Ecole Centrale Casablanca

Apprentissage Statistique et Réseaux de Neurones

TP 1 : Régression linéaire

Auteur :
Ibrahim RESMOUKI

25 octobre 2020

Introduction

Dans le cadre du cours d'"**Apprentissage statistique et réseaux de neurones**", j'ai été amené à étudier dans ce premeir TP, le modèle de régression linéaire. Le TP est structuré en deux parties, une première où je genère mes propres données et où je compare le calcul effectué par les fonctions prédéfinies de R et le calcul manuel à travers les formules vues en cours ; puis une deuxième partie où je travaille sur des données réelles.

1 Génération du modèle

Je commence par simuler mes variables, représentées par ma matrice X qui est une matrice de taille (n,p), générée à partir d'une loi normale centrée ou exponentielle, suivant si je veux que ma matrice contienne des valeurs positives et négatives ou que positives.Puis, je "force" ma varible cible Y à vérifier l'équation :

$$Y = X\beta + \epsilon$$

. Il faut aussi que X soit de pkein rang i.e. $rg(X) = p$ et que le vecteur $(1, 1, .., 1) \in \mathbb{R}^n$ ne soit pas dans l'image de X. Pour cela, le rang de X doit être égale à p et le rang de la matrice formée par X à laquelle $(1, 1, .., 1)$ est accolé doit être égale à p + 1. En effet, il faut et il suffit que la famille de vecteurs de X et $(1, 1, .., 1)$ soit libre.

Ici, ϵ représente un bruit blanc gaussien de variance 2 et $\beta = (-2, 7)$ représente les coefficients de notre régression linéaire.

Une première intution nous pousse à vouloir visualiser les matrices que nous avons créer. En suivant les directives du TP, on commence par subdiviser uniformément l'axe des abssices en prenant :

$$x = \frac{max(X_i) - min(X_i)}{n + max(X_i)} * [1 : n]$$

En effet, cela permet d'avoir $n = 100$ valeurs comprises entre $min(X_i$ et $max(X_i)$. Les figures suivantes permettent de visualiser Y en fonction des colonnes de X suivant si X a été générée par une loi normale centrée ou exponentielle.

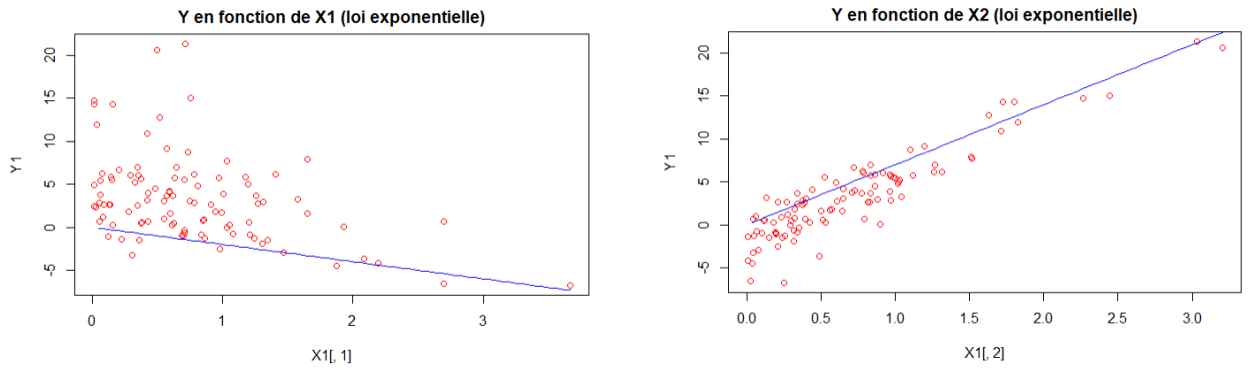

FIGURE 1 – Plot de Y par rapport à X_1 *et* X_2 (loi exponentielle)

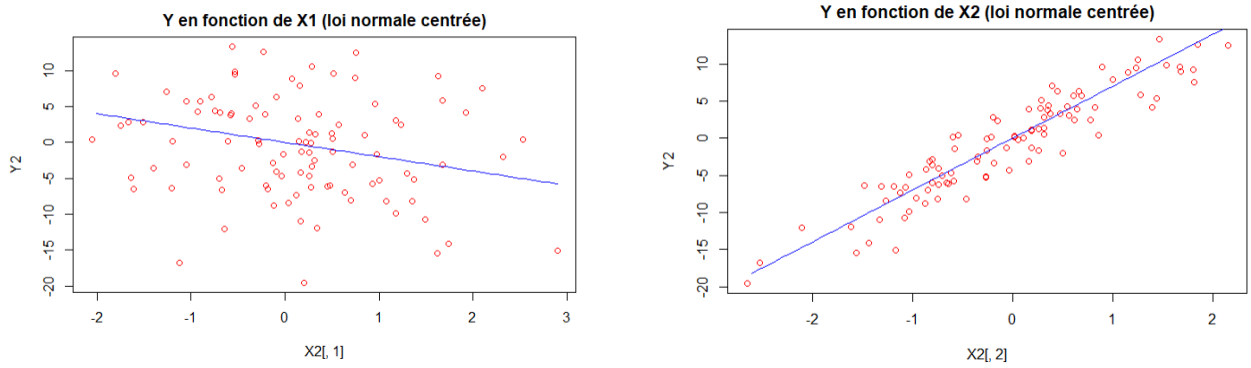

FIGURE 2 – Plot de Y par rapport à X_1 *et* X_2 (loi normale centrée)

Intérprétation des figures 1 et 2

Il ne faut pas bien évidemment oublier le faire que notre régression linéaire consiste en un plan séparateur et non pas une droite. La visualisation de Y en fonction de chacune des colonnes nous donne une vision projetée, ce qui explique la première figure où on remarque un décalage de la droit tracée par rapport aux données.

