Exercice Kolmogorov-Smirnov

Rachid Sahli

22 janvier 2025

Exercice 1

On travaille sur le jeu de données Recensement.

```
head(data) # Aperçu des données
```

```
##   AGE SEXE REGION STAT_MARI SAL_HOR SYNDICAT CATEGORIE NIV_ETUDES NB_PERS
## 1  58    F     NE         C   13.25      non         5         43       2
## 2  40    M      W         M   12.50      non         7         38       2
## 3  29    M      S         C   14.00      non         5         42       2
## 4  59    M     NE         D   10.60      oui         3         39       4
## 5  51    M      W         M   13.00      non         3         35       8
## 6  19    M     NW         C    7.00      non         3         39       6
##   NB_ENF REV_FOYER
## 1      0        11
## 2      0         7
## 3      0        15
## 4      1         7
## 5      1        15
## 6      0        16
```

Représentation sur un même graphique l'estimation du salaire horaire chez les hommes est chez les femmes (on utilisera l'estimateur à noyau avec le choix bw.bcv).

Rappel

Estimateur à noyau pour la densité de probabilité : L'idée est de lisser l'histogramme des données en utilisant des noyaux (des fonctions qui servent à attribuer un poids à chaque point de données en fonction de sa distance à un point donnée).

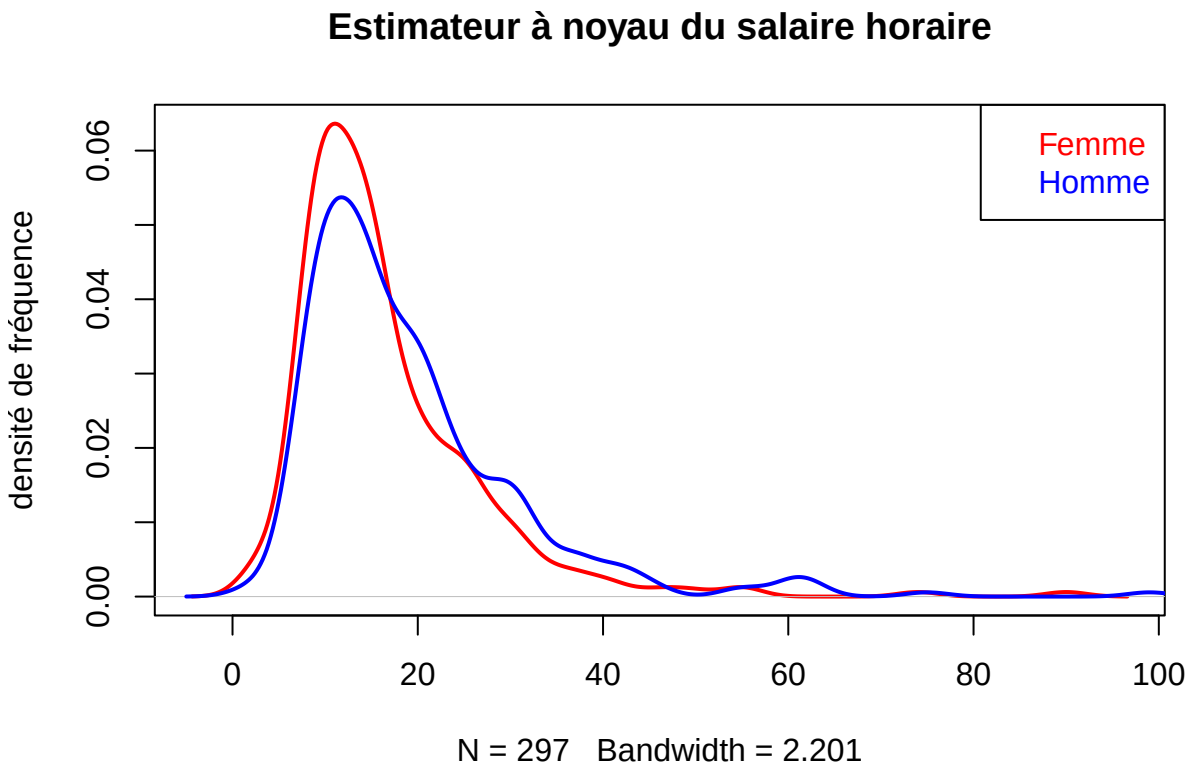
L'estimation de la densité d'un point x à l'aide de l'estimateur à noyau est donnée par la formule suivante :

