Practical 1 - Simple Linear Regression - Partial Correction

Julien Chiquet & Guillem Rigaill & Anastase Alexandre Charantonis

September 22th, 2016

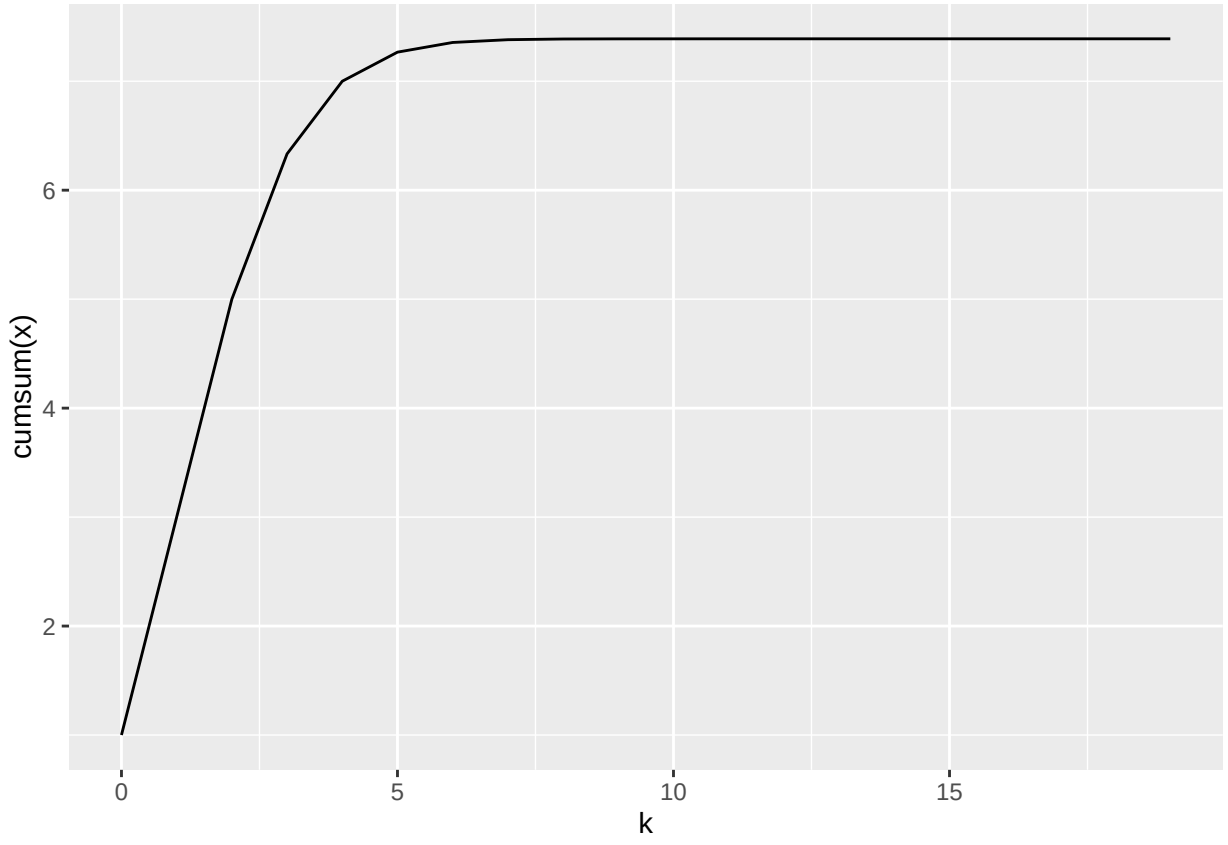
Quick refresh of `R` basics

1. **Vector Manipulation**. We remind you that

$$e^x = \sum_{k \geq 0} \frac{x^k}{k!}.$$

   Insert in a vector named `exp2` the 20 first elements of this sequence. Remove any values below? 10^{-8}. Use them to give an estimation of e^2 and compare that value to `exp(2)`.

```
k <- 0:19
x <- 2^k/factorial(k)
qplot(k,cumsum(x), geom="line")
```

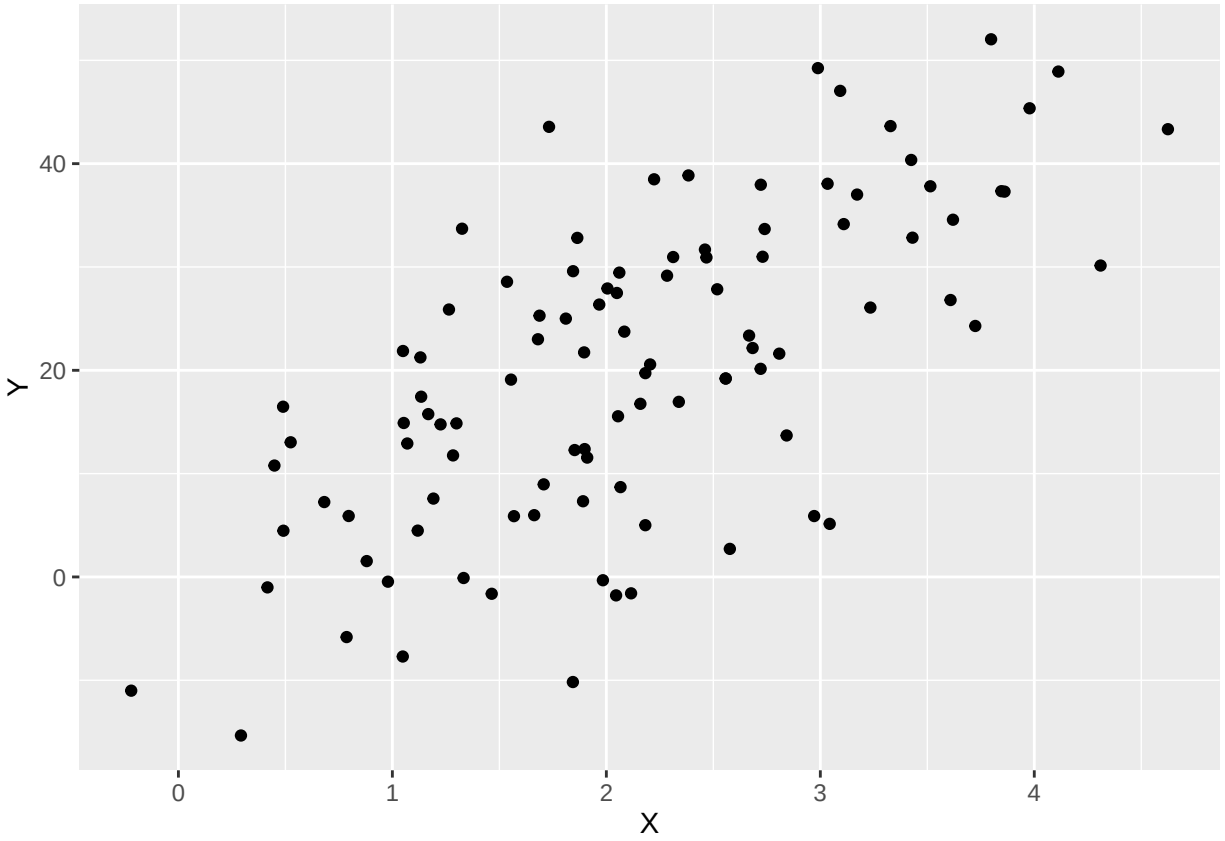



```
exp2.hat <- sum(x[x > 1e-8])
print(exp(2)-exp2.hat)
```

```
## [1] 3.546512e-09
```

2. **Data Simulation**. Using the `rnorm` function, (type `?rnorm` to get help on this function) generate a vector X containing 100 samples of a normal distribution with a mean value of 2 and a variance of 1. Generate another vector, named Y of the same size obtained by multiplying X by 9.8 and adding a Gausssian noise with a standard deviation of 10.

```
n <- 100
X <- rnorm(n,mean=2,sd=1)
Y <- 9.8 * X + rnorm(n,sd=10)
qplot(X,Y)
```



3. **Read and write a data file**. Put the vectors X and Y in a `data.frame`. Save that `data.frame` with the `write.table` command. Read the obtained table with the `read.table` command. Compare the matrix you obtained to the initial matrix.

```
donnees <- data.frame(X=X,Y=Y)
write.table(donnees, file="my_file.csv")
mes.donnees <- read.table(file="my_file.csv")
head(mes.donnees)
```

```
##          X         Y
## 1 1.167822 15.760950
## 2 3.619945 34.572413
## 3 3.171889 37.007347
```

```
## 4 1.464294 -1.629876
## 5 2.808136 21.610046
## 6 1.554435 19.098091
```

The elements of the table correspond to each other but the names of the variable in which the table is contained in not stored with the `write.table` *command.*

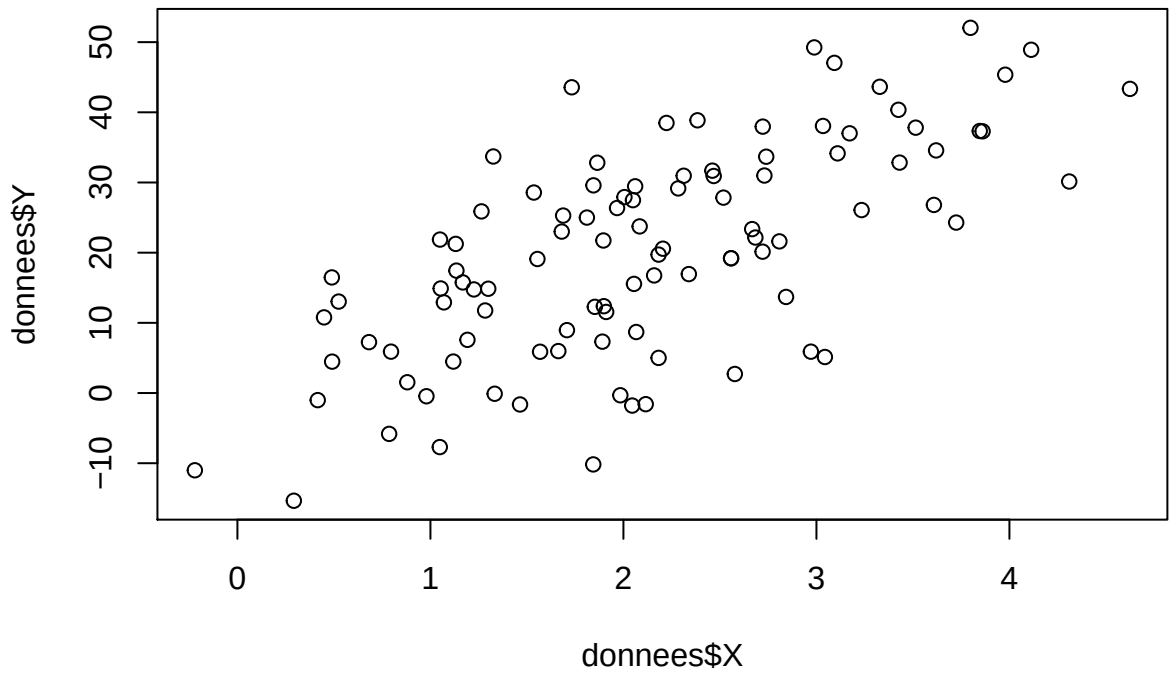
4. **Read and write a table in the RData format**. Put the vectors X et Y in a `data.frame`. Save the `data.frame` using the `save` command. Read it with the `load` command. Compare the matrix obtained to the initial one.

```
save(donnees, file="my_file.csv")
load(file="my_file.csv")
```

