vrije Universiteit *amsterdam*

School of Business and Economics

Exam:	**Solution Sample Exam Data Analysis 1**
Code:	-
Examinator:	-
Co-reader:	-
Date:	-
Time:	-
Duration:	2 hours
Calculator allowed:	Yes
Graphical calculator allowed:	No
Number of questions:	3
Type of questions:	Open
Answer in:	English
Credit score:	100 credits counts for a 10
Grades:	-
Inspection:	-
Number of pages:	-

Good luck!

(This page is intentionally left blank.)

Question 1 (30/100 points)

(a) We can obtain the sample mean $\bar{x}$ as follows

$$\bar{x} = \frac{1}{6}(5.2 + 2.7 + 2.2 - 0.7 + 0.3 - 1.8) = 1.32.$$

Instead, for the median we first order the data

$$-1.8, \quad -0.7, \quad 0.3, \quad 2.2, \quad 2.7, \quad 5.2,$$

finally, we obtain the median m as

$$m = \frac{0.3 + 2.2}{2} = 1.25.$$

(b) An advantage of the sample median compared to the sample mean is that the median is robust against outliers. Instead, an advantage of the mean is that the mean has an easy interpretation: *everybody knows what an average is*.

(c) The quantile q_α is the value such that about $(1 - \alpha) \times 100\%$ of the data points are larger of this vale and about $\alpha \times 100\%$ of the data points are lower. We can say that about 90% of the probability distribution of the variable is between 12.7 and 23.5. In other words, given the observed data, we can expect an observation from this variable to be between 12.7 and 23.5 with 90% probability.

(d) The R function `oper_vec` takes as input 2 numeric vectors of the same length. In particular, the function checks if the 2 vectors have the same length and if they are numeric vectors. If one or more of these conditions is not satisfied, the function returns an error message': *incorrect argument*. In case the arguments are correct, the function uses an `ifelse` conditional execution to obtain a vector that contains in each entry the maximum between the corresponding entries of the original vectors. Finally, the function gives this vector as output.

Question 2 (40/100 points)

(a) (i) The OLS estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ are given by

$$\hat{\beta}_1 = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{s_{xy}}{s_x^2}, \quad \text{and} \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}.$$

Therefore, we obtain that $\hat{\beta}_1$ is given by

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{7.3}{6.8} = 1.07,$$

and $\hat{\beta}_0$ is given by

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 2.3 - 1.07 \times (-1.3) = 3.69.$$

(ii) In the above regression model the R^2 is equal to the square of the sample correlation, that is, $R^2 = r_{xy}^2$. The sample correlation is

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{7.3}{\sqrt{6.8 \times 11.6}} = 0.82.$$

Therefore

$$R^2 = r_{xy}^2 = 0.82^2 = 0.67.$$

The interpretation is that the variable x is able to explain the 67% of the variability of the the variable y.

(iii) The total sum of squares TSS is

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 = (n-1)s_y^2 = 212 \times 11.6 = 2459.2.$$

Furthermore, we know that

$$R^2 = \frac{ESS}{TSS},$$

therefore the explained sum of squared ESS is

$$ESS = TSS \times R^2 = 2459.2 \times 0.67 = 1647.7.$$

Finally, we have that $TSS = ESS + RSS$. Therefore the residual sum of squares RSS is

$$RSS = TSS - ESS = 2459.2 - 1647.7 = 811.5.$$

(b) The regression model is

$$y_i = \beta_0 + \beta_1 x_i + \beta_3 x_i^3 + u_i.$$

Therefore the intercept is included in the model. R includes the intercept by default. To remove the intercept we need to add "-1" as regressor in the above code. The code therefore becomes

```
> x3 <- x^3
> reg <- lm(y~-1+x+x3)
```

(c) *See the solution of the exercises of week 2.*

Question 3 (30/100 points)

(a) The ML rule constructs the discriminant regions R_1 and R_2 in such a way to select the population with the highest density function. In particular, the regions R_1 and R_2 are

$$R_1 = \{x : f_1(x) > f_2(x)\}, \quad R_2 = \{x : f_1(x) < f_2(x)\},$$

where $f_1(x)$ is the density of the first population Π_1 and $f_2(x)$ is the density of the second population Π_2. Instead, the *ECM* rule selects the regions in such a way to minimize the Expected Cost of Misclassification. It can be shown that the regions are

$$R_1 = \left\{x : \frac{f_1(x)}{f_2(x)} > \frac{C(1|2)}{C(2|1)}\right\}, \quad R_2 = \left\{x : \frac{f_1(x)}{f_2(x)} < \frac{C(1|2)}{C(2|1)}\right\},$$

where $C(1|2)$ and $C(2|1)$ are the misclassification costs. In general the ML rule and *ECM* rule are different. However, they are equivalent when the misclassification costs are the same $C(1|2) = C(2|1)$.

(b) *See the solution of the exercises of week 3.*