MAST20005/MAST90058: Week 6 Problems

1. To analyse the data from a brain study, we use the regression model: $Y_i = \alpha + \beta x_i + \varepsilon_i$, $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$, $i = 1, \dots, n$. The response is the brain weight (on a log-scale) for $n = 62$ terrestrial mammals, while the predictor is the body weight (also on a log-scale). Consider the following (partial) R output:

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.13479
x                0.75169
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 0.6943 on 60 degrees of freedom
Multiple R-squared:  0.9208,    Adjusted R-squared:  0.9195
F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

   Recall that $\mathrm{var}(\hat\beta) = \sigma^2/K$, where $K = \sum_i (x_i - \bar x)^2$. Find a 95% confidence interval for β. You may use the following information:

```
> sd(x)
[1] 3.123128
> qt(c(0.999, 0.99, 0.975, 0.95), df = 60)
[1] 3.23 2.39 2.00 1.67
```


2. Consider random variables X_1, X_2, X_3 having joint density $f(x_1, x_2, x_3)$. Suppose that

$$\mathbb{E}(X_3 \mid X_1 = x_1, X_2 = x_2) = \alpha + \beta_1(x_1 - \mu_1) + \beta_2(x_2 - \mu_2)$$

   where $\mu_i = \mathbb{E}(X_i)$. Show that:

   (a) $\alpha = \mu_3$,

   (b) both of:

$$\beta_1 = \frac{\sigma_{13}\sigma_2^2 - \sigma_{12}\sigma_{23}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}, \quad \beta_2 = \frac{\sigma_{23}\sigma_1^2 - \sigma_{12}\sigma_{13}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}$$

   where $\sigma_i^2 = \mathrm{var}(X_i)$ and $\sigma_{ij} = \mathrm{cov}(X_i, X_j)$.

3. Consider the simple linear model $Y = \alpha_0 + \beta(x_i - \bar x) + \varepsilon$ where $\varepsilon \sim \mathrm{N}(0, \sigma^2)$.

   (a) Show that

$$\sum_{i=1}^{n} [Y_i - \alpha_0 - \beta(x_i - \bar x)]^2 = n(\hat\alpha_0 - \alpha_0)^2 + (\hat\beta - \beta)^2 \sum_{i=1}^{n}(x_i - \bar x)^2 + \sum_{i=1}^{n}[Y_i - \hat\alpha_0 - \hat\beta(x_i - \bar x)]^2$$

   (b) For an appropriate value of c (which one?), show that the endpoints for a $100\cdot(1-\gamma)\%$ confidence interval for α_0 are:

$$\hat\alpha_0 \pm c\frac{\hat\sigma}{\sqrt{n}}.$$

   (c) Letting F^{-1} be the inverse cdf of χ^2_{n-2}, show that a $100\cdot(1-\gamma)\%$ confidence interval for σ^2 is:

$$\left(\frac{(n-2)\hat\sigma^2}{F^{-1}(1-\gamma/2)}, \frac{(n-2)\hat\sigma^2}{F^{-1}(\gamma/2)}\right).$$

4. Explain why the model $\mu(x) = \beta_1 e^{\beta_2 x}$ is not a linear model.

5. To fit the quadratic curve $y = \beta_1 + \beta_2 x + \beta_3 x^2$ to a set of points, we minimise

$$h(\beta_1, \beta_2, \beta_3) = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i - \beta_3 x_i^2)^2.$$

   By setting the three first partial derivatives of h with respect to β_1, β_2 and β_3 to zero, show that β_1, β_2 and β_3 satisfy the normal equations:

$$\begin{aligned}
\sum y_i &= \beta_1 n + \beta_2 \sum x_i + \beta_3 \sum x_i^2 \\
\sum x_i y_i &= \beta_1 \sum x_i + \beta_2 \sum x_i^2 + \beta_3 \sum x_i^3 \\
\sum x_i^2 y_i &= \beta_1 \sum x_i^2 + \beta_2 \sum x_i^3 + \beta_3 \sum x_i^4
\end{aligned}$$

6. (a) Show that:

$$\begin{aligned}
\sum_{i=1}^{n} (x_i - \bar{x}) y_i &= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i (y_i - \bar{y}) \\
&= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i
\end{aligned}$$

   (b) Prove the following identity for the sum of squared residuals:

$$d^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \frac{\left[\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

7. The following table gives the leaf area, y, of a particular type of tree at age x years.

x	8	11	17	20	23	26
y	14.8	17.3	20.8	24.4	29.3	35.0
	9.0		23.7	28.9		33.4
	11.0			27.8		37.8

   Some useful statistics:

$$\begin{array}{lll}
n = 13 & \sum x_i y_i = 6282.3 & \sum (x_i - \bar{x})(y_i - \bar{y}) = 741.1 \\
\sum x_i = 230 & \sum x_i^2 = 4648 & \sum (x_i - \bar{x})^2 = 578.8 \\
\sum y_i = 313.2 & \sum y_i^2 = 8546.0 & \sum (y_i - \bar{y})^2 = 1000.2
\end{array}$$

   Using a simple linear regression model, calculate the following:

   (a) Estimates of all of the parameters

   (b) Standard errors for all of the regression coefficients

   (c) A 95% confidence interval for the expectation of Y when $x = 18$

   (d) A 95% prediction interval for Y when $x = 18$