Statistical Analysis and Document Mining

TD1: Multiple linear regression

January 2022

Exercise 1

The dataset `swiss` available in `R` contains the following information :

```
Description:

     Standardized fertility measure and socio-economic indicators for
     each of 47 French-speaking provinces of Switzerland at about 1888.

Format:

     A data frame with 47 observations on 6 variables, _each_ of which
     is in percent, i.e., in [0,100].

       [,1]  Fertility         Ig, ‘common standardized
                               fertility measure’
       [,2]  Agriculture       % of males involved in agriculture
                               as occupation
       [,3]  Examination       % draftees receiving highest mark
                               on army examination
       [,4]  Education         % education beyond primary school for draftees.
       [,5]  Catholic          % ‘catholic’ (as opposed to ‘protestant’).
       [,6]  Infant.Mortality  live births who live less than 1
                               year.

     All variables but ‘Fertility’ give proportions of the
     population.
```

We want to study the effect of these 5 socio-economic indicators on the fertility measure.

1. Firstly, we apply a multiple regression:

```
Call:
lm(formula = Fertility ~ . , data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max
-15.2743  -5.2617   0.5032   4.1198  15.3213

Coefficients:
```

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       66.91518   10.70604   6.250 1.91e-07 ***
Agriculture       -0.17211    0.07030  -2.448  0.01873 *
Examination       -0.25801    0.25388  -1.016  0.31546
Education         -0.87094    0.18303  -4.758 2.43e-05 ***
Catholic           0.10412    0.03526   2.953  0.00519 **
Infant.Mortality  1.07705    0.38172   2.822  0.00734 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared: 0.7067,     Adjusted R-squared: 0.671
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

- Is there a relationship between `Examination` and fertility measure? In which direction?
- Is there a relationship between `Education` and fertility measure? In which direction?

2. The simple regression, using the variable `Examination` gives :

```
Call:
lm(formula = Fertility ~ Examination, data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max
-25.9375  -6.0044  -0.3393   7.9239  19.7399

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  86.8185     3.2576  26.651  < 2e-16 ***
Examination  -1.0113     0.1782  -5.675 9.45e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.642 on 45 degrees of freedom
Multiple R-squared: 0.4172,     Adjusted R-squared: 0.4042
F-statistic: 32.21 on 1 and 45 DF,  p-value: 9.45e-07
```

Is this result in adequation with the result performed by the multiple regression?
Hint: the correlation matrix for this dataset is

```
> cor(swiss)
                  Fertility Agriculture Examination   Education   Catholic Infant.Mortality
Fertility         1.0000000  0.35307918  -0.6458827 -0.66378886  0.4636847       0.41655603
Agriculture       0.3530792  1.00000000  -0.6865422 -0.63952252  0.4010951      -0.06085861
Examination      -0.6458827 -0.68654221   1.0000000  0.69841530 -0.5727418      -0.11402160
Education        -0.6637889 -0.63952252   0.6984153  1.00000000 -0.1538589      -0.09932185
Catholic          0.4636847  0.40109505  -0.5727418 -0.15385892  1.0000000       0.17549591
Infant.Mortality  0.4165560 -0.06085861  -0.1140216 -0.09932185  0.1754959       1.00000000
```

Exercise 2

The dataset `mtcars` available in `R` contains the following information :

```
Description:

     The data was extracted from the 1974 _Motor Trend_ US magazine,
     and comprises fuel consumption and 10 aspects of automobile design
     and performance for 32 automobiles (1973-74 models).

Format:

     A data frame with 32 observations on 11 variables.

       [, 1]  mpg   Miles/(US) gallon
       [, 2]  cyl   Number of cylinders
       [, 3]  disp  Displacement (cu.in.)
       [, 4]  hp    Gross horsepower
       [, 5]  drat  Rear axle ratio
       [, 6]  wt    Weight (lb/1000)
       [, 7]  qsec  1/4 mile time
       [, 8]  vs    V/S
       [, 9]  am    Transmission (0 = automatic, 1 = manual)
       [,10]  gear  Number of forward gears
       [,11]  carb  Number of carburetors
```

We would like to study the influence of these ten factors on the fuel consumption. To do so, we consider two statistical analysis with `R`:

