Regression linéaire

Introduction

Ce document est une application du modèle de regression linéaire sur deux variable constitué aléatoirement. X est un ensemble de nombre de 1 jusqu'a 12 et y est constitué d'un ensemble de chiffre. Le but est d'établir un modèle de regression linéaire entre ces deux variable.

Les données utilisées

```
library("tidyverse")
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.4.1     v purrr   1.0.1
v tibble  3.1.8     v dplyr   1.1.0
v tidyr   1.3.0     v stringr 1.5.0
v readr   2.1.4     v forcats 1.0.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```
library("car")
```

```
Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:dplyr':
```

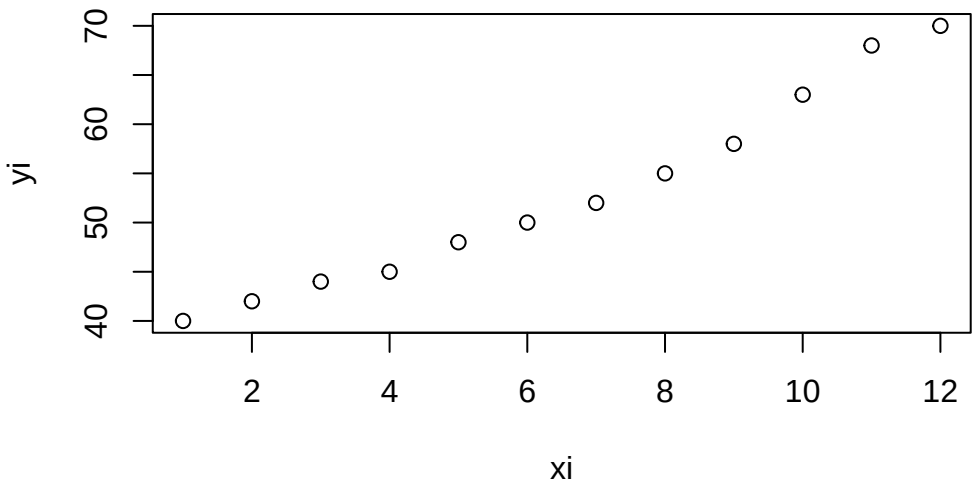
```
    recode

The following object is masked from 'package:purrr':

    some
```

```
library("corrplot")
```

```
corrplot 0.92 loaded
```

```
xi <- c(1:12)
yi <- c(40, 42, 44, 45, 48, 50, 52, 55, 58, 63, 68, 70)
```

```
plot(xi, yi)
```



En se référant á nuage de point, le relation entre x et y semble linéraire, on ne peux se baser sur de simple graphe pour affirmé cela, calculons le coéficient de corrélation.

```
r<-cor(xi,yi)
```

ce dernier est de 0.98, le coéfficient de corrélation est une mésure que associe une valeur a la corrélation linéaire qui existe entre deux variable. notre valeur est proche de 1, sans doute qu'il y a une forte corrélation entre ces deux variable en se basant toujours sur l'échatillons collectés. Vous être plus confiant de cette valeur, fesons un test sur cette valeur

```r
cor.test(xi,yi)
```

```
    Pearson's product-moment correlation

data:  xi and yi
t = 17.396, df = 10, p-value = 8.355e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9417075 0.9956089
sample estimates:
      cor
0.9838759
```

On fait un test sous l'hypothèse que la corrélation est nulle, le résultat du test donne une valeur p qui est très faible ce qui revient a dire qu'on commettra une erreur en rejétant l'hypothèse nulle qui est que la corrélation entre ces deux variable est différent de zero. En ajout, on est 95% confiant que la valeur réelle du corrélation se trouve entre 0.94 et 0.99.

```
Warning: `data_frame()` was deprecated in tibble 1.1.0.
i Please use `tibble()` instead.
```

