Question 1 (33/100 points)

(a) **[8 points]** Consider the following data points

$$7.7; \quad -3.5; \quad 8.6; \quad 2.8; \quad -1.7; \quad -1.2; \quad 0.5$$

Obtain the 1st, 2nd and 3rd quartile of this dataset

Answer:
After sorting the data points from the smallest to largest, we immediately see that 1st quartile is $Q_1 = -1.7$, the 2nd is $Q_2 = 0.5$, and the 3rd is $Q_3 = 7.7$.

(b) **[15 points]** You have obtained the skewness and kurtosis of a certain variable. The skewness is $\gamma_1 = -2.35$ and the kurtosis is $\gamma_2 = 7.33$. A colleague of yours makes the following statements:
(i) *"The variable is leptokurtic and the left tail is heavier than the right tail"*.
(ii) *"There are outliers on the left tail of the distribution but not on the right tail"*.
(iii) *"I expect the mean to be larger than the median because the median is robust to outliers"*.
For each statement, say whether you agree or not. Justify your answers.

Answer:
(i) The statement is true. The kurtosis larger than 3 indicates that the distribution is leptokurtic and the negative skewness suggests that the left tail is heavier than the right tail.
(ii) It is true that the large kurtosis and negative skewness indicate presence of outliers on the left tail. On the other hand, it is not necessarily true that there are no outliers on the right tail. The kurtosis does not provide information on whether outliers are on the left or right tail and the negative skewness only suggests that the left tail is heavier. However, there may be outliers on the right tail as well.
(iii) The statement is not true. The left tail is heavier than the right tail. This means that there should be "more" outliers on the left tail and therefore the mean should be lower than the median since the mean will be affected by these outliers and instead the median will be robust.

(c) **[10 points]** The data frame `grades` contains the variables `name`, `math1`, `math2` and `math_final`. The variable `name` contains the names of the students, the variable `math1` contains the grades of the 1st part of a Math exam, the variable `math2` contains the grades of the 2nd part of the Math exam, and the variable `math_final` contains the final grade, which is missing. The dataset is given below:

```
> grades
  names math1 math2 math_final
```

```
1   Bob   6.5   4.5       NA
2  Lucy   5.5   8.5       NA
3   Eve   5.0   7.5       NA
4  Mark   7.0   5.0       NA
```

The following R code is given:

```
n <- length(grades$names)
k <- 1
repeat{
  grades$math_final[k] <- 0.3*grades$math1[k]+0.7*grades$math2[k]
  k <- k+1
  if(k>n){break}
}
```

Explain briefly what the R code is doing. What will be contained in `math_final` after the *repeat loop*?

The teacher of the course wants to set a minimum grade of 5.5 for each part of the exam in order to pass the course. She wants the final grade to be equal to the minimum between the grades of the two parts if at least one of the two parts has a grade that is lower than 5.5. How would you adjust the code to account for this? Sketch the code and explain what it does.

Answer:

The code stores the number of rows in the dataset in the R object `n`. Then, the *repeat loop* computes the final grade of each student as a weighted average between the 1st and 2nd part of the exam (weight 0.3 on the 1st part and 0.7 on the 2nd). The loop is stopped when the index `k` is larger than the number of rows in the dataset. After the loop, the vector `math_final` contains $(5.1, 7.6, 6.75, 5.6)$.

To set the minimum grade of the 5.5, the code can be adjusted as follows

```
repeat{
  grades$math_final[k] <- 0.3*grades$math1[k]+0.7*grades$math2[k]
  if(grades$math1[k]<5.5 | grades$math2[k]<5.5) {
   grades$math_final[k] <- min(grades$math1[k],grades$math2[k])
  }
  k <- k+1
  if(k>n){break}
}
```

Question 2 (34/100 points)

(a) **[10 points]** You have a dataset with three variables: x, y and z. The sample means, variances and covariances are obtained:

$$\bar{x} = -6.9, \quad \bar{y} = 0.7, \quad \bar{z} = 1.2; \qquad s_x^2 = 6.3, \quad s_y^2 = 1.0, \quad s_z^2 = 9.1;$$

$$s_{xy} = 1.3, \quad s_{xz} = 3.5, \quad s_{yz} = -2.6.$$

Say whether you agree or not with the following statements. Justify your answers.

