ANALYSE DES DONNÉES

LUU THI THUY NGA

Table des matières

1 Introduction

Dans ce travail nous essayron de trouver le modèle qui explique le mieux les dépenses par habitant consacrées à l'enseignement public dans un État aux Etats Unis avec la base de données tirées de Chatterjee et Price (1977, p.108). Pour ceci nous utiliserons une base de données avec 50 observations sur les 6 variables suivantes.

État : État

Region Région (1 = nord-est, 2 = centre nord, 3 = sud, 4 = ouest)

X1 Nombre d'habitants pour mille résidant dans les zones urbaines en 1970

X2 Revenu personnel par habitant en 1973

X3 Nombre de résidents pour mille de moins de 18 ans en 1974

Y Dépenses par habitant consacrées à l'enseignement public dans un État, projetées pour 1975

Dans ce travail nous verrons donc comment les variations du : nombre d'habitants pour mille résidant dans les zones urbaines, revenu personnel par habitant, nombre de résidents pour mille de moins de 18 dans les années précédentes ainsi que l'état et la région où il se situe feront varier les dépenses par habitant consacrées à l'enseignement public dans un État. Y sera donc notre Variable Dépendante sur tout le long de ce travail.

```
data <-read_excel("education.xlsx")
#data <-data.frame(debut)
```

```
data=data[!duplicated(data),]
attach(data)
data[is.na(data),] <- 0
```

2 Visualisation de la base de données en vérifiant les variables

2.1 Charger de la base de données et observer les premiers lignes de la base de données

```
data[1:5,] %>% kable("latex",booktabs=T, caption = "L'extrait de la base des données") %>%   kable_styli
```

Tab. 1 : L'extrait de la base des données

etat	region	X1	X2	X3	Y
ME	1	508	3944	325	235
NH	1	564	4578	323	231
VT	1	322	4011	328	270
MA	1	846	5233	305	261
RI	1	871	4780	303	300

2.2 Vérifions si toutes les variables sont bonnes :

```
sum(is.na(data))
```

```
## [1] 0
```

Nous avons bien 6 variables comme noté dans le sujet

```
colnames(data)
```

```
## [1] "etat"   "region" "X1"     "X2"     "X3"     "Y"
```

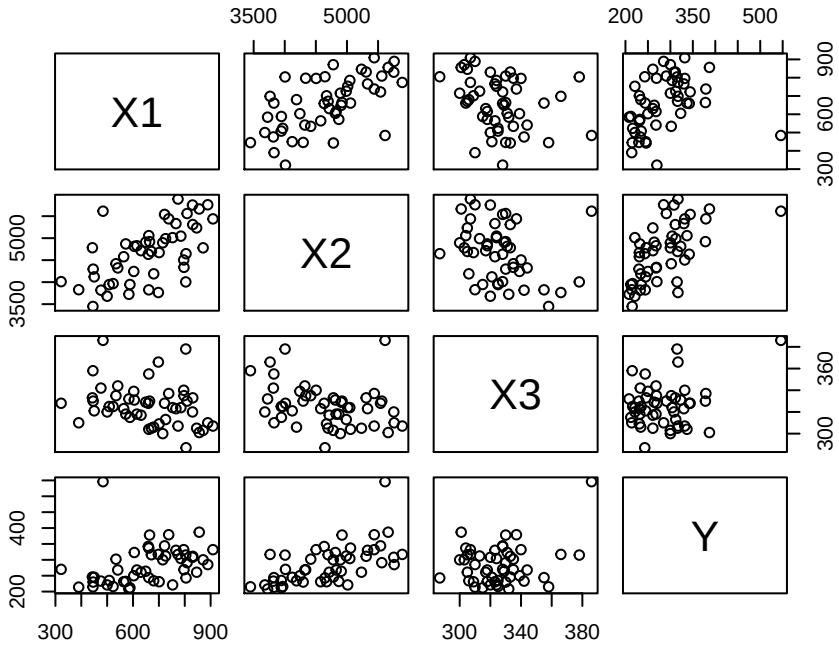
Observons les caractéristiques des variables dans la base de données

```
str(data)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    50 obs. of  6 variables:
##  $ etat  : chr  "ME" "NH" "VT" "MA" ...
##  $ region: num  1 1 1 1 1 1 1 1 2 ...
##  $ X1    : num  508 564 322 846 871 774 856 889 715 753 ...
##  $ X2    : num  3944 4578 4011 5233 4780 ...
##  $ X3    : num  325 323 328 305 303 307 301 310 300 324 ...
##  $ Y     : num  235 231 270 261 300 317 387 285 300 221 ...
```

```
data$region=as.factor(data$region)
data$etat=as.factor(data$etat)
attach(data)
```

```
plot(data[,3:6]) # Permet d'établir directement une relation entre toutes les variables d'un data.frame
```



On observe la relation entre les variables quantitatives de notre tableau

Observer le sommaire des variables quantitatives par régions

```
quanti=data[, lapply(data, is.numeric) == TRUE]
table_one <- tableby(region ~ ., data = quanti, test=FALSE,
                     numeric.stats=c("min", "max", "mean", "median","sd"),
                     control=tableby.control(digits=2L))


summary(table_one,text=T)%>% kable(format = "latex",booktabs=T, caption = "Sommaire des variables quanti
```

Tab. 2 : Sommaire des variables quantitatives par région

	1 (N=9)	2 (N=12)	3 (N=16)	4 (N=13)	Total (N=50)
X1					
- Min	322.00	443.00	390.00	484.00	322.00
- Max	889.00	830.00	805.00	909.00	909.00
- Mean	705.00	644.25	601.31	707.15	657.80
- Median	774.00	660.00	595.50	726.00	662.50
- SD	198.22	114.86	130.81	132.08	145.02
X2					
- Min	3944.00	4296.00	3448.00	3764.00	3448.00
- Max	5889.00	5753.00	5540.00	5613.00	5889.00
- Mean	4972.33	4930.83	4209.75	4806.08	4675.12
- Median	4894.00	4888.50	4043.50	4813.00	4706.00
- SD	723.22	375.64	587.20	590.03	644.51
X3					
- Min	300.00	304.00	287.00	305.00	287.00
- Max	328.00	337.00	358.00	386.00	386.00
- Mean	311.33	323.33	325.81	337.85	325.74
- Median	307.00	326.00	324.00	333.00	324.50
- SD	10.99	9.87	18.00	25.41	19.42
Y					
- Min	231.00	221.00	208.00	268.00	208.00
- Max	387.00	379.00	344.00	546.00	546.00
- Mean	287.33	286.33	246.19	328.38	284.60
- Median	285.00	266.00	238.50	315.00	269.50
- SD	47.49	59.58	39.89	67.56	61.34

La moyenne des variables est la plus petite dans les zones du sud.

