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.

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

A reasonable model might be

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

with \( V_i \)s independent, \( i = 1, \ldots, n \).
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.

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) State whether each explanatory variable is qualitative or quantitative.

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

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

Integers from 0 to 100 (?)

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

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

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 \(X\) and the vector \(\beta\).

It can be written in the usual form with \(\beta = [\mu]\) and \(X' = [1 1 \cdots 1]\).

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

\[
\gamma_0 + \gamma_{11}(x_i - \gamma_1)^2 = \gamma_0 + \gamma_{11}(x_i^2 - 2\gamma_1x_i + \gamma_1^2) = \beta_0 + \beta_1x_i + \beta_{11}x_i^2,
\]

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. 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.
This can be written in the usual form, with $\beta' = [\mu \, \tau_2]$ and

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

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

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. 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 $V(Y_i)$).

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 $E(Y) = X\beta$ and $V(Y) = \sigma^2 G$, where

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

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

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

and so

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

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

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 $\sigma^2$ for each $i$. This is again a linear
model, which can be written

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