Modèles Linéaires

TD 2 : Régression linéaire simple
Licence 3 MIASHS

Guillaume Metzler
Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France
guillaume.metzler@univ-lyon2.fr

Résumé

Dans la précédente fiche, nous avons étudié trois façons différentes d'obtenir les solutions à notre problème de régression linéaire simple.
Dans la présente fiche, nous allons maintenant nous intéresser à la propriété de ces estimateurs et tester l'influence de la variable explicative sur la variable à expliquer.
Plus précisément :

- on étudiera le biais et la variance de la pente du modèle, donné par $\hat{\beta}_1$
- on étudiera le biais et la variance de l'ordonnée à l'origine du modèle $\hat{\beta}_0$
- on construira des intervalles de confiance sur ces derniers
- on testera la significativité de la pente du modèle.

Modèle de régression simple

On rappelle que le modèle linéaire gaussien simple s'écrit sous la forme

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

où Y est la variable à expliquer, X est la variable explicative et ε est une variable aléatoire représentant les erreurs du modèle que l'on supposera normalement distribuée avec une variance inconnue égale à σ^2
L'estimation des paramètres se fait à l'aide d'un échantillon $S = \{(y_i, x_i)\}_{i=1}^n$.

Dans ce TD, on se concentre sur l'étude des paramètres du modèle et nous appliquerons ensuite cela sur les données présentées ci-dessous, pour obtenir la droite de régression associée, présentée en Figure 1.

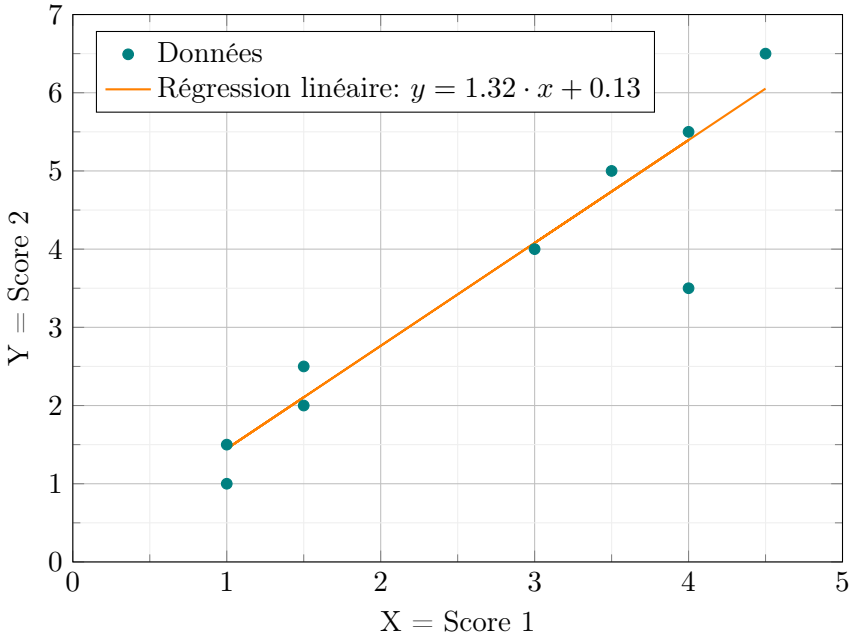


FIGURE 1 – Application de la régression linéaire simple gaussien sur les données présentées dans la table associée. On cherche alors à expliquer le score obtenu au deuxième examen en fonction du score obtenu au premier examen.

Y : Score examen 2	3.5	4	5	1	2	1.5	2.5	5.5	6	6.5
X : Score examen 1	4	3	3.5	1	1.5	1	1.5	4	3.5	4.5

On rappelle que les paramètres du modèle de régression sont données par les relations

$$\hat{\beta}_1 = \frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[X]} \quad \text{et} \quad \hat{\beta}_0 = \mathbb{E}[Y] - \hat{\beta}_1 \, \mathbb{E}[X].$$

On utilisera également le fait que la variance des erreurs σ^2 peut être estimée par

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Régression linéaire avec R

Pour effectuer la régression linéaire avec R, vous pouvez utiliser le code suivant

```r
# Pour charger un jeu de données
data = read.csv("../data/reglin.csv", sep=";")
# Régression linéaire
mymodel = lm(Y~X,data)
```

Un résumé statistiques de la régression linéaire peut être obtenu à l'aide de la commande

```r
summary(mymodel)
```

Notre objectif sera d'expliquer comment ces valeurs sont obtenues au fur et à mesure des séances.
On pourra extraire les coefficient de la régression à l'aide de la commande

```r
coeff = mymodel$coefficients
```