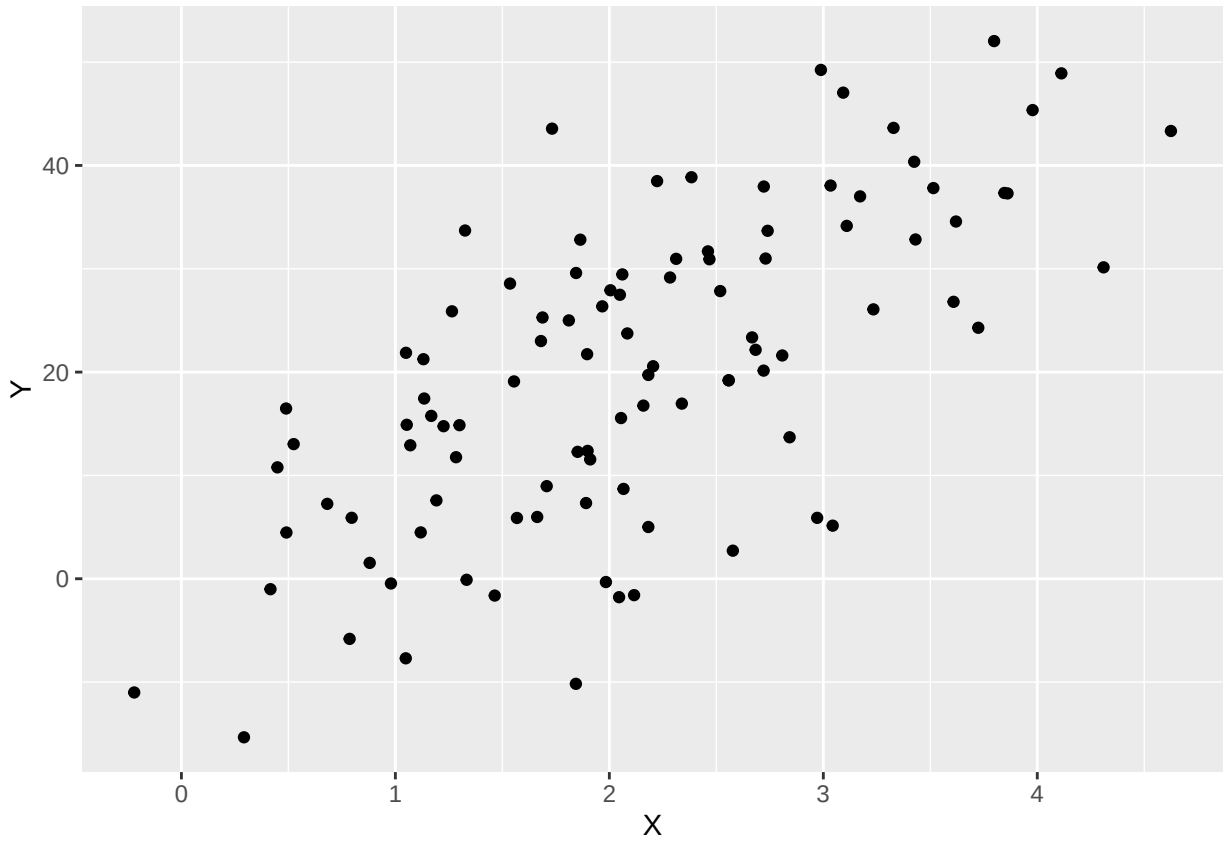
With the `save` *command, both the name and the data included in* `donnees` *are stored (in a binary format, i.e. compressed).*

5. **Scatter plot**. Trace the scatter plot of Y versus X, first using the `plot` command then using the help of the `ggplot2` library.

```
plot(donnees$X,donnees$Y)
```

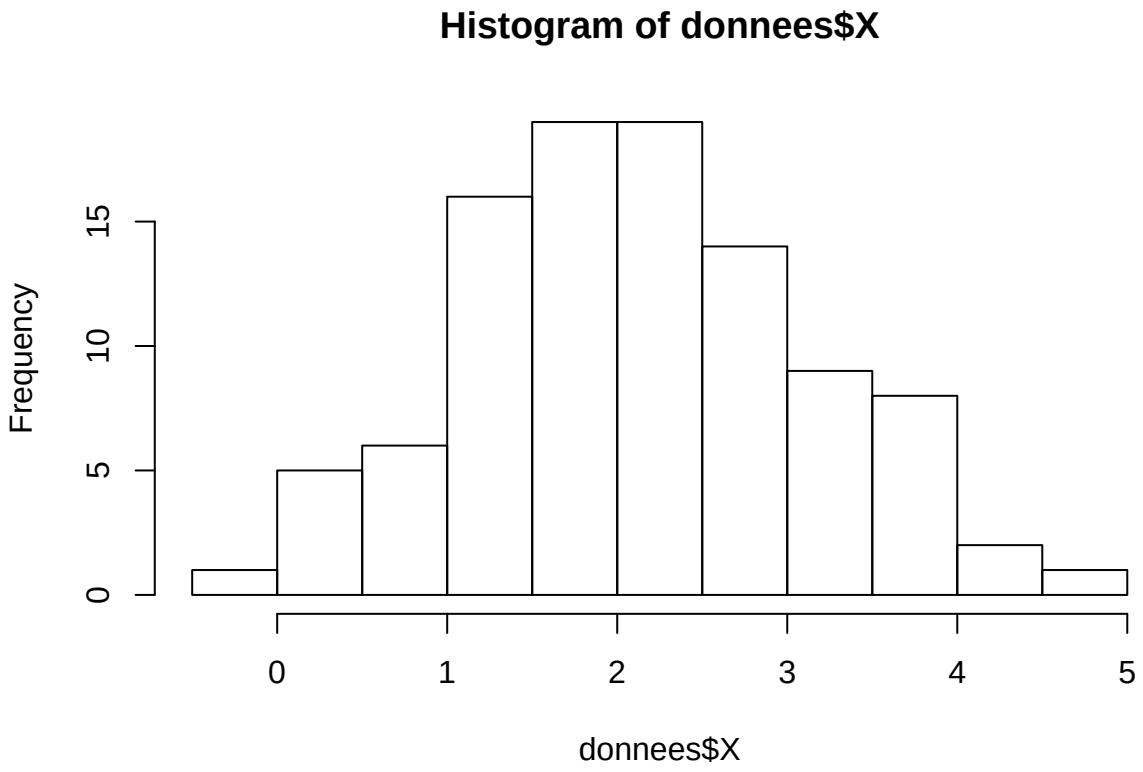



```
ggplot(donnees, aes(x=X,y=Y)) + geom_point()  ## or qplot(donnees$X,donnees$Y)
```

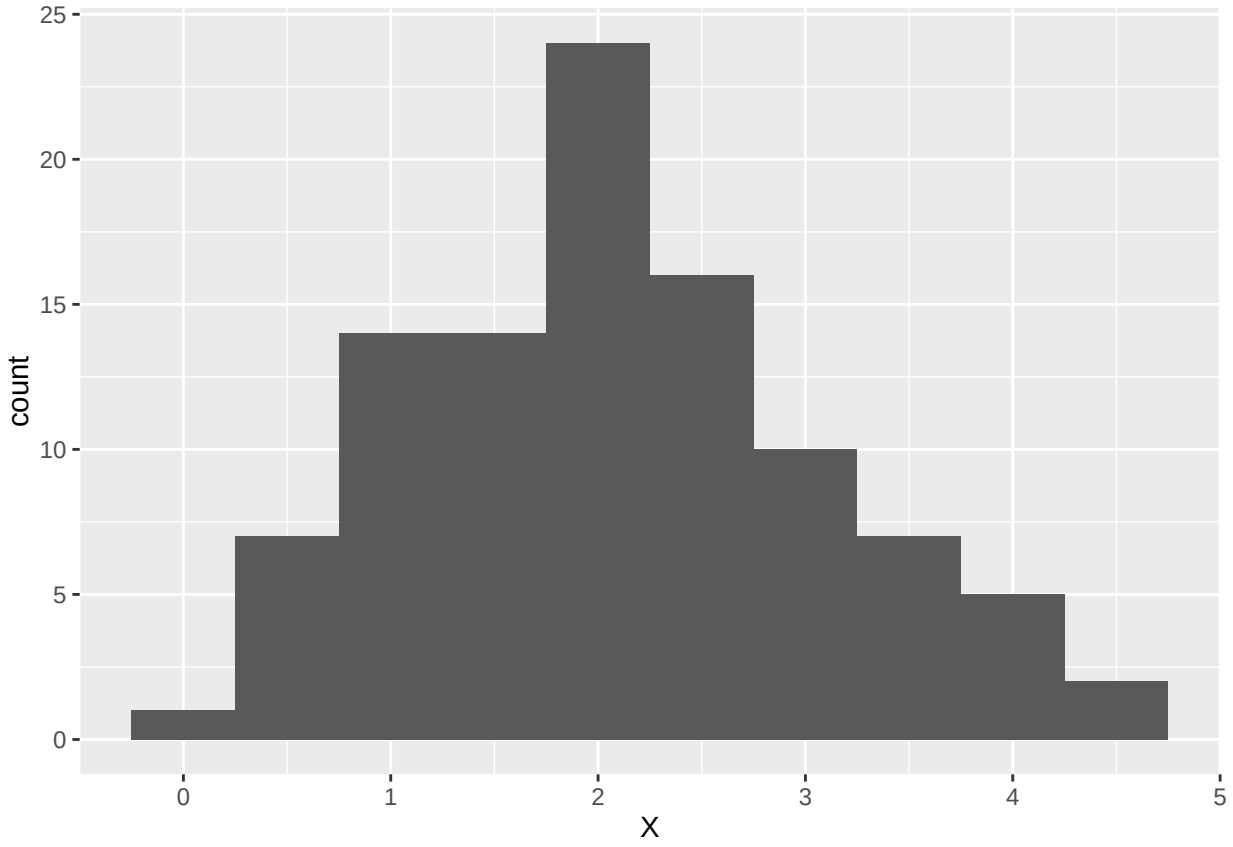


6. **Histogram**. Trace the histogram of X.

```
hist(donnees$X)
```

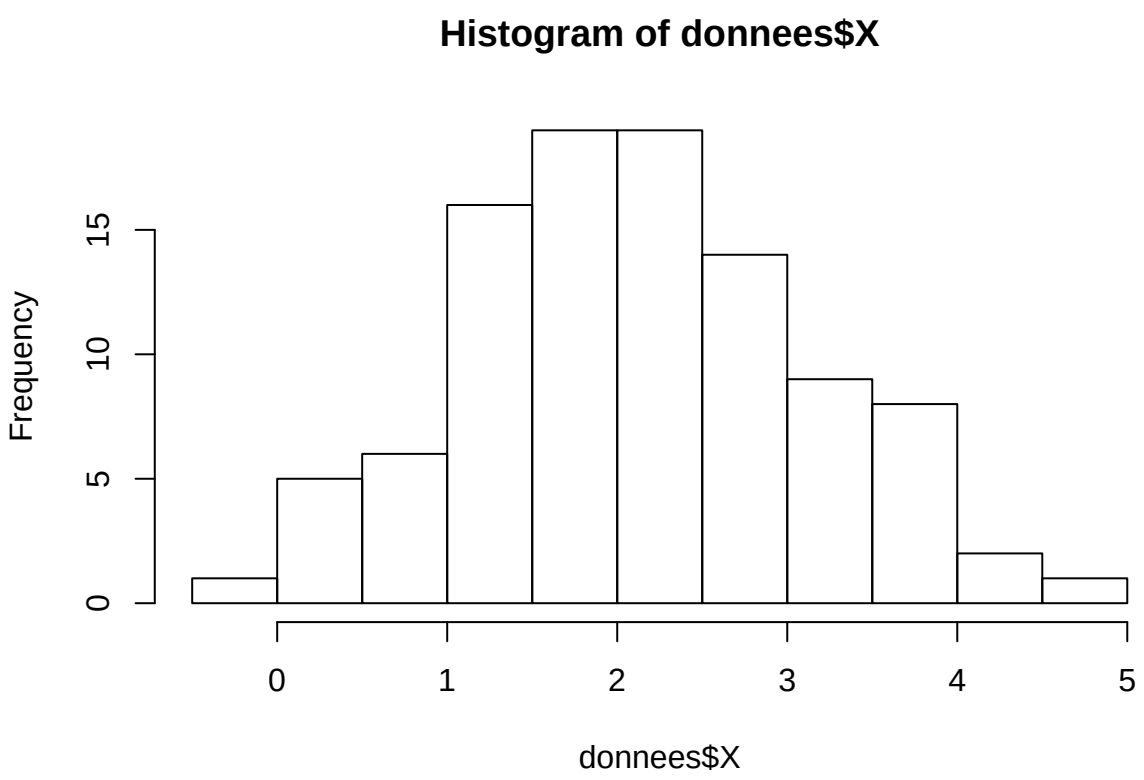



```
ggplot(donnees, aes(x=X)) + geom_histogram(binwidth=.5)
```