Calcul de $\hat{\beta}$

Je continue le TP avec X générée à partir de la loi exponentielle. J'essaye à présent d'estimer la valeur de β à partir de la formule vue en cours :

$$\hat{\beta} = (X^t X)^{-1} X^t Y \tag{1}$$

J'obtiens alors $\hat{\beta} = (-1.89, 6.85)$. L'estimation du paramètre n'est pas parfaite même si l'équation (1) fourni une solution analytique pour trouver $\hat{\beta}$. Cela vient principalement du bruit gaussien ϵ que nous avons rajouté.

Fonction prédéfinie lm

En utilisant la fonction *lm* on trouve que $\hat{\beta} = (0.06, -2.36, 7.16)$. La fonction prend en compte l'existence d'une potentielle ordonnée à l'origine β_0. Sa valeur est proche de 0 et $P(|T_0| > t) = 0.81$. On ne peut donc pas rejeter l'hypothèse $H_0 : \beta_0 = 0$.

Dans la suite du TP, $X \in \mathbb{R}^{nx(p+1)}$ puisqu'on accole à X le vecteur colonne $(1, 1, .., 1) \in \mathbb{R}^n$ qui permet de rendre compte de l'ordonnée à l'origine β_0.En recalculant $\hat{\beta}$ on obtient les mêmes valeurs que celles de la fonction lm.

Ècart-type du bruit gaussien

$\hat{\sigma}$ est un estimateur de l'écart-type du bruit gaussien en effet, son équation est :

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

Avec :

$$||Y - \hat{Y}|| = ||X\beta + \epsilon - X\hat{\beta}||$$

Où $\hat{Y}$ est le vecteur des valeurs prévues
Nous obtenons $\hat{\sigma} = 1.30925$. Le résultat est en effet très proche de $\sigma = 1.4$.

Coefficient de corrélation empirique

R^2 est le coefficient de corrélation empirique entre les valeurs prévues $\hat{Y}$ et les valeurs observées Y, donné par :

$$R^2 = \frac{||\hat{Y} - \bar{Y}||^2}{||Y - \bar{Y}||^2}$$

Ce coefficient prend au maximum 1 et au minimum 0. Plus il est élevé plus les prédicitions du modèle $\hat{Y}$ sont corrélées avec Y. J'obtiens $R^2 = 0.93$. Le modèle est en effet très performant puisque nous avons "forcer" nos données à suivre l'équation $Y = X\beta + \epsilon$.

Estimateur de l écart-type de l'estimateur $\hat{\beta}$

$\hat{\sigma_\beta}$ est un estimateur de l'écart-type de $\hat{\beta}$, c'est un vecteur de $\mathbb{R}^p$, donné par :

$$\hat{\sigma_{\beta_j}} = \hat{\sigma^2}[(X^tX)^{-1}]_{jj}$$

Notre grandeur simulée est $\hat{\sigma_\beta} = (0.03, 0.03, 0.05)$, même valeur donnée par *summary(res)*. A noter que l'expression de $\hat{\sigma_\beta}$ fait intervenir $(X^tX)^{-1}$. Notre estimateur deviendrait donc très imprécis si la matrice X est mal conditionnée !

2 Données réelles

Dans cette seconde partie, nous allons utiliser deux jeux de données réelles : **Boston Housing Data** et **Forest Fires** ; sur lesquelles nous allons appliquer la régression linéaire.

Boston Housing Data

Cette base de données décrit les villes de la banlieue de Boston en prenant 14 différentes features. Elle contient 507 observations.
Nous modélisons donc le nombre de crimes par habitant en fonction des différentes informations sur les villes, et nous appliquons la régression linéaires sur nos données.
À l'aide la fonction *lm*, nous effectuons une régression linéaire du taux de criminalité par habitant par ville (**CRIM**) en fonction du reste des informations de notre base de données.
Nous récapitulons les résultats de cette régression dans le tableau suivant :

Critère	**Description**
R^2	0.4536
Intercept	17.13
$Pr(>\|t\|)$ de l'*Intercept*	0.018218
PTRATIO	-0.266
$Pr(>\|t\|)$ de *PTRATIO*	0.153686
p-value du test de Fisher	$< 2.2.10^{-16}$

Ayant obtenu un $R^2 = 0.4536$ très faible, il est clair que notre modèle n'est pas adapté pour notre modélisation.
On ne peut pas non plus dire que le β_j, associé au ratio élèves/enseignants est significativement non nul, car on voit que $Pr(> |t|) > 0.05$.

Forest Fires

La seconde base de données contient différentes grandeurs physiques et indices forêt météo (**FWI**) relevées lors d'incendies.
Nous cherchons ici à expliquer la superficie brûlée en fonctions des autres données relevées.
Dans un premier temps, nous effectuons la régression linéaire de $area$ en fonction des autres valeurs numériques de la base de données. Nous obtenons une $p - value$ inférieure à 0.01 pour deux facteurs : DMC et DC.
Quand nous effectuons la régression sur $log(1 + area)$, il ne ressort que le facteur DMC.
On peut donc conclure que c'est le facteur DMC qui est le plus important.