TP : Linear regression

- Linear regression model -

In first part (Steps 1–8) of this tutorial we consider the following fixed-design one-dimensional ($p = 1$) linear regression model :

$$Y_i = \beta_0 + \beta_1 \; x_i + \epsilon_i \,, \quad \epsilon_i \sim \mathcal{N}(0, \sigma) \quad \text{i.i.d.} \,, \quad i = 1, ..., n \,.$$

Being a particular but simply interpretable case it facilitates intuitive understanding and enables easy two-dimensional visualization. (Later, in Steps 9–13, we regard a higher-dimensional case.) ML estimators for $\hat{\beta}_0$ and $\hat{\beta}_1$ are :

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

with their respective variances being :

$$\mathbb{V}[\hat{\beta}_0] = \hat{\sigma}^2 \Big(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \Big) \quad \text{and} \quad \mathbb{V}[\hat{\beta}_1] = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \,,$$

where $\hat{\sigma}^2$ is the unbiased estimator for σ^2 :

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \big(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \big)^2 \,.$$

Further, one can show that for any $j = 0, ..., p$ it holds

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\mathbb{V}[\hat{\beta}_j]}} \sim t\big(n - (p+1)\big) \,.$$

This allows to construct a Student-t test with the null hypothesis being $\mathcal{H}_0 \,:\, \beta_j = 0$ in which one should reject $\mathcal{H}_0$ at the level α if :

$$\frac{\hat{\beta}_j}{\sqrt{\mathbb{V}[\hat{\beta}_j]}} \notin [-t_{1-\frac{\alpha}{2}}^{\left(n-(p+1)\right)}; t_{1-\frac{\alpha}{2}}^{\left(n-(p+1)\right)}] \,,$$

where $t_{1-\frac{\alpha}{2}}^{\left(n-(p+1)\right)}$ is the quantile of the Student-t distribution with $\big(n - (p+1)\big)$ degrees of freedom at the level $1 - \frac{\alpha}{2}$. For a statistical test, the p-value is the probability that, under $\mathcal{H}_0$, the test statistic takes the value at least as extreme as its observed value.

Using the estimated coefficients one obtains the regression line $(\hat{\beta}_0 + \hat{\beta}_1 \; x)$, which is the mean of Y for any $x \in \mathbb{R}$. This is called the point prediction. Usually, one is also interesting in the inferential information, *i.e.* in a confidence interval on $\hat{\beta}_0 + \hat{\beta}_1 x$ and on Y (the last one is often called prediction interval as it reflect the variance of Y as well). One can show that

$$\frac{(\hat{\beta}_0 + \hat{\beta}_1 x) - (\beta_0 + \beta_1 x)}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}} \sim t(n-2) \quad \text{and} \quad \frac{Y - (\hat{\beta}_0 + \hat{\beta}_1 x)}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}} \sim t(n-2) \,.$$

This allows for construction of confidence and prediction intervals (at the level α) :

$$\mathrm{CI}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \,\pm\, t_{1-\frac{\alpha}{2}}^{(n-2)} \, \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}} \,,$$

$$\mathrm{PI}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \,\pm\, t_{1-\frac{\alpha}{2}}^{(n-2)} \, \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}} \,.$$

The second part (Steps 9–13) of the tutorial deals with the general linear regression model (in the obvious notation) :

$$Y_i = \beta_0 + \beta_1\ x_{i,1} + \cdots + \beta_p\ x_{i,p} + \epsilon_i\,, \quad \epsilon_i \sim \mathcal{N}(0,\sigma) \quad \text{i.i.d.}\,, \quad i = 1, ..., n\,.$$

Here, we shortly generalize those of the above mentioned material relevant for performing this tutorial. Denote the design matrix $X = \big(x_1, ..., x_n\big)^\top$ with $x_i = (1, x_{i,1}, ..., x_{i,p})^\top$ and let $C = \hat\sigma^2 (X^\top X)^{-1}$ where

$$\hat\sigma^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} \big(Y_i - (\hat\beta_0 + \hat\beta_1 x_{i,1} + \cdots + \hat\beta_p x_{i,p})\big)^2\,.$$