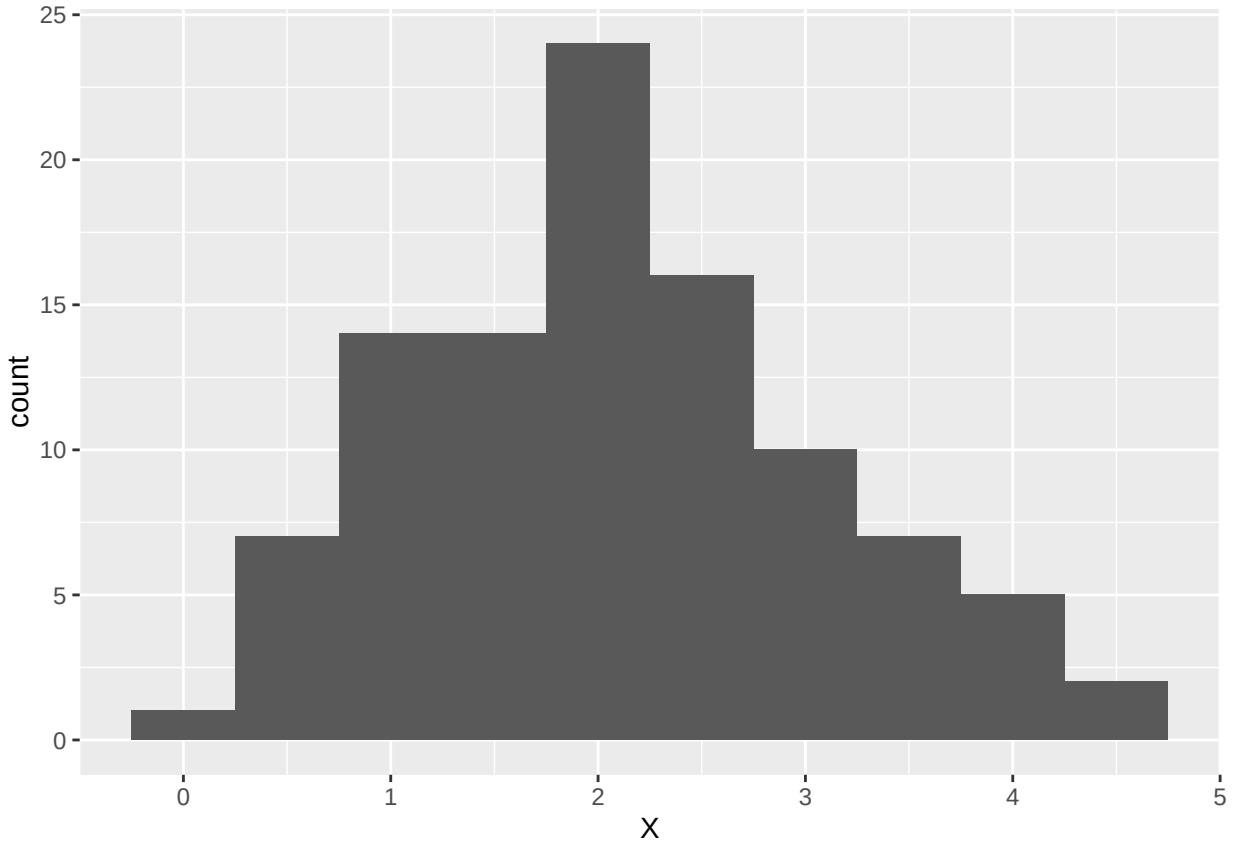


7. **“For”" loop**. A random variable follows the χ^2 distribution but we do not know its degreees of freedom. We want to estimate this degree of freedom. Using the empirical mean of n samplings of this random variable. Use `R` to evaluate the quality of this estimation for $n = 3$ and $n = 100$. Use a `for` loop.

```
hist(donnees$X)
```

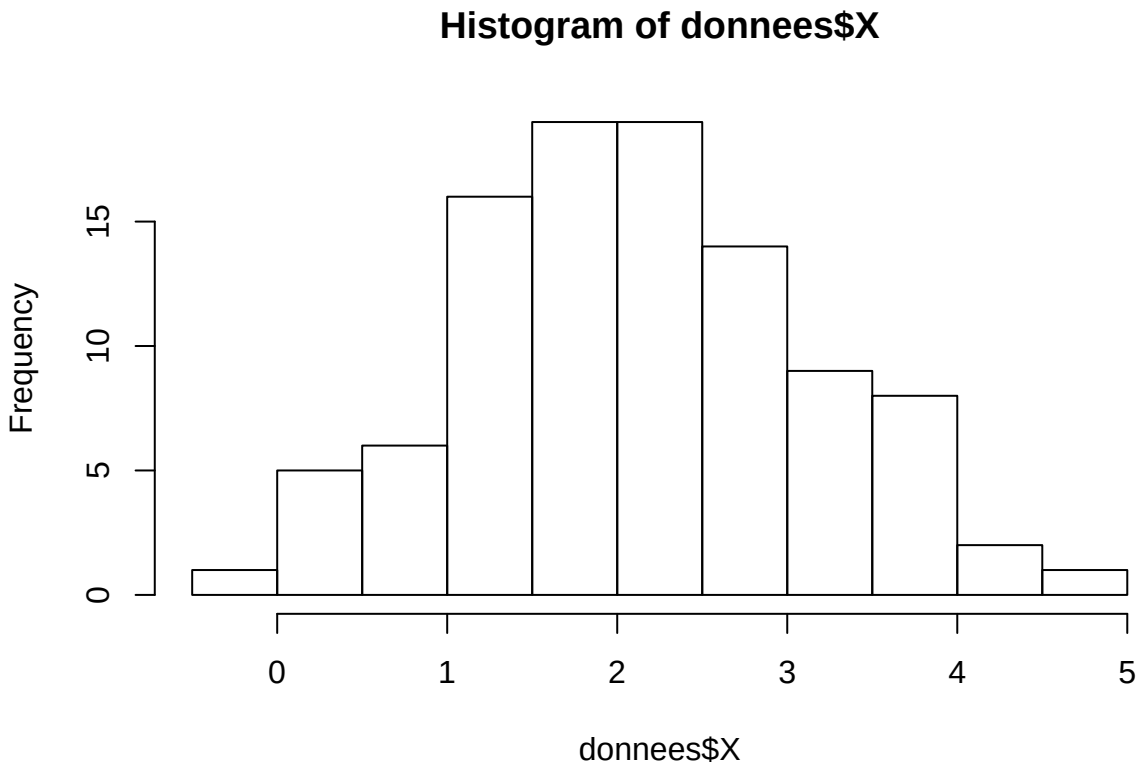



```
ggplot(donnees, aes(x=X)) + geom_histogram(binwidth=.5)
```

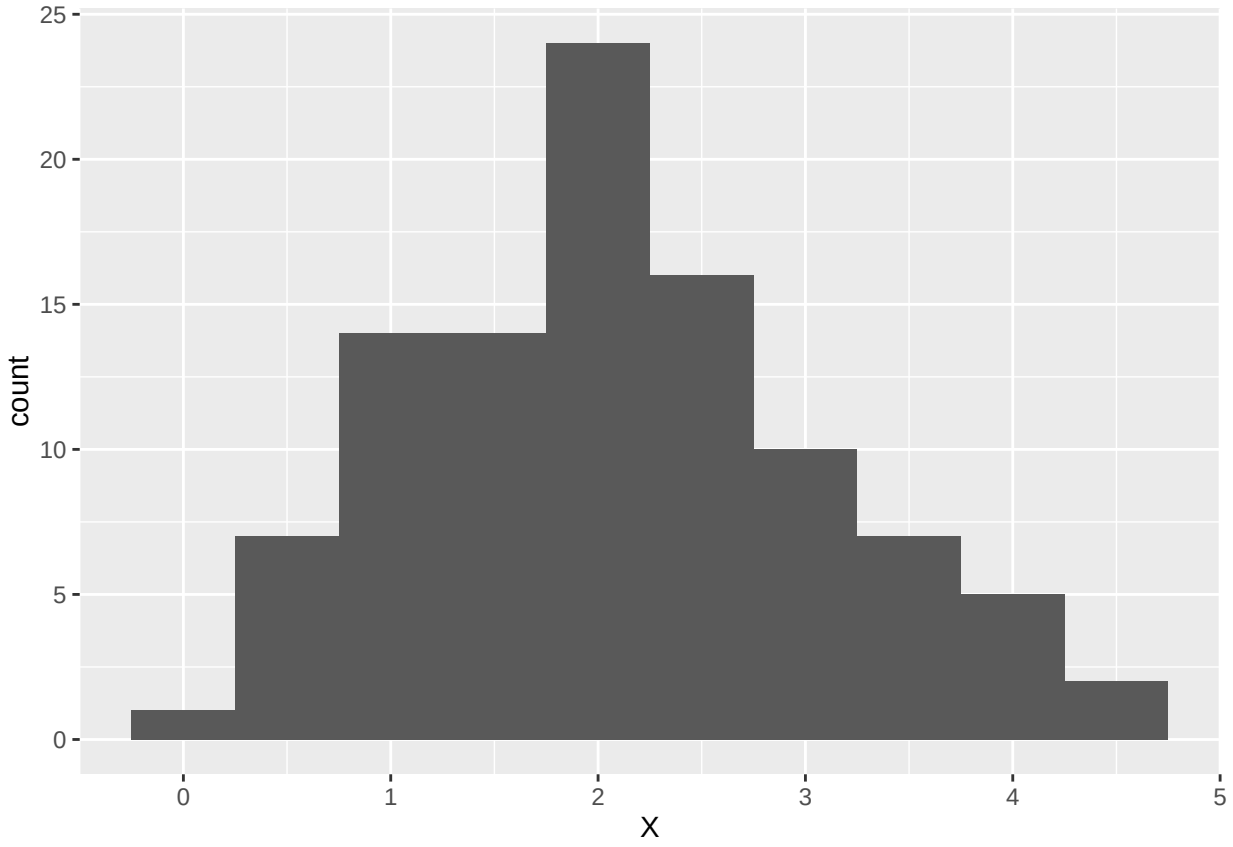


8. **Alternative "For" loop** (faster, more elegant, more adapted to `R`). Do the same thing with the `sapply` function.

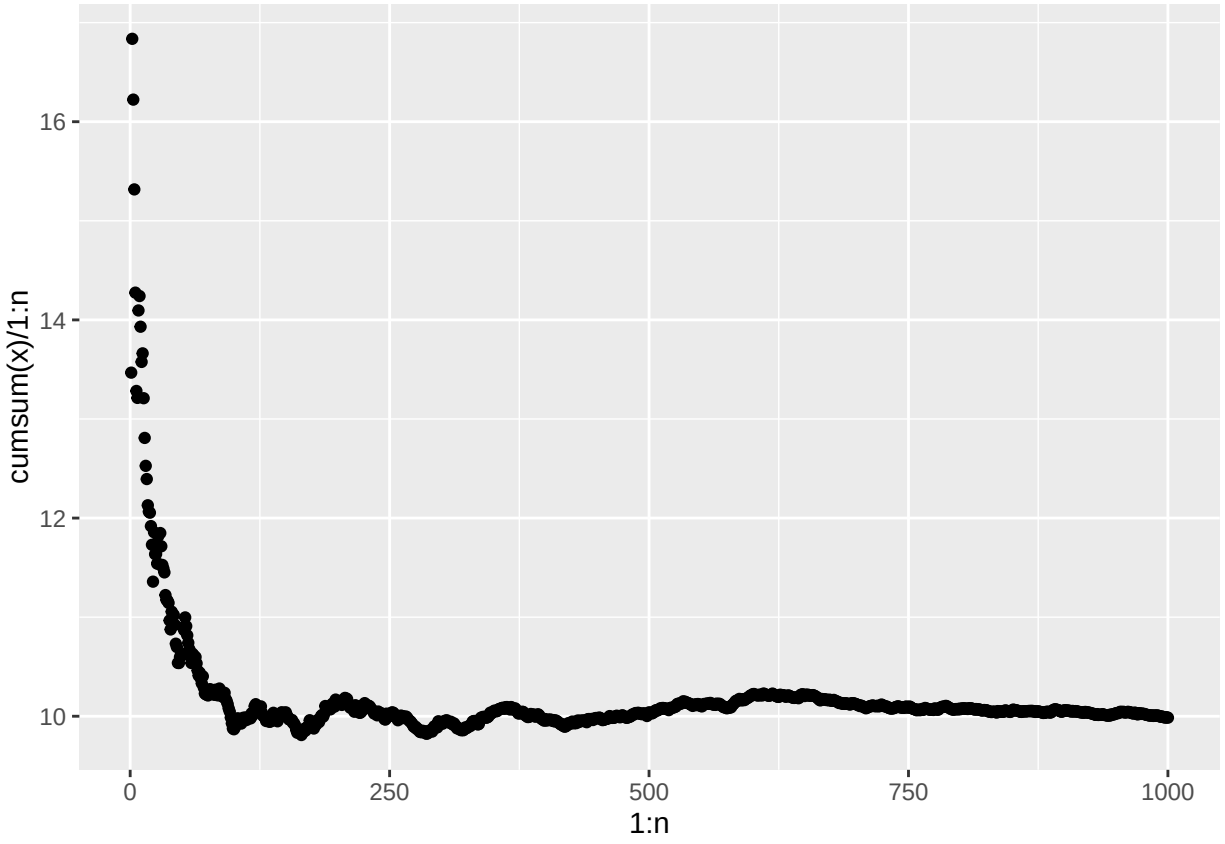
```
hist(donnees$X)
```




```
ggplot(donnees, aes(x=X)) + geom_histogram(binwidth=.5)
```



```r
n <- 1000
true.df <- 10
x <- rchisq(n,true.df)
qplot(1:n,cumsum(x)/1:n)
```



