Régression Linéaire

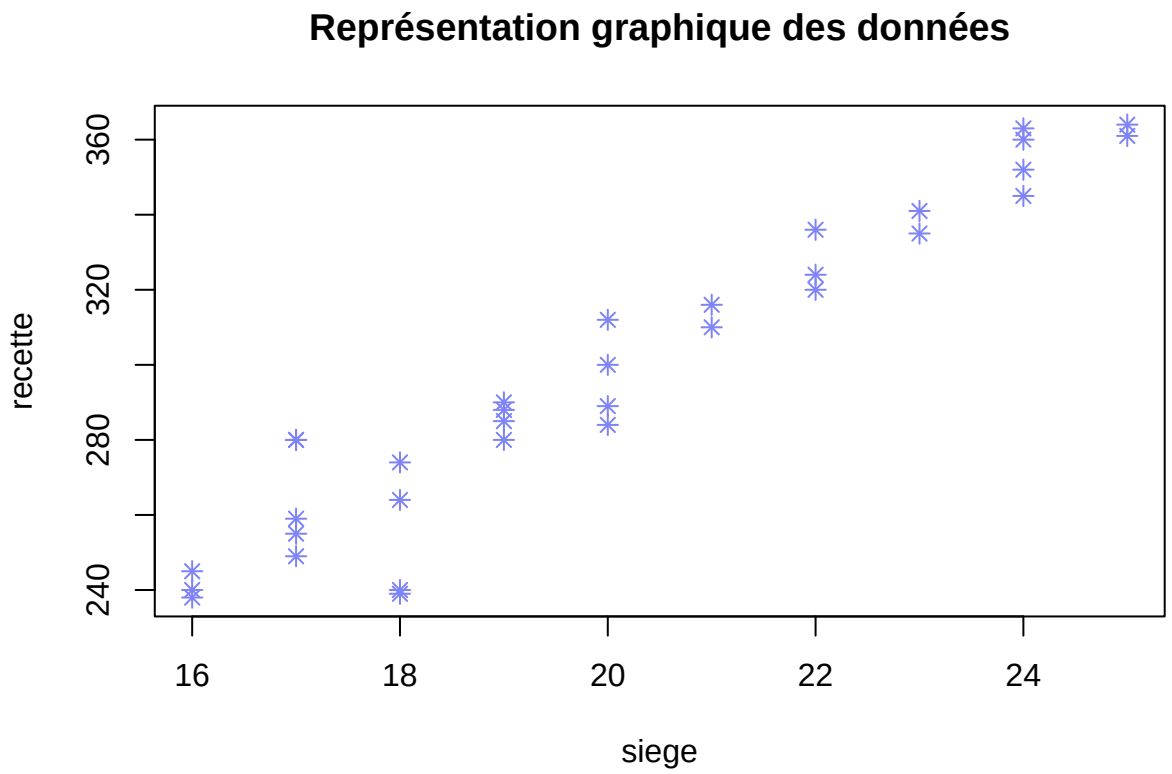
Guillaume Metzler

11/23/2021

```
library(readxl)
Recette_Bar <- read_excel("Recette_Bar.xlsx")
data <- Recette_Bar[,-1]
colnames(data)= c("siege","recette")
```

On va d'abord regarder ce que donne nos données, en faisant uen représentation graphique sur laquelle le montant des recettes est donné en fonction du nombre de siège dans le bar.

```
plot(data, main = "Représentation graphique des données", pch = 8, col = "#7e83f7")
```



Première constatation très simple, le nombre de siège semble avoir un impact sur le montant des recette générés par le bar, *{i.e.} plus le nombre de siège occupé est élevé plus les recette sont élevées.*

On cherche maintenant à construire un modèle qui approximera le mieux cette tendance que l'on observe les données, à l'aide d'un modèle linéaire simple. Des commandes permettent de faire cela simplement à l'aide du logiciel. C'est ce que nous allosn faire dans un premier temps, puis nous chercherons à retrouver les paramètres du modèle "à la main".

