Question 1

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

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

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

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

Solution to Question 1

Part (a):

By definition, if we write $\mathrm{E}\,(X) = \mu_X$ and $\mathrm{E}\,(Y) = \mu_Y$, then using the linearity of expectation,

$$\begin{aligned}
\mathrm{Cov}\,(X,Y) &= \mathrm{E}[(X-\mu_X)\,(Y-\mu_Y)] \\
&= \mathrm{E}[XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y] \\
&= \mathrm{E}[XY] - \mu_X \mathrm{E}[Y] - \mu_Y \mathrm{E}[X] + \mathrm{E}[\mu_X\mu_Y] \\
&= \mathrm{E}[XY] - \mu_X\mu_Y - \mu_Y\mu_X + \mu_X\mu_Y \\
&= \mathrm{E}[XY] - \mu_X\mu_Y \\
&= \mathrm{E}\,(XY) - \mathrm{E}\,(X)\,\mathrm{E}\,(Y)
\end{aligned}$$

Part (b):

We use Part (a).

$$\mathrm{Cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \mathrm{E}\left[\left(\sum_{i=1}^{n} a_i X_i\right)\left(\sum_{j=1}^{m} b_j Y_j\right)\right] - \mathrm{E}\left[\sum_{i=1}^{n} a_i X_i\right]\mathrm{E}\left[\sum_{j=1}^{m} b_j Y_j\right]$$

$$= \mathrm{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j X_i Y_j\right] - \mathrm{E}\left[\sum_{i=1}^{n} a_i X_i\right]\mathrm{E}\left[\sum_{j=1}^{m} b_j Y_j\right]$$

and using the linearity of expectation,

$$\mathrm{Cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \mathrm{E}\,[X_i Y_j] - \left(\sum_{i=1}^{n} a_i \mathrm{E}\,[X_i]\right)\left(\sum_{j=1}^{m} b_j \mathrm{E}\,[Y_j]\right)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \mathrm{E}\,[X_i Y_j] - \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \mathrm{E}\,[X_i]\,\mathrm{E}\,[Y_j]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \left(\mathrm{E}\,[X_i Y_j] - \mathrm{E}\,[X_i]\,\mathrm{E}\,[Y_j]\right)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \mathrm{Cov}\,(X_i, Y_j,)$$

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 $x < q_{0.25} - k\text{IQR}$ or $x > q_{0.75} + k\text{IQR}$, 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?

Solution to Question 2

Part (a):

Sorting the data,

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

One finds the median as the 6th value (since there are 11 values), and therefore the lower quartile is at index $(1+6)/2 = 3.5$. This means that $q_{0.25}$ is the average of the 3rd and 4th order statistics (ordered values), i.e. $q_{0.25} = (52 + 65)/2 = 117/2 = 58.5$.

The upper quartile is computed similarly; it is 3.5 units away from the largest values, i.e. the average of 103 and 124. Therefore $q_{0.75} = 227/2 = 113.5$.

The IQR is therefore $113.5 - 58.5 = 55$.

Part (b):

Using Tukey's criterion for outliers, the lower limit is $58.5 - 1.5(55) = -24$ and the upper limit is $113.5 + 1.5(55) = 196$. Therefore, according to Tukey's criterion, the value 287 is an outlier.

Part (c):

Whether or not you take action in this case is a personal choice. The transaction with value 287 is an outlier, so if one were strictly following the criterion one woudl take action. On the other hand, 287 does not seem to be very different to the others. If one does not take action, it is risking that a fraudulent transaction goes through and the customer (or the company) loses money. If action is taken, and the transaction is blocked or delayed, this could have consequences for the customer if this is not a fradulent transaction. The point is - it is a trade-off, and finding the optimal decision rule while balancing these aspects is not easy.

Part (d):

There is no right answer to this question, and it is meant to make you think of your own possible outlier detection algorithm. For example, one approach would be to consider a threshold based on standard deviations from the mean. Denoting the sample mean of this data set by $\overline{x}$, one can compute $\overline{x} \approx 100$. The sample standard deviation is $s \approx 72$. One option would be to set the upper limit as $\overline{x} + 10s \approx 820$.