Simple linear regression: Northern pike and DDT

DDT (dichlorodiphenyltrichloroethane) is a relativelly potent pesticide. It has a high level of toxicity and non-biodegradable nature. It can accumulate in the liver and certain other tissues. We study here the effect of its accumulation on the Northern pike (*Brochet* in French).

1. Preliminaries

a) Import the data from 'Brochet.txt'.

```r
brochets <- read.table(file="Brochet.txt", row.names=1, header=TRUE)
```

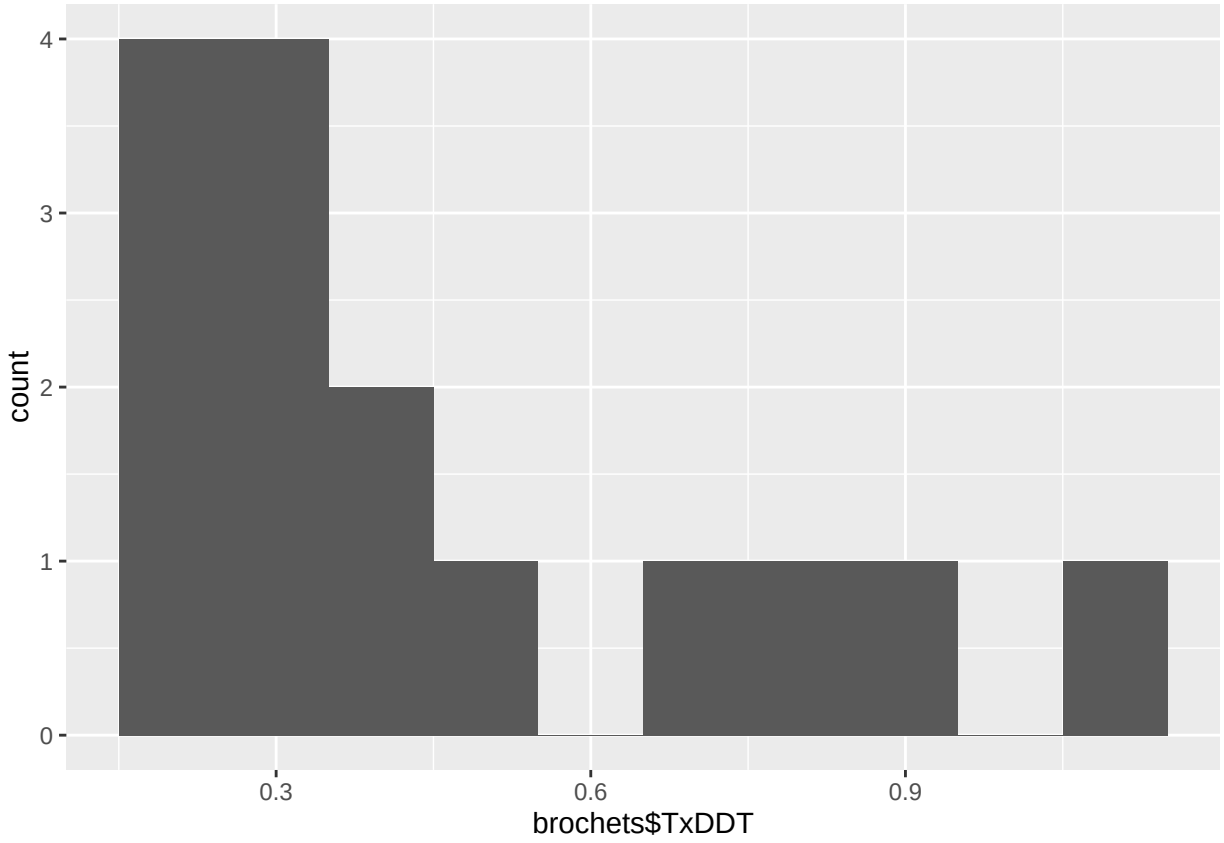
b) Calculate the mean, median and variance of the age of the Northern pike and of their accumulation of DDT.

```r
summary(brochets)
```

```
##       Age         TxDDT
##  Min.   :2   Min.   :0.180
##  1st Qu.:3   1st Qu.:0.265
##  Median :4   Median :0.330
##  Mean   :4   Mean   :0.450
##  3rd Qu.:5   3rd Qu.:0.590
##  Max.   :6   Max.   :1.100
```

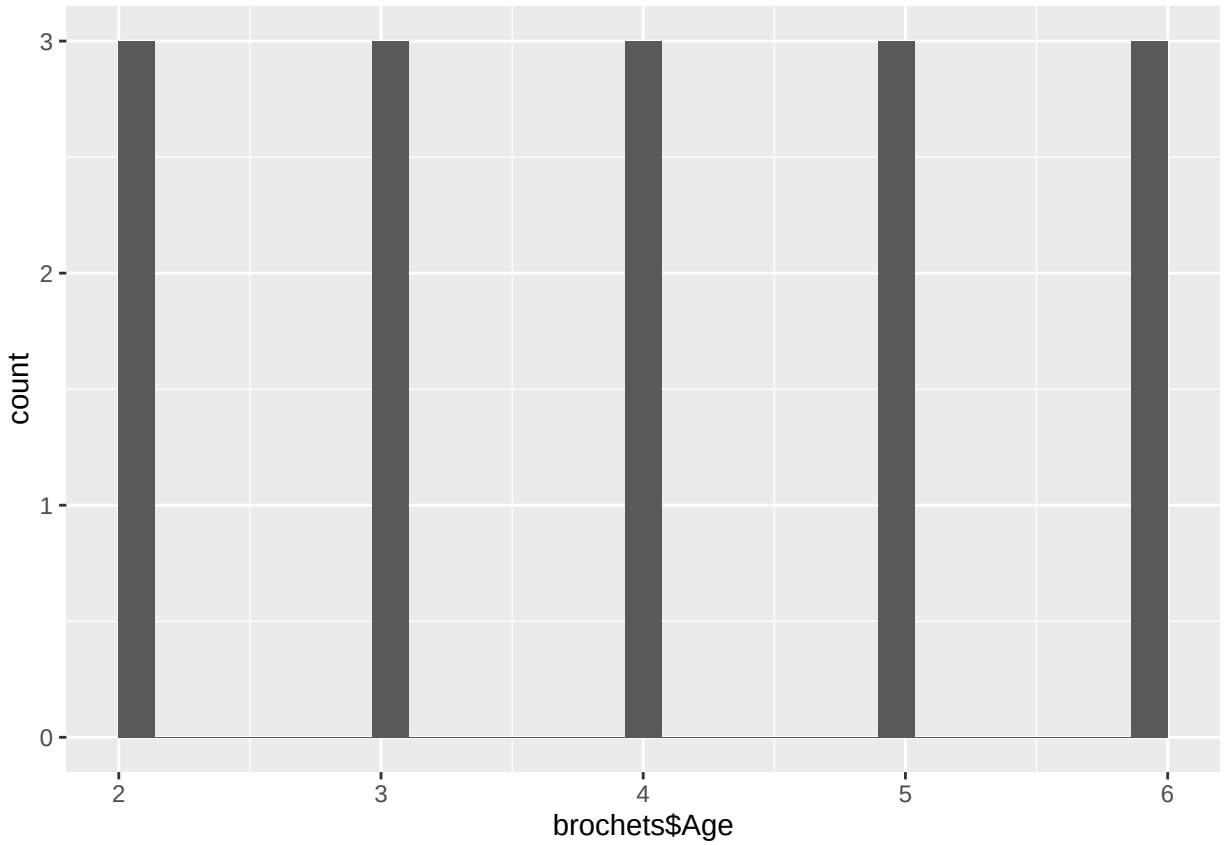
c) Plot the histograms of the age of the Northern pike and of their accumulation of DDT.

```
qplot(brochets$TxDDT, geom="histogram", binwidth=.1)
```




```
qplot(brochets$Age, geom="histogram")
```

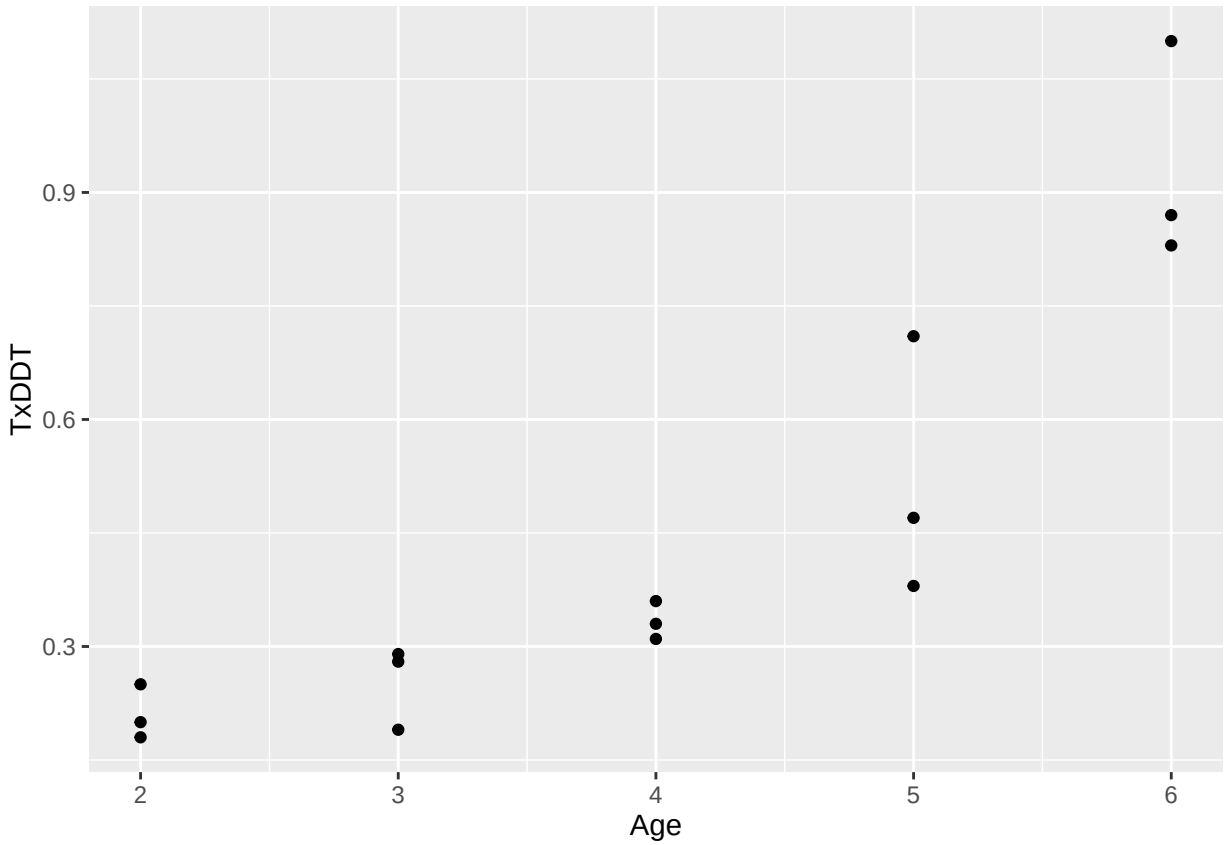
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The distribution by age group is homogenious.

d) Trace the scatter plot of the accumulation of DDT in the Northern pike versus their age.