effectuons tous d'abord un test pour savoir si les données, je veux dire les variables x et y proviennent d'une distribution normal. Pour cela on va effectuer deux test, le test de Kolmogorom smirnov et celui de shapiro.

```r
ks.test(yi,"pnorm")
```

```
    Exact one-sample Kolmogorov-Smirnov test

data:  yi
D = 1, p-value = 3.331e-16
alternative hypothesis: two-sided
```

```r
ks.test(xi,"pnorm")
```

```
    Exact one-sample Kolmogorov-Smirnov test

data:  xi
D = 0.89392, p-value = 4.063e-12
alternative hypothesis: two-sided
```

```
shapiro.test(xi)
```

```
    Shapiro-Wilk normality test

data:  xi
W = 0.9669, p-value = 0.8757
```

```
shapiro.test(yi)
```

```
    Shapiro-Wilk normality test

data:  yi
W = 0.93739, p-value = 0.465
```

Modèle de regression linéaire

Concevons le modèle de regression linéaire. Pour cela on va utiliser la fonction lm offerte par r en donnant la variable dependante yi et la variable indépendante xi. Un résumé du modèle nous fournis les détails suivant:

```
 regxy <- lm(yi ~ xi)
```

```
summary(regxy)
```

```

Call:
lm(formula = yi ~ xi)
```

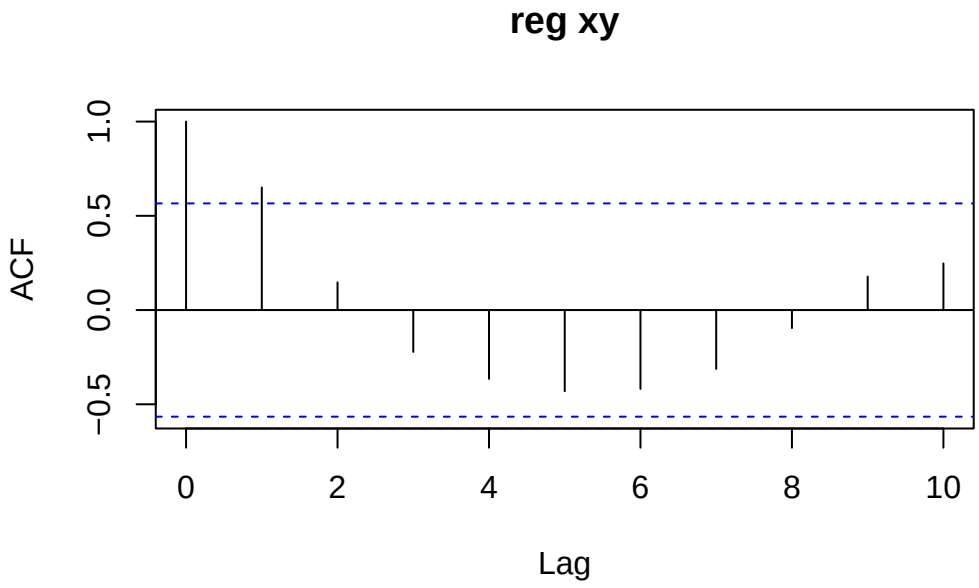
```
Residuals:
    Min      1Q  Median      3Q     Max
-2.2890 -1.6029 -0.1614  1.5728  2.7319

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.0758     1.1612    30.2 3.70e-11 ***
xi            2.7448     0.1578    17.4 8.35e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.887 on 10 degrees of freedom
Multiple R-squared:  0.968, Adjusted R-squared:  0.9648
F-statistic: 302.6 on 1 and 10 DF,  p-value: 8.355e-09
```

Commencons par les résidues du modèle. ces résidues se rangent entre l'intervale de -2.2890 et 2.7319 avec un médian de -0.1614. ces resultat montrent qu'il y a pas une grande variance entre les résidues. En ajout, la valeur p des paramètre du modèle sont très faible, une est de 3.70e-11 et l'autre 8.35e-09 ce qui nous conduit a ne pas rejeter les estimations du modèle. Pour aller plus loin etudions les résidues du modèles, afin de valider le modèle.

