Régression linéaire exercice sur les céréals

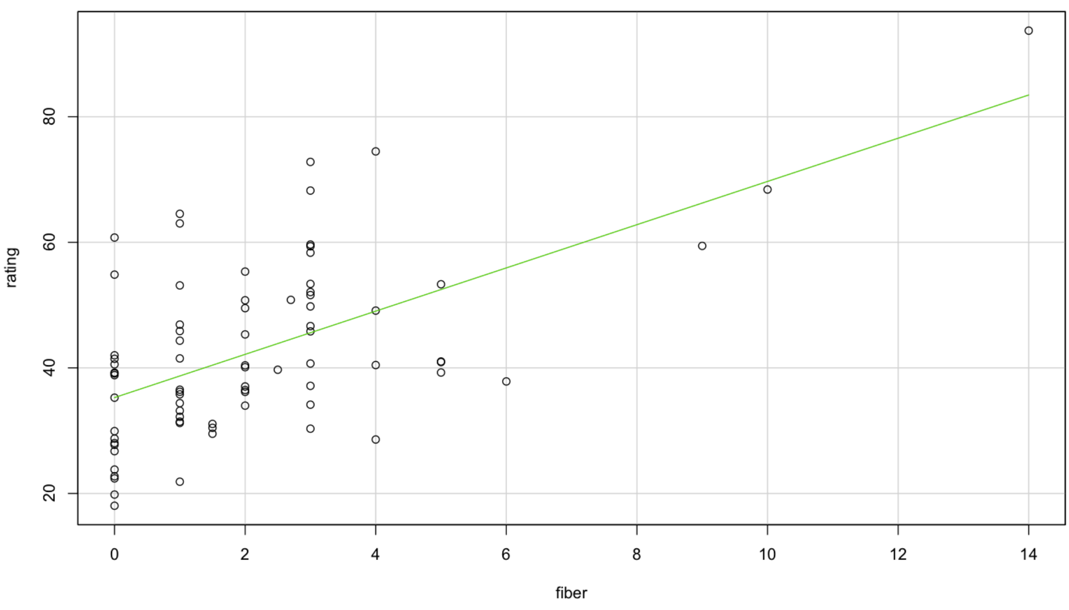
Rappels

SCT = somme(Yi – Ybar)^2 -> nos observations – moyenne
SCE = Somme(Yi - ^yi)^2 -> nos observation – regression linéaire
SCR = somme(yi – ybar)^2 -> regression - moyenne