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

Le noyeau nous permet de tracer deux densités qu'on peut superposer.

```
h_bcd_f = bw.bcv(data$SAL_HOR[data$SEXE == "F"],
                 lower = 0.01, upper = 4) # Selection des fenetres de l'estimateur a noyau
h_bcd_m = bw.bcv(data$SAL_HOR[data$SEXE == "M"],
                 lower = 0.01, upper = 4)
plot(density(data$SAL_HOR[data$SEXE == "F"], bw = h_bcd_f), lwd=2, col="red",
```

```
     main="Estimateur à noyau du salaire horaire",
     ylab="densité de fréquence") # On trace le graph
lines(density(data$SAL_HOR[data$SEXE == "M"], bw = h_bcd_m), col = "blue", lwd = 2)
legend("topright", legend = c("Femme","Homme"), text.col = c("red","blue"))
```



On observe sur le graphique ci-dessus que le mode est presque atteint au même endroit pour les deux sexes. Cependant, la dispersion chez les hommes est plus importante que chez les femmes. Les hommes ont donc tendance à avoir des salaires plus élevés. Il nous faut maintenant réaliser un test de Kolmogorov-Smirnov (KS), car on observe des différences au niveau des deux courbes de densités.

Test de l'égalité des distibutions du salaire horaire chez les hommes et chez les femmes à l'aide de la commande ks.test

Nous réalisons le test de KS pour voir si les différences sont significatives. C'est un test à deux échantillons. On sait que R fait un test asymptotique car n $>$ 100 dans les deux groupes.

Les deux hypothèses du test sont :

$$H_0 : F_X = F_Y$$

contre

$$H_1 : F_X \neq F_Y$$

```
ks.test(data$SAL_HOR[data$SEXE == "F"],data$SAL_HOR[data$SEXE == "M"])
```

```
## Warning in ks.test.default(data$SAL_HOR[data$SEXE == "F"],
## data$SAL_HOR[data$SEXE == : p-value will be approximate in the presence of ties
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
```

```
## data:  data$SAL_HOR[data$SEXE == "F"] and data$SAL_HOR[data$SEXE == "M"]
## D = 0.12766, p-value = 0.01519
## alternative hypothesis: two-sided
```

Le *warning : p-value will be approximate in the presence of ties* nous indique qu'il y a des valeurs identiques dans les données (ex aequo). Les p-valeurs sont donc approximés en fonction des ex aequo et cela peut avoir une influence sur l'exactitude du test.

Utilisation de la fonction tiebreak du package robusTest.

On utilise le package robusTest qui permet de réaliser des test robustes

```
# Import library -----
library(robusTest)

Sal <- tiebreak(data$SAL_HOR) # Résolution des ex aequo dans un vecteur

ks.test(Sal[data$SEXE == "F"],Sal[data$SEXE == "M"])
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  Sal[data$SEXE == "F"] and Sal[data$SEXE == "M"]
## D = 0.14085, p-value = 0.005259
## alternative hypothesis: two-sided
```

La p-valeur observé est de l'ordre de 0.009. Elle est inférieure au seuil $\alpha = 5\%$, on rejette donc H_0. Il y a bien une différence significative entre le salaire horaire des hommes et des femmes au niveau de risque 5%.

Les variables Sexe et Salaire Horaires sont elles indépendantes ?

Nous pouvons appliquer le test de KS a deux échantillons, dans deux cas de figures. Il est paricullièrement utile lorsque les deux échantillons sont définis par une variable qualitative (var auxiliaire). Par exemple, dans le cadre de l'analyse des salaires horaires, on peut parler de densités conditionnelles du salaire horaire en fonction du sexe. En d'autres termes, les groupes sont définis par une variable qualitative, ici le sexe, et nous comparons deux densités conditionnelles.

Sous l'hypothèse nulle (H_0), on suppose que les deux distributions sont identiques, ce qui implique que les deux groupes sont indépendants l'un de l'autre. Si les distributions sont identiques, cela signifie qu'il n'y a pas de différence significative entre les deux groupes, et donc qu'ils suivent la même loi.

