Statistical Inference and Data analysis - Take Home 2 - 2019

General instructions:

- This homework is due on **Thursday December 12, 2019, 4 pm**. Please place a hardcopy (printout) of your homework in the mailbox of Worku Biyadgie Ewnetu (mailbox 02.32) at the Department of Mathematics. R-codes should be included in an Appendix of your homework.
- You are also requested to send a pdf containing your homework, together with a separate and executable '`.r`' file that contains your R-code. These files should be named as follows `firstname.lastname.pdf` and `firstname.lastname.r`. Send both files by E-mail to Clément Cerovecki (clement.cerovecki@kuleuven.be) and Worku Ewnetu (workubiyadgie.ewnetu@kuleuven.be). The two files need to be received also before the above strict deadline.
- The homework is only complete when you submitted a hard copy of the homework (mail-box) and you have sent the pdf and the executable R-code file (by E-mail).
- Homeworks that come in too late get a zero mark.
- If you do have a serious problem with the homework (e.g. you really do not understand the assignment), then you can contact Clément Cerovecki. However, this should really be an exception.

Further instructions:

The aim of this take home is to apply the theory and to perfect your understanding of it. You don't need to use any specific packages on `R`, i.e. it is enough to use basic vector and matrix controls such as `c()`,`matrix()`, `%*%`, basic graphical tools, the functions `qf`, `qt`, `qnorm` and `rnorm`. Every computation can be done by using the formulas that have been presented in class (and that you can find as well in the lecture notes).

We always use the following standard matrix notations i.e. $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{Y} \in \mathbb{R}^{n\times 1}$ contains the response variable, $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the design matrix, $\boldsymbol{\beta} = (\beta_0, \dots \beta_{p-1})^T \in \mathbb{R}^{p\times 1}$ is the regression parameter, and $\boldsymbol{\varepsilon} \in \mathbb{R}^{n\times 1}$ contains the errors. In the following, you can find further details for some specific questions:

Exercise 1. (b) An exact test is expected here.
(d) An asymptotic test is expected here.
(e) You can decide yourself whether Gaussian assumption was relevant or not. Consequently you would be able to compute an exact or an asymptotic confidence interval.

Exercise 3. Cubic regression, means that $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3$, for $i = 1, \dots, n$.

Exercise 4. The 3 questions are independent to each other.

Exercise 1. The file `ex1.txt` contains an 60×2 dimensional matrix whose first column is the explanatory variable, and second column is the variable of interest.
(a) Fit a linear model, assuming that the strong Gaussian assumption is relevant.
(b) Test whether $\beta_1 = 0$ at the level $\alpha = 0.01$.
(c) Explain what is a Q–Q plot and apply it to the residuals.
(d) Perform a test that do not requires normality of the errors.
(e) Determine a 99% confidence region for $\widehat{\boldsymbol{\beta}}$.

Exercise 2. The file `ex2.txt` contains an 80×2 dimensional matrix whose first column is the explanatory variable X_i, and second column is the variable of interest Y_i for $i = 1, \dots n$.
(a) Compute the ordinary least squares $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$.
Suppose we further know that the errors are correlated and satisfy the following equation:

$$\varepsilon_i = \rho \varepsilon_{i-1} + \eta_{i1}, \quad \text{for } i = 1, \dots, n \tag{1}$$

where $(\eta_n)_{n \in \mathbb{Z}}$ are i.i.d. standard Gaussian and $\rho = 0.8$.
(b) Use (1) to compute the variance of $\boldsymbol{\varepsilon}$.
(c) Transform the model in such a way that the errors are non longer correlated. Compute the ordinary least squares for this new model and compare it to the one obtained in (a).

Exercise 3. The file `ex3.txt` contains an 120×2 dimensional matrix whose first column is the fixed design x_i, and second column is the variable of interest Y_i for $i = 1, \dots n$. We know that Y follows a cubic regression model with respect to x.
(a) Compute the ordinary least squares $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$.
Suppose we further know that the errors are Gaussian but heteroscedastic as follows:

$$\sigma(x) = \begin{cases} x^2 & \text{if } x \in [0, 4/3] \\ 4(x-2)^2 & \text{if } x \in [4/3, 2] \end{cases}$$

(b) Compute the weighted least square estimator $\widehat{\boldsymbol{\beta}}_{\text{WLS}}$.
(c) Compare it to $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ and with true parameter i.e. $\boldsymbol{\beta} = (0.5, 1, -2, 1)^T$.
(d) Determine the distribution of $\widehat{\boldsymbol{\beta}}_{\text{WLS}}$.

Exercise 4. Let $\mathbf{X}$ be a 3-dimensional Gaussian vector with parameters

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1.25 & 1.50 & 0.5 \\ 1.50 & 5.25 & 3.5 \\ 0.50 & 3.50 & 3.0 \end{pmatrix}.$$

(a) Produce $n = 200$ simulations of $\mathbf{X}$.
(b) Compute $P(X_1 > 1 | X_2 = 1, X_3 = -2)$ and $P(X_1 > 1 | X_2 + X_3 = -1)$.
(c) Let $\mathbf{Y} = (X_1, X_2)^T$. Represent graphically the density contours that comprise 95% of the probability mass of $\mathbf{Y}$.