Problems 1 (revision)

Exercise 1. (Maximum likelihood)

Let $Y_1, \dots, Y_n$ be independent replicates following the distributions $Y_i \sim \mathcal{N}\left(\mu, \sigma^2/w_i\right)$, with known weights $w_i > 0$, $i = 1, \dots, n$.

a) Prove that the log-likelihood satisfies

$$\ell(\mu, \sigma^2) = -\frac{1}{2}\left\{\sum_{i=1}^n \log\left(2\pi\frac{\sigma^2}{w_i}\right) + \frac{1}{\sigma^2}\sum_{i=1}^n w_i(y_i-\mu)^2\right\},$$

and that the maximum likelihood estimators for μ and σ^2 are

$$\hat{\mu} = \bar{y}_w = \frac{1}{\bar{w}}\sum_{i=1}^n w_i y_i \text{ and } \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n w_i(y_i - \bar{y}_w)^2,$$

with $\bar{w} = \sum_{i=1}^n w_i$.

We now suppose that $Y_1, \dots, Y_n$ are independent replicates $Y_i \sim \mathcal{N}\left(X_i\beta, \sigma^2/w_i\right)$, with known weights $w_i > 0$, $X \in \mathbb{R}^{n\times p}$, X_i is the i-th row of X, and a vector of unknown parameters $\beta \in \mathbb{R}^p$.

b) Prove that the maximum likelihood estimator for β is

$$\hat{\beta} = \left(X^TWX\right)^{-1}X^TWY,$$

and find the bias and variance of this estimator.

c) Find an unbiased estimator of σ^2.

Exercise 2 (Interpreting an R output)**.** We adjust the model $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \varepsilon$ to $n = 13$ measures of cement properties. We get the following table :

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 48.19363    3.91330  12.315 6.17e-07 ***
x1           1.69589    0.20458   8.290 1.66e-05 ***
x2           0.65691    0.04423  14.851 1.23e-07 ***
x3           0.25002    0.18471   1.354    0.209
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

a) Explain in detail how the columns “`t value`” and “`Pr(>|t|)`” are computed. What does they mean ? Comment on the observed values.

b) Knowing that $\widehat{\text{corr}}(\hat{\beta}_2, \hat{\beta}_3) = -0.08911$, what is the p-value for the null hypothesis $\beta_2 - \beta_3 = 0$? With 5% significance level, is it possible to reject the null hypothesis ?

Reminder : $S^2c^{\mathrm{T}}\left(X^{\mathrm{T}}X\right)^{-1}c = \left\{\widehat{\mathrm{SE}}\left(\hat{\beta}_2\right)\right\}^2 + \left\{\widehat{\mathrm{SE}}\left(\hat{\beta}_3\right)\right\}^2 - 2\,\widehat{\text{corr}}\left(\hat{\beta}_2, \hat{\beta}_3\right)\widehat{\mathrm{SE}}\left(\hat{\beta}_2\right)\widehat{\mathrm{SE}}\left(\hat{\beta}_3\right)$, with $c = (0, 0, 1, -1)^{\mathrm{T}}$.

Exercise 3 (Automatic model selection)**.** We consider the same dataset on cement properties as in Exercice 2. The residual sum of squares (RSS) and Mallows C_p (not all of them) for the models with *intercept* are :

Model	RSS	C_p	Model	RSS	C_p	Model	RSS	C_p
- - - -	2715.8	442.58	1 2 - -	57.9		1 2 3 -	48.1	
			1 - 3 -	1227.1	197.94	1 2 - 4	48.0	
1 - - -	1265.7	202.39	1 - - 4	74.8	5.49	1 - 3 4	50.8	
- 2 - -	906.3		- 2 3 -	415.4	62.38	- 2 3 4	73.8	7.325
- - 3 -	1939.4	314.90	- 2 - 4	868.9	138.12			
- - - 4	883.9	138.62	- - 3 4	175.7	22.34	1 2 3 4	47.9	5

a) Use *forward selection* and *backward elimination* to choose a model for this dataset with a 5% confidence level. Use the F-test statistic

$$F = \frac{\mathrm{RSS}\left(\hat{\beta}_{\mathrm{L}}\right) - \mathrm{RSS}\left(\hat{\beta}_{\mathrm{L}\cup\{\mathrm{j}\}}\right)}{\mathrm{RSS}\left(\hat{\beta}_{\mathrm{full}}\right)/(13-5)},$$

to determine if the addition of the jth variable is significant.

b) Another selection criteria is Mallows C_p :

$$C_p = \frac{\mathrm{RSS}_p}{s^2} + 2p - n.$$

Note that here s^2 is the variance estimator of the full model.

i) How do we use this criterion ? Compute the missing C_p.

ii) What are the selected models with Mallows criterion, using *forward selection*, and then *backward elimination* ? What is the overall best model ?