Autocarelation

L'autocorrelation est le calcule de cal covariance croiser d'une variable. étant données que les résidues de l'estimation son basé sur un échantillons, ces dernier varient d'un échantillons a un autre. On va alors supposer que chaque résidue est une variable aléatoire dépendant du temps, comme ca on calcule la covariance des résidues dans le temps. Ici est le l'affichage du résultat avec la fonction autocorrelation fonction.

```r
acf(residuals(regxy), main="reg xy")
```



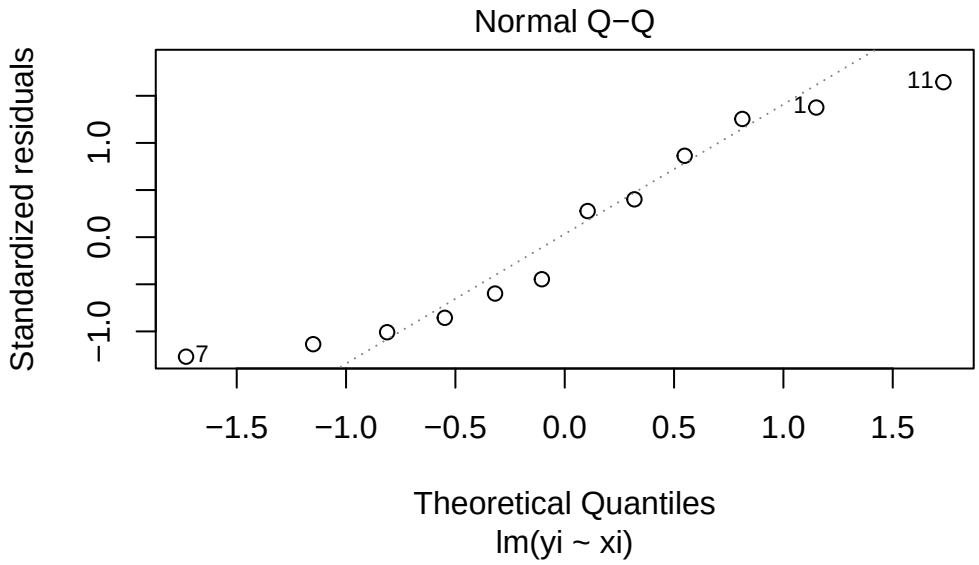
Ceci est un test d'autocorrelation qui á été proposé par **Durbin-Watson**. Ce test permet de déterminer si les résidue d'une regression linéaire son corrélé ou pas? L'hypothèse null est qu'il y a pas de autocorrélation et la D-W statistique est la statistique utilisé pour faire le test. ssi on prend 0.05 et 0.09 comme seuil de tolérance de notre test, en théorie l'hypothèse null est validé ca le D-W test est dans l'intervale de 0.09 et 4-0.09. Les résidues ne sont pas corrélés.

```
durbinWatsonTest (regxy)
```

```
 lag Autocorrelation D-W Statistic p-value
   1       0.6505047     0.4546369       0
 Alternative hypothesis: rho != 0
```

Une autre manière de validé le modèle est de test la normalité des rèsidues. si les résidues sont normale, les modèles sera alors validé. Il y a plusieurs manière d'effectué le test de normalité sur un variable, dans ce cadre nous allons utilisé le shapiro.test. Tous d'abord le QQ plot nous indique avec une tolérance la normalité du résidue. Si on se réfère au test de shapiro avec une tolérance de 0.05, la valeur p est supérieur a cette valeur alpha, ce serait un erreur de rejeter l'hypthèse nulle.

