Theory of Linear Models
Exercises 1

13 January 2020

1. Biochemical theory implies that, for simple enzymes, their reaction rate should follow Michaelis-Menten kinetics, i.e.

$$v = \frac{V_0[S]}{K_m + [S]},$$

where v is the initial reaction rate, $[S]$ is the substrate concentration and V_0 and K_m are unknown parameters, which have to be estimated experimentally.

A biochemist carries out an experiment in which she makes four separate runs at each of five different substrate concentrations and records the initial reaction rate. The data obtained do not exactly follow the (deterministic) Michaelis-Menten model.

(a) List as many reasons as you can think of why the data do not exactly follow the model implied by theory.

(b) Suggest a statistical model which might be useful for the data from such an experiment.

2. Consider scores in the LTCC exam as the response variable.

(a) list as many explanatory variables as you can think of which might be related to the response.

(b) State whether each explanatory variable is qualitative or quantitative.

(c) What is the set of possible values for the response variable?

(d) Suggest a suitable distribution for the response variable.

3. Assume $Y_i \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, n$, with all random variables independent. Show how this can be written as a linear model, i.e. specify the matrix $\mathbf{X}$ and the vector $\boldsymbol{\beta}$.

4. Show that the model $E(Y_i) = \gamma_0 + \gamma_{11}(x_i - \gamma_1)^2$; $V(\mathbf{Y}) = \sigma^2\mathbf{I}$ is a linear model.

5. Consider the model $Y_{gi} \sim N(\mu_g, \sigma^2)$, for data from two groups $g = 1, 2$, with $\mu_g = \mu + \tau_g$, $\tau_1 = -\tau_2$ and all random variables independent.

   (a) Write this as a linear model.

   (b) Draw a sketch to show how a histogram of such data would look (for large n).

6. In a study of the short-term effects of pollution, data were collected from n sites on the concentration of SO_2 in rainfall, along with several explanatory variables, $x_1, \ldots, x_q$. However, due to a misunderstanding, some sites collected rainfall for 24 hours and some for 48 hours.

   (a) Suggest a suitable model for these data (consider $V(Y_i)$).

   (b) Show how this can be written as a linear model.

7. Derive the appropriate linear model for a randomized complete block design, i.e. starting from the deterministic model $y_{i(r)} = u_i + t_r$ for t treatments in $n = bt$ experimental units, derive the appropriate stochastic model under randomization, where randomization is carried out as follows:

   - the experimental units are divided into b blocks of t units each;
   - each treatment is assigned to one unit in each block;
   - within each block independently, units are randomized to unit labels.

8. In the notation used in lectures, prove that

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

   i.e. show that the cross-product term is zero.

9. Find the normal equations for the model

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}; \quad V(\mathbf{Y}) = \sigma^2\mathbf{G},$$

   where $\mathbf{G}$ is a known nonsingular matrix.

10. Find the least squares estimator of μ in the model $E(Y_i) = \mu;\ V(Y_i) = \sigma^2;\ i = 1, \ldots, n$, where all random variables are independent.