Réalisez des modélisations de données performantes : activités

V. Lefieux

ACTIVITE P2

```
library(ggplot2) # Pour les graphiques ggplot
```

On importe le fichier *arbres.txt* qui contient 138 données sur des épicéas parisiens (source : https://opendata.paris.fr/explore/dataset/les-arbres/table/).

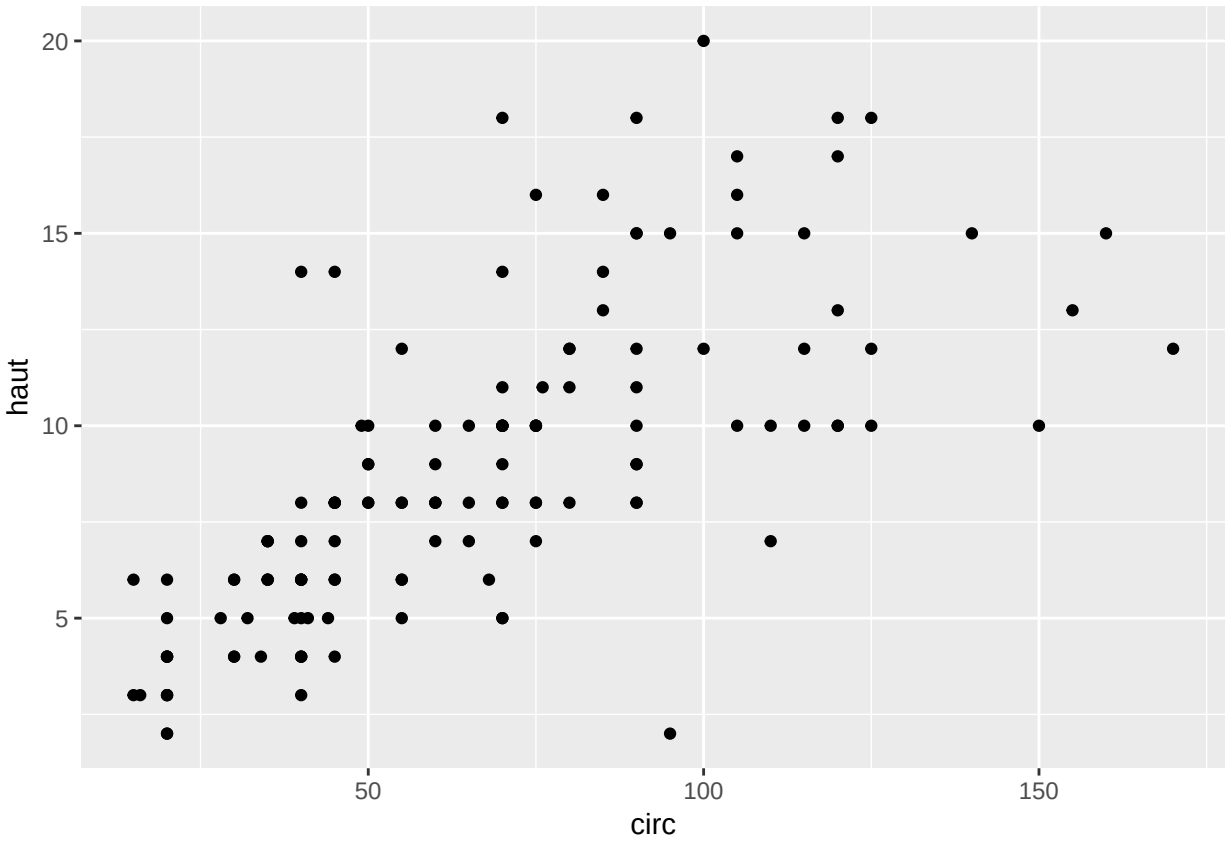
On y trouve les 2 variables suivantes :

- circ : circonférence de l'arbre (en cm),
- haut : hauteur de l'arbre (en m).

```
setwd("/Users/vincentlefieux/Dropbox/DocsACADEMIQUE/PolysSlides/Reg_RegLog_AnVar/Activites_dataset/")
arbres <- read.table("arbres.txt",header=TRUE,sep=";",dec=",")
```

On représente le nuage de points (circ,hauteur) :

```
ggplot(arbres,aes(x=circ,y=haut))+
  geom_point()+
  xlab("circ")+
  ylab("haut")
```



On constate que le nuage de points n'est pas très éloigné d'une droite, on lance la régression linéaire simple :

```
reg_simp <- lm(haut~circ,data=arbres)
```

On obtient le coefficient de détermination ainsi que les paramètres et leur tests de significativité :