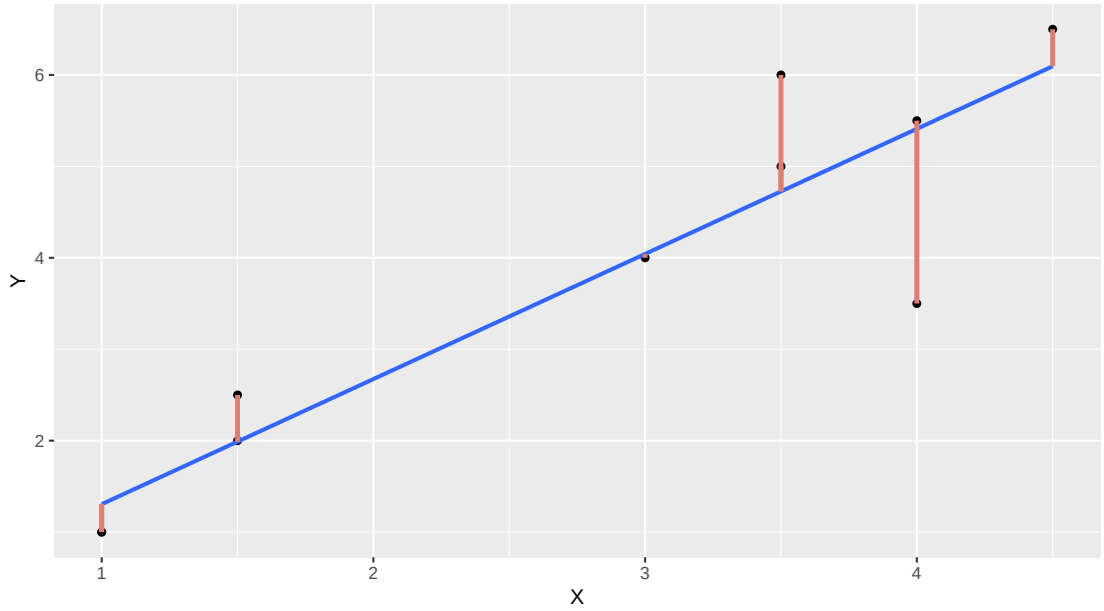
De la même façon, nous pourrons obtenir les résidus de la régression comme suit :

```r
mymodel$residuals
```

On pourra enfin représenter nos données, la droite de régression, ainsi que les résidus graphiquement

```r
# Graphical Representation of the Model and Residuals

    library(ggplot2)
    ggplot(data, aes(x=X, y=Y)) +
    geom_point() +
    geom_smooth(method=lm,se=FALSE) +
    geom_segment(aes(x = X, y = Y, xend = X,
                     yend = coeff[1] + coeff[2]*X,
                     col = "Residuals"),
                     col = "#DF7D72", lwd= 1.2, data = data)
```



Etude des propriétés de la pente du modèle

On cherche à étudier le biais et la variance de cet estimateur.

1. Justifier que pour tout entier $i \in [\![1, n]\!]$, nous avons

$$\mathbb{E}[y_i] = \beta_0 + \beta_1 x_i.$$

2. En déduire la relation

$$\mathbb{E}[\bar{y}] = \beta_0 + \beta_1 \bar{x},$$

où $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ et $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

3. A l'aide de l'expression de $\hat{\beta}_1$, montrer que ce dernier est un estimateur sans biais de β_1.
4. A l'aide des relations suivantes

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{et} \quad \bar{y} = \beta_0 + \beta_1 \bar{x}.$$

Montrer que l'on a

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

5. Dans la suite, on posera $\omega_i = \dfrac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$. En déduire que la variance de l'estimateur $\hat{\beta}_1$ est égale à

$$\mathrm{Var}[\hat{\beta}_1] = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Etude des propriétés de l'ordonnée à l'origine du modèle

Nous reprenons le même travail avec l'ordonnée à l'origine cette fois-ci.

1. Montrer que l'on

$$\hat{\beta}_0 = \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x}.$$

2. En déduire que $\hat{\beta}_0$ est un estimateur sans biais de β_0.
3. Déterminer la variance de $\hat{\beta}_0$.

Intervalles de confiances

On peut montrer que les variables aléatoires suivantes

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\mathrm{Var}[\hat{\beta}_0]}} \quad \text{et} \quad \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\mathrm{Var}[\hat{\beta}_1]}}$$

suivent une loi de Student à $n - 2$ degrés de libertés, la valeur 2 étant liée aux nombres de paramètres du modèle.

1. Donner l'expression des intervalles de confiance de niveau $1 - \alpha$ des paramètres β_0 et β_1.
2. Effectuer l'application numérique en prenant $\alpha = 0.05$ et en utilisant les données de l'énoncé.

Test de la significativité du modèle

On cherche enfin à savoir si le modèle appris est significatif. Pour cela on va tester la significativité de la pente du modèle et regarder si cette dernière est significativement différente de 0.
Cette vérification nous permettra d'affirmer que la variable X permet, au moins en partie, d'expliquer les valeurs de la variable aléatoire Y.

1. A l'aide de l'intervalle de confiance précédemment construit, peut-on dire que la pente du modèle est significative ?

2. On peut également procéder au test statistique suivant :

$$H_0 : \beta_1 = 0 \text{ versus } H_1 : \beta_1 \neq 0.$$

Après avoir déterminé la valeur de la statistique de test, déterminer la p-value associée au test et conclure au risque d'erreur $\alpha = 0.05$.