(i) "*The relationship between x and z is stronger than the relationship between x and y.*"
(ii) "*The simple linear regression model $y = \beta_0 + \beta_1 x + u$ will produce better predictions of y than the simple linear regression model $y = \beta_0 + \beta_1 z + u$.*"

Answer:
(i) The statement is not true. The sample correlation between x and y is 0.47 instead the sample correlation between x and y is 0.52. Therefore, the sample statistics may actually indicate that the linear relationship between x and y is weaker than the relationship between x and y. The sample covariance does not provide and indication of how strong the relationship is.
(ii) The statement is not true. The sample correlation between y and z is -0.86. Therefore, the linear relationship between y and z is stronger than the one between x and y. The model $y = \beta_0 + \beta_1 x + u$ has $R^2 = 0.27$ and the model $y = \beta_0 + \beta_1 z + u$ has $R^2 = 0.74$. This meas that the 2nd model explains more of the variability of y and therefore it is expected to produce better predictions.

(b) **[14 points]** Available is a dataset with 2 variables and $n = 16$ observations for each of the 2 variables. Consider a linear regression model of the form $y_i = \beta_0 + \beta_1 x_i + u_i$. The OLS estimates of β_0 and β_1, the R^2 and the standard error of the regression (SER) are obtained:

$$\hat{\beta}_0 = 2.3, \quad \hat{\beta}_1 = -1.7, \quad R^2 = 0.85, \quad SER = 3.5.$$

(i) Obtain a prediction of y given $x = -2.0$.
(ii) Obtain the total sum of squares (TSS), the residuals sum of squares (RSS) and the explained sum of squares (ESS) of the regression.
(iii) Obtain the sample correlation r_{xy} between x and y.

Answer:
(i) The prediction is $2.3 - 1.7 \times (-2.0) = 5.7$.
(ii) First, we obtain the the RSS form the SER. We know that $SER = \sqrt{RSS/(n-2)}$. Therefore, we obtain

$$RSS = (n-2) \times SER^2 = 14 \times 3.5^2 = 171.5.$$

Next, we obtain the TSS as follows

$$TSS = RSS/(1 - R^2) = 171.5/0.15 = 1143.3.$$

Finally, we obtain the ESS as follows

$$ESS = TSS - RSS = 1143.3 - 171.5 = 971.8.$$

(iii) We can obtain the sample correlation from the equation $R^2 = r_{xy}^2$. We have that $|r_{xy}| = \sqrt{R^2} = 0.92$. The sign of the sample correlation is negative since $\hat{\beta}_1$ is negative. Therefore, we obtain $r_{xy} = -0.92$

(c) **[10 points]** A colleague of yours has estimated the following regression models using a variable of interest y_i and 3 regressors, $x_{1,i}$ $x_{2,i}$ and $x_{3,i}$, $i = 1, \ldots, n$.

(1) $y_i = \beta_0 + \beta_1 x_{1,i} + u_i$.
(2) $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$.
(3) $y_i = \beta_0 + \beta_2 x_{2,i} + \beta_3 x_{3,i} + u_i$.

Your colleague makes the following 2 statements:
(i) "*The adjusted R^2 (adj-R^2) of model (1) is larger than the adj-R^2 of model (2). Therefore, model (1) is better than model (2). This also means that there is no relationship between y and x_2.*"
(ii) "*I have obtained that the R^2 of model (1) is larger than the R^2 of model (3). There must be an error since model (3) has more variables than model (1) and therefore its R^2 must be larger.*"

For each statement, say whether you agree or not. Justify your answers.