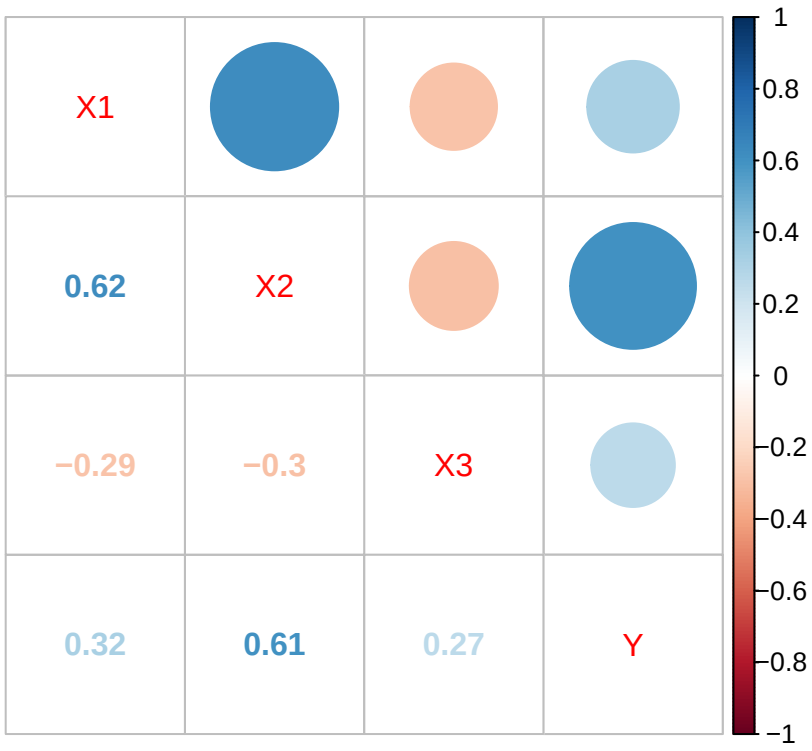
L'écart-type de la variable Y est le plus grand à "ouest", ce qui indique une majeure dispersion de cette variable.

La variable X2 est la plus dispersée.

La tranche de dépenses par habitant consacrées à l'enseignement public est plus grande au sud (entre 208 à 564).

2.3 Corrélation

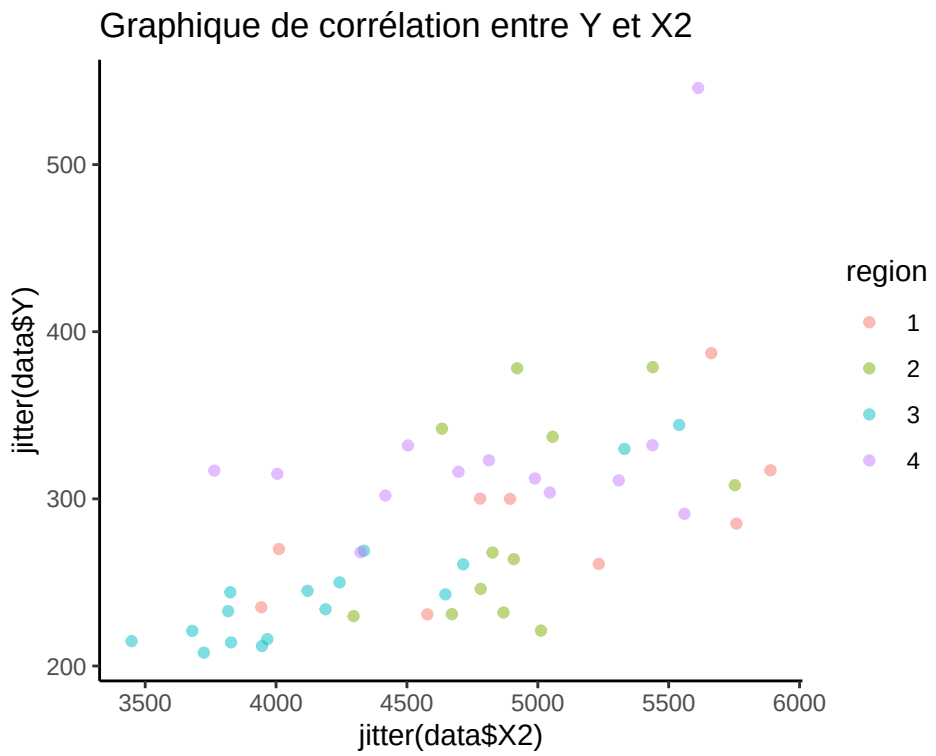
```
mcor<-cor(data[,3:6])
corrplot.mixed(mcor)
```



La corrélation entre X2 et Y, et entre X2 et X1, est forte. Il y a alors peut-être une problème de colinéarité.

2.3.1 Graphique de corrélation

```
ggplot(data = data, mapping = aes(x = jitter(data$X2), y = jitter(data$Y))) +
  geom_point(mapping = aes(colour = region), alpha = 0.5)+
  scale_fill_discrete("Region", labels = c("nord-est", "centre nord", "sud", "ouest"))+
  ggtitle("Graphique de corrélation entre Y et X2")
```



Le graphique montre qu'il y a une corrélation croissante entre les deux variables.