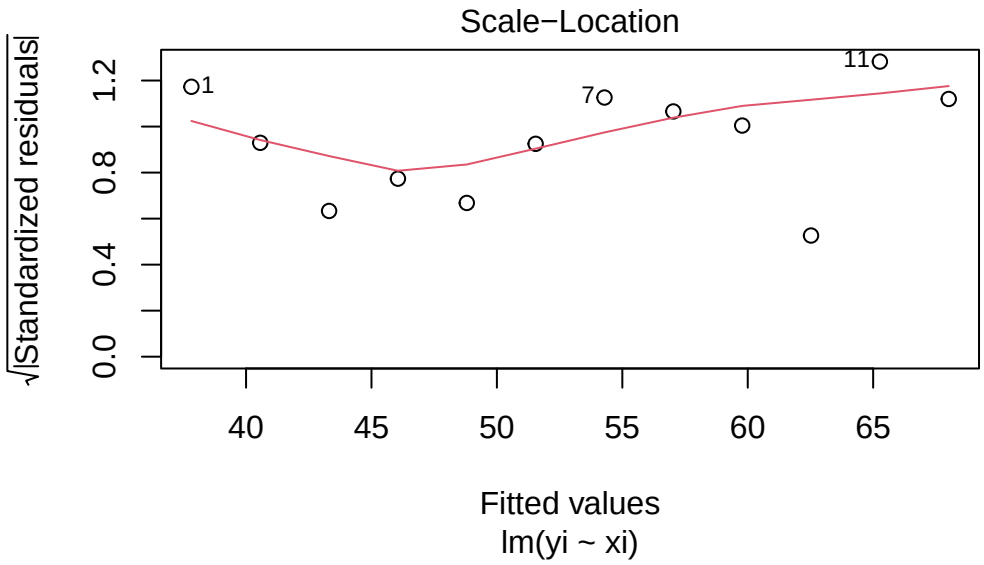
```
plot(regxy,2)
```




```
shapiro.test(residuals(regxy))
```

```
    Shapiro-Wilk normality test

data:  residuals(regxy)
W = 0.91467, p-value = 0.2448
```

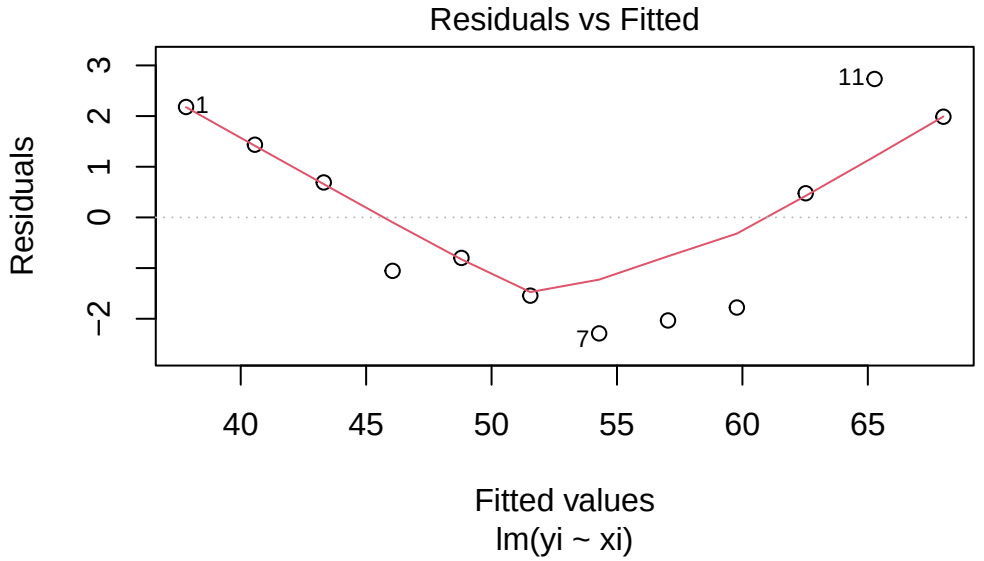
```
plot(regxy, 3)
```




```
ncvTest(regxy)
```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.377958, Df = 1, p = 0.5387
```

```
plot(regxy,1)
```



```r
confint(regxy)
```

```
                2.5 %    97.5 %
(Intercept) 32.488339 37.663176
xi           2.393194  3.096316
```