Answer:
(i) It is true that if the adj-R^2 of model (1) is larger than the adj-R^2 of model (2), then model (1) can be considered better. This also suggest that conditional on x_1, there is no linear relationship between y and x_2. However, there may be some nonlinear relationship between y and x_2.
(ii) I do not agree with the statement. The R^2 of model (1) does not need to be larger than the R^2 of model (2). Model (2) does not include the variable x_1. Therefore, the result can be due to the fact that x_1 has a stronger linear relationship with y.

Question 3 (33/100 points)

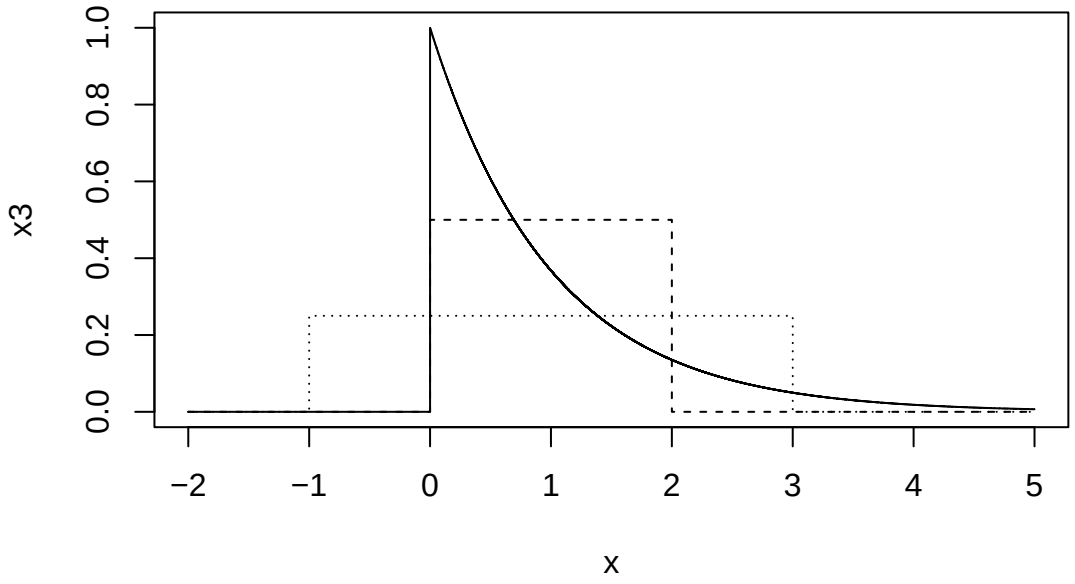
(a) **[15 points]** We have an observation x that we want to classify as a member of any of the three populations Π_1, Π_2 and Π_3. We know that population Π_1 has an exponential distribution[1] with rate $\lambda = 1$, population Π_2 has a uniform distribution between 0 and 2 (i.e. $f_2(x) \sim U(0,2)$) and population Π_3 has a uniform distribution between -1 and 3 (i.e. $f_3(x) \sim U(-1,3)$).

(i) Obtain the discriminant regions R_1, R_2 and R_3 based on the Maximum Likelihood (ML) discriminant rule. Draw a graph of the densities $f_1(x)$, $f_2(x)$ and $f_3(x)$ of the three populations.

(ii) Obtain the probabilities of correct classification p_{11}, p_{22} and p_{33} of the ML rule.

Answer:

(i) First, we obtain a plot of the densities To obtain the first region, we obtain the set



of points such that $f_1(x) > f_2(x)$ and $f_1(x) > f_3(x)$. The exponential density is a decreasing function in the positive real line. Furthermore, we have $f_1(x) = f_2(x)$ if

$$\exp(-x) = 1/2 \quad \Leftrightarrow \quad x = \log(2),$$

[1]The probability density function of an exponential distribution with rate $\lambda > 0$ is

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0$$

and $f_1(x) = f_3(x)$ if

$$\exp(-x) = 1/4 \quad \Leftrightarrow \quad x = \log(4).$$

Note that $\log(4) < 2$. Therefore, we obtain that $R_1 = (0, \log(2)) \cup (3, \infty)$, $R_2 = (\log(2), 2)$, and $R_3 = (-1, 0) \cup (2, 3)$.