2.3.2 Test de corrélation

Faisons un test avec les hypothhèses :

$H_0 : t = 0$ Pas de corrélation entre les variables Y et X2

$H_1 : t \neq 0$ Présence de corrélation entre les variables

```
cor.test(data$X2,data$Y, method = c("pearson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$X2 and data$Y
## t = 5.3098, df = 48, p-value = 2.785e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3972106 0.7582618
## sample estimates:
##       cor
## 0.6083027
```

Selon les résultats du test, le coefficient de corrélation entre les variables Y et $X2$ est de 0.6083027 ce qui est cohérent avec la matrice de corrélation. Le *p_value* est inférieur à 0.05, donc nous pouvons rejetter l'hypothèse nulle, il y a donc une corrélation forte entre les deux variables.

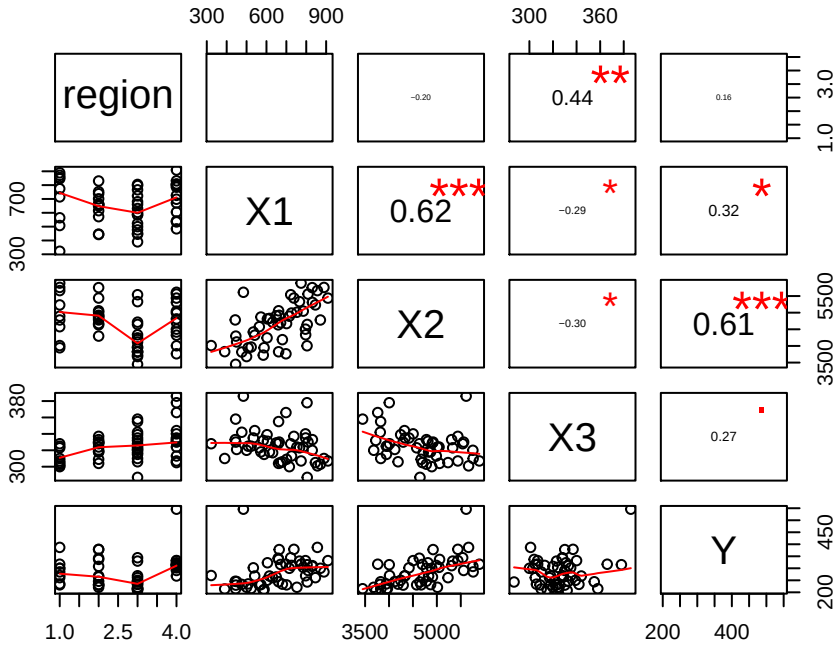
Le coefficient de corrélation de pearson mesure une corrélation linéaire entre deux variables.

```
panel.cor_simple <- function(x, y, digits=2, prefix="", cex.cor)
    {
        usr <- par("usr"); on.exit(par(usr))
```

```
        par(usr = c(0, 1, 0, 1))
        r <- cor(x, y)
        txt <- format(c(r, 0.123456789), digits=digits)[1]
        txt <- paste(prefix, txt, sep="")
        if(missing(cex.cor)) cex <- 0.8/strwidth(txt)

        test <- cor.test(x,y)
        # borrowed from printCoefmat
        Signif <- symnum(test$p.value, corr = FALSE, na = FALSE,
                      cutpoints = c(0, 0.001, 0.01, 0.05, 0.1, 1),
                      symbols = c("***", "**", "*", ".", " "))

        text(0.5, 0.5, txt, cex = cex * abs(r))
        text(.8, .8, Signif, cex=cex, col=2)
    }
pairs(data[,2:6], lower.panel=panel.smooth, upper.panel=panel.cor_simple)
```



Dans la matrice de corrélation on voit qu'il y a une corrélation positive entre X2 et X1, entre X3 et région, et entre X2 et Y. C'est-à-dire que la droite de corrélation entre ces variables a une pente positive et linéaire.

2.3.3 Matrice de corrélation et p-value

```
rcorr(as.matrix(data[,2:6]))
```

```
##        region    X1    X2    X3    Y
## region   1.00 -0.01 -0.20  0.44 0.16
## X1      -0.01  1.00  0.62 -0.29 0.32
## X2      -0.20  0.62  1.00 -0.30 0.61
## X3       0.44 -0.29 -0.30  1.00 0.27
```

```
## Y        0.16  0.32  0.61  0.27 1.00
##
## n= 50
##
##
## P
##        region X1     X2     X3     Y
## region        0.9669 0.1631 0.0013 0.2779
## X1     0.9669        0.0000 0.0437 0.0225
## X2     0.1631 0.0000        0.0362 0.0000
## X3     0.0013 0.0437 0.0362        0.0595
## Y      0.2779 0.0225 0.0000 0.0595
```

Pour les valeurs de *p-value* faibles, on peut dire que les corrélations entre les variables sont fortement significatives.

Dans la matrice de corrélation on voit qu'il y a une corrélation positive entre X2 et X1 et entre X2 et Y. C'est-à-dire que la droite de corrélation entre ces variables a une pente positive et linéaire. On observe aussi une corrélation positive significative entre X3 et Y.

3 Choix des variables

3.1 Les variables qualitatives :

On essaie maintenant d'expliquer Y par la variable qualitative "Etat". Le modèle est de la forme :