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

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

prove that

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

Solution to Question 3

Part (a):

$$\begin{aligned}
\mathrm{Var}(\widehat{\epsilon}_i) &= \mathrm{Var}(Y_i - (\widehat{\boldsymbol{\beta}}_0 + \widehat{\boldsymbol{\beta}}_1 x_i)) \\
&= \mathrm{Var}(Y_i) + \mathrm{Var}(\widehat{\boldsymbol{\beta}}_0 + \widehat{\boldsymbol{\beta}}_1 x_i) - 2\mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_0 + \widehat{\boldsymbol{\beta}}_1 x_i) \\
&= \mathrm{Var}(Y_i) + \mathrm{Var}(\widehat{\boldsymbol{\beta}}_0 + \widehat{\boldsymbol{\beta}}_1 x_i) - 2\mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_0) - 2x_i \mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_1) \\
&= \mathrm{Var}(Y_i) + \Big[\mathrm{Var}(\widehat{\boldsymbol{\beta}}_0) + \mathrm{Var}(\widehat{\boldsymbol{\beta}}_1 x_i) + 2\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\beta}}_1 x_i) \Big] - 2\mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_0) - 2x_i \mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_1) \\
&= \mathrm{Var}(Y_i) + \mathrm{Var}(\widehat{\boldsymbol{\beta}}_0) + x_i^2 \mathrm{Var}(\widehat{\boldsymbol{\beta}}_1) + 2x_i \mathrm{Cov}(\widehat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\beta}}_1) - 2\mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_0) - 2x_i \mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_1)
\end{aligned}$$

Part (b):

$$\begin{aligned}\mathrm{Cov}(Y_i, \widehat{\beta}_1) &= \mathrm{Cov}\left(Y_i, \frac{1}{S_{xx}} \sum_{j=1}^{n} (x_j - \overline{x})\, Y_j\right) \\ &= \frac{1}{S_{xx}} \sum_{j=1}^{n} (x_j - \overline{x})\, \mathrm{Cov}\,(Y_i, Y_j) \\ &= \frac{1}{S_{xx}} (x_i - \overline{x})\, \mathrm{Var}\,(Y_i) \\ &= \left(\frac{x_i - \overline{x}}{S_{xx}}\right) \sigma^2\end{aligned}$$

Part (c):

$$\begin{aligned}\mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_0) &= \mathrm{Cov}\left(Y_i, \sum_{j=1}^{n} \left(\frac{1}{n} - \frac{(x_j - \overline{x})\,\overline{x}}{S_{xx}}\right) Y_j\right) \\ &= \sum_{j=1}^{n} \left(\frac{1}{n} - \frac{(x_j - \overline{x})\,\overline{x}}{S_{xx}}\right) \mathrm{Cov}\,(Y_i, Y_j) \\ &= \left(\frac{1}{n} - \frac{(x_i - \overline{x})\,\overline{x}}{S_{xx}}\right) \mathrm{Var}\,(Y_i) \\ &= \left(\frac{1}{n} - \frac{(x_i - \overline{x})\,\overline{x}}{S_{xx}}\right) \sigma^2\end{aligned}$$

Part (d):

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

$$= n \cdot \left\{\frac{n-2}{n}\right\} + \frac{1}{S_{xx}} \left(n \cdot \left\{\frac{1}{n} \sum_{j=1}^{n} x_j^2\right\} + \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i\overline{x} - 2\sum_{i=1}^{n} (x_i - \overline{x})^2\right)$$

$$= (n-2) + \frac{1}{S_{xx}} \left(\sum_{j=1}^{n} x_j^2 + \sum_{i=1}^{n} x_i^2 - 2n\overline{x}^2 - 2S_{xx}\right)$$

$$= (n-2) + \frac{2}{S_{xx}} \left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2 - S_{xx}\right)$$

$$= (n-2) + \frac{2}{S_{xx}} (S_{xx} - S_{xx})$$

$$= n - 2$$

This implies

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

as required