Statistical Modelling and Estimation
Exercises 1 Solutions

20 January 2020

1. (a) The biochemical theory is unlikely to be exact, although it might be a good approximation, the setting of the substrate concentration will not be exact, all other variables which are supposed to be held constant will not be exactly constant and the reaction rate will not be measured exactly.

   (b) A reasonable model might be

$$V_i \sim N\left(\frac{V_0[S]_i}{K_m + [S]_i}, \sigma^2\right),$$

   with V_is independent, $i = 1, \ldots, n$.

2. (a) The main point of this question is that you should list not just potential causal variables, such as having previously done an MSc in Statistics and time spent studying, but also variables which might be correlated, but not causally related, e.g. age, place of residence. There is almost no upper limit on the number of variables you could list.

   (b) Having done an MSc and place of residence are qualitative; time spent studying and age are quantitative.

   (c) Integers from 0 to 100.

   (d) A suitable distribution might be Binomial(100,π), but a normal distribution should be a reasonable approximation.

3. It can be written in the usual form with $\boldsymbol{\beta} = [\mu]$ and $\mathbf{X}' = [1\ 1\ \cdots\ 1]$.

4.

$$\begin{aligned}\gamma_0 + \gamma_{11}(x_i - \gamma_1)^2 &= \gamma_0 + \gamma_{11}(x_i^2 - 2\gamma_1 x_i + \gamma_1^2) \\ &= \beta_0 + \beta_1 x_i + \beta_{11} x_i^2,\end{aligned}$$

where $\beta_0 = \gamma_0 + \gamma_1^2\gamma_{11}$, $\beta_1 = -2\gamma_1\gamma_{11}$ and $\beta_{11} = \gamma_{11}$, which is a linear model.

5. (a) This can be written in the usual form, with $\boldsymbol{\beta}' = [\mu\ \tau_2]$ and

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix}.$$

(b) The histogram would be bimodal. The point is that we should not expect a plot of the raw data to look as if they come from a normal distribution.

6. (a) We would expect the variance of the SO_2 concentration to be smaller from the sites which collected rain for 48 hours. It *might* be reasonable to assume that their variance is divided by 2. Hence the model would have $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ and $V(\mathbf{Y}) = \sigma^2\mathbf{G}$, where

$$\mathbf{G} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\mathbf{I} \end{bmatrix}.$$

(b)

$$\mathbf{G}^{1/2} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \frac{1}{\sqrt{2}}\mathbf{I} \end{bmatrix}$$

and so

$$\mathbf{G}^{-1/2} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sqrt{2}\mathbf{I} \end{bmatrix}.$$

Then, as in the lecture notes $E(\mathbf{Z}) = \mathbf{U}\boldsymbol{\beta}$ and $V(\mathbf{Z}) = \sigma^2\mathbf{I}$, where $\mathbf{Z} = \mathbf{G}^{-1/2}\mathbf{Y}$ and $\mathbf{U} = \mathbf{G}^{-1/2}\mathbf{X}$.

7. Model under randomization, for unit j in block i receiving treatment r, is

$$Y_{ij} = \mu + t_r + \sum_k \delta_{i;jk} d_{ik},$$

where the mean of d_{ik} in block i is not 0, but b_i, so we can write

$$Y_{ij} = \mu + t_r + b_i + \sum_k \delta_{i;jk} e_{ik},$$

where e_{ik} have mean 0 and variance σ^2 for each i. This is again a linear model, which can be written

$$Y_{ij} = \mu + t_r + b_i + \epsilon_{ij}.$$

8. The cross-product term is

$$\begin{aligned} CPT &= (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= 2(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= 2(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\{\mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}\} \\ &= 2(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\{\mathbf{X}' - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\}\mathbf{Y} \\ &= 2(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\mathbf{X}' - \mathbf{X}')\mathbf{Y} \\ &= 0 \end{aligned}$$

9. Writing the model, as in lectures, as

$$E(\mathbf{Z}) = \mathbf{U}\boldsymbol{\beta}; \quad V(\mathbf{Z}) = \sigma^2\mathbf{I},$$

where $\mathbf{Z} = \mathbf{G}^{-1/2}\mathbf{Y}$ and $\mathbf{U} = \mathbf{G}^{-1/2}\mathbf{X}$, the normal equations are

$$\begin{aligned} \mathbf{U}'\mathbf{U}\boldsymbol{\beta} &= \mathbf{U}'\mathbf{Z} \\ \Rightarrow \mathbf{X}'\mathbf{G}^{-1/2\prime}\mathbf{G}^{-1/2}\mathbf{X}\boldsymbol{\beta} &= \mathbf{X}'\mathbf{G}^{-1/2\prime}\mathbf{G}^{-1/2}\mathbf{Y} \\ \Rightarrow \mathbf{X}'\mathbf{G}^{-1}\mathbf{X}\boldsymbol{\beta} &= \mathbf{X}'\mathbf{G}^{-1}\mathbf{Y}, \end{aligned}$$

since $\mathbf{G}$ is symmetric.

10. From Exercises 1, $\boldsymbol{\beta} = [\mu]$ $\mathbf{X}' = [1\ 1\ \cdots\ 1]$, so $\mathbf{X}'\mathbf{X} = n$ and $\mathbf{X}'\mathbf{Y} = \sum_{i=1}^{n} Y_i$. Hence the least squares estimator of μ is $\hat{\mu} = \bar{Y}$.