$$Y = \beta_0 + \beta_1 etat_i + \epsilon_i$$

```
mod=lm(Y~etat,data=data)
summary(mod)
```

```
##
## Call:
## lm(formula = Y ~ etat, data = data)
##
## Residuals:
## ALL 50 residuals are 0: no residual degrees of freedom!
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)      311         NA      NA       NA
## etatAL          -103         NA      NA       NA
## etatAR           -90         NA      NA       NA
## etatAZ            21         NA      NA       NA
## etatCA            21         NA      NA       NA
## etatCO            -7         NA      NA       NA
## etatCT             6         NA      NA       NA
## etatDE            33         NA      NA       NA
## etatDY           -95         NA      NA       NA
## etatFL           -68         NA      NA       NA
## etatGA           -61         NA      NA       NA
## etatHI           235         NA      NA       NA
## etatIA           -79         NA      NA       NA
## etatID           -43         NA      NA       NA
```

```
## etatIL            -3         NA      NA       NA
## etatIN           -47         NA      NA       NA
## etatKS            26         NA      NA       NA
## etatLA           -67         NA      NA       NA
## etatMA           -50         NA      NA       NA
## etatMD            19         NA      NA       NA
## etatME           -76         NA      NA       NA
## etatMI            68         NA      NA       NA
## etatMN            67         NA      NA       NA
## etatMO           -80         NA      NA       NA
## etatMS           -96         NA      NA       NA
## etatMT            -9         NA      NA       NA
## etatNB           -43         NA      NA       NA
## etatNC           -66         NA      NA       NA
## etatND           -65         NA      NA       NA
## etatNH           -80         NA      NA       NA
## etatNJ           -26         NA      NA       NA
## etatNM             6         NA      NA       NA
## etatNV           -20         NA      NA       NA
## etatNY            76         NA      NA       NA
## etatOH           -90         NA      NA       NA
## etatOK           -77         NA      NA       NA
## etatOR             5         NA      NA       NA
## etatPA           -11         NA      NA       NA
## etatRI           -11         NA      NA       NA
## etatSC           -78         NA      NA       NA
## etatSD           -81         NA      NA       NA
## etatTN           -99         NA      NA       NA
## etatTX           -42         NA      NA       NA
## etatUT             4         NA      NA       NA
## etatVA           -50         NA      NA       NA
## etatVT           -41         NA      NA       NA
## etatWA             1         NA      NA       NA
## etatWI            31         NA      NA       NA
## etatWV           -97         NA      NA       NA
## etatWY            12         NA      NA       NA
##
## Residual standard error: NaN on 0 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:    NaN
## F-statistic:   NaN on 49 and 0 DF,  p-value: NA
```

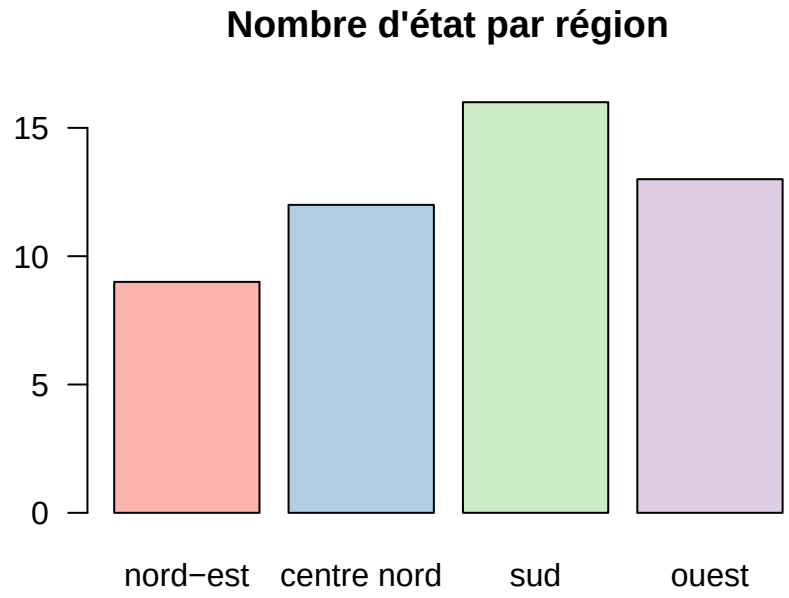
On obtient un R2 qui est égale à 1, c'est-à-dire, juste le variable "état" peut expliquer tout Y. En outre, pour les variables qualitatives, il nous reste "région" et les états sont compris dans les régions (comme le résumé dans le tableau et graphique ci-dessous). Pour observer la variation de Y avec les autres variables, on peut exclure la variable "état" dans nos modèles.

```
quali<-rowSums(table(data$region,data$etat))
names(quali)=c("nord-est","centre nord", "sud", "ouest")
quali %>% kable(format = "latex",booktabs=T, caption = "Nombre des états par région") %>%  kable_styli
```

Tab. 3 : Nombre des états par région

	x
nord-est	9
centre nord	12
sud	16
ouest	13

```
barplot(quali,col=brewer.pal(n = 4, name = "Pastel1"),
        horiz=F, las=1,main="Nombre d'état par région")
```



3.2 Choix des variables avec la nouvelle base de donnée

Avant de commencer les tests des différents modèles de régression, nous allons d'abord choisir les variables pertinentes.

Le critère d'information d'Akaike (Le critère AIC) est un indicateur qui permet de reconnaitre la pertinence des variables, il mesure la qualité de l'ajustement, plus on rajoute de paramètres plus on perd de degré de liberté, ainsi plus l'AIC est faible, mieux c'est.

3.2.1 Que les variables quantitatives

```
dataquanti=data[,-c(1,2)]
null = lm(Y~1, data = dataquanti)
full = lm(Y~., data = dataquanti)
step(null, data=dataquanti, scope = list(lower = null, upper = full), direction = "forward")
```

```
## Start:  AIC=412.63
```

```
## Y ~ 1
##
##        Df Sum of Sq    RSS    AIC
## + X2    1     68222 116146 391.53
## + X1    1     19130 165238 409.16
## + X3    1     13285 171083 410.89
## <none>              184368 412.63
##
## Step:  AIC=391.53
## Y ~ X2
##
##        Df Sum of Sq    RSS    AIC
## + X3    1     40787  75359 371.90
## <none>              116146 391.53
## + X1    1       950 115196 393.12
##
## Step:  AIC=371.9
## Y ~ X2 + X3
##
##        Df Sum of Sq   RSS    AIC
## <none>              75359 371.90
## + X1    1    11.302 75348 373.89
```

```
##
## Call:
## lm(formula = Y ~ X2 + X3, data = dataquanti)
##
## Coefficients:
## (Intercept)           X2           X3
##  -557.89121      0.07182      1.55561
```

D'après le critère d'information d'Akaike, les variables X1 et X2 sont des variables importantes qu'on doit garder dans nos modèles.