Grâce au test, réalisé précedemment, on sait qu'au niveau de risque 5, les variables ne sont pas indépendantes.

Test de l'égalité des esperances

À présent, nous testons l'égalité des espérances dans les deux groupes. En effet, bien que les deux distributions soient significativement différentes, les espérances peuvent être égales. Nous utilisons pour cela un test de Welch.

```
t.test(Sal[data$SEXE == "F"],Sal[data$SEXE == "M"])
```

```
##
##  Welch Two Sample t-test
##
## data:  Sal[data$SEXE == "F"] and Sal[data$SEXE == "M"]
## t = -2.8084, df = 585.38, p-value = 0.005144
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -4.3682754 -0.7729057
## sample estimates:
## mean of x mean of y
##  16.60232  19.17291
```

Le salaire horaire moyen des femmes est de 16,6 $, tandis que celui des hommes s'élève à 19,1 $. Avec une p-valeur inférieur au seuil $\alpha = 5$, la différence des espérances est significativement différente. On rejette H_0 (les moyennes des salaires des femmes et des hommes sont égales).

Le test de Welch aurait suffit à dire que les distributions sont différentes, cela car si les espérances sont différentes, les distributions le sont et vice-versa.

Exercice 2 : Réalisation du même exercice avec des groupes différents

Dans cet exercice, nous réalisons le même exercice que le précédent. Cependant, nous comparons les groupes "syndiqués" et "non syndiqués" de la variable SYNDICAT.

```
kable(prop.table(table(data$SYNDICAT))*100)
```

Var1	Freq
non	82.80467
oui	17.19533

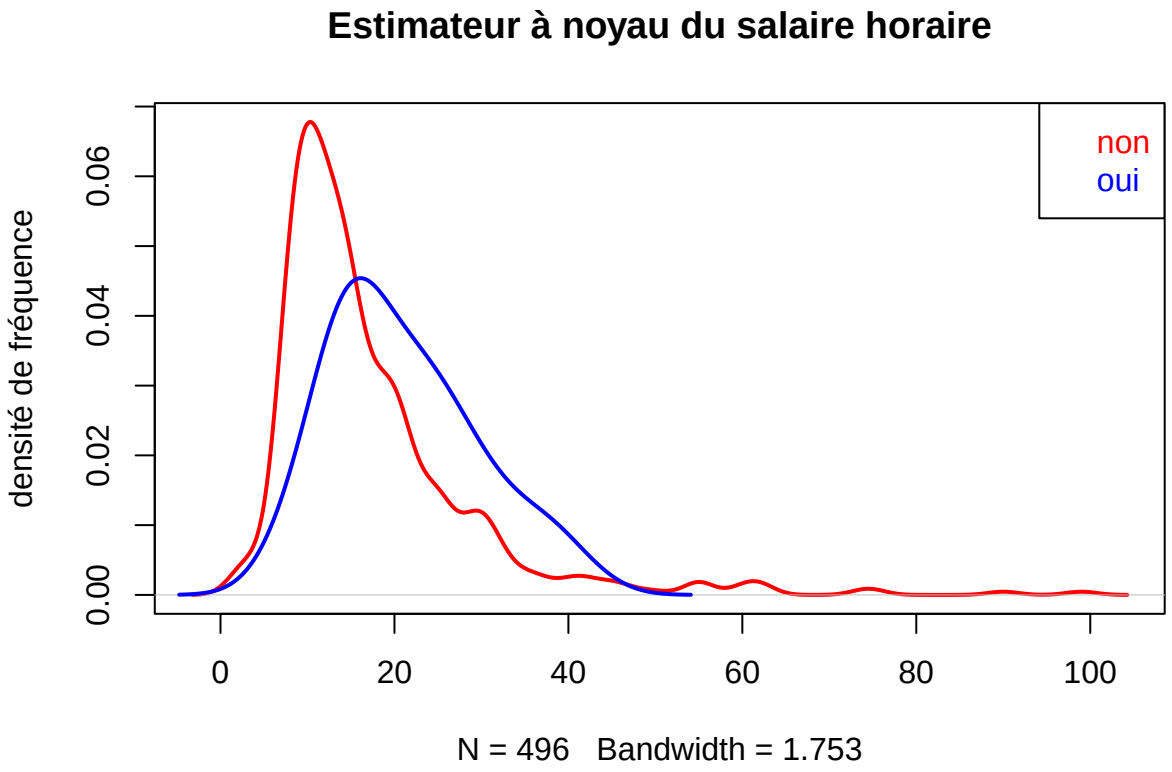
On observe un fort déséquilibre entre les deux groupes. Nous savons donc que nous allons obtenir une très faible p-valeur et qu'il y a une différence significative entre les deux groupes.