Pour cela, on va chercher à estimer les paramètres à l'aide d'un modèle dit Gaussien, qui va prendre la forme suivante

$$Y = aX + b + \varepsilon,$$

où Y représente la variable réponse, c'est-à-dire le montant des recettes dans notre cas, X la variable explicative, c'est-à-dire celle qui va nous servir à expliquer les valeurs de Y, c'est le nombre de sièges. Les paramètres a et b sint les paramètres de notre droite et ε est ce que l'on appelle un "bruit blanc".

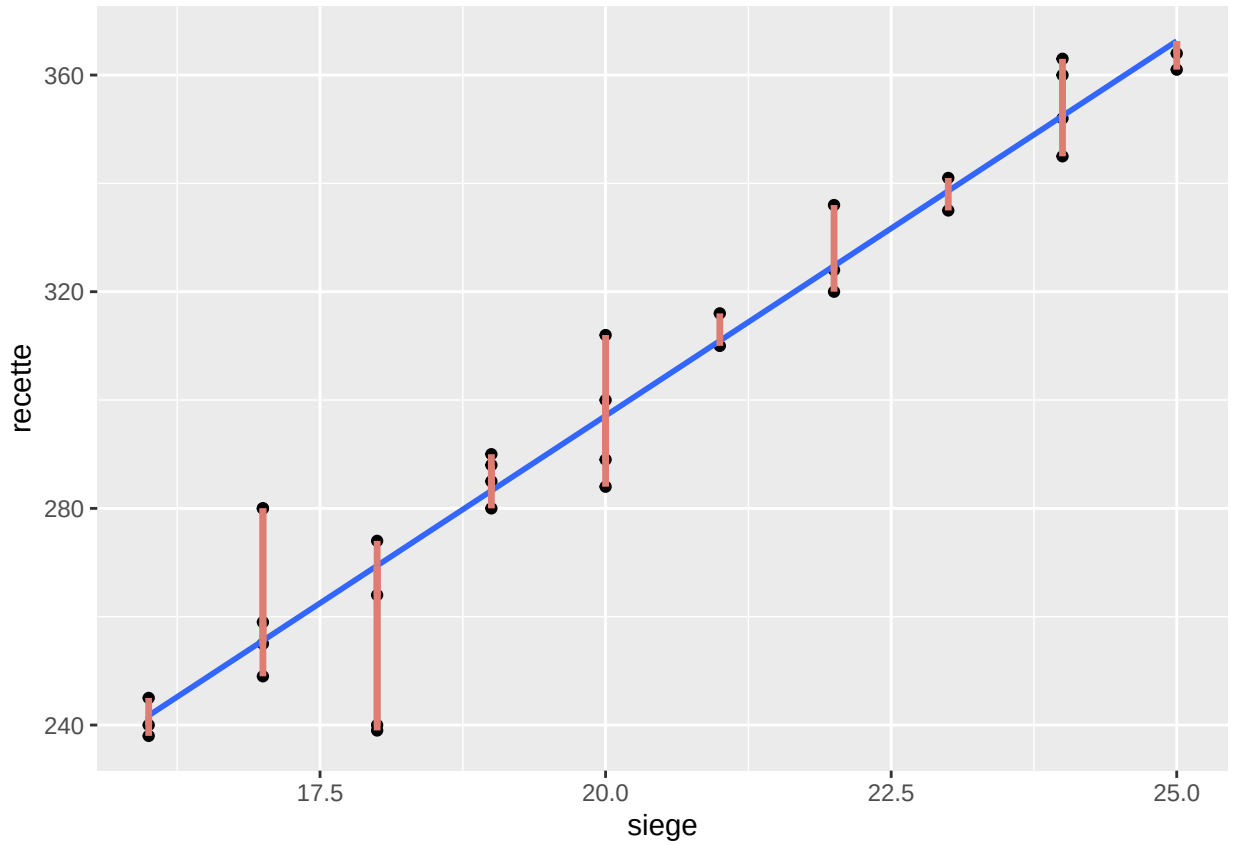
On peut alors estimer les paramètres de la droite qui estime le mieu le nuage de points de la façon suivante et ajouter cette droite sur notre graphique précédent

```
# Estimation des paramètres du modèle
my_lm <- lm(recette~siege, data = data)
coeff <- my_lm$coefficients
coeff
```

```
## (Intercept)       siege
##    20.34648    13.83747
```

```
# Représentation graphique de la droite de régression
library(ggplot2)
ggplot(data, aes(x=siege, y=recette)) +
geom_point() +
geom_smooth(method=lm,se=FALSE) +
geom_segment(aes(x = siege, y = recette, xend = siege,
                 yend = coeff[1] + coeff[2]*siege, col = "Residuals"),
             col = "#DF7D72", lwd= 1.2, data = data)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



L'estimation effectuée nous donne les valeurs suivantes pour notre modèle

$$a = 13.83747 \quad et \quad b = 20.34648.$$

On va essayer de retrouver cela à la "main" en faisant intervenir des quantités statistiques connues. On commence simplement par rappeler que la régression linéaire consiste à minimiser l'erreur quadratique entre la vraie valeur de y et les valeurs prédites par le modèle. On cherche donc à minimiser la quantité suivante

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \iff \sum_{i=1}^{n} (y_i - (ax_i + b))^2 .$$

Sans faire les calculs, on admettra que les valeurs de a et b qui permettent de minimiser cette quantité sont données par les relations suivantes

$$a = \frac{COV(X,Y)}{VAR(X)} \quad et \quad b = \mathbb{E}[Y] - \frac{COV(X,Y)}{VAR(X)}\mathbb{E}[X] = \mathbb{E}[Y] - a\mathbb{E}[X].$$

Il nous faudra donc calculer ces différentes quantités afin de vérifier que l'on retrouve bien des valeurs identiques à celle prédite par notre logiciel. Vérifions cela de suite. On rappelle les définitions

de la moyenne

$$\bar{x} = \mathbb{E}[X] = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

de la variance

$$s^2 = Var[X] = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

et de la covariance

$$cov = Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