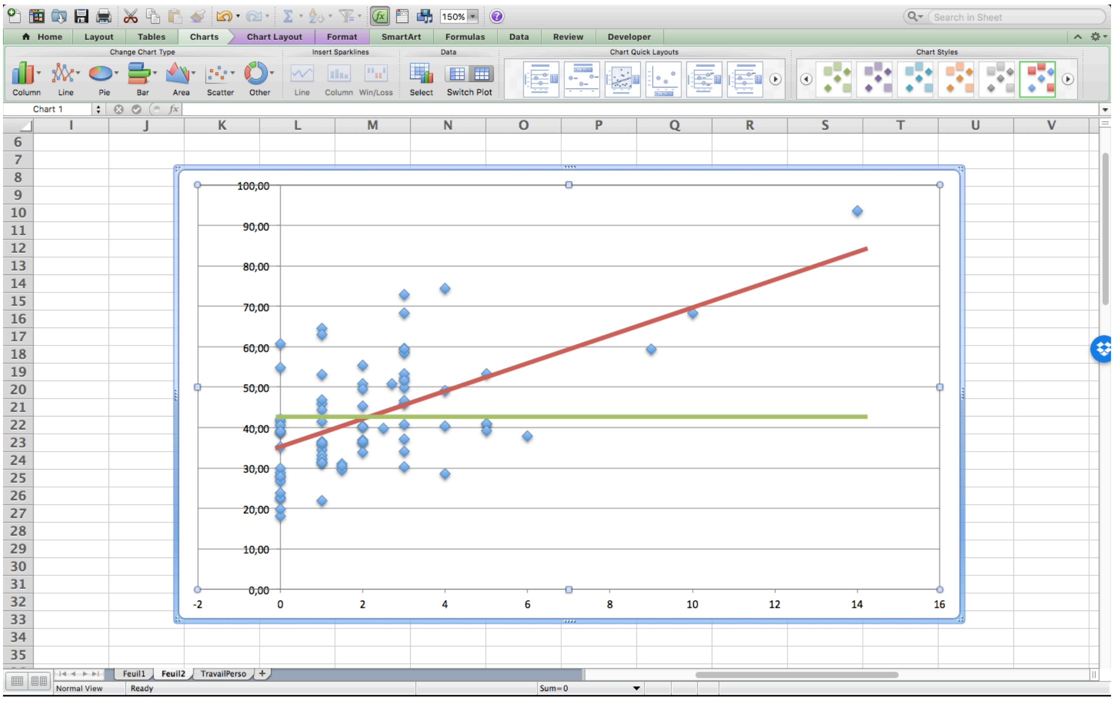
Variable explicative : Fibers

Dans R :

```
Call:
lm(formula = rating ~ fiber, data = cereal)

Residuals:
Min      1Q  Median      3Q     Max
-20.436  -8.159  -2.037   6.491  27.216

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.2566     1.7674  19.948  < 2e-16 ***
fiber         3.4430     0.5524   6.233 2.45e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.48 on 75 degrees of freedom
Multiple R-squared:  0.3412,	Adjusted R-squared:  0.3325
F-statistic: 38.85 on 1 and 75 DF,  p-value: 2.445e-08
```



Dans Excel:



Observations du tests

Rappel théorique

Bêta0 représente notre ‘b’ le coéfficiant additionner dans le calcul de notre équation de notre droite de régression (il représente l’ordonné à l’origine). Donc si une céréal contient 0(b) fibre j’aurai un taux de 35 par exemple.

Bêta1 lui est notre ‘a’ de cette même équation et représente le degré de la pente de la droite de régression. Donc si j’ai a qui vaut 2,4 ; A chaque fois que j’augmente de 1 mes fibres, j’augmente mon taux nutritif de 2,4.

R représente notre coefficient de corrélation linéaire et donc r^2 lui est le taux d’explication de notre variable par rapport à ce qu’on cherche à expliquer.

Concrètement

R^2 dans notre exercice avec le ratio et les fibres vaut 34% ce qui signifie que la variable des fibres explique 34% le taux nutritionnel des céréals.

Test de validation du modèle

Rappel

"H_0: variance expliquée par la régression = variance des résidus (càd la relation linéaire est non significative) »

Le test F permet de vérifier :

H0 : VarianceExpliquée / VarianceNonExpliquée = 1
H1 : VarianceExpliquée / VarianceNonExpliquée > 1

En gros verifier que ce qu'on explique est significativement > que ce qu'on explique pas. Si p valeur < 5% on rejette H0.

Si on rejette H0 on dit que l'on a expliquer significativement +.

Les tests T permettent de vérifier :

Pour rappel, le test t a pour but de déterminer si la valeur d'espérance μ d'une population de distribution normale et d'écart-type inconnu est égale à une valeur déterminée μ_0. Pour ce faire, nous tirons de cette population un échantillon de taille n dont on calcule la moyenne $\bar{x}$ et d'écart-type s.

H0 : bêta0 = 0
H1 : Bêta0 != 0 *// si notre Bêta est significativement dans Y = Ax+B != 0 et donc est important dans le calcul car Y = Ax + 0 sert a rien.*

H0 : bêta1 = 0
H1 : Bêta1 != 0 *//Si je rejette pas h0 on a pas besoin de faire la regression car pas intéressante si y = 0x + B.*

Bêta1 et bêta0 deviendront après, si notre modèle ici est bon, des estimateurs des paramètres de la régression. Car le but c'est de pouvoir appliquer notre échantillon à la population.

Concrètement

Test F :

```
F test to compare two variances

F-statistic: 38.85 on 1 and 75 DF,  p-value: 2.445e-08
```

Notre p-value est < 5% ce qui signifie que l'on rejette H0 et donc que l'on peut dire que « la régression est significative dans son ensemble ». taux de fibre explique significativement notre taux nutritionnel.

Test T pour bêt0 et bêta1:

```
Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.2566    1.7674  19.948  < 2e-16 ***
fiber        3.4430    0.5524   6.233 2.45e-08 ***
```

notre p-value que ce soit pour bêta0 et bêta1 est < 5% ce qui signifie que l'on rejette   H0 ; Pour ces valeurs cela signifique qu'elles ont leur utilité dans ce modèle et que ce dernier peut donc être représentatif de notre population.