On représente graphiquement l'estimation du salaire horaire dans les deux groupes avec l'estimateur à noyau.

```
h_bcd_f = bw.bcv(data$SAL_HOR[data$SYNDICAT == "non"],
                 lower = 0.01, upper = 4) # Choisir les fenetres de l'estimateur a noyau
h_bcd_m = bw.bcv(data$SAL_HOR[data$SYNDICAT == "oui"],
                 lower = 0.01, upper = 4)
```

```
## Warning in bw.bcv(data$SAL_HOR[data$SYNDICAT == "oui"], lower = 0.01, upper =
## 4): minimum occurred at one end of the range
```

```
plot(density(data$SAL_HOR[data$SYNDICAT == "non"], bw = h_bcd_f), lwd=2, col="red", main="Estimateur à n
     ylab="densité de fréquence") # On trace le graph
lines(density(data$SAL_HOR[data$SYNDICAT == "oui"], bw = h_bcd_m), col = "blue", lwd = 2)
legend("topright", legend = c("non","oui"), text.col = c("red","blue"))
```



Il y a une différence notable entre le salaire horaire des deux groupes. Les personnes non syndiquées ont tendance à avoir un salaire horaire plus élevé que les personnes syndiquées.

On effectue le test de comparaison des deux distributions avec le test de KS.

```
ks.test(data$SAL_HOR[data$SYNDICAT == "oui"],data$SAL_HOR[data$SYNDICAT == "non"])
```

```
## Warning in ks.test.default(data$SAL_HOR[data$SYNDICAT == "oui"],
## data$SAL_HOR[data$SYNDICAT == : p-value will be approximate in the presence of
## ties
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  data$SAL_HOR[data$SYNDICAT == "oui"] and data$SAL_HOR[data$SYNDICAT == "non"]
## D = 0.328, p-value = 2.143e-08
## alternative hypothesis: two-sided
```

La différence entre les deux distributions est hautement significative. La p-valeur = 2.143e-08, elle est très proche de 0.

Puis, nous testons l'égalité des espérances entre les deux groupes, comme précédemment.

```
t.test(data$SAL_HOR[data$SYNDICAT == "oui"],data$SAL_HOR[data$SYNDICAT == "non"])
```

```
##
##  Welch Two Sample t-test
##
## data:  data$SAL_HOR[data$SYNDICAT == "oui"] and data$SAL_HOR[data$SYNDICAT == "non"]
## t = 3.8964, df = 193.65, p-value = 0.0001344
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.888225 5.759207
```

```
## sample estimates:
## mean of x mean of y
##  21.06456  17.24085
```

La p-valeur est également très faible, cependant, elle ne l'est pas autant que celle du test de KS. Les moyennes nous indiquent une différence entre les deux groupes.

En somme, les distributions et les esperances sont significativement différente.

La p-valeur est tres petite aussi, mais pas autant que celui du K-s. La moyenne dans le groupe des sindique est de tant et dans le groupe

Il n'y a pas de doute ici au niveau des distributions et des esperances, elles sont significativement différente.

Exercice 3 : Réalisation du même exercice selon deux autres groupes différents

Ici, nous créons deux niveaux d'études dans une nouvelle variable (NIV) à partir de la variable NIV_ETUDES :