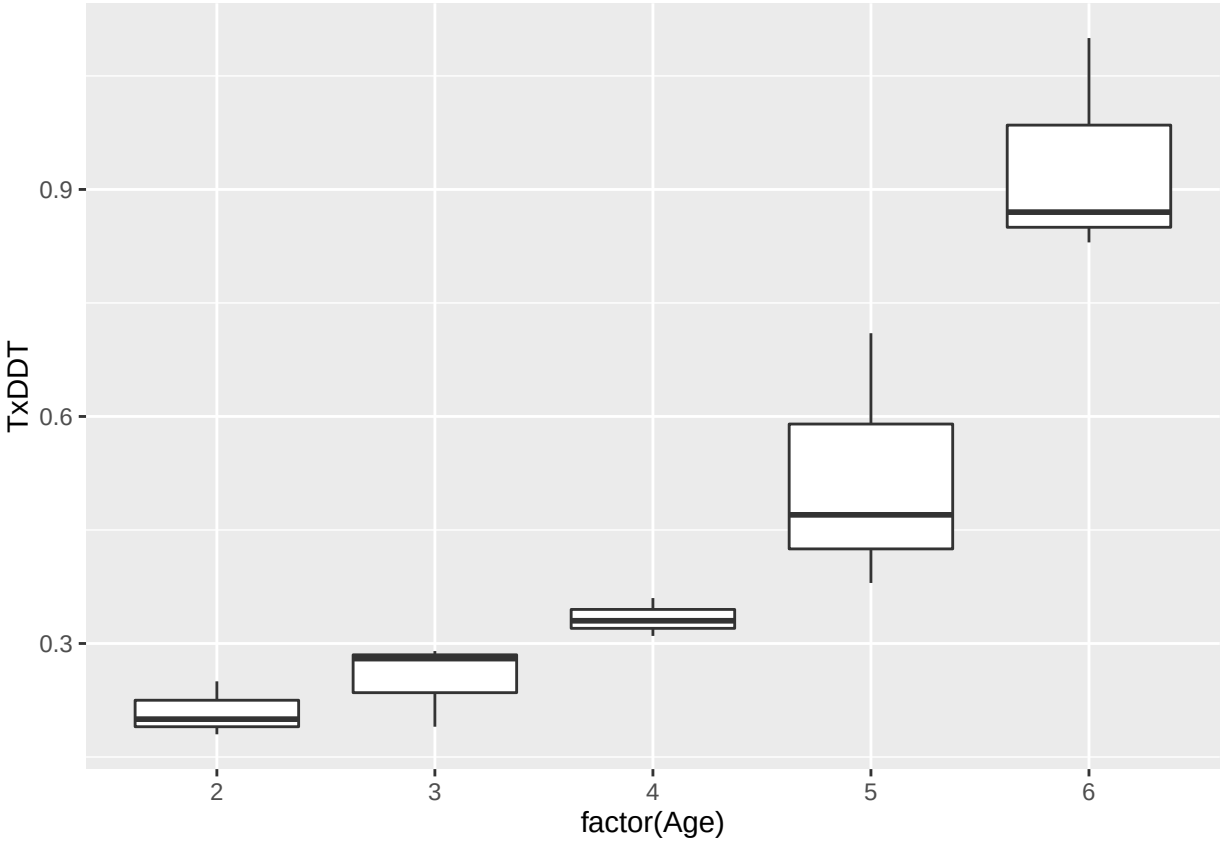
```
ggplot(brochets, aes(Age, TxDDT)) + geom_point()
```



This scatter plot seems to indicate that there exists a strong link between the age and the concentration of DTT in the Northern pike. It could lead to a hypothesis that there is a relation between these two variables, but is not enough to draw conslusions on the type and signifiance of this relationship. Such a hypothesis needs to be confirmed through statistical tests.

e) Make a bloxplot of the accumulation of DDT in the Northern pike versus their age. What do you observe?

```
ggplot(brochets, aes(factor(Age), TxDDT)) + geom_boxplot()
```



This graph firther confirms the existance of a relation between the two variables but indicates that the variance of DDT concentration is not constant in each age group. This is further seen in the next example.

f) Calulate the variance of the concentration of DDT in each age group.

```
with(brochets, tapply(TxDDT, Age, var))
```

```
##            2            3            4            5            6
## 0.0013000000 0.0030333333 0.0006333333 0.0291000000 0.0212333333
```

2. A first model

a) Create a linear regression model to explain the concentration of DDT in a Northern pike based on its age.

We define as (X_i, Y_i) the couple (Age,TxDDT) of the i^{th} individual. We can define the model as:

$$Y_i = \beta_0 + X_i\beta_1 + \varepsilon_i$$

b) Use `R` to estimate the parameters of this model. First apply the formulation given in the lecture, then use the `lm` function. Calculate the slope, bias and residual variance.

```r
## Calculation based on the lecture
n <- nrow(brochets)
beta1 <- with(brochets, cov(Age,TxDDT)/var(Age))
beta0 <- with(brochets, mean(TxDDT) - mean(Age) * beta1)
sigma.hat <- sqrt(with(brochets, sum((TxDDT - beta0 - beta1*Age)^2))/(n-2))
cat("\nbeta0, beta1, sigma.hat: ", c(beta0,beta1,sigma.hat))
```

```
##
## beta0, beta1, sigma.hat:  -0.2353333 0.1713333 0.1454824
```

```r
## Calculation using the lm function
M1 <- lm(TxDDT~Age,brochets)
cat("\nbeta0, beta1, sigma.hat: ", c(coefficients(M1), sqrt(sum(residuals(M1)^2)/(n-2))))
```

```
##
## beta0, beta1, sigma.hat:  -0.2353333 0.1713333 0.1454824
```

c) Test the parameters of the model. Make a analysis of variance. Calculate `manually` the Fisher score and the r^2 coefficient.

```r
## Fisher and $r^2$ scores
SCR <- sum(residuals(M1)^2)
SCM <- with(brochets, sum((fitted(M1) - mean(TxDDT))^2))
SCT <- with(brochets, sum((TxDDT - mean(TxDDT))^2))
f <- (SCM/1) / (SCR/(n-2))
r2 <- SCM/SCT

## We find the same values as those obtained through annova
anova(M1)
```

```
## Analysis of Variance Table
##
## Response: TxDDT
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Age        1 0.88065 0.88065  41.609 2.165e-05 ***
## Residuals 13 0.27515 0.02117
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The information obtained by using the ***summary*** *command is even richer: the student test is performed on each regression coefficient, along with the estimator's value, standard deviation and its $p-value$. This analysis is conclusive as to the significativaty of the slope, but with a slight error at the origin of the axes since a new-born Northern pike's concentration should be quasi-null.*

We also can see that the r^2 statistic indicates that the model is able to explain 75% of the total variability of the concentration of DDT in the Northern pike based on its age.

```r
summary(M1)
```

```
##
## Call:
```

```
## lm(formula = TxDDT ~ Age, data = brochets)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24133 -0.10500  0.01133  0.08300  0.30733
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.23533    0.11269  -2.088    0.057 .
## Age          0.17133    0.02656   6.450 2.16e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1455 on 13 degrees of freedom
## Multiple R-squared:  0.7619, Adjusted R-squared:  0.7436
## F-statistic: 41.61 on 1 and 13 DF,  p-value: 2.165e-05
```

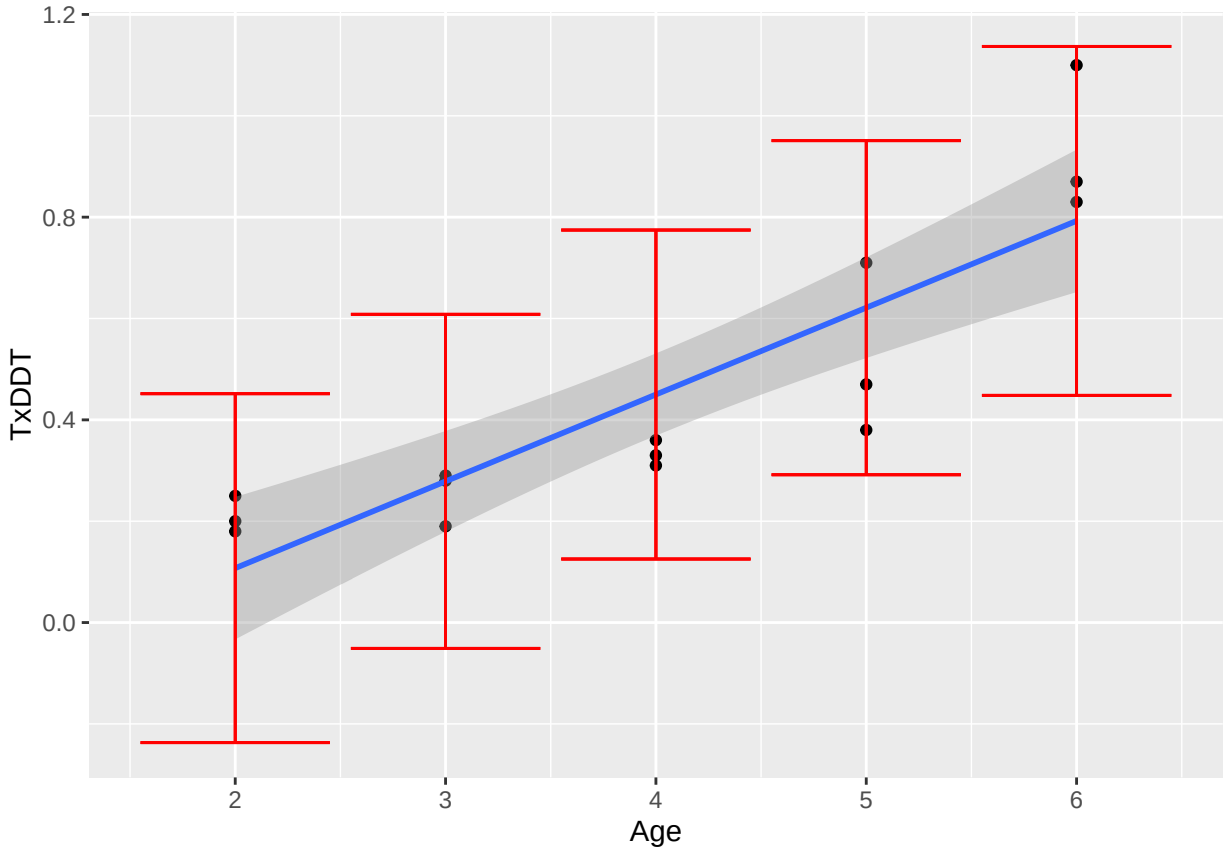
d) Plot the linear regression. Add confidence intervals. Calculate those first by applying the formulas from in the lecture and then by using the `predict`command.

ggplot includes the confidence interval on the predicted values with the prediction interval we add a posteriori.