Exercise 4 (Graphical diagnostics)**.**

a) Figure 1 shows standardized residuals for four different datasets. For each case, discuss the fit and quickly explain how we could fix possible models misspecifications.

b) Figure 2 shows four Gaussian Q-Q plots. None of the datasets follow a Gaussian law but distributions with

i) heavier tails than Gaussian law ;

ii) lighter tails than Gaussian law ;

iii) a positive skewness ;

iv) a negative skewness.

Match every case i)–iv) with a Q-Q in Figure 2. Answers must be justified.

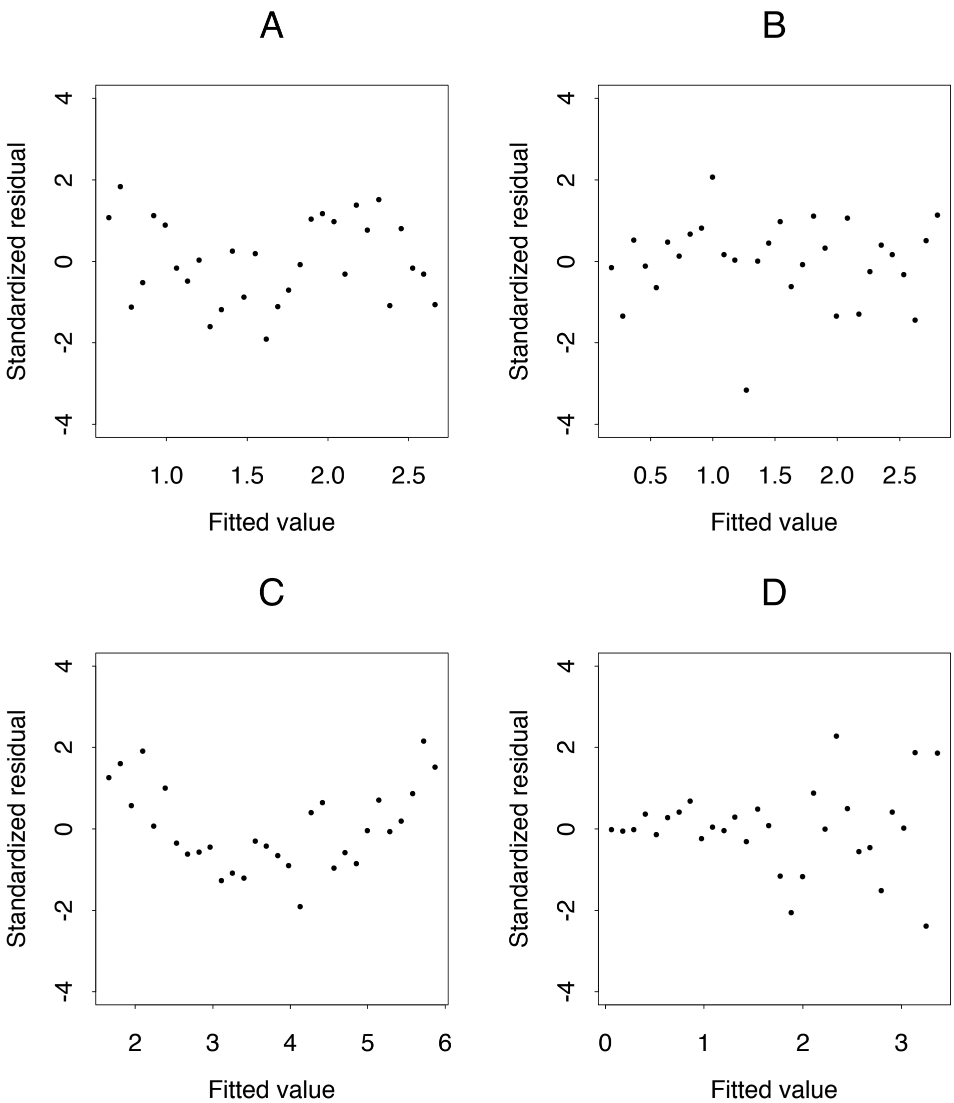

FIGURE 1 – Standardized residuals for four Gaussian linear models.