3.2.2 Avec variables " région"

```
fit <- lm(Y~X1+X2+X3+region,data=data)
step <- stepAIC(fit, direction="both")
```

```
## Start:  AIC=375.03
## Y ~ X1 + X2 + X3 + region
##
##          Df Sum of Sq    RSS    AIC
## - X1      1       671  69042 373.52
## - region  3      6976  75348 373.89
## <none>                 68371 375.03
## - X3      1     21112  89483 386.49
## - X2      1     48443 116814 399.82
##
## Step:  AIC=373.52
## Y ~ X2 + X3 + region
##
##          Df Sum of Sq    RSS    AIC
## - region  3      6317  75359 371.90
## <none>                 69042 373.52
```

```
## + X1      1       671  68371 375.03
## - X3      1     23716  92758 386.29
## - X2      1     60875 129917 403.13
##
## Step:  AIC=371.9
## Y ~ X2 + X3
##
##          Df Sum of Sq    RSS    AIC
## <none>                 75359 371.90
## + region  3      6317  69042 373.52
## + X1      1        11  75348 373.89
## - X3      1     40787 116146 391.53
## - X2      1     95725 171083 410.89
```

```
step$anova %>% kable(format = "latex",booktabs=T, caption = "Table du critère d'information d'Akaike") %
```

TAB. 4 : Table du critère d'information d'Akaike

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	43	68371.10	375.0341
- X1	1	671.2195	44	69042.32	373.5226
- region	3	6316.5672	47	75358.88	371.8997

Même avec l'ajout d'une variable qualitative, l'importance de X2 et X3 est gardée. En plus, on ne peut pas exclure la variable "région", parce qu'on a une corrélation forte entre "région" et "X3" qu'on a fait au début.

4 Réaliser les modèles régression linéaire :

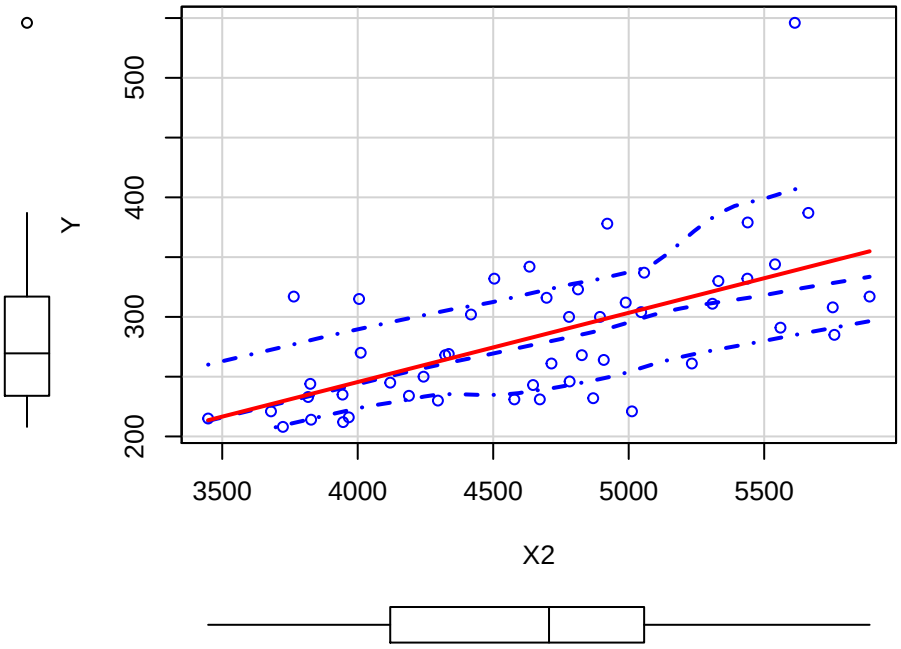
4.1 Premier modèle de régression

Etudier la relation entre X2 et Y :

En observant les coefficients de corrélation et le critère AIC, on obtient une forte relation entre X2 et Y. On réalise maintenant une modèle linéaire simple entre Y - variable expliquée et X2 - variable qui a la plus forte corrélation avec Y.

Droite de régression linéaire avec son intervalle de confiance

```
scatterplot(Y~X2, data=data, regLine=list(method=lm, lty=1, lwd=2, col="red"))
```



Courbe de moyenne (milieu, bleu), les courbes d'intervalle de confiance (en bleu) et une régression obtenue par les moindres carrés (en rouge ici). On voit que la regression est entre la borne inférieure et supérieure.

On voit bien qu'il y a une corrélation positive entre X2 et Y.

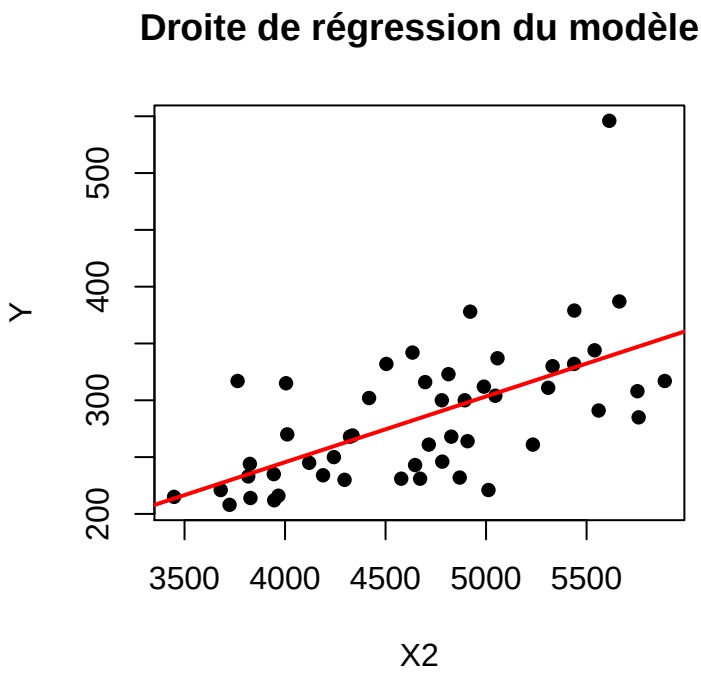
On vois qu'il y a quelques points, au moins 20%, qui sont en dehors de l'intervalle de confiance. Ceci viendrait à dire que X2 n'explique pas parfaitement Y.

