Prelims Statistics and Data Analysis – Sheet 4 TT 2019

1. In the model $Y_i = \alpha + \beta x_i + \epsilon_i$, $i = 1, \ldots, n$, where $E(\epsilon_i) = 0$, show that the least squares estimator of β is

$$\widehat{\beta} = \frac{n \sum x_i Y_i - (\sum x_i)(\sum Y_i)}{n \sum x_i^2 - (\sum x_i)^2}.$$

Show that $\widehat{\beta}$ is unbiased for β. Under what additional assumptions is $\widehat{\beta}$ the maximum likelihood estimator of β?

2. Suppose $x_1, \ldots, x_n$ are known constants and that $Y_1, \ldots, Y_n$ satisfy the 'regression through the origin' model $Y_i = \beta x_i + \epsilon_i$, where the ϵ_i are independent $N(0, \sigma^2)$ random variables. Show that the maximum likelihood estimator of β is $\widehat{\beta} = \sum x_i Y_i / \sum x_i^2$. What is the distribution of $\widehat{\beta}$?

Suppose we have data giving the distance, in miles, by road (y_i) and in a straight line (x_i) for several different journeys. Why might we prefer to consider the model above to the model $Y_i = \alpha + \beta x_i + \epsilon_i$?

Assuming the 'regression through the origin' model, if the straight-line distance between two locations is 12 miles, how would you use the model to predict the expected distance by road?

How could we find a 95% confidence interval for this expected distance?

3. (a) Suppose $Y_1, \ldots, Y_n$ satisfy

$$Y_i = a + b(x_i - \overline{x}) + \epsilon_i \qquad (1)$$

where the ϵ_i are independent $N(0, \sigma^2)$ and the constants x_i are not all equal.

Find the maximum likelihood estimators $\widehat{a}$ and $\widehat{b}$. Show that $\widehat{a}$ and $\widehat{b}$ are unbiased for a and b, respectively, and find their variances.

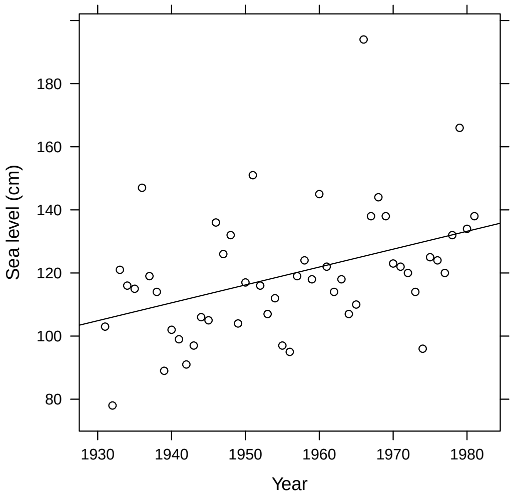
Assuming σ^2 is known, show how the distribution of $\widehat{b}$ can be used to construct a 95% confidence interval for b.

(b) The plot below (data from Davison (2003)) shows annual maximum sea levels in Venice for 1931–1981. Consider model (1) and also the second model

$$Y_i = \alpha + \beta x_i + \epsilon_i. \qquad (2)$$

Give an interpretation in words of the estimates $\widehat{a} = 119.6$ and $\widehat{b} = 0.567$ for model (1), and $\widehat{\alpha} = -989.4$ and $\widehat{\beta} = 0.567$ for model (2).

From the point of view of interpreting the model, do you prefer model (1) or (2)?



4. (Optional: using R or Matlab)

R: work through Rdemo-2. It looks briefly at a few small datasets, check that you understand the approach taken. Most/all of the material involving residuals will be in lectures in week 3, you may want to wait until then to do some parts.

For the two Olympics datasets, fit simple models and hence predict the winning times in 2012 and 2016. How do these compare to the actual winning times?

Matlab: you will first need to download and read-in each dataset, see the final page of Rdemo-2.