Tutorial Problems

Question 1

Suppose the linear regression model is given by $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1 \ldots, n \geq 2$. Assume that $\varepsilon_i \sim NID(0, 1)$, that is assumptions (A1)-(A4) hold. Because of convenience the scale of the x values is changed (e.g. from inches to centimeters) and the transformed explanatory values $z = x/\tau$ are used instead. Write the new model as

$$Y_i = \gamma_0 + \gamma_1 z_i + \varepsilon_i.$$

(a) Represent estimates of γ_0 and γ_1 in terms of $\hat{\beta}_0$ and $\hat{\beta}_1$. (Lecture 3)

(b) Show that r^2 is invariant. (Lecture 3)

Question 2

Show that the F-test statistic for testing $H_0 : \beta_1 = 0$,

$$F = \frac{\hat{\beta}_1^{\,2} S_{XX}}{\hat{\sigma}^2},$$

can be written as

$$F = \frac{r^2(n-2)}{1-r^2},$$

where r is the coefficient of correlation between x and Y. (Assumed knowledge)

Question 3

A vector of random variables $X = (X_1, X_2, X_3)^\top$ has covariance matrix

$$\Sigma = \begin{pmatrix} 9 & -4 & 1 \\ -4 & 25 & 0 \\ 1 & 0 & 2 \end{pmatrix}.$$

(a) Find the correlation coefficient between X_1 and X_2. (Assumed knowledge)

(b) Find the variance of $Y = -2X_1 + 3X_2$. Write Y as $a^\top X$ and show that the variance is $a^\top \Sigma a$. (Assumed knowledge + Lecture 5)

(c) What is the standard deviation of X_3? (Assumed knowledge)

Question 4

In Lecture 3, you saw that a simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

assuming errors $\epsilon_i \sim NID(0, \sigma^2)$ have the joint density evaluated at (the observed values) $\boldsymbol{y}^\top = (y_1, \ldots, y_n)$ as

$$
\begin{aligned}
f(\boldsymbol{y}; \beta_0, \beta_1, \sigma) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_1-\beta_0-\beta_1 x_1)^2}{2\sigma^2}} \times \cdots \times \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_n-\beta_0-\beta_1 x_n)^2}{2\sigma^2}} \\
&= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\beta_0-\beta_1 x_i)^2}. \qquad (1)
\end{aligned}
$$

(a) Write the log-likelihood, $\ell(\beta_0, \beta_1; \boldsymbol{y}, \sigma) = \log f(\boldsymbol{y}; \beta_0, \beta_1, \sigma)$.
(b) Find β_0 and β_1 that maximises ℓ assuming σ is a known fixed value.

Computer Problems

For the following questions, use the `olympic.txt` dataset that consists of the winning heights or distances (in inches) for the High Jump, Discus and Long Jump events at the Olympics up to 1996. (Lecture 3 & 4)

Question 1

(a) Store the `olympic.txt` dataset in R as the data frame `olympic`. **Hint**: the dataset is tab delimited (`sep="\t"`).

(b) Describe, and where possible explain, any unusual features about `olympic`.

(c) Create a new data frame `olympicMetric` that has measurements in metres, by using the conversion 1 m = 39.3701 inches, and the full year (e.g. 1900 rather than 0). Show the first 6 rows of the `olympicMetric`. **Hint**: The Olympics were held every 4 years except for 1916, 1940, and 1944 due to war.

You should use `olympicMetric` for the next two questions.

Question 2

(a) Plot the first 20 values of LongJump (x_i) against the first 20 values of HighJump (y_i). Briefly comment on the pattern.

(b) Fit the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim NID(0, \sigma^2)$$

for $i = 1, \ldots, 20$ using

```
olympicLm <- lm(HighJump ~ LongJump, data = olympicMetric)
```

(c) Find the regression point estimates for the parameters (β_0, β_1, σ^2) using a `summary` output of `olympicLm`.

(d) Check the model assumptions using the graphical diagnostic plots from the lectures, and a scatter plot with associated fitted regression line in red.

Question 3

This question relates to material in Lecture 5.

(a) Construct 95% confidence intervals for the parameters of the model.

(b) Find point estimates for missing values of HighJump.

(c) Is it more appropriate to construct prediction intervals or confidence intervals for missing values of HighJump? Explain your answer.

(d) Construct the appropriate intervals, based on your answer to the previous part, at the 99% level for missing values of HighJump.