```
inter.pred <- data.frame(predict(M1,interval="prediction")[, -1])
```

```
## Warning in predict.lm(M1, interval = "prediction"): predictions on current data refer to _future_ res
```

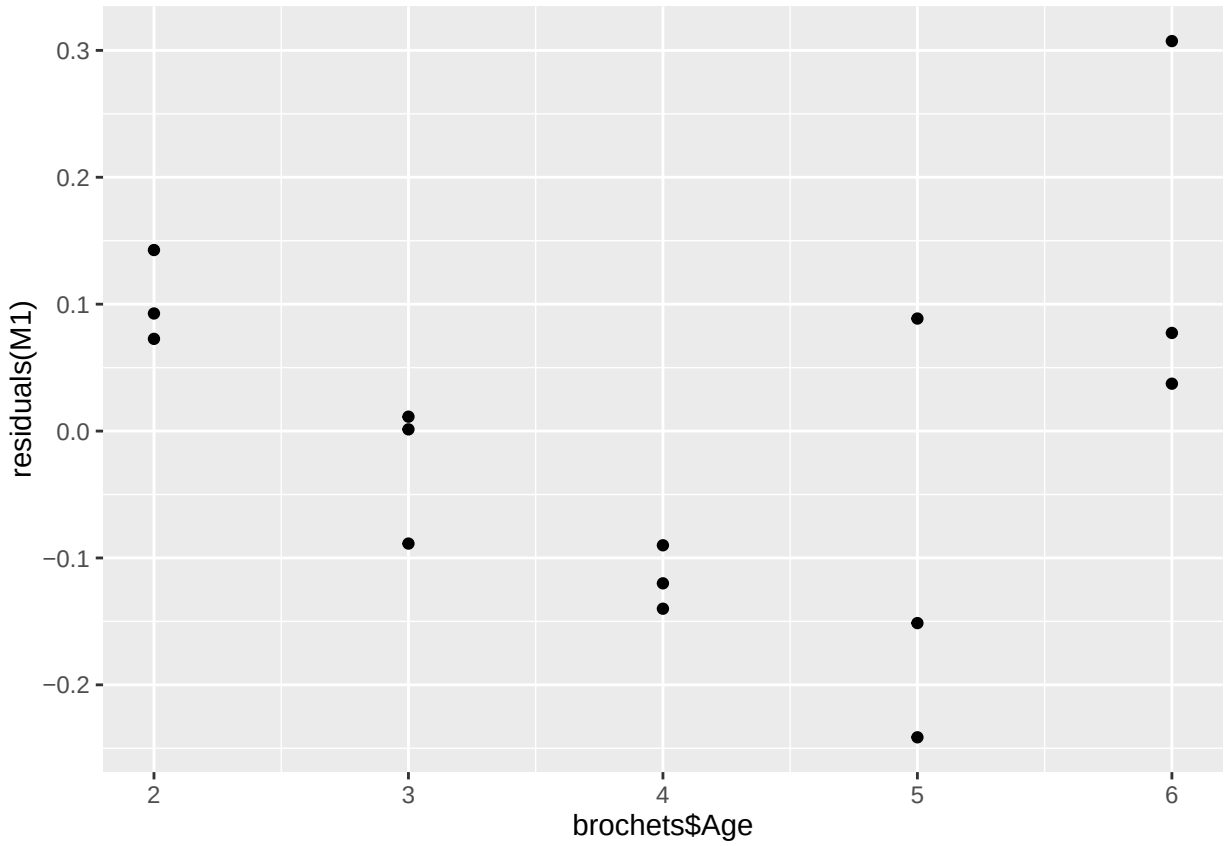
```
ggplot(cbind(brochets, inter.pred), aes(Age, TxDDT)) +
  geom_point() + stat_smooth(method="lm", formula=y~x) +
  geom_errorbar(aes(x=Age, ymin=lwr, ymax=upr), colour="red")
```




```
## Manually we obtain the same results
Y.hat <- beta0 + beta1*brochets$Age
alpha <- 0.05
T <- with(brochets, (Age-mean(Age))^2/sum((Age-mean(Age))^2))
lower <- Y.hat - qt(1-alpha/2, df=n-2) * sigma.hat * sqrt(1 + 1/n + T)
upper <- Y.hat + qt(1-alpha/2, df=n-2) * sigma.hat * sqrt(1 + 1/n + T)
```

e) Plot a graph og the residues to evaluate the pertience of your model and perform ###use diagnostics###. You can also use the `plot` function of `R` applied to the outputs of the `lm` function.

```
qplot(brochets$Age, residuals(M1))
```




```
shapiro.test(residuals(M1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(M1)
## W = 0.96554, p-value = 0.7874
```

Problem: strong exponential trend in the residuals. The normality tests however are conclusive as to the estimated values.

3. Square function Modeling

a) Create a new linear regression model to explain the concentration of DDT in a Northern pike based on the square of its age.

We define as (X_i, Y_i) the couple (Age,TxDDT) of the i^{th} individual. We can define the model as:

$$Y_i = \beta_0 + X_i^2 \beta_1 + \varepsilon_i$$

b) Use `R` to estimate the parameters of this model.

```
M2 <- lm(TxDDT~I(Age^2), brochets)
```

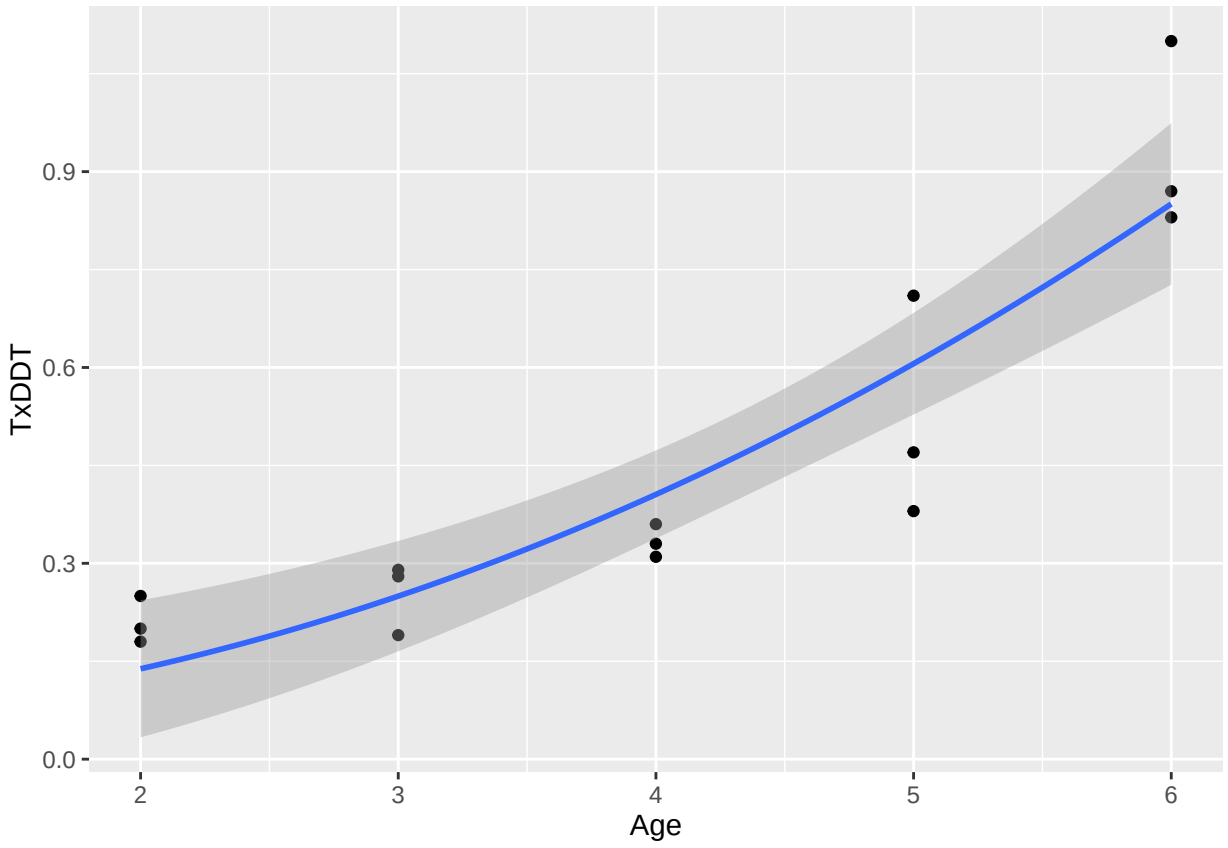
c) Test the model's parameters and performs and analysis of variation.

```
summary(M2)
```

```
##
## Call:
## lm(formula = TxDDT ~ I(Age^2), data = brochets)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.22577 -0.06761  0.01945  0.05154  0.24945
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.049450   0.057317   0.863    0.404
## I(Age^2)    0.022253   0.002688   8.280 1.53e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.119 on 13 degrees of freedom
## Multiple R-squared:  0.8406, Adjusted R-squared:  0.8283
## F-statistic: 68.55 on 1 and 13 DF,  p-value: 1.532e-06
```

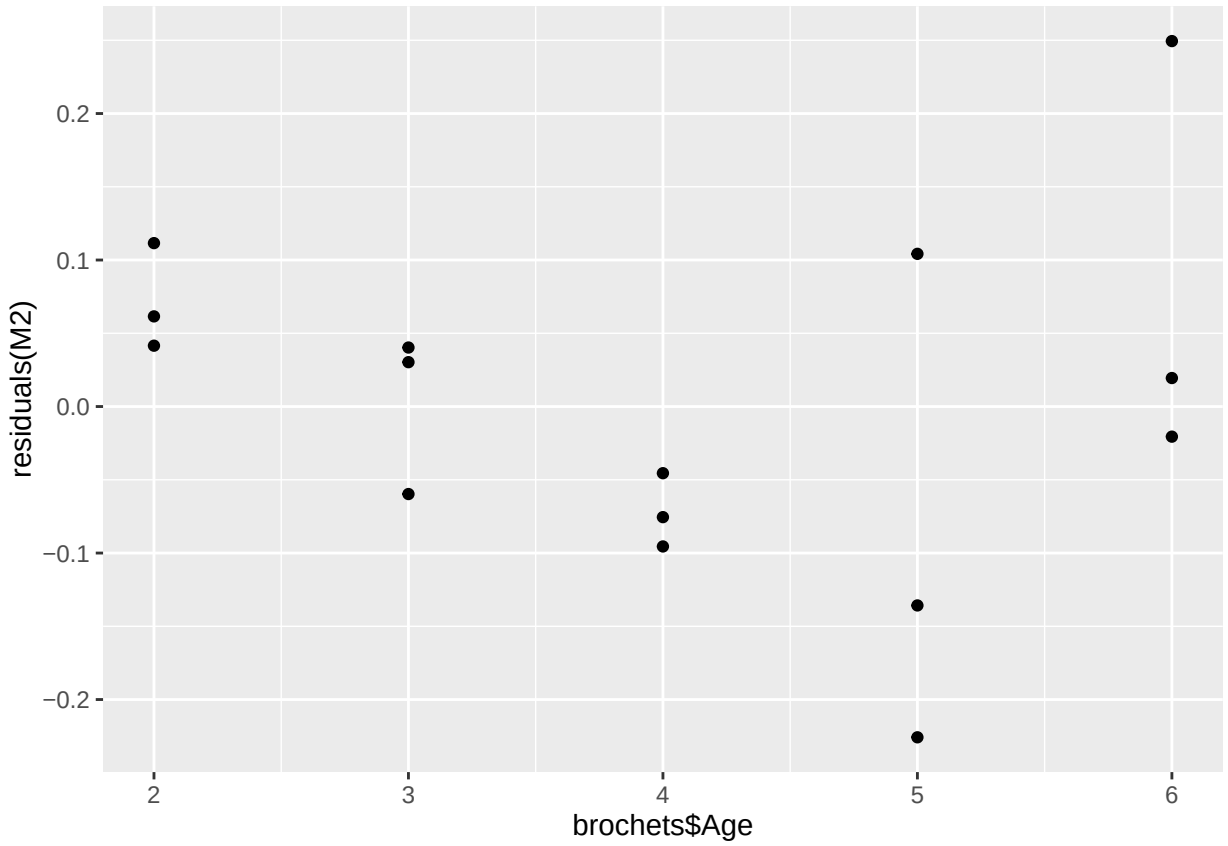
d) Plot the regression. You can use the `geom_smooth` function of `ggplot`.

```
ggplot(brochets, aes(Age, TxDDT)) + geom_point() + stat_smooth(method="lm", formula=y~I(x^2))
```



e) Perform a diagnostic test of the model.