- Inférieur à niveau bac
- Supérieur à niveau bac

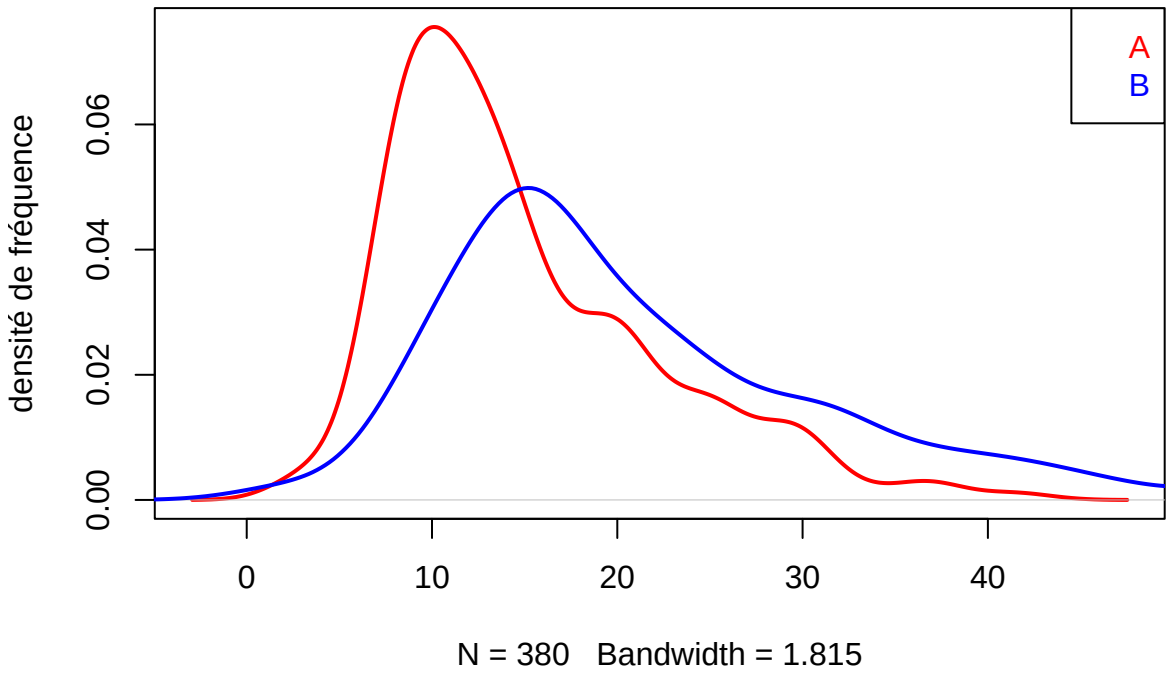
```
data$NIV[data$NIV_ETUDES <= 40] = "A"
data$NIV[data$NIV_ETUDES > 40] = "B"

table(data$NIV)
```

```
##
##   A   B
## 380 219
```

On observe également un déséquilibre important entre les deux niveaux d'études.

On estime les densité des deux groupes à l'aide du noyau. On les superpose sur le graphique suivant.

```
h_bcd_a = bw.bcv(data$SAL_HOR[data$NIV == "A"],
                 lower = 0.01, upper = 4) # Choisir les fenetres de l'estimateur a noyau
h_bcd_b= bw.bcv(data$SAL_HOR[data$NIV == "B"],
                 lower = 0.01, upper = 4)
plot(density(data$SAL_HOR[data$NIV == "A"], bw = h_bcd_a), lwd=2, col="red",
     main="Estimateur à noyau du salaire horaire",
     ylab="densité de fréquence") # On trace le graph
lines(density(data$SAL_HOR[data$NIV == "B"], bw = h_bcd_b), col = "blue", lwd = 2)
legend("topright", legend = c("A","B"), text.col = c("red","blue"))
```



On observe des différences entre les deux courbes. On voit que les personnes ayant un niveau d'études, ont des salaires plus élevées que ceux ayant un nievau supérieur au bac. On réalise un test de KS pour tester l'égalité des distributions.

```
Sal <- tiebreak(data$SAL_HOR) # Résolution des ex aequo dans un vecteur

ks.test(Sal[data$NIV == "A"],Sal[data$NIV == "B"])
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  Sal[data$NIV == "A"] and Sal[data$NIV == "B"]
## D = 0.34065, p-value = 1.987e-14
## alternative hypothesis: two-sided
```

La p-valeur est très faible et inférieur au seuil $\alpha = 5\%$, on rejette l'hypothèse H_0. Il y a donc une différence significative du salaire horaire entre les deux groupes de niveau d'études. On peut aussi en déduire que les variables salaires horaires et niveau d'étude ne sont pas indépendantes.