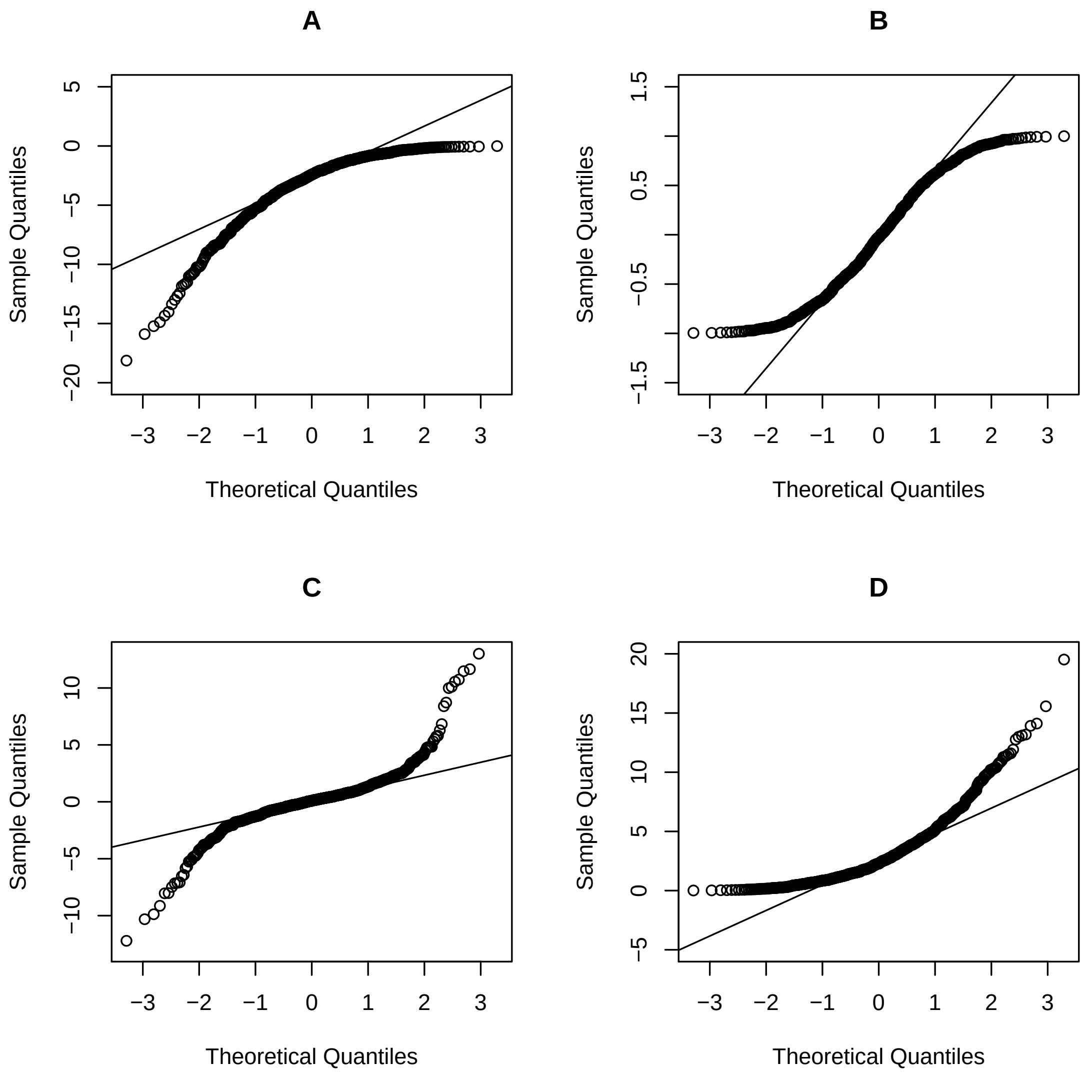

Figure 2 – Four Gaussian Q-Q plots with non-Gaussian distributed data.