Réaliser notre première régression linéaire de forme :

$$Y_i = \beta_1 + \beta_2 X2_i$$

L'affichage du résultat

```
mod <- lm(Y~X2, data=data)
# Droite de régression du modèle
plot(Y~X2,pch=16,data=data, main="Droite de régression du modèle")
abline(mod,col="red",lwd=2)
```



Description de la régression :

```
stargazer(mod, type= "latex",title="Modèle de régression entre Y et X2" ,table.placement="H", header=FA
```

TAB. 5 : Modèle de régression entre Y et X2

	Dependent variable :
	Y
X2	0.058^{***}
	(0.011)
Constant	13.936
	(51.447)
Observations	50
R^2	0.370
Adjusted R^2	0.357
Residual Std. Error	49.191 (df = 48)
F Statistic	28.194^{***} (df = 1 ; 48)
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01

le R^2 (coefficient de détermination)

Si $R^2 = 1$, le nuage est très allongé, X et Y tracent une droite.

Si $R^2 = 0$, le nuage se disperse dans tous les sens et une régression linéaire ne permettra pas de faire des prédictions.

```
cor(data$Y,data$X2)^2
```

```
## [1] 0.3700322
```

Ici, R^2 près de 0, on peut dire que la variable X2 n'est pas trop forte pour expliquer variable Y, il faut ajouter plus de variables pour expliquer le modèle.

4.2 Formule d'après le proposition du critère AIC

Réaliser notre première régression linéaire de forme :

$$Y_i = \beta_1 + \beta_2 X2_i + \beta_3 X3_i$$

```
mod1 <- lm(Y~X2+X3, data=data)
```

Description de la régression :

```
stargazer(mod1, type= "latex",title="Résumé du modèle 2" ,table.placement="H",header=FALSE)
```

Tab. 6 : Résumé du modèle 2

	Dependent variable :
	Y
X2	0.072^{***}
	(0.009)
X3	1.556^{***}
	(0.308)
Constant	-557.891^{***}
	(120.864)
Observations	50
R^2	0.591
Adjusted R^2	0.574
Residual Std. Error	40.042 (df = 47)
F Statistic	33.994^{***} (df = 2 ; 47)
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01

On observe qu'avec l'ajout de la variable X3 on augmente R^2.

4.3 Formule avec toutes les variables :

4.3.1 Variable qualitative

On suppose le prochain modèle de la forme :

$$Y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \beta_3 X3_i + \beta_4 region_i + \epsilon_i$$

```
mod2<-lm(Y~X1+X2+X3+region,data = data)
```

```
stargazer(mod2, type= "latex",title="Résumé du modèle 2" ,table.placement="H",header=FALSE)
```

TAB. 7 : Résumé du modèle 2

	Dependent variable :
	Y
X1	−0.035
	(0.053)
X2	0.072***
	(0.013)
X3	1.301***
	(0.357)
region2	−15.727
	(18.163)
region3	−8.640
	(18.539)
region4	18.597
	(19.688)
Constant	−451.675***
	(139.539)
Observations	50
R^2	0.629
Adjusted R^2	0.577
Residual Std. Error	39.875 (df = 43)
F Statistic	12.159*** (df = 6 ; 43)
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01

On observe qu'avec l'ajout de la variable qualitative "région", on augmente R^2 .

4.3.2 Variable dummy

En terme des variables qualitatives, il existe aussi une variable qui s'appelle dummy.

Par rapport au dernier test, la variable "région" n'est pas toujours une variable importante.

Le but de l'ajout est pour observer s'il y a une modification de cette variable qualitative, est ce que les coefficients prennent le sens plus précis.

Créer une variable dummy par supposer qu'un 1 pour ce qui vient du sud_ouest, et 0 pour les autres zones.

```
#Création une variable dummy sud_ouest
data$region=as.numeric(data$region)
data$sud_ouest <- ifelse(data$region >2 , 1, 0)
data1=data[,-1]
head(data1) %>% kable(format = "latex",booktabs=T, caption = "L'extrait de la nouvelle base de donnée")
```

TAB. 8 : L'extrait de la nouvelle base de donnée

region	X1	X2	X3	Y	sud_ouest
1	508	3944	325	235	0
1	564	4578	323	231	0
1	322	4011	328	270	0
1	846	5233	305	261	0
1	871	4780	303	300	0
1	774	5889	307	317	0

```
mod10<-lm(Y~X1+X2+X3+sud_ouest,data = data1)
```

```
stargazer(mod10, type= "latex",title="Résumé du modèle 2 avec une modification de la variable qualitati
```

TAB. 9 : Résumé du modèle 2 avec une modification de la variable qualitative

	Dependent variable :
	Y
X1	−0.018
	(0.053)
X2	0.077***
	(0.013)
X3	1.456***
	(0.329)
sud_ouest	13.334
	(13.316)
Constant	−546.535***
	(123.598)
Observations	50
R^2	0.600
Adjusted R^2	0.565
Residual Std. Error	40.471 (df = 45)
F Statistic	16.891*** (df = 4; 45)
Note :	*p<0.1; **p<0.05; ***p<0.01

On essaie maintenant de comparer les modèles qui sont avec le variables "région" et une variable dummy "sud_ouest"

```
stargazer(mod1,mod10, type= "latex",title="Comparer les deux modèles" ,table.placement="H",header=FALSE
```

Tab. 10 : Comparer les deux modèles

	Dependent variable :	
	Y	
	(1)	(2)
X1		−0.018
		(0.053)
X2	0.072***	0.077***
	(0.009)	(0.013)
X3	1.556***	1.456***
	(0.308)	(0.329)
sud_ouest		13.334
		(13.316)
Constant	−557.891***	−546.535***
	(120.864)	(123.598)
Observations	50	50
R^2	0.591	0.600
Adjusted R^2	0.574	0.565
Residual Std. Error	40.042 (df = 47)	40.471 (df = 45)
F Statistic	33.994*** (df = 2 ; 47)	16.891*** (df = 4 ; 45)
Note :		*p<0.1 ; **p<0.05 ; ***p<0.01

L'introduction de la variable dummy n'aide pas à améliorer la qualité du modèle, donc on peut garder la variable "région" comme une variable qualitative dans le modèle