```
## Still not a white noise.
qplot(brochets$Age, residuals(M2))
```



4. Logarithmic transformation model

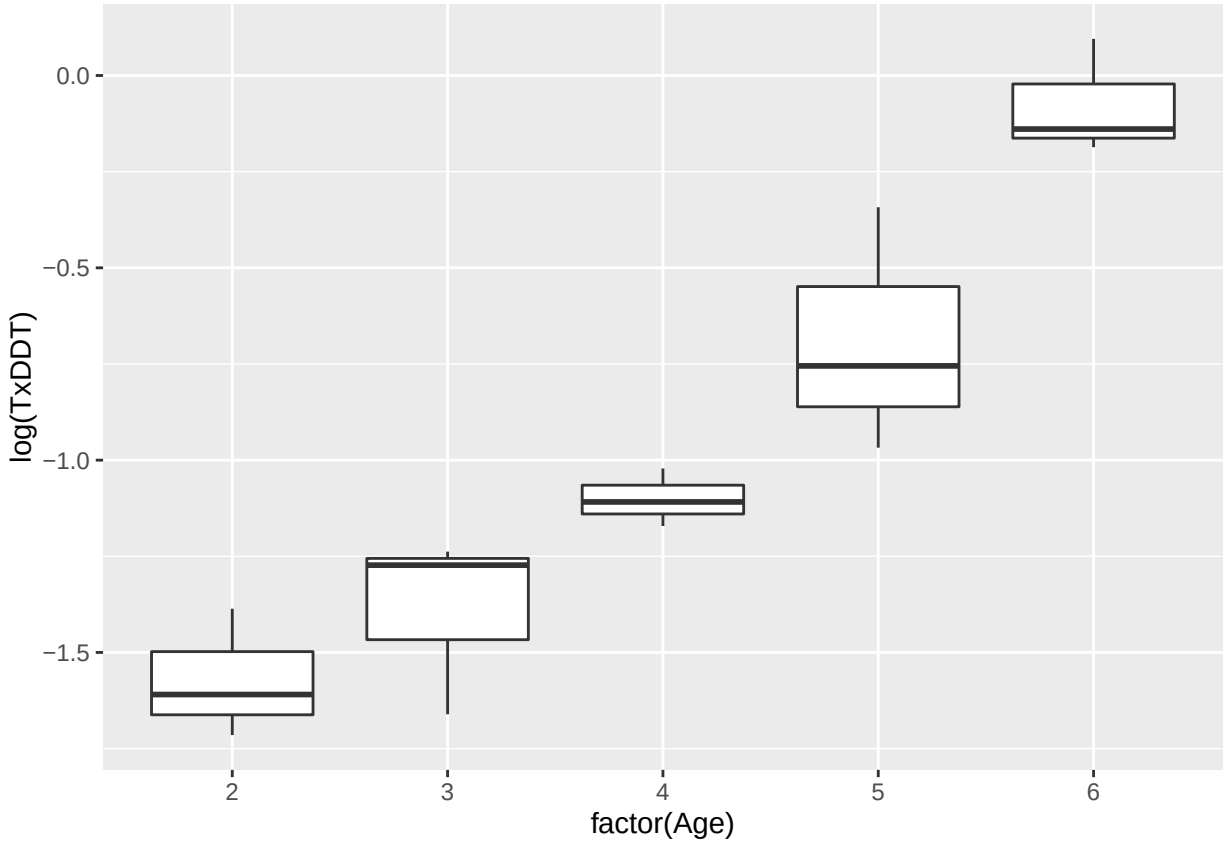
a) Create a new linear regression model to explain the log of the concentration of DDT in a Northern pike based on its age.

We define as (X_i, Y_i) the couple (Age,TxDDT) of the i^{th} individual. We can define the model as:

$$log(Y)_i = \beta_0 + X_i\beta_1 + \varepsilon_i$$

b) Trace the box-plot of the log of the DTT concentration in Northern pike as a function of its age. What do you notice?

```
ggplot(brochets, aes(factor(Age), log(TxDDT))) + geom_boxplot()
```



The tendency now seems to be linear between the two variables and the variance in each bin seems homogenious.

c) Calculate the variance of the log of the DDT concentration in each age group.

```
with(brochets, tapply(log(TxDDT), factor(Age), var))
```

```
##           2           3           4           5           6
## 0.028134801 0.055066904 0.005639919 0.101017900 0.022760074
```

d) Use R to estimate the parameters of this model.

```
M3 <- lm(log(TxDDT)~Age,brochets)
```

e) Test the model's parameters and perform an analysis of variance.

```
summary(M3)
```

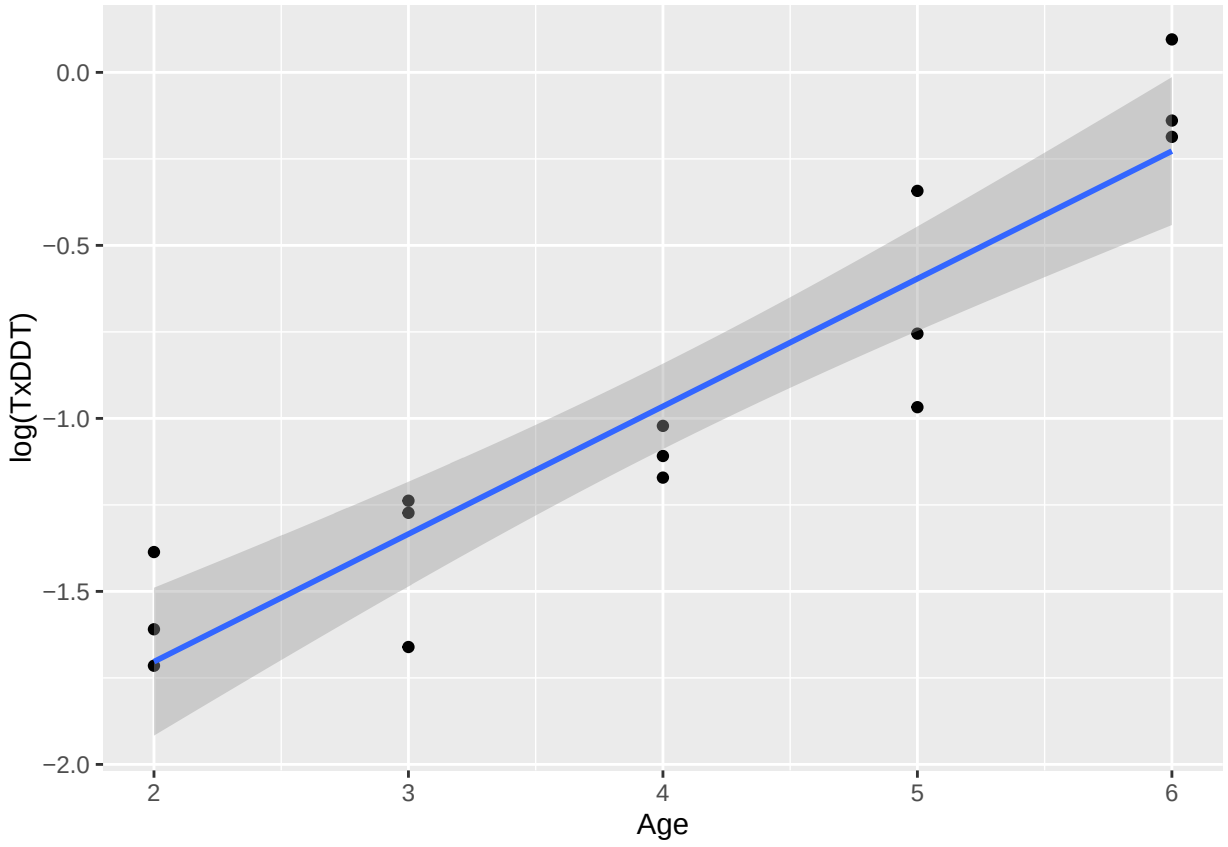
```
##
## Call:
## lm(formula = log(TxDDT) ~ Age, data = brochets)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37122 -0.15103  0.04114  0.09496  0.32278
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.44086    0.17135 -14.245 2.61e-09 ***
## Age          0.36890    0.04039   9.134 5.09e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2212 on 13 degrees of freedom
## Multiple R-squared:  0.8652, Adjusted R-squared:  0.8548
## F-statistic: 83.43 on 1 and 13 DF,  p-value: 5.092e-07
```

The log transformations increases the significativity of the model. Having performed a log transformation the values cannot cross the X=0 axis.

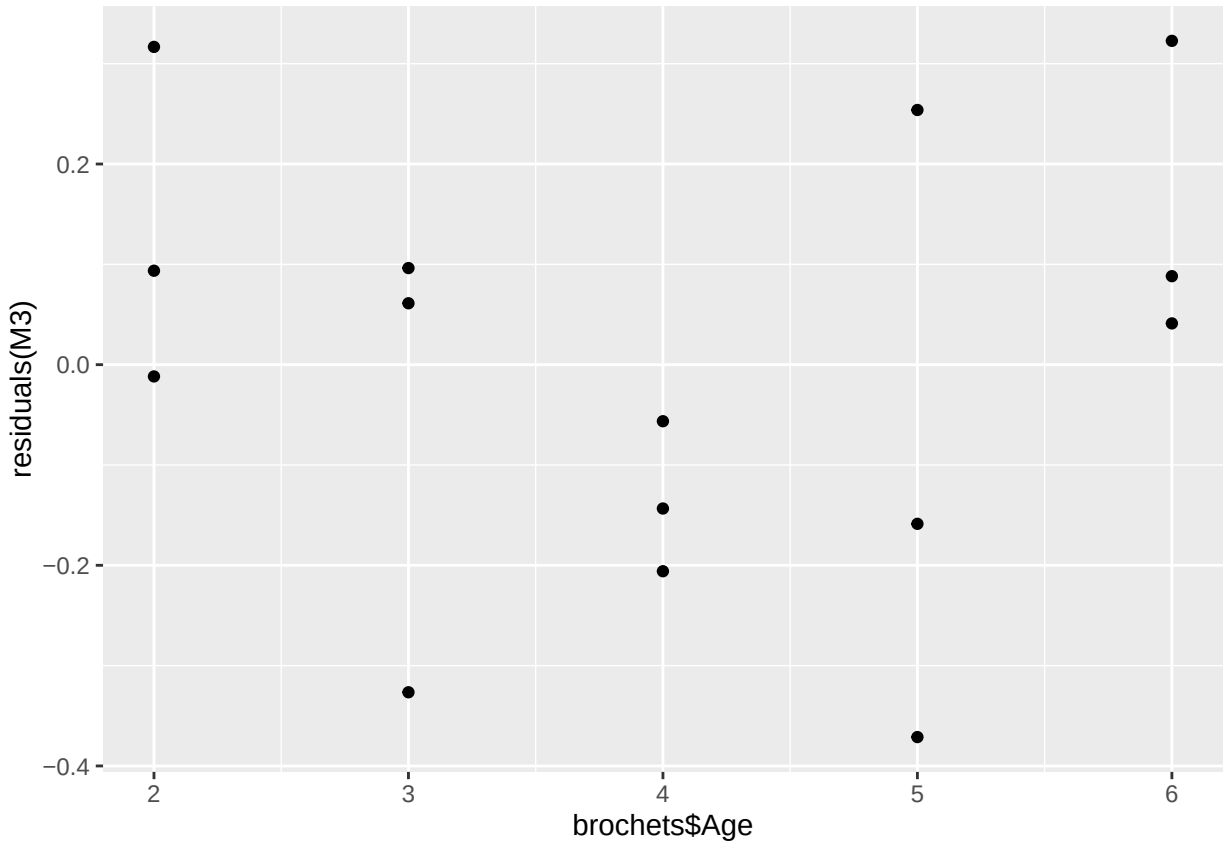
f) Plot the regression line. You can use the `geom_smooth` function of `ggplot`.

```
ggplot(brochets, aes(Age, log(TxDDT))) + geom_point() + stat_smooth(method="lm", formula=y~x)
```



g) Perform a model diagnostic.

```
qplot(brochets$Age, residuals(M3))
```



It is still better but the results can be improved upon by using multi-linear regression.

5. Towards a multi-linear regression

a) Create a new linear regression model to explain the log of the concentration of DDT in a Northern pike based on its age and the square of its age.

We define as (X_i, Y_i) the couple (Age,TxDDT) of the i^{th} individual. We can define the model as:

$$log(Y)_i = \beta_0 + X_i\beta_1 + X_i^2\beta_2 + \varepsilon_i$$

b) Use `R` to estimate the parameters of this model.

