ST102 Outline solutions to Exercise 19

1.* (a) We minimise:

$$S = \sum_{i=1}^{2} \varepsilon_i^2 = (y_1 - \alpha - \beta)^2 + (y_2 + \alpha - \beta)^2$$

by partially differentiating with respect to α and β. We have:

$$\frac{\partial S}{\partial \alpha} = -2(y_1 - \alpha - \beta) + 2(y_2 + \alpha - \beta)$$

and:

$$\frac{\partial S}{\partial \beta} = -2(y_1 - \alpha - \beta) - 2(y_2 + \alpha - \beta).$$

Equating the partial derivatives to zero leads to:

$$\widehat{\alpha} = \frac{y_1 - y_2}{2}$$

and:

$$\widehat{\beta} = \frac{y_1 + y_2}{2}.$$

We find that:

$$\mathrm{E}(\widehat{\alpha}) = \frac{\alpha + \beta + \alpha - \beta}{2} = \alpha$$

and:

$$\mathrm{E}(\widehat{\beta}) = \frac{\alpha + \beta - \alpha + \beta}{2} = \beta$$

showing that both are unbiased estimators.

(b) The variance of $\widehat{\alpha}$ is calculated by taking into account that ε_1 and ε_2 are independent and normally distributed with a common variance of σ^2.
For the variance of $\widehat{\alpha}$ we have:

$$\mathrm{Var}(\widehat{\alpha}) = \frac{\mathrm{Var}(y_1) + \mathrm{Var}(y_2)}{4} = \frac{\mathrm{Var}(\varepsilon_1) + \mathrm{Var}(\varepsilon_2)}{4} = \frac{\sigma^2}{2}.$$

2. We first estimate the slope:

$$\widehat{\beta}_1 = \frac{\sum\limits_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum\limits_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum\limits_{i=1}^{n} x_i^2 - n\bar{x}^2} = -0.0187$$

and then estimate the intercept to be $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} = 62.2628$. Hence the fitted regression line is $\widehat{y} = 65.2628 - 0.0187x$, which may be interpreted as follows: a one cent increase in the price of gasoline provides an decrease in sales revenue of \$18,700.

3.* (a) For the least squares estimator we need to minimise:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \theta x_i^2)^2.$$

Differentiating with respect to θ, we have:

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \sum_{i=1}^{n} (y_i - \theta x_i^2)^2 = -2 \sum_{i=1}^{n} x_i^2 (y_i - \theta x_i^2).$$

Equating to zero, we obtain (after dividing by -2):

$$\sum_{i=1}^{n} x_i^2 (y_i - \widehat{\theta} x_i^2) = 0.$$

Rearranging gives:

$$\widehat{\theta} = \frac{\sum_{i=1}^{n} y_i x_i^2}{\sum_{i=1}^{n} x_i^4}.$$

(b) Given that the ε_is are independent and identically distributed, the likelihood function is:

$$\begin{aligned} L(\theta, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \theta x_i^2)^2\right) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta x_i^2)^2\right). \end{aligned}$$

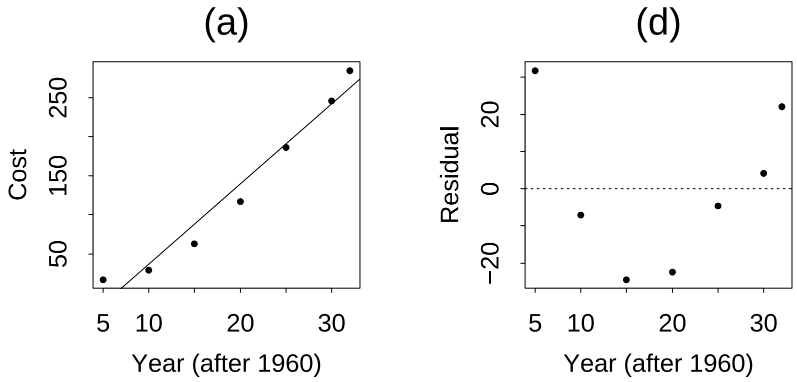
The log-likelihood function is:

$$l(\theta, \sigma^2) = \log L(\theta, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta x_i^2)^2$$

and maximising this is the same as the minimisation in (a). Therefore, the maximum likelihood estimator of θ is the same as the least squares estimator:

$$\widehat{\theta} = \frac{\sum_{i=1}^{n} y_i x_i^2}{\sum_{i=1}^{n} x_i^4}.$$

4. (a) Here is the plot, together with the fitted regression line obtained from part (b). The residual plot for part (d) is also given.



(b) We obtain $n = 7$, $\sum_i x_i = 137$, $\sum_i y_i = 945.4$, $\sum_i x_i^2 = 3{,}299$, $\sum_i y_i^2 = 195{,}715.2$ and $\sum_i x_i y_i = 24{,}855.7$. Therefore, the estimate of β_1 is:

$$\widehat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i x_i y_i - \sum_i x_i \sum_j y_j / n}{\sum_i x_i^2 - (\sum_i x_i)^2 / n} = 10.28.$$

The estimate of β_0 is:

$$\widehat{\beta}_0 = \frac{1}{n} \sum_i y_i - \widehat{\beta}_1 \frac{1}{n} \sum_i x_i = -66.22.$$

Hence the fitted model is $\widehat{y} = -66.22 + 10.28x$. This indicates that social security cost increases by about \$10.28 billion every year.

(c) We first have to estimate σ using $\widehat{\sigma} = \left(\sum_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 / (n-2) \right)^{1/2} = 23.22$. Note that:

$$\sum (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 = \sum y_i^2 + n\widehat{\beta}_0^2 + \widehat{\beta}_1^2 \sum x_i^2 - 2\widehat{\beta}_0 \sum y_i - 2\widehat{\beta}_1 \sum x_i y_i + 2\widehat{\beta}_0 \widehat{\beta}_1 \sum x_i.$$

The standard error of $\widehat{\beta}_1$ is then estimated:

$$\text{E.S.E.}(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{(\sum_i (x_i - \bar{x})^2)^{1/2}} = 0.93.$$

The test statistic for testing $\text{H}_0 : \beta_1 = 0$ vs. $\text{H}_1 : \beta_1 > 0$ is:

$$T = \frac{\widehat{\beta}_1}{\text{E.S.E.}(\widehat{\beta}_1)} \sim t_{n-2} = t_5$$

under H_0. Since $t = 10.28/0.93 = 11.05 > 2.015 = t_{0.05,5}$, we reject the null hypothesis at the 5% significance level. Therefore, there is strong evidence indicating that social security costs increase over time.

(d) The residual plot shows a clear non-random pattern, indicating inadequacy of the linear model. Looking at the original data plot in (a), we would think about applying a log-transformation, i.e. let $z = \log(y)$, in order to accommodate the non-linear relationship between y and x.

Note that, unfortunately, the non-linearity in this dataset cannot be removed using a simple log-transformation. Fitting a linear regression of $z = \log(y)$ on x, results in the estimated model $\widehat{z} = 2.434 + 0.106x$, with $\widehat{\sigma} = 0.173$. The plots below indicate that there is still some pattern in the residuals.