Question 1

(a) Prove that for any two random variables X and Y,

$$\mathrm{Cov}\left(X,Y\right)=\mathrm{E}\left(XY\right)-\mathrm{E}\left(X\right)\mathrm{E}\left(Y\right)$$

(b) Prove that for random variables $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ that

$$\mathrm{Cov}\left(\sum_{i=1}^{n}a_iX_i, \sum_{j=1}^{m}b_jY_j\right)=\sum_{i=1}^{n}\sum_{j=1}^{m}a_ib_j\mathrm{Cov}\left(X_i,Y_j\right)$$

Question 2

Suppose the following 11 values are the transaction amounts (in £) of online purchases for a particular credit card customer in a given month.

$$45,\ 81,\ 52,\ 23,\ 147,\ 92,\ 76,\ 124,\ 287,\ 103,\ 65$$

Tukey's criterion states that, given the lower quartile $q_{0.25}$, the upper quartile $q_{0.75}$ and the interquartile range IQR, if a value x is either $x < q_{0.25} - k\mathrm{IQR}$ or $x > q_{0.75} + k\mathrm{IQR}$, for $k = 1.5$, then x is considered to be an outlier.

(a) Compute the lower and upper quartiles, and the interquartile range for this dataset.

(b) According to Tukey's criterion, are any of these transaction amounts outliers?

(c) If any of the transactions is an outlier, would you take any action? What could be the consequences of (i) inaction (doing nothing) or (ii) taking action (preventing the transaction from going through)?

(d) If you were designing your own fraud detector for this customer (not using Tukey's criterion) for the next month, how high would a value need to be for you to decide that a value is anomalous and potentially fraudulent? In other words, at what value would you set the threshold?

Question 3

Recall that given the random variables $Y_1, Y_2, \ldots, Y_n$ and the observations $x_1, x_2, \ldots, x_n$, the estimators $\widehat{\boldsymbol{\beta}}_0$ and $\widehat{\boldsymbol{\beta}}_1$ are defined by

$$\widehat{\boldsymbol{\beta}}_0 = \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(x_i - \overline{x})\,\overline{x}}{S_{xx}} \right) Y_i$$

$$\widehat{\boldsymbol{\beta}}_1 = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x})\, Y_i$$

where $S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$.

(a) Show that

$$\mathrm{Var}(\widehat{\epsilon}_i) = \mathrm{Var}(Y_i) + \mathrm{Var}(\widehat{\boldsymbol{\beta}}_0) + x_i^2 \mathrm{Var}(\widehat{\boldsymbol{\beta}}_1) + 2x_i \mathrm{Cov}(\widehat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\beta}}_1) - 2\mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_0) - 2x_i \mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_1).$$

(b) Show that

$$\mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_0) = \left(\frac{1}{n} - \frac{(x_i - \overline{x})\,\overline{x}}{S_{xx}} \right) \sigma^2.$$

(c) Show that

$$\mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_1) = \left(\frac{x_i - \overline{x}}{S_{xx}} \right) \sigma^2.$$

(d) Given that

$$\sum_{i=1}^{n} \mathrm{Var}\left(\widehat{\epsilon}_i\right) = \sigma^2 \sum_{i=1}^{n} \left[\frac{n-2}{n} + \frac{1}{S_{xx}} \left(\frac{1}{n} \sum_{j=1}^{n} x_j^2 + x_i^2 - 2x_i \overline{x} - 2\left(x_i - \overline{x}\right)^2 \right) \right],$$

prove that

$$\frac{1}{n} \sum_{i=1}^{n} \mathrm{Var}\left(\widehat{\epsilon}_i\right) = \left(\frac{n-2}{n} \right) \sigma^2.$$