Enfin, nous comparons les espérances des deux groupes avec le test ci-dessous.

```
t.test(data$SAL_HOR[data$NIV == "A"],data$SAL_HOR[data$NIV == "B"])
```

```
##
##  Welch Two Sample t-test
##
## data:  data$SAL_HOR[data$NIV == "A"] and data$SAL_HOR[data$NIV == "B"]
## t = -7.6226, df = 279.5, p-value = 3.891e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.180229  -6.001437
## sample estimates:
```

```
## mean of x mean of y
##  14.94026  23.03110
```

Il y a une différence de salaires très élevés entre les deux groupes.

Exercice 4 : Refaire le même exercice avec deux groupes différents.

Ici, nous créeons deux classes d'âge à partir de la variable AGE :

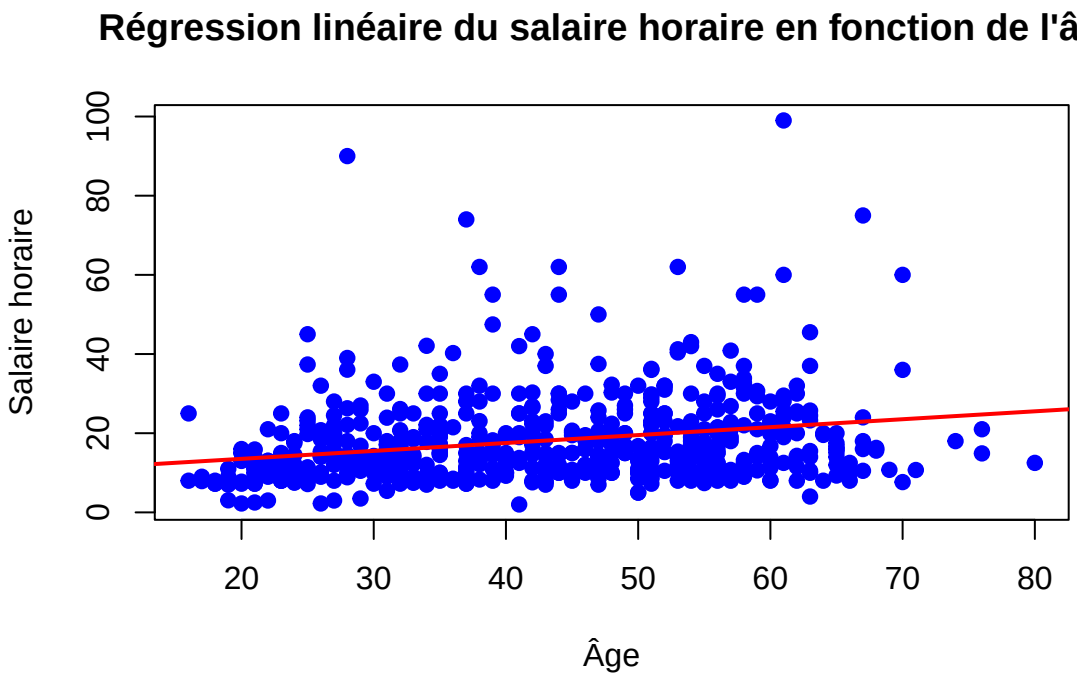
- Inférieur à 40 ans
- Supérieur à 40 ans