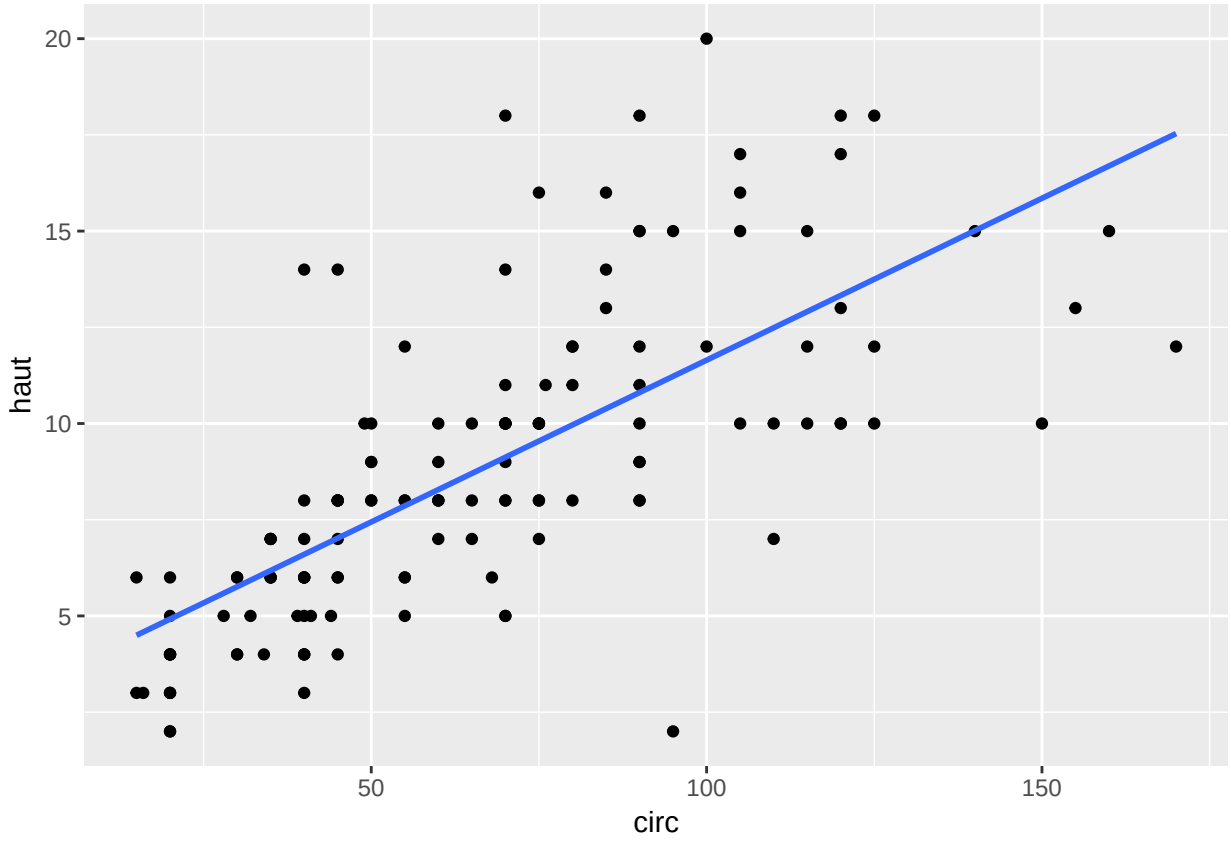
```
summary(reg_simp)
```

```
##
## Call:
## lm(formula = haut ~ circ, data = arbres)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2259 -1.7937 -0.2306  1.0708  8.8769
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.23529    0.53646   6.031 1.45e-08 ***
## circ         0.08411    0.00722  11.649  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.86 on 136 degrees of freedom
## Multiple R-squared:  0.4995, Adjusted R-squared:  0.4958
## F-statistic: 135.7 on 1 and 136 DF,  p-value: < 2.2e-16
```

- On constate que le coefficient de détermination vaut environ 0.5, ce qui n'est pas très élevé.
- On rejette la nullité des paramètres au niveau de test 5%.

Il est possible de tracer la droite de régression :

```
ggplot(arbres,aes(x=circ,y=haut))+
  geom_point()+
  stat_smooth(method="lm",se=FALSE)+
  xlab("circ")+
  ylab("haut")
```



ACTIVITE P3

On importe le fichier *arbres.txt* :

```
setwd("/Users/vincentlefieux/Dropbox/DocsACADEMIQUE/PolysSlides/Reg_RegLog_AnVar/Activites_dataset/")
arbres <- read.table("arbres.txt",header=TRUE,sep=";",dec=",")
```

On effectue une régresion linéaire multiple de haut sur circ et la racine carrée de circ:

```
arbres$circ_sqrt <- sqrt(arbres$circ)

reg_multi <- lm(haut~circ+circ_sqrt,data=arbres)
summary(reg_multi)
```

```
##
## Call:
## lm(formula = haut ~ circ + circ_sqrt, data = arbres)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -9.5360 -1.7847 -0.1174  0.9435  8.2153
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.20826    2.88404  -2.153  0.03313 *
## circ        -0.06594    0.04562  -1.446  0.15063
## circ_sqrt    2.46322    0.74005   3.328  0.00113 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.759 on 135 degrees of freedom
## Multiple R-squared:  0.5374, Adjusted R-squared:  0.5306
## F-statistic: 78.42 on 2 and 135 DF,  p-value: < 2.2e-16
```