(ii) The correct classification probabilities are

$$p_{11} = \int_{R_1} f_1(x)dx = \int_0^{\log(2)} e^{-x}dx + \int_3^{\infty} e^{-x}dx = 1 - \frac{1}{2} + e^{-3},$$

$$p_{22} = \int_{R_2} f_2(x)dx = (2 - \log(2)) \times \frac{1}{2} = 1 - \frac{\log(2)}{2},$$

$$p_{33} = \int_{R_3} f_3(x)dx = (2) \times \frac{1}{4} = \frac{1}{2}.$$

(b) **[8 points]** Consider the ML discriminant rule with two normal populations with means μ_1 and μ_2, $\mu_1 > \mu_2$, and the same variance σ^2. The missclassification probabilities are given by

$$p_{12} = p_{21} = \Phi\left(-\frac{\mu_1 - \mu_2}{2\sigma}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Discuss how the means μ_1 and μ_2 and the variance σ^2 of the normal distributions affect the missclassification probabilities p_{12} and p_{21}.

Answer:

The missclassification probabilities are large if the means μ_1 and μ_2 are close to each other and small if they are far apart. The extreme case where $\mu_1 \approx \mu_2$ leads to missclassification probabilities equal to 0.5. This makes sense since it will be harder to discriminated between the two populations when they have a similar mean. When the variance is large, the missclassification probabilities will be large. Instead, they will be small when the variance is small. This also makes sense since a large variance indicates that the densities are widely dispersed along the real line and therefore it will be harder to discriminate between them.

(c) **[10 points]** Assume we have two normal[2] populations Π_1 and Π_2 with means equal to zero and different variances σ_1^2 and σ_2^2, $\sigma_1^2 > \sigma_2^2$. More specifically, we have $f_1(x) \sim N(0, \sigma_1^2)$ and $f_2(x) \sim N(0, \sigma_2^2)$. Derive the discriminant regions R_1 and R_2 of the ML discriminant rule.

[2]The probability density function of a normal $N(\mu, \sigma^2)$ is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Answer:

The discriminant region R_1 is the set of values such that $f_1(x) > f_2(x)$. We obtain

$$\begin{aligned}
& \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma_1^2}\right) > \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma_2^2}\right) \\
\Leftrightarrow \quad & -\frac{1}{2}\log(\sigma_1^2) - \frac{x^2}{2\sigma_1^2} > -\frac{1}{2}\log(\sigma_2^2) - \frac{x^2}{2\sigma_2^2} \\
\Leftrightarrow \quad & x^2\left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right) > \log(\sigma_1^2) - \log(\sigma_2^2) \\
\text{(since } \sigma_1 > \sigma_2\text{)} \quad \Leftrightarrow \quad & x^2 > \log(\sigma_1^2/\sigma_2^2)\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 - \sigma_2^2}.
\end{aligned}$$

Therefore, the discriminant region R_1 is

$$R_1 = \left(-\infty,\ -\sqrt{\log(\sigma_1^2/\sigma_2^2)\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 - \sigma_2^2}}\right) \cup \left(\sqrt{\log(\sigma_1^2/\sigma_2^2)\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 - \sigma_2^2}},\ \infty\right),$$

and the discriminant region R_2 is

$$R_2 = \left(-\sqrt{\log(\sigma_1^2/\sigma_2^2)\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 - \sigma_2^2}},\ \sqrt{\log(\sigma_1^2/\sigma_2^2)\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 - \sigma_2^2}}\right).$$