Exercice 5

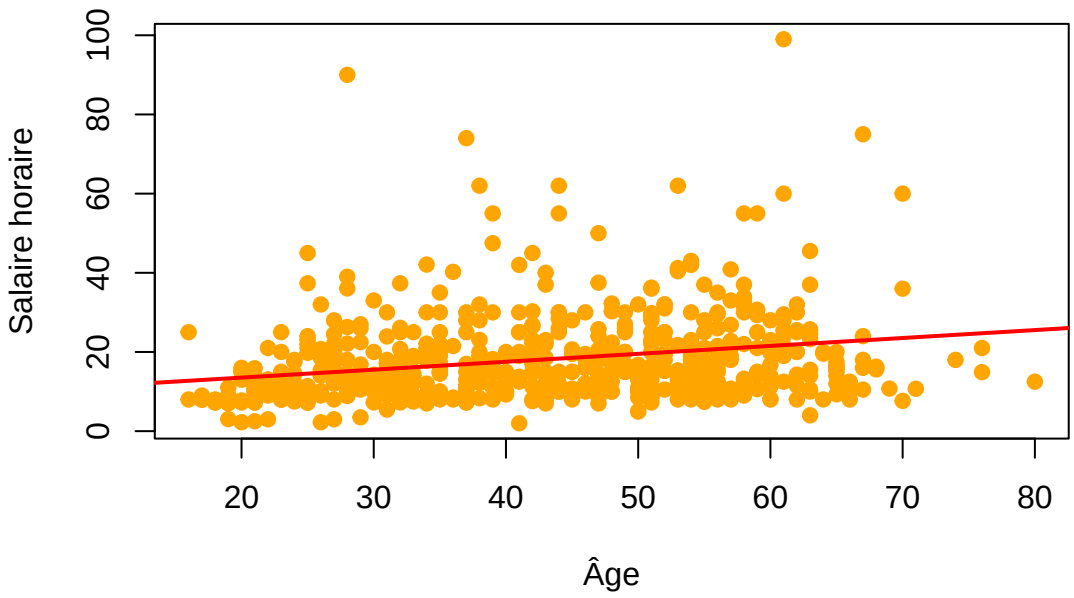
Courbe de régression du salaire horaire en fonction de l'âge

Nous traçons la courbe de régression du salaire horaire en fonction de l'âge avec une partition de taille 8.

```r
# Calcul sans la fonction lm()

# Moyennes
moyenne_age <- mean(data$AGE)
moyenne_salaire <- mean(data$SAL_HOR)

# Somme des produits des écarts
numerateur <- sum((data$AGE - moyenne_age) * (data$SAL_HOR - moyenne_salaire))

# Somme des carrés des écarts pour l'âge
denominateur <- sum((data$AGE - moyenne_age)^2)

# Pente (beta1)
beta_1 <- numerateur / denominateur

# Intercept (beta0)
beta_0 <- moyenne_salaire - beta_1 * moyenne_age

plot(data$AGE, data$SAL_HOR, main = "Régression linéaire du salaire horaire en fonction de l'âge",
     xlab = "Âge", ylab = "Salaire horaire", pch = 19, col = "blue")

# Ajouter la ligne de régression manuellement
abline(a = beta_0, b = beta_1, col = "red", lwd = 2)
```




```r
# Calcul avec la fonction lm()

modele <- lm(SAL_HOR ~ AGE, data = data)

plot(data$AGE, data$SAL_HOR, main = "Régression du salaire horaire en fonction de l'âge",
     xlab = "Âge", ylab = "Salaire horaire", pch = 19, col = "orange")

abline(modele, col = "red", lwd = 2)

title(main = "Régression du salaire horaire en fonction de l'âge")
```



La relation entre le salaire horaire et l'âge est positive. Plus la personne vieillit, plus son salaire horaire tend à augmenter. Pour chaque année supplémentaire d'âge, on s'attend à ce que le salaire horaire augmente de 0.20 $. Par exemple, si une personne a 30 ans, un changement à 31 ans pourra augmenter son salaire horaire de 0.20 $, selon ce modèle linéaire.

Test de l'indépendance de l'âge et du salaire horaire à l'aide de la commande indeptest du package robusTest.

On test l'indépendance de l'âge et du salaire horaire.

```
indeptest(data$AGE,data$SAL_HOR, ties.break = "random")
```

```
##
## Robust independence test for two continuous variables
##
## t = 1.7618, p-value <1e-4
##
## Ties were detected in the dataset and they were randomly broken
```

Le message *Warning : The data contains ties! Use ties.break='random'*, nous indique qui il y a encore des valeurs identiques dans nos données. Pour corriger cela, on utilise le paramètre *ties.break = TRUE*. La p-valeur obtenue à l'issue de ce test est très faible, les deux variables sont donc statistiquement dépendantes. On rejette l'hypothèse nulle d'indépendance. On constate bien une relation significative entre ces deux variables.

Exercice 6

On travaille maintenant sur le jeu de données Eucalyptus.