Then one can show that

$$\begin{aligned}
&\mathbb{V}[\hat\beta_j] = C_{jj}\,,\\
&\mathrm{CI}(x_\cdot) = \hat\beta_0 + \hat\beta_1\ x_{\cdot,1} + \cdots + \hat\beta_p\ x_{\cdot,p} \pm t_{1-\frac{\alpha}{2}}^{\big(n-(p+1)\big)} \sqrt{\hat\sigma^2 x_\cdot^\top (X^\top X)^{-1} x_\cdot}\,,\\
&\mathrm{PI}(x_\cdot) = \hat\beta_0 + \hat\beta_1\ x_{\cdot,1} + \cdots + \hat\beta_p\ x_{\cdot,p} \pm t_{1-\frac{\alpha}{2}}^{\big(n-(p+1)\big)} \sqrt{\hat\sigma^2 \big(1 + x_\cdot^\top (X^\top X)^{-1} x_\cdot\big)}\,.
\end{aligned}$$

Analysis of the investment data If needed, a tutorial on `pandas` can be helpful : http://pandas.pydata.org/pandas-docs/stable/tutorials.html Let us use the **Investment Data Set** [1] downloadable from the institution web site as a CSV (blank separated) file **"`invest.txt`"**. Before going to the Steps of this tutorial, please take a look at the preceding it preliminary part, which suggests a declarative overview of several notions connected to the linear regression model.

1) Import the data from the file "`invest.txt`" and print them in a readable form, *e.g.* a table containing first 5 observations.
2) Plot the data with Gross National Product (GNP, column "gnp") being the abscissa and Investment (column "invest") being the ordinate.

NOTE : When working with monetary data, one often resorts to a logarithm transform to account for inequality of scale. First, transform the two above mentioned columns, GNP and Investment, via logarithm ; further in this exercise we will be working with these columns log-transformed. In Steps 3–6, all the calculations should be done with elementary arithmetic and not involving existing libraries for running linear regression.

3) For the linear regression of Investment on GNP, estimate the intercept and the slope and their standard deviations, as well as the determination coefficient for the ordinary least squares. Output them in a readable form.
4) Test the significance of the slope using the Student-t test. Report the value of the test statistic and the p-value.
5) For the GNP value 1000, estimate the necessary Investment, provide confidence and prediction intervals for the 90% level.
6) On a plot with logarithmic axes (GNP as abscissa and Investment as ordinate), plot the data, the estimated regression line, the confidence and prediction intervals for all values of log(GNP) between its minimum and maximum in the data set.
7) Estimate the intercept, the slope, and the determination coefficient and predict the necessary Investment for the GNP value 1000 using existing functionality. The class `LinearRegression()` from `sklearn.linear_model` is suggested but not obligatory ; any other available implementation of the linear regression can be used instead. Report the estimated values and make sure that those calculated 'by hand' (Steps 3 and 5) coincide with the ones obtained using existing implementation.
8) On a plot with logarithmic axes (GNP as abscissa and Investment as ordinate), plot the data, the regression line and the predicted point (in a different color). The graphic should coincide with the corresponding elements from the one in Step 6.

NOTE : Further, consider an additional explanatory variable, namely Interest (column "interest", without a logarithmic transform). In Steps 9–12, all the calculations should be done with elementary arithmetic and not involving existing libraries for running linear regression. (Use function `inv` from `numpy.linalg` for inversion of a matrix (= two-dimensional `numpy`-array) and function `eig` from the same package for calculating its eigenvalues.)

1. See Greene (2012) - *Econometric Analysis*, Prentice Hall, Upper Saddle River, NJ.

9) For the linear regression of Investment on GNP and Interest, compute the associated Gram matrix. Is it of full rank ?

10) For the linear regression of Investment on GNP and Interest, estimate the three regression coefficients and their standard deviations, as well as the determination coefficient for the ordinary least squares. Additionally, test significance of each coefficient using the Student-t test. Report the regression coefficients, corresponding to them p-values, and the determination coefficient in a readable form. Discuss significance of the estimated regression coefficients.

11) For the values of GNP = 1000 and Interest = 10, predict the log(Investment) and provide confident and prediction intervals at the 99.9% level.

12) On a same 3D-plot with axes being log(GNP), Interest, and log(Invetment), draw data points, their predictions, regression plane and surfaces of the 99.9% confidence intervals for each pair of values of log(GNP) and Interest between their minimum and maximum.

13) Estimate the regression and determination coefficients and predict the necessary log(Investment) for GNP = 1000 and Interest = 10 using existing functionality. Again, the class `LinearRegression()` from `sklearn.linear_model` is suggested but not obligatory ; any other available implementation of the linear regression can be used instead. Report the estimated values and make sure that those calculated 'by hand' (Steps 10 and 11) coincide with the ones obtained using existing implementation.