La variable *haut* n'est pas significative au niveau de test 5%, on la retire donc :

```
reg_multi <- lm(haut~circ_sqrt,data=arbres)
summary(reg_multi)
```

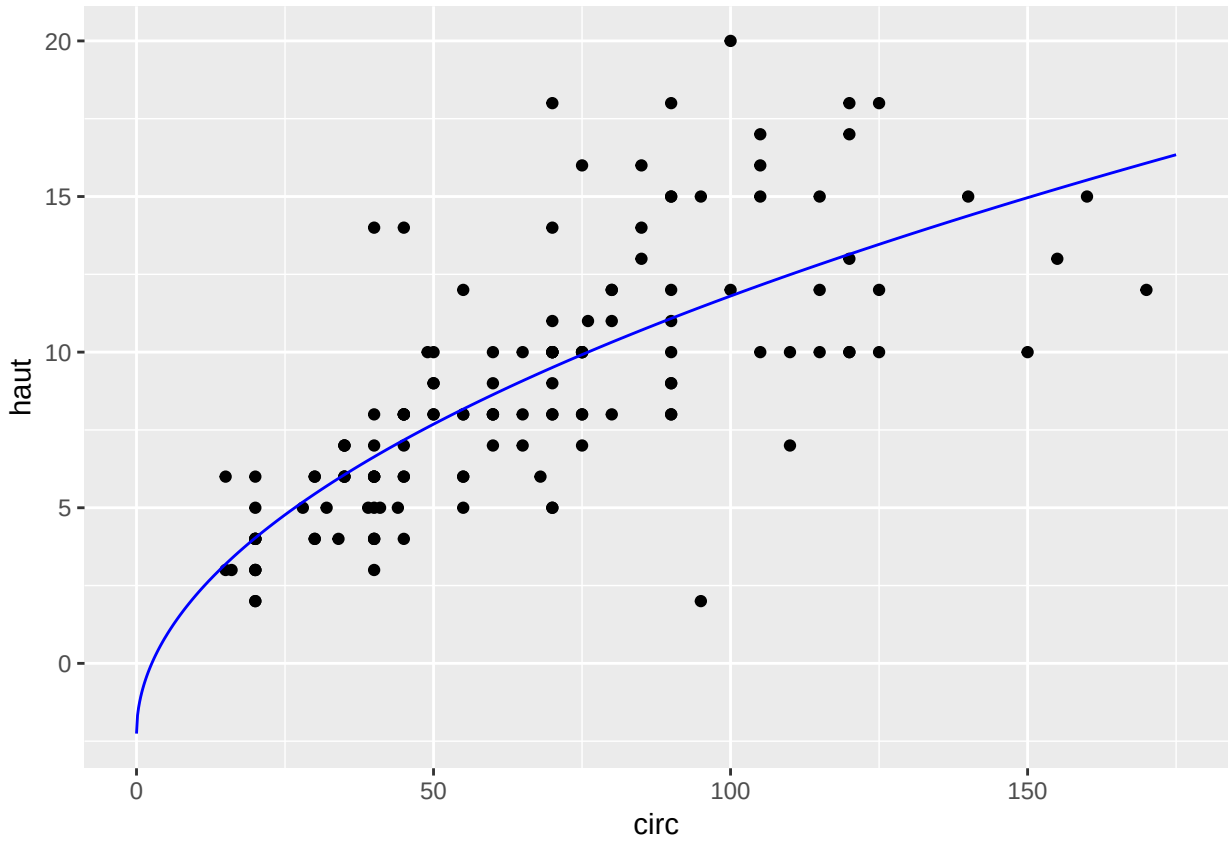
```
##
## Call:
## lm(formula = haut ~ circ_sqrt, data = arbres)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4472 -1.8761 -0.1134  0.9389  8.4934
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.2570     0.9233  -2.444   0.0158 *
## circ_sqrt     1.4060     0.1135  12.390   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.77 on 136 degrees of freedom
## Multiple R-squared:  0.5303, Adjusted R-squared:  0.5268
## F-statistic: 153.5 on 1 and 136 DF,  p-value: < 2.2e-16
```

- On constate que le coefficient de détermination vaut environ 0.98, ce qui est réellement meilleur.
- On rejette la nullité des paramètres au niveau de test 5%.

Il est possible de tracer la "courbe" de régression :

```
circ_prev <- seq(0,175,len=1000)
haut_prev <- reg_multi$coefficients[1]+reg_multi$coefficients[2]*sqrt(circ_prev)
fct_reg <- data.frame(circ_prev=circ_prev,haut_prev=haut_prev)

ggplot()+
  geom_point(data=arbres,aes(x=circ,y=haut))+
  geom_line(data=fct_reg,aes(x=circ_prev,y=haut_prev),col="blue")+
  stat_smooth(method="lm",se=FALSE)+
  xlab("circ")+
  ylab("haut")
```


haut
circ
20
15
10
5
0
0
50
100
150