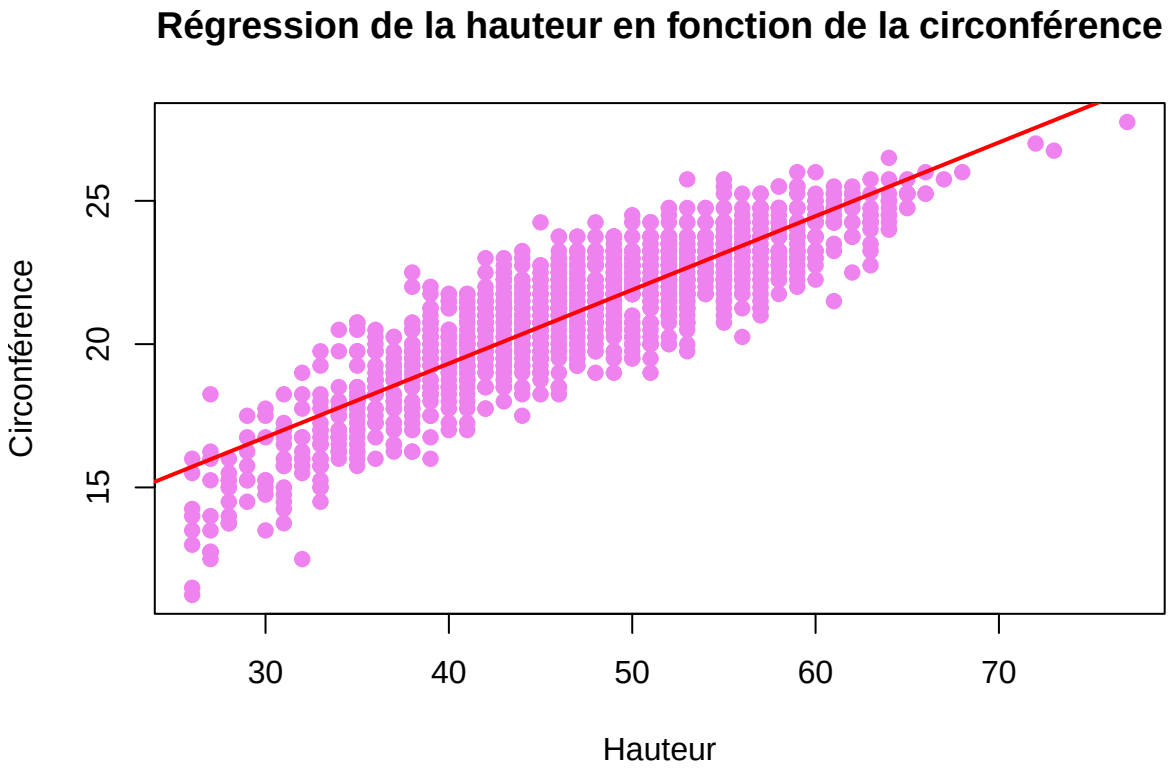
```
head(euca) # Aperçu des données
```

```
##      ht circ bloc  clone
## 1 18.25   36    1 L2-123
## 2 19.75   42    1 L2-123
## 3 16.50   33    1 L2-123
## 4 18.25   39    1 L2-123
## 5 19.50   43    1 L2-123
## 6 16.25   34    1 L2-123
```

Courbe de régression de la hauteur en fonction de la circonférence (partition taille 16)

Nous traçons la courbe de régression de la hauteur en fonction de la circonférence.

```
modele_euca <- lm(ht ~ circ, data = euca)

plot(euca$circ, euca$ht, main = "Régression de la hauteur en fonction de la circonférence",
     xlab = "Hauteur", ylab = "Circonférence", pch = 19, col = "violet")

abline(modele_euca, col = "red", lwd = 2)
```



La régression montre une relation positive entre la circonférence et la hauteur. À mesure que la circonférence augmente, la hauteur estimée augmente également. Le modèle est linéaire.

Test de l'indépendance de la hauteur et de la circonférence.

```
indeptest(euca$ht,euca$circ, ties.break = "random")
```

```
##
## Robust independence test for two continuous variables
##
## t = 6.7263, p-value <1e-4
##
## Ties were detected in the dataset and they were randomly broken
```

La p-value très faible, cela indique que le test fournit une preuve très forte contre l'hypothèse nulle d'indépendance. Il existe alors une relation significative entre ces deux variables. Elles sont dépendantes.