Question 1 (30/100 points)

(a) Find the sample variance of the following 3 data points:

$$1.0\ ; \quad 3.3\ ; \quad 2.6$$

(b) For a certain variable you have obtained that the skewness is -2.2 and the kurtosis is 12.5. What can you say about the distribution of the observations? Would you expect to have some outliers?

(c) Consider the following R code

```
> x <- c(1, 5, 7, 3, 2)
> z <- x[x>=3]
```

What is in the R object `z`? Explain briefly what the R code is doing.

(d) The following R code with a `for` loop is given

```
> v <- 1:5
>
> for(i in 1:5){
+   if(v[i]==4) {break}
+   v[i] <- v[i]-1
+ }
```

What is in the R object `v` after running the `for` loop? Explain briefly what the R code is doing.

Solution:

(a) We obtain that $\bar{x} = 2.3$ and $s^2 = 1.39$

(b) As skewness equal to -2.2 indicates that the distribution is not symmetric about its mean. In particular, we expect the left tail to be heavier than the right tail. Instead a kurtosis equal to 12.5 indicates that the tails of the distribution are heavier than the tails of a normal distribution. We say that the distribution is heavy-tailed or leptokurtic. We expect to have outliers because outliers are extreme observations and therefore by definition they lead to high kurtosis.

(c) The R object `z` is a vector containing the numbers 5, 7 and 3. The code is taking the original vector `x`, selecting the values that are greater than or equal to 3 and storing them into the vector `z`.

(d) The R object `v` will contain the following values: 0, 1, 2, 4 and 5. The R code creates the `v` that contains numbers from 1 to 5. The the loop takes each element and detract 1 until the loop gets top 4. At that point the command `break` is triggered and the loop stops.

Question 2 (40/100 points)

(a) Available is a dataset with 2 variables and $n = 100$ 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 ($\hat{\beta}_0$ and $\hat{\beta}_1$), the RSS and the TSS are obtained:

$$\hat{\beta}_0 = 2.6, \quad \hat{\beta}_1 = 1.5, \quad RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 308.6, \quad TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 558.6.$$

(i) Interpret the coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

(ii) Obtain a prediction for the variable y when the observed x is equal to 3.5.

(iii) Obtain the R^2 and the standard error of the regression (SRE).

(b) A colleague of yours has estimated the linear regression model $y_i = \beta_0 + \beta_1 x_i + u_i$ using a certain dataset. She claims that the adjusted-R^2 (R^2_{Adj}) obtained from the regression is negative. Is this possible? Why? What can you say about the relationship between the variable y_i and x_i?

(c) Consider the regression model without intercept given by $y_i = \beta_1 x_i + u_i$.

(i) Show that the OLS estimate of β_1 is

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

by setting the derivative of the sum of squares to zero.

(ii) Show that, in general, the equality $TSS = ESS + RSS$ is no longer true in the regression model without the intercept.

Solution:

(a) (i) The estimated intercept $\hat{\beta}_0$ indicates that the regression line is 2.6 when the observed x is zero. This may not have a meaningful interpretation and in this case it is unclear since it is not explained what the variables x and y are. Instead, the slope $\hat{\beta}_1$ indicates that a unit increase in x leads to an expected increase of 1.5 in y.

(ii) The prediction is given by $\hat{y} = 2.6 + 1.5 \times 3.5 = 7.85$.

(iii) The R^2 is

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{308.6}{558.6} = 0.45.$$

The standard SRE is

$$SRE = \sqrt{\frac{RSS}{n-2}} = \sqrt{\frac{308.6}{98}} = 1.77.$$

(b) The adjusted-R^2 can be negative. Note that the R^2 is always between 0 and 1 (when β_0 is included) and the adjusted-R^2 is equal to the R^2 minus a penalty term that depends on the number of regressors. Therefore it can be negative when the R^2 is close to zero. In this case the interpretation is that the variable x (as included in the model) does not explain a relevant proportion of the variability of y. Therefore we could consider the model with only the intercept. This means that there is not a linear relationship between x and y. However, as we have seen in class, it may be that there is a nonlinear relationship between x and y and this could be captured including powers of x.

