Ensimag 2nd year **Statistical Analysis and Document Mining**
MSIAM 1st year

Worksheet n°2

Exercise 1. Apply the results on multiple linear regression (with p predictors) to find as a particular case the usual results of simple linear regression (with 1 predictor):

1. The least squares estimators of β_1, β_0 and σ^2 are:

$$\widehat{\beta}_1 = \frac{C_{xY}}{s_x^2},\ \widehat{\beta}_0 = \bar{Y}_n - \frac{C_{xY}}{s_x^2}\bar{x}_n,\ \text{ and } \widehat{\sigma}^2 = \frac{n}{n-2}S^2_{Y|x} = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{\beta}_1 x_i - \widehat{\beta}_0)^2$$

2. The variance and covariance of the estimators of β_1 and β_0 are:

$$Var[\widehat{\beta}_1] = \frac{\sigma^2}{ns_x^2},\ Var[\widehat{\beta}_0] = \frac{\sigma^2}{n}\Big(1 + \frac{\bar{x}_n^2}{s_x^2}\Big) \text{ and } Cov(\widehat{\beta}_1, \widehat{\beta}_0) = -\frac{\sigma^2\bar{x}_n}{ns_x^2}$$

Indications. The notations are given in the slides "Preliminaries 2".

1. Write the simple linear model as $Y = X\beta + \varepsilon$, with $X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$.

   Compute $X^TX = n\begin{pmatrix} 1 & \bar{x}_n \\ \bar{x}_n & s_x^2 + \bar{x}_n^2 \end{pmatrix}$ and $(X^TX)^{-1} = \frac{1}{ns_x^2}\begin{pmatrix} s_x^2 + \bar{x}_n^2 & -\bar{x}_n \\ -\bar{x}_n & 1 \end{pmatrix}$.

   The OLSE of β_0 and β_1 are given by $\hat{\beta} = (X^TX)^{-1}X^TY$. The result for $\widehat{\sigma}^2$ is direct.

2. Just apply $K_{\hat{\beta}} = \sigma^2(X^TX)^{-1}$.

Exercice 2. 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 the 5 socio-economic indicators on the fertility measure.

1. First, 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 only 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 provided by the multiple regression?

Indications. This exercise is done in a regular classroom without computers, so the students cannot use `R`. Bring your own laptop and run the Exercise 2.2.R file.

First, make some simple manipulations on the swiss dataset. Second, apply the multiple regression. Explain the output of `lm` and use formulas to find again the estimations of the β_j's and σ^2, the total, residual and explained variances, the coefficient of determination, confidence intervals for the β_j's, the F-statistic and the p-value of the F-test.

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

   $\beta_{Exam} = -0.258$, $\beta_{Educ} = -0.871$. These 2 values are negative, so the first conclusion is that there exists a negative relationship between Examination and Fertility, and also between Education and Fertility. This is confirmed by the coefficients of correlation: `cor(Fertility, Examination) = - 0.646`, `cor(Fertility, Education) = -0.664`. Moreover, `cor(Examination, Education) = 0.698`. These two variables are positively correlated, it is logical.

   The p-values of the tests of "$\beta_j = 0$" vs "$\beta_j \neq 0$" are 0.315 for Examination and $2.43\,10^{-5}$ for Education. So "$\beta_{Educ} = 0$" is strongly rejected but "$\beta_{Exam} = 0$" is not. It is consistent with the fact that the only confidence interval including 0 is for Examination. Therefore, if we consider all the 5 variables, the variable Examination can be removed. This is because it is highly correlated to the other variables.

2. Is this result in adequation with the result provided by the multiple regression?

   $\beta_{Exam} = -1.011$ is still negative, so a negative relationship is still observed between Fertility and Examination. But, this time, the p-value of the test of "$\beta_{Exam} = 0$" is very small ($9.45\,10^{-7}$), so the assumption "$\beta_{Exam} = 0$" is strongly rejected. This result seems contradictory with the previous one.

   In fact, both tests are not comparable because they are not related to the same model. In the first one, not rejecting "$\beta_{Exam} = 0$" means that the 4 other variables have a significant enough predictive power on Fertility. In the second one, not rejecting "$\beta_{Exam} = 0$" would mean that the model with no predictor $Y = \beta_0 + \varepsilon$ would be acceptable. This is clearly not the case.

Exercise 3. 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).

     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
```

3 multiple regressions are performed.

```
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

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

> Xp = mtcars$carb - mtcars$gear
> summary(lm(formula = mtcars$mpg ~  Xp+  mtcars$drat))

Call:
lm(formula = mtcars$mpg ~ Xp + mtcars$drat)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5755 -1.8971  0.1869  1.3460  5.8875

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.7837     3.8960   0.201  0.84197
Xp           -2.3185     0.3748  -6.187 9.56e-07 ***
mtcars$drat   4.8041     1.1082   4.335  0.00016 ***
---
Signif. codes:  0 ~***~ 0.001 ~**~ 0.01 ~*~ 0.05 ~.~ 0.1 ~ ~ 1

Residual standard error: 2.995 on 29 degrees of freedom
Multiple R-squared: 0.7689,     Adjusted R-squared: 0.753
F-statistic: 48.26 on 2 and 29 DF,  p-value: 5.941e-10
```

1. Has the number of cylinders an effect on the fuel consumption? In which direction?

   Indications. $\hat{\beta}_{cyl} = -0.114$, so the number of cylinders seems to have a slightly negative effect on the number of cylinders. But this value is close to 0. 0 belongs to the confidence interval for $\hat{\beta}_{cyl}$ or, equivalently, the null hypothesis of the test of $\beta_{cyl} = 0$ is not rejected since the p-value 0.916 is very large. So it is not possible to conclude that the number of cylinders has a significant effect.

2. Has any predictor an effect ?

   Indications. The same conclusion holds for all predictors since all the p-values are high. Except for wt, they are all greater than 23%. So none of the hypotheses $\beta_j = 0$ will be rejected. However, the global model is considered as significant since the p-value of the F-test is very small ($3.79\,10^{-7}$). This means that each of the predictors considered alone has no significant effect, it is the conjonction of the predictors taken altogether which has a significant effect. Here again, this is linked to the correlation between the predictors. For instance, `cor(cyl,disp) = 0.902` and `cor(hp,vs)=-0.723`.

3. To decrease the fuel consumption of a car, is it more efficient to delete one carburetor (`carb`), or to increase the number of forward gears (`gear`) ?

   Indications. $\hat{\beta}_{carb} = -0.119 < 0$, so decreasing carb will increase mpg and reduce the fuel consumption. Same reasoning for $\hat{\beta}_{gear} = 0.655 > 0$. Since carb and gear are integers and $\hat{\beta}_{gear} > |\hat{\beta}_{carb}|$, it will be more efficient to add a gear than to delete one carburetor.

   The second regression wants to explain mpg using only carb, gear and drat. In fact, drat has been kept because, from cross-validation, it is the most significant variable. The conclusion on the estimates $\hat{\beta}_{carb}$ and $\hat{\beta}_{gear}$ is the same as before.

   The third regression wants to explain mpg using carb-gear and drat. $\hat{\beta}_{carb-gear} = -2.32 < 0$, so in order to increase mpg, one has to decrease carb-gear, i.e. to reduce gear or increase carb. Same conclusion as before.

   In this model, it is possible to test $\beta_{carb-gear} = 0$, which is equivalent to testing $\beta_{gear} = \beta_{carb}$ in the previous model. The p-value of this test is very small ($9.56\,10^{-7}$), so this hypothesis is rejected. Therefore, reducing gear and increasing carb are not equivalent and it is better to add a gear.