I.4 Let x denote the number of years of formal education and let Y denote an individual's income at age 30. Assume that simple linear regression is applicable and consider this data:

Formal education (years)	Income ($000)
8	8
12	15
14	16
16	20
16	25
20	40

a) Write down **y**, **X** and $\boldsymbol{\theta}$ for this data.

$$\mathbf{y} = \begin{bmatrix} 8 \\ 15 \\ 16 \\ 20 \\ 25 \\ 40 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 8 \\ 1 & 12 \\ 1 & 14 \\ 1 & 16 \\ 1 & 16 \\ 1 & 20 \end{bmatrix} \text{ and } \boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$

b) Use the R function `lm` to find $\hat{\boldsymbol{\theta}}$. What is the equation for the estimated expected value?

```
> attach(Incomes.dat)
> Income.lm <- lm(Income ~ Education, Incomes.dat)
> summary(Income.lm)

Call:
lm(formula = Income ~ Education, data = Incomes.dat)

Residuals:
     1      2      3      4      5      6 
 3.344  0.232 -3.824 -4.880  0.120  5.008 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.5680     7.0141  -2.220  0.09066
Education     2.5280     0.4736   5.338  0.00593

Residual standard error: 4.323 on 4 degrees of freedom
Multiple R-Squared: 0.8769,     Adjusted R-squared: 0.8461 
F-statistic: 28.49 on 1 and 4 DF,  p-value: 0.005934
```

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} -15.568 \\ 2.528 \end{bmatrix}$$

The equation for the estimated expected value is $\widehat{E[Y]} = -15.568 + 2.528x$

c) You are given that

$$\mathbf{Q}_M = \begin{bmatrix} 0.648 & 0.344 & 0.192 & 0.040 & 0.040 & -0.264 \\ 0.344 & 0.232 & 0.176 & 0.120 & 0.120 & 0.008 \\ 0.192 & 0.176 & 0.168 & 0.160 & 0.160 & 0.144 \\ 0.040 & 0.120 & 0.160 & 0.200 & 0.200 & 0.280 \\ 0.040 & 0.120 & 0.160 & 0.200 & 0.200 & 0.280 \\ -0.264 & 0.008 & 0.144 & 0.280 & 0.280 & 0.552 \end{bmatrix}$$

Compute the fitted values by calculating $\mathbf{Q}_M\mathbf{y}$.

$$\mathbf{Q}_M\mathbf{y} = \begin{bmatrix} 0.648 & 0.344 & 0.192 & 0.040 & 0.040 & -0.264 \\ 0.344 & 0.232 & 0.176 & 0.120 & 0.120 & 0.008 \\ 0.192 & 0.176 & 0.168 & 0.160 & 0.160 & 0.144 \\ 0.040 & 0.120 & 0.160 & 0.200 & 0.200 & 0.280 \\ 0.040 & 0.120 & 0.160 & 0.200 & 0.200 & 0.280 \\ -0.264 & 0.008 & 0.144 & 0.280 & 0.280 & 0.552 \end{bmatrix} \begin{bmatrix} 8 \\ 15 \\ 16 \\ 20 \\ 25 \\ 40 \end{bmatrix}$$

$$= \begin{bmatrix} 4.656 \\ 14.768 \\ 19.824 \\ 24.880 \\ 24.880 \\ 34.992 \end{bmatrix}$$

Use the equation for the estimated expected value to verify the first fitted value.

$$\widehat{E[Y]} = -15.568 + 2.528 \times 8 = 4.656$$

Use the fitted values to compute the residuals.

$$\hat{\boldsymbol{\varepsilon}} = \begin{bmatrix} 8 \\ 15 \\ 16 \\ 20 \\ 25 \\ 40 \end{bmatrix} - \begin{bmatrix} 4.656 \\ 14.768 \\ 19.824 \\ 24.880 \\ 24.880 \\ 34.992 \end{bmatrix} = \begin{bmatrix} 3.344 \\ 0.232 \\ -3.824 \\ -4.880 \\ 0.120 \\ 5.008 \end{bmatrix}$$

d) Use the R function `aov` to obtain the ANOVA table for testing that the slope is zero given that the intercept is in the model.

```
> Income.aov <- aov(Income ~ Education, Incomes.dat)
> summary(Income.aov)
            Df Sum Sq Mean Sq F value   Pr(>F)
Education    1 532.57  532.57  28.492 0.005934
Residuals    4  74.77   18.69
```

What is the corrected total SSq for this analysis?

The corrected total SSq is $532.5653 + 74.7680 = 607.3333$.

Verify that the Residual SSq is the sum of the squares of the residuals.

The sum of squares of the elements of $\hat{\boldsymbol{\varepsilon}} = \begin{bmatrix} 3.344 \\ 0.232 \\ -3.824 \\ -4.880 \\ 0.120 \\ 5.008 \end{bmatrix}$ *is 74.768.*

e) What model best describes this data?

As the Education term is significant ($p = 0.0059$), the model that includes its coefficient is better than one that does not. So the model for the data would be $E[Y_i] = -15.568 + 2.528x_i$ *with* $E[\varepsilon_i] = 0$, $\mathrm{var}[\varepsilon_i] = \sigma^2$ *and* $\mathrm{cov}[\varepsilon_i, \varepsilon_j] = 0,\ i \neq j$.