4.4 Modèle avec les relations entre les variables explicatives

Retour dans le tableau de corrélation, on peut conclure qu'il y a des corrélations fortes entre deux couples $X1 - X2$ et $X3 - region$. On observe maintenant un nouveau modèle avec l'ajout de ces relations :

$$Y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \beta_3 X3_i + \beta_4 region_i + \beta_5 X1_i * X2_i + \beta_6 X3_i * region_i + \epsilon_i$$

```
mod3<-lm(Y~X1+X2+X3+region+X2*X1+X3*region,data = data)
```

```
stargazer(mod3, type= "latex",title="Résumé du modèle 3" ,table.placement="H",header=FALSE)
```

Tab. 11 : Résumé du modèle 3

	Dependent variable :
	Y
X1	0.345
	(0.307)
X2	0.132^{***}
	(0.044)
X3	-1.400
	(1.209)
region	-238.033^{**}
	(107.766)
X1 :X2	-0.0001
	(0.0001)
X3 :region	0.768^{**}
	(0.339)
Constant	134.451
	(457.969)
Observations	50
R^2	0.666
Adjusted R^2	0.619
Residual Std. Error	37.858 (df = 43)
F Statistic	14.273^{***} (df = 6 ; 43)
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01

L'ajout des interactions entre variables augmente le R^2 du modèle.

4.5 Modèle log-niveau

Testons le modèle avec un modèle log-niveau pour observer s'il y a une amélioration par l'ajout des fonctions de log :

```
mod4<-lm(log(Y)~X1+X2+X3+region+X2*X1+X3*region,data = data)
```

```
stargazer(mod4, type= "latex",title="Résumé du modèle 4" ,table.placement="H",header=FALSE)
```

Tab. 12 : Résumé du modèle 4

	Dependent variable :
	log(Y)
X1	0.001
	(0.001)
X2	0.0004***
	(0.0001)
X3	−0.003
	(0.004)
region	−0.631*
	(0.353)
X1 :X2	−0.00000
	(0.00000)
X3 :region	0.002*
	(0.001)
Constant	4.838***
	(1.502)
Observations	50
R^2	0.653
Adjusted R^2	0.604
Residual Std. Error	0.124 (df = 43)
F Statistic	13.464*** (df = 6 ; 43)
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01

Comparer le modèle niveau-niveau et log-niveau

```
stargazer(mod3,mod4, type= "latex",title="Comparer les deux modèles" ,table.placement="H",header=FALSE)
```

Tab. 13 : Comparer les deux modèles

	Dependent variable :	
	Y	log(Y)
	(1)	(2)
X1	0.345	0.001
	(0.307)	(0.001)
X2	0.132***	0.0004***
	(0.044)	(0.0001)
X3	−1.400	−0.003
	(1.209)	(0.004)
region	−238.033**	−0.631*
	(107.766)	(0.353)
X1 :X2	−0.0001	−0.00000
	(0.0001)	(0.00000)
X3 :region	0.768**	0.002*
	(0.339)	(0.001)
Constant	134.451	4.838***
	(457.969)	(1.502)
Observations	50	50
R^2	0.666	0.653
Adjusted R^2	0.619	0.604
Residual Std. Error (df = 43)	37.858	0.124
F Statistic (df = 6 ; 43)	14.273***	13.464***
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01	

En comparant les deux modèles, l'ajout un variable de logarithme fait une légère baisse de R2. On garde alors le modèle niveau-niveau.

4.6 Comparaison tous les modèles

$$Y_i = \beta_1 + \beta_2 X2_i + \beta_3 X3_i$$
$$Y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \beta_3 X3_i + \beta_4 region_i + \epsilon_i$$

$$Y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \beta_3 X3_i + \beta_4 region_i + \beta_5 X1_i * X2_i + \beta_6 X3_i * region_i + \epsilon_i$$

Comparer tous les modèles et choisir le dernier modèle à garder

```
stargazer(mod1,mod2,mod3, type= "latex",title="Comparaison des modèles" ,table.placement="H",header=FALS
```

Tab. 14 : Comparaison des modèles

	Dependent variable :		
		Y	
	(1)	(2)	(3)
X1		−0.035	0.345
		(0.053)	(0.307)
X2	0.072***	0.072***	0.132***
	(0.009)	(0.013)	(0.044)
X3	1.556***	1.301***	−1.400
	(0.308)	(0.357)	(1.209)
region2		−15.727	
		(18.163)	
region3		−8.640	
		(18.539)	
region4		18.597	
		(19.688)	
region			−238.033**
			(107.766)
X1 :X2			−0.0001
			(0.0001)
X3 :region			0.768**
			(0.339)
Constant	−557.891***	−451.675***	134.451
	(120.864)	(139.539)	(457.969)
Observations	50	50	50
R^2	0.591	0.629	0.666
Adjusted R^2	0.574	0.577	0.619
Residual Std. Error	40.042 (df = 47)	39.875 (df = 43)	37.858 (df = 43)
F Statistic	33.994*** (df = 2 ; 47)	12.159*** (df = 6 ; 43)	14.273*** (df = 6 ; 43)
Note :			*p<0.1 ; **p<0.05 ; ***p<0.01

Pour notre modèle final :

- On garde le modèle niveau-niveau
- On conclut que pour variable qualitative on va garder région
- On décide de ne garder comme interaction entre variables que X2 :X3
- Les variables quantitatives à garder sont X2 et X3

5 Contrôle le modèle de régression :

Le modèle final est de la forme :

$$Y_i = \beta_0 + \beta_1 X2_i + \beta_2 X3_i + \beta_3 region_i + \beta_4 X3_i * region_i + \epsilon_i$$

5.1 Interprétation

5.1.1 Test de significativité du modèle

Pour ce test, on effectue des hypothèses :

H_0 : absence de significativité globale de coefficients

H_1 : au moins un des coefficients est différents à 0

```
mod6<-lm(Y~X2+X3+region+X3*region,data = data)

summary(mod6)
```

```
##
## Call:
## lm(formula = Y ~ X2 + X3 + region + X3 * region, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.624 -23.570  -5.049  20.475  88.981
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.687e+02  3.768e+02   0.713   0.4795
## X2           6.672e-02  9.313e-03   7.164 5.83e-09 ***
## X3          -1.019e+00  1.133e+00  -0.900   0.3731
## region      -2.218e+02  1.024e+02  -2.166   0.0356 *
## X3:region    7.153e-01  3.202e-01   2.234   0.0305 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.34 on 45 degrees of freedom
## Multiple R-squared:  0.6412, Adjusted R-squared:  0.6094
## F-statistic: 20.11 on 4 and 45 DF,  p-value: 1.483e-09
```

Notre résultat est :F-statistic : 20.11 on 4 and 45 DF, p-value : 1.483e-09

On obtient ici une petite valeur de p-value (qui est inférieur à 5%), donc on peut rejeter l'hypothèse nulle du test. Le modèle est significatif.