```
M4 <- lm(log(TxDDT)~Age + I(Age^2), brochets)
```

c) Test the parameters of the model. Perform an analysis of variance to compare the 3 models M2, M3, M4 (log, + Age, + square of the Age).

```
summary(M4)
```

```
##
## Call:
## lm(formula = log(TxDDT) ~ Age + I(Age^2), data = brochets)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30079 -0.09397 -0.04723  0.14917  0.32431
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.45485    0.43331  -3.358   0.0057 **
## Age         -0.19454    0.23561  -0.826   0.4251
## I(Age^2)     0.07043    0.02913   2.417   0.0325 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1888 on 12 degrees of freedom
## Multiple R-squared:  0.9093, Adjusted R-squared:  0.8942
## F-statistic: 60.18 on 2 and 12 DF,  p-value: 5.553e-07
```

```
anova(M3, M4)
```

```
## Analysis of Variance Table
##
## Model 1: log(TxDDT) ~ Age
## Model 2: log(TxDDT) ~ Age + I(Age^2)
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1     13 0.63614
## 2     12 0.42781  1   0.20833 5.8438 0.03247 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

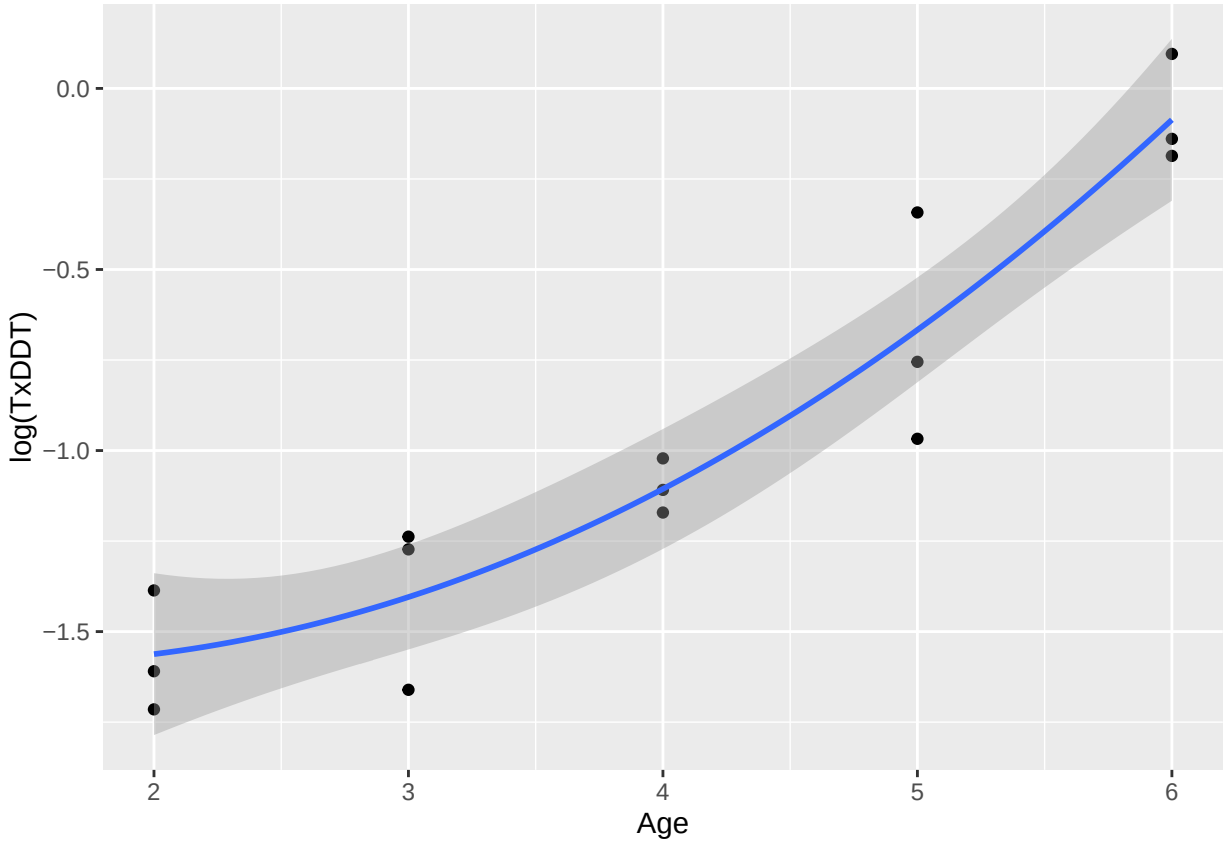
The R^2 is better than before (90% of the variability is explained), even though the age coefficient is not that significant. We can note that the anova between the M3 and M4 models corresonds to the Student test of the last parameter of the model. The Fisher test in the summary command corresponds to the anova between the model with simple log and the model with age and the quare of the age, as seen below.

```
anova(lm(log(TxDDT)~1, brochets), M4)
```

```
## Analysis of Variance Table
##
## Model 1: log(TxDDT) ~ 1
## Model 2: log(TxDDT) ~ Age + I(Age^2)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     14 4.7187
## 2     12 0.4278  2    4.2909 60.181 5.553e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

d) Plot the regression curve.

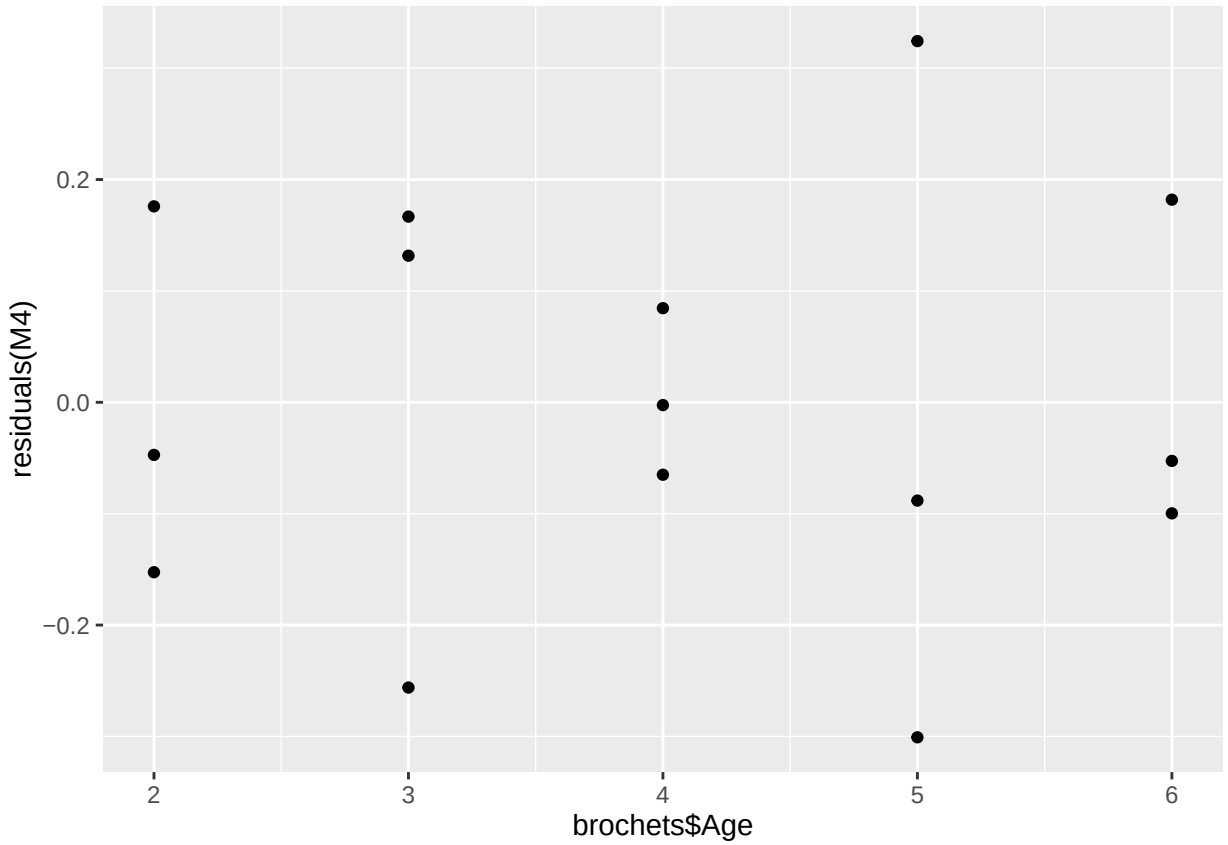
```
ggplot(brochets, aes(Age, log(TxDDT))) + geom_point() + stat_smooth(method="lm", formula=y~x + I(x^2))
```



e) Validate your hypothesis. Use R to evaluate the relevance of the model. What are your thoughts?

It is much better than before: there is not significant trend.

```
qplot(brochets$Age, residuals(M4))
```

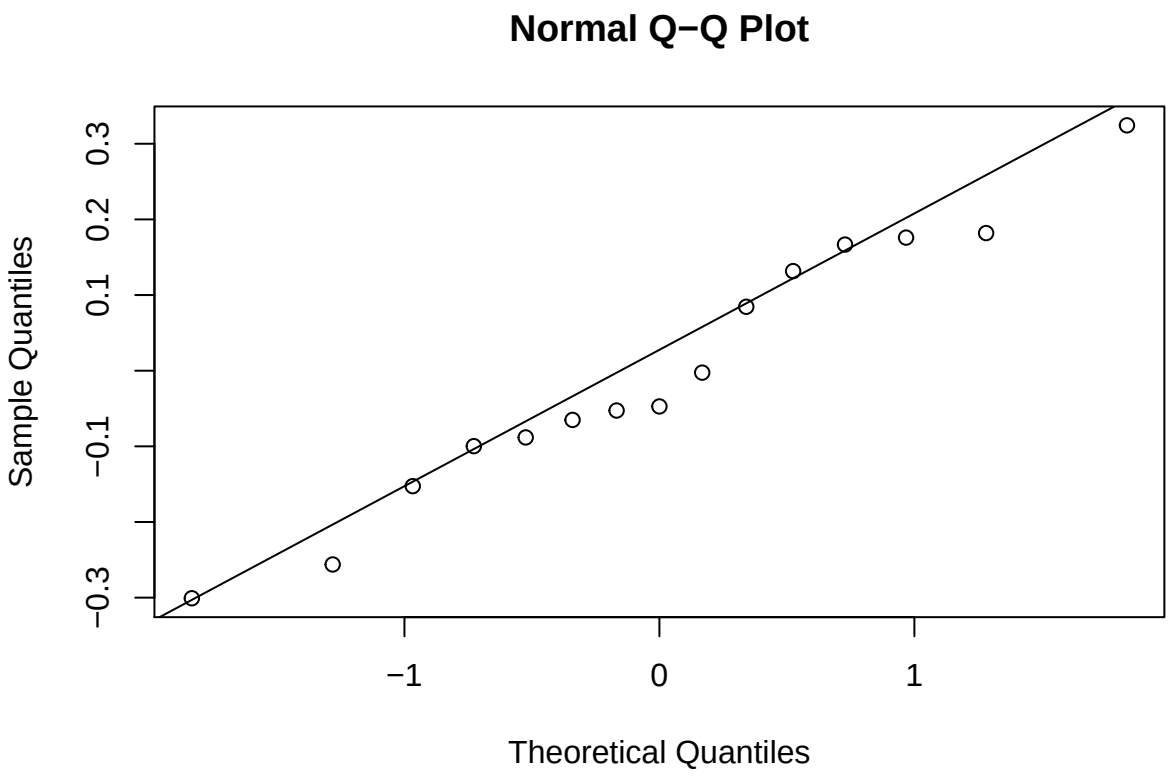


The indepence and normality tests seem to be conclusive:

```
## normamlity
shapiro.test(residuals(M4))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(M4)
## W = 0.96903, p-value = 0.8434
```

```
qqnorm(residuals(M4))
qqline(residuals(M4))
```




```
## independence
library(car)
durbinWatsonTest(M4)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.2090444      2.389629    0.82
##  Alternative hypothesis: rho != 0
```