1. We first do a multiple linear regression considering all features:

```
Call:
lm(formula = mpg ~ . , data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4506 -1.6044 -0.1196  1.2193  4.6271

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.30337   18.71788   0.657   0.5181
cyl         -0.11144    1.04502  -0.107   0.9161
disp         0.01334    0.01786   0.747   0.4635
hp          -0.02148    0.02177  -0.987   0.3350
drat         0.78711    1.63537   0.481   0.6353
wt          -3.71530    1.89441  -1.961   0.0633 .
qsec         0.82104    0.73084   1.123   0.2739
vs           0.31776    2.10451   0.151   0.8814
am           2.52023    2.05665   1.225   0.2340
gear         0.65541    1.49326   0.439   0.6652
carb        -0.19942    0.82875  -0.241   0.8122
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared: 0.869,      Adjusted R-squared: 0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

2. Then, we do a new regression with fewer predictors:

```
Call:
lm(formula = mpg ~ carb + gear + drat, data = mtcars)

Residuals:
   Min     1Q Median     3Q    Max
-8.333 -1.802  0.369  1.543  6.122

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.7848     3.8829   0.202  0.84129
carb         -2.3866     0.3786  -6.303 8.13e-07 ***
gear          3.5144     1.1553   3.042  0.00506 **
drat          3.6309     1.5395   2.358  0.02557 *
---
Signif. codes:  0 ~***~ 0.001 ~**~ 0.01 ~*~ 0.05 ~.~ 0.1 ~ ~ 1

Residual standard error: 2.985 on 28 degrees of freedom
Multiple R-squared: 0.7784,     Adjusted R-squared: 0.7547
F-statistic: 32.79 on 3 and 28 DF,  p-value: 2.656e-09
```

- Has the number of cylinders an effect on the fuel consumption? In which direction?
- Has any predictor an effect and in which way ?

Exercise 3

Let $\hat{\beta}$ the estimated coefficients obtained by multiple regression:

$$\hat{\beta} = \arg\min_{\beta} \left[\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right]$$

Consider now a new estimator $\hat{\beta}$ constrained to satisfy the relation:

$$\sum_{j=1}^{p} \hat{\beta}_j^2 \leqslant t$$

where t is a fixed positive real number. Using Lagrange multipliers, this is equivalent to write:

$$\hat{\beta} = \arg\min_{\beta} \left[\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right]$$

where λ is a fixed positive real number. (there is a bijection between values of t and λ).

1. Why is it interesting to introduce λ (or t) ?
2. Write the new estimator $\hat{\beta}$.
3. How could we choose λ?
4. Is the new estimator equivariant to scale?
5. Compare the bias and variance of the new estimator to those from the initial one.

Exercise 4

Prove Proposition 1. As a bonus, prove its Corollary.

Proposition 1 *Let U be some Gaussian vector with covariance matrix $\sigma^2 I$, E be some linear subspace, Π_E be some orthogonal projector on E (assuming the canonical dot product).*
Then $\Pi_E U$ and $\Pi_{E^\perp} U$ are two independent Gaussian vectors.

Corollary 1 *As a consequence, in the linear regression model $Y = X\beta + \varepsilon$,*

- *$X\hat{\beta}$ and $Y - X\hat{\beta}$ are independent,*
- *$\hat{\beta}$ and $\hat{\sigma}$ are independent.*

Hints: we remind that the MLE is $\hat{\beta} = (XX^T)^{-1}X^TY$ and that the minimal variance unbiased estimator of σ^2 is $\dfrac{n}{n-p-1}\displaystyle\sum_{i=1}^{n}\hat{\varepsilon}_i^2$.

Exercise 5:

Prove Proposition 2. As a bonus, prove its Corollary.

Proposition 2 *Let U be some* centred *Gaussian vector with covariance matrix $\sigma^2 I$, E be some linear subspace, $r = dim(E)$, Π_E be some orthogonal projector on E (assuming the canonical dot product).*
Then $\frac{1}{\sigma^2}\|\Pi_E U\|^2$ has a χ^2_r distribution.

Corollary 2 *As a consequence, in the linear regression model $Y = X\beta + \varepsilon$,*

$$\frac{1}{\sigma^2}\|Y - X\hat{\beta}\|^2 = \frac{n-p-1}{\sigma^2}\hat{\sigma}^2$$

has a χ^2_{n-p-1} distribution if X is a matrix in $\mathbb{R}^{n\times(p+1)}$ of full rank.
Moreover, $\dfrac{(n-p-1)\|X(\hat{\beta}-\beta)\|^2}{(p+1)\|Y - X\hat{\beta}\|^2}$ has a Fisher distribution $\mathcal{F}(p+1, n-p-1)$.

Hints: diagonalize matrix Π_E and use the basis that makes Π_E diagonal to compute $\|\Pi_E U\|^2$ and its distribution. Note that this should involve some linear isometry.