ISEP
IF.1105 - IF.1205 - IF.2301 - Data Science
January 2023

Linear Regression

1 Part I : Theoretical questions

The simple linear regression allows to explain a quantitative variable Y by only one regressor X :

$$Y = \beta_0 + \beta_1 X + \epsilon \qquad \text{with} \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

The variable X is assumed to be deterministic whereas Y and ϵ are random

To estimate the parameters of the model β_0 and β_1 and σ, we have n observations $(y_1, x_1), \ldots, (y_n, x_n)$. The variables ϵ_i are assumed to be independent and follow the same law. The least squares estimators of β_0 and β_1 are defined by :

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

the estimator of σ^2 is given by $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-2}$.

1. What is the law of the random variables $(\hat{\beta}_0, \hat{\beta}_1)$?
2. Are the variables Y_i independent ? Are they identically distributed ? Justify your answer
3. Give a practical example where this model can be used. What does the target variable Y represent ? How about the regressor X ?

2 Part II : Practical applications

2.1 Simple Linear Regression

2.1.1 Preliminary study

In this section, you will fit a linear model to predict the ozone content according to other meteorological variables. In moodle, you will find a text file named *ozone.txt*. This dataset file contains the following variables :

— the maximum content of ozone measured during the day (`maxO3`),
— the temperature at noon (`T12`),
— the nebulosity at noon (`Ne12`) and
— the maximum content of ozone the day before (`maxO3v`).

Exercise 1. Open the text file and save it in your working directory. Import the data into a data frame that you will name `oz`.

- In R use the read.table command. You can use the command attach(oz) to be able to use directly the variable names.
- In Python, use the function read_csv() function from the pandas library.

How many observations does this dataset contain ? How many variables are there ?

Exercise 2. Calculate descriptive statistics for all the variables in the dataset :

- In R you can use the function summary() to produce a descriptive statistics of each variable. In addition, use the plot() function to plot the scatter plot of the variables.
- In Python, the command oz.describe() allows to calculate descriptive statistics. In addition, the function scatter_matrix() of the *pandas* library.

1. What are the values of the mean, the quartiles, the maximum and minimum of the maximum content of ozone during the day ? In your own words interpret the scatter plot.

2. Now, you will calculate the correlation matrix for all the features in the dataset.
   Which variable is correlated the most with the ozone content maxO3 ? Interpret the correlation coefficient.

- In R, use the function cor() to calculate the empirical correlation matrix.
- In Python, the command oz.corr() allows to calculate the empirical correlation matrix.

2.1.2 Fit of the model and goodness of fit

Exercise 3. In this exercise you will fit a simple linear model with only one regressor, the *nubelosity*, to predict the *ozone content*.

- In R Using the lm() function to fit the model : maxO3 $= \beta_0 + \beta_1$Ne12 $+ \epsilon$. You will denote the output of the lm function oz.regsimple.
- In Python, you will use the module statsmodels, a module that supports specifying models using R-style formulas and pandas DataFrames :

```
import statsmodels.api as sm
lm = sm.OLS.from_formula('maxO3 ~ Ne12', oz)
oz_regsimple = lm.fit()
print oz_regsimple.summary()
```

What are the coefficients estimates ? Interpret coefficient estimate $\hat{\beta}_1$.
In addition you can draw the scatter plot between maxO3 and Ne12 and add the regression line just fitted. To this end, in R you can use the plot() and the abline() functions. In Python You can use the function to Use the abline command to add to the scatter plot. Whereas in Python you can use the code matplotlib library :

```
import matplotlib.pyplot as plt
plt.scatter(oz.Ne12, oz.maxO3)
plt.plot([x,x], [x,x], 'k-', color = 'r') #plot 2 points of the fitted line
```

Exercise 4. Confidence intervals (CI) for the parameters

Give the general expression of a $(1-\alpha)$ confidence interval for the parameter β_1. Calculate the 90% confidence interval for this coefficient. Interpret the results.

- In R the confint() command allows to calculate confidence intervals for the model parameters.

- In Python you can use the command `oz_regsimple.conf_int(alpha= )`

Exercise 5. You can obtain a summary report of the fitting by running :

- the command `summary(oz.regsimple)` in R.
- the command `oz_regmult.summary()` in Python and printing the output.

You will obtain the following table :

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	$\hat{\beta}_0$	$\hat{\sigma}_{\hat{\beta}_0}$	statistic t_{β_0}	p-value of the test $H_0 : \beta_0 = 0$
Ne12	$\hat{\beta}_1$	$\hat{\sigma}_{\hat{\beta}_1}$	statistic T_{β_1}	p-value of the test $H_0 : \beta_1 = 0$

1. What can you say about the estimated variances of the estimated coefficients ?
2. Elaborate the zero slope hypothesis test and conclude if there is a relationship between the between the maximum content of ozone the day before maxO3 and the nebulosity Ne12. Is β_1 significantly non zero ?
3. Interpret the value of R^2. What can you say about the quality of the model ?

2.2 Multiple Linear regression

In this section you will perform linear regression with more than one regressor and compare this model to that obtained by fitting simple linear regression.

Exercise 6. Fit the regression model on two predictors : the nebulosity Ne12 and the maximum content of ozone the day before maxO3v.

$$\texttt{maxO3} = \beta_0 + \beta_1 \texttt{Ne12} + \beta_2 \texttt{maxO3v} + \epsilon$$

You can use the same functions used for simple linear regression with the formula :

`'maxO3 ∼ Ne12 + maxO3v'`

You will call the output of the function `oz.regmult`.

What are the coefficient estimates ? Give an interpretation to these coefficients. Perform the zero slope hypothesis test.

2.2.1 Making predictions

Exercise 7. We would like to predict the ozone content for tomorrow. Today the maximal ozone content is 80 and the weather forecast estimates the nebulosity tomorrow at noon will be 6.

What is the predicted value of maxO3 ? Use the simple and the multiple regression models previously fitted. Compare both results.

- In R the `predict.lm()` function predicts the estimated values of the target variable for new values of the regressors. It takes as input value at least an object of type `lm()` and a data frame with the new values. To create a new dataframe you can run the command `new.data <- data.frame(Ne12=6,maxO3v=80)`.
- In Python use the `predict()` function after creating a dataframe with the new values using the `DataFrame()` function of the `pandas` library.

2.3 Coefficient of determination R^2

In the lecture we defined the coefficient of determination R^2 and the adjusted coefficient of determination $\bar{R}^2$ to evaluate the goodness of fit of our model.

Exercise 8.

1. What do the coefficients R^2 and $\bar{R}^2$ measure ? In your opinion, which one is more adapted to compare the models `oz.regsimple` to the `oz.regmult` ?

2. The summary report gives you the R^2 and the $\bar{R}^2$ coefficients of your models.
   Which model would you choose to predict the ozone content for tomorrow ?