```
# Calcul de la moyenne

moy_x <- mean(data$siege)
moy_y <- mean(data$recette)

# Calcul de la variance

var_x <- var(data$siege)
var_y <- var(data$recette)

# Calcul de la covariance

cov_xy <- cov(data$siege,data$recette)
```

On peut alors recalculer la valeeur des coefficients de notre droite

```
# Pour la valeur de a

cov_xy/var_x
```

```
## [1] 13.83747
```

```
# Pour la valeur de b

moy_y - (cov_xy/var_x)*moy_x
```

```
## [1] 20.34648
```

On retrouve bien les valeurs estimées par notre fonction. On peut enfin calculer le coefficient de corrélation, noté ρ, entre nos deux variables à partir de nos données

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var[X]Var[Y]}}$$

On trouve alors une valeur de

```
rho = cov_xy/sqrt(var_x*var_y)
rho
```

```
## [1] 0.9608969
```

qui coïncide bien avec celle estimée par le logiciel. On peut ensuite calculer le coefficient de corrélation entre les deux variables. Pour rappel, ce dernier est défini par

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

```
cor(data$siege,data$recette)
```

```
## [1] 0.9608969
```

La corrélation entre les deux variables est donc positive et on pourrait affirmer que les deux variables sont fortement corrélées. Il reste à savoir si cette corrélation est significative !

Pour cela on peut calculer le coefficient de détermination et vérifier que sa valeur est proche de 1. *Ce coefficient de détermination, noté* r^2*, n'est rien d'autre que le carré de la corrélation linéaire entre les deux variables.*

```
cor(data$siege,data$recette)^2
```

```
## [1] 0.9233229
```