Exercise 5 (Models with factors).

In R, the general formula for a model is

```
response~expression
```

where the left-hand side, `reponse`, can be missing, the right-hand side, `expression`, is a collection of terms joints by operators, and the full formula is similar to an arithmetic expression. Let

$$y = \begin{pmatrix} 217 \\ 143 \\ 186 \\ 121 \\ 157 \\ 143 \end{pmatrix}, \quad X = \begin{pmatrix} 152 & 1 & 1 \\ 93 & 1 & 2 \\ 127 & 1 & 3 \\ 109 & 2 & 1 \\ 141 & 2 & 2 \\ 136 & 2 & 3 \end{pmatrix},$$

and `x`, `a`, `b` denotes the columns of $X = [x, a, b]$.

a) A *factor* is a variable which represents a categorical observation (command `as.factor()` in R). For instance, if `a` is a factor, then `y~a` represents the model

$$y_j = \beta_0 + \alpha_1 + \varepsilon_j, \quad j = 1, 2, 3; \qquad y_j = \beta_0 + \alpha_2 + \varepsilon_j, \quad j = 4, 5, 6,$$

where β_0, α_1 et α_2 are unknown parameters. Formally, we use indicator functions :

$$y_j = \beta_0 + \alpha_1 I_{(a_j = \text{“1”})} + \alpha_2 I_{(a_j = \text{“2”})} + \varepsilon_j, \tag{1}$$

where $I_E = 1$ is E is true, and 0 otherwise. Note that the values “1” and “2” in the vector `a` does not represent the numbers 1 and 2, but categories, groups, classes or levels. For instance, `a` could represent “1” = “regular food regime”, and “2” = “food regime with growth inhibitors”.
Suppose that `a` and `b` are factors :

I. Give the design matrix for the model (1), as well as the vector of parameters.

II. Note that this matrix is *not* full rank. What is the consequence on the parameters estimation ?

III. Erase the column corresponding to α_1 to make the matrix full rank. What is now the interpretation of the parameters β_0 and α_2 ?

IV. When the model includes a constant β_0, R supresses automatically the first level of every factors. Give the design matrix for the following models :

(i) `y~a`, (ii) `y~a+b`, (iii) `y~x+a-1` (iv) `y~b+x-1`.

b) Suppose that `a` and `b` are (again) factors : an *interaction* component is represented as `a:x` or `a:b`. For instance, `y~a:x` stands for the model

$$y_j = \beta_0 + \alpha_1 x_j + \varepsilon_j, \quad j = 1, 2, 3; \qquad y_j = \beta_0 + \alpha_2 x_j + \varepsilon_j, \quad j = 4, 5, 6;$$

which can also be written

$$y_j = \beta_0 + \alpha_1 I_{(a_j = \text{“}1'')} x_j + \alpha_2 I_{(a_j = \text{“}2'')} x_j + \varepsilon_j$$

with indicator functions, i.e. a model with different slopes for the groups “1” and “2”, but with a common intercept.
Similarly, the expression `y~a:b` represents the model

$$y_j = \beta_0 + \alpha_j + \varepsilon_j, \quad j = 1, \ldots, 6;$$

which can also be written

$$y_j = \beta_0 + \sum_{i=1}^{2} \sum_{l=1}^{3} \gamma_{i,l} I_{(a_j = \text{“}i'')} I_{(b_j = \text{“}l'')} + \varepsilon_j.$$

It's a model with different *intercepts* for every level combinations of `a` et `b`.

Find the design matrices for the following models :

(i) `y~a:x`, (ii) `y~a:b` (iii) `y~a+b:x`, (iv) `y~a+a:b:x`.

Precise which matrices have linearly independent columns.

Indication : You can check you answers using the following command lines :

```
> y <- c(217,143,186,121,157,143)
> X <- matrix(c(152,93,127,109,141,136,1,1,1,2,2,2,1,2,3,1,2,3),6,3)
> df <- data.frame(y = y, x = X[,1], a = as.factor(X[,2]), b = as.factor(X[,3]))
> model.matrix(reponse~expression, data = df)
```