(c) (i) See solution of exercises week 2.

(ii) We have that

$$\begin{aligned} TSS &= \sum(y_i - \bar{y})^2 = \sum(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\ &= \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \bar{y})^2 + 2\sum(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\ &= RSS + ESS + 2\sum \hat{u}_i(\hat{y}_i - \bar{y}). \end{aligned}$$

Therefore the results is proved if we can show that in general $\sum \hat{u}_i(\hat{y}_i - \bar{y}) \neq 0$. it is immediate to see that

$$\sum \hat{u}_i(\hat{y}_i - \bar{y}) = \sum \hat{u}_i\hat{y}_i - \bar{y}\sum \hat{u}_i = \hat{\beta}_1 \sum \hat{u}_i x_i - \bar{y}\sum \hat{u}_i = \bar{y}\sum \hat{u}_i \neq 0,$$

where $\hat{\beta}_1 \sum \hat{u}_i x_i = 0$ from the first order condition obtained in the previous point. Furthermore $\bar{y}\sum \hat{u}_i \neq 0$ in general because without the intercept there is no first order condition ensuring that the sum of the residuals is zero.

Question 3 (30/100 points)

(a) Consider the following *confusion matrix* containing the number of misclassified and correctly classified observations for the populations Π_1 and Π_2.

		True membership Π_1	**True membership** Π_2
Predicted	Π_1	$n_{11} = 125$	$n_{12} = 21$
	Π_2	$n_{21} = 13$	$n_{22} = 174$

Obtain the estimated probabilities of misclassification $\hat{p}_{12}$ and $\hat{p}_{21}$ and the apparent error rate (APER).

(b) We have an observation x that we want to classify as a member of either population Π_1 or Π_2. We know that the populations Π_1 and Π_2 have an exponential distribution with rates $\lambda_1 = 1$ and $\lambda_2 = 2$, respectively. Note that the density function of an exponential distribution with rate $\lambda > 0$ is

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

(i) Obtain the discriminant regions R_1 and R_2 based on the Maximum Likelihood (ML) discriminant rule.
(ii) Obtain the misclassification probabilities p_{12} and p_{12}.
(iii) Assume that $C(1|2) = 2C(2|1)$, that is, the misclassification cost $C(1|2)$ is 2 times the misclassification cost $C(2|1)$. How would you expect the regions R_1 and R_2 obtained from the ECM discriminant rule to differ from the ones obtained from the ML rule? Justify your answer. *Do not calculate the ECM discriminant regions!*

Solution:

(a) The estimated probabilities of misclassification $\hat{p}_{12}$ and $\hat{p}_{21}$ are

$$\hat{p}_{12} = \frac{n_{12}}{n_{12} + n_{22}} = \frac{21}{21 + 174} = 0.108,$$

and

$$\hat{p}_{12} = \frac{n_{12}}{n_{21} + n_{11}} = \frac{13}{13 + 125} = 0.094.$$

The APER is

$$APER = \frac{n_{12} + n_{21}}{n_{21} + n_{11} + n_{12} + n_{22}} = \frac{13 + 21}{13 + 21 + 125 + 174} = 0.102.$$

(b) (i) The densities are $f_1(x) = e^{-x}$ and $f_1(x) = 2e^{-2x}$. Therefore we obtain that

$$f_1(x) > f_2(x) \Leftrightarrow x > \log(2).$$

Therefore $R_1 = (\log(2), +\infty)$ and $R_2 = (0, \log(2))$.
(ii) The misclassification probabilities are

$$p_{12} = \int_{\log(2)}^{\infty} 2e^{-2x} = \frac{1}{4},$$

and

$$p_{21} = \int_{0}^{\log(2)} e^{-x} = \frac{1}{2}.$$

(iii) We expect the region R_2 of the ECM rule to be bigger than the region R_2 for the ML rule, and vice versa for R_1. This should be the case because we afford an higher cost when we misclassify an observation in R_1 compared to when we misclassify an observation in R_2. Therefore we expect R_2 to be larger because by definition the ECM rule minimizes the ECM.

End of the exam!