5.1.2 Interprétation des coefficients :

Significativité / Test de Student

D'après le résumé du modèle, on peut conclure que tous les coefficients sont significatives sauf $X3$. Mais dans notre modèle, il y a une relation entre X3 et region, donc on ne peut pas supprimer $X3$ du modèle même s'il n'aide pas à expliquer Y dans le modèle.

Magnitude du coefficient :

X2 : L'augmentation de 1 *dollar* dans le revenu personnel par habitant en 1973 va faire augmenter $6.672e{-}02$ *dollar* dans les dépenses par habitant consacrées à l'enseignement public dans un État. C'est bien expliqué

dans la réalité : les impôts sont indexés sur le revenu donc si celui-ci augmente alors la collecte d'impôts va aussi augmenter. L'Etat aura alors un plus grand budget et pourra dépenser plus dans l'enseignement public.

X3 : L'augmentation de 1 *personne* de résidents pour mille de moins de 18 ans en 1974 va faire diminuer $1.019e+00$ dollars dans les dépenses par habitant consacrées à l'enseignement public dans un État. Ceci pourrait être expliqué du fait que la population active baisse, donc moins de personnes imposables, ce qui entraîne une baisse du budget de l'Etat.

Région : L'augmentation d' 1 *unité* de la variable région va faire diminuer $2.218e+02$ *dollars* dans les dépenses par habitant consacrées à l'enseignement public dans un État. Ça va expliquer une réalité aux Etats-Unis : les états de l'ouest américain investissent moins en Education publique.

X3 :Région : L'augmentation d'1 *unité* de la variable X3 :region va faire augmenter de $7.153e-01$ *dollars* les dépenses par habitant consacrées à l'enseignement public dans un État. Ça va expliquer une réalité aux Etats-Unis : plus on avance vers l'ouest des états unis, plus les états vont investir dans l'éducation publique en fonction du nombre d'habitants par mille de moins de 18 ans.

5.1.3 Interprétation du modèle

Qualité du modèle

On regarde maintenant le coefficient de détermination R2, ici, 64.12% de variation dans la variable Y qui est expliquée par des variations dans les variables explicatives.

Plus cette valeur est proche de 1, plus l'adéquation entre le modèle et les données sont forte. Mais on va prendre en compte aussi ici une valeur de R2 ajusté qui est égale à 60.94% pour éviter l'influence par l'ajout trop de variables explicatives dans la régression.

Intervalle de confiance des variables explicatives du modèle

```
confint(mod6) %>% kable(format = "latex",booktabs=T, caption = "Intervalle de confiance") %>%   kable_s
```

TAB. 15 : Intervalle de confiance

	2.5 %	97.5 %
(Intercept)	-490.2019481	1027.5247512
X2	0.0479617	0.0854782
X3	-3.3014188	1.2627647
region	-428.0574096	-15.6025732
X3 :region	0.0703372	1.3601688

5.2 Test Shapiro-Wilk

Interprétation du test :

H_0 : Les résidus du modèle suivent une loi Normale

H_1 : Les résidus du modèle ne suivent pas une loi Normale

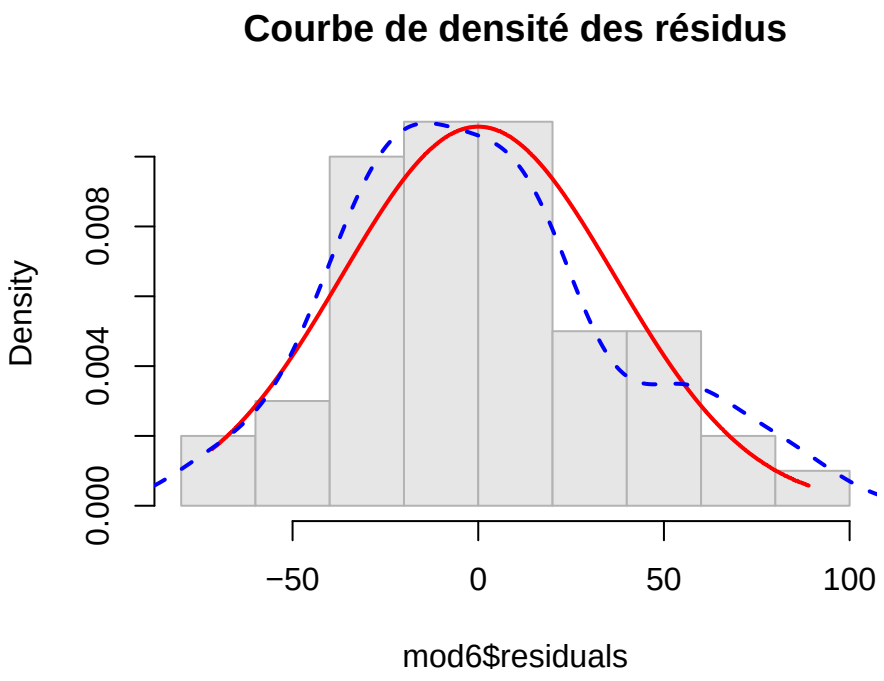
On garde l'hypothèse nulle cause d'une valeur de $p-value = 0.2305 > 0.05$. Le risque de rejeter l'hypothèse nulle alors qu'elle est vraie est de 23.05%.

```
shapiro.test(residuals(mod6))
```

```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  residuals(mod6)
## W = 0.96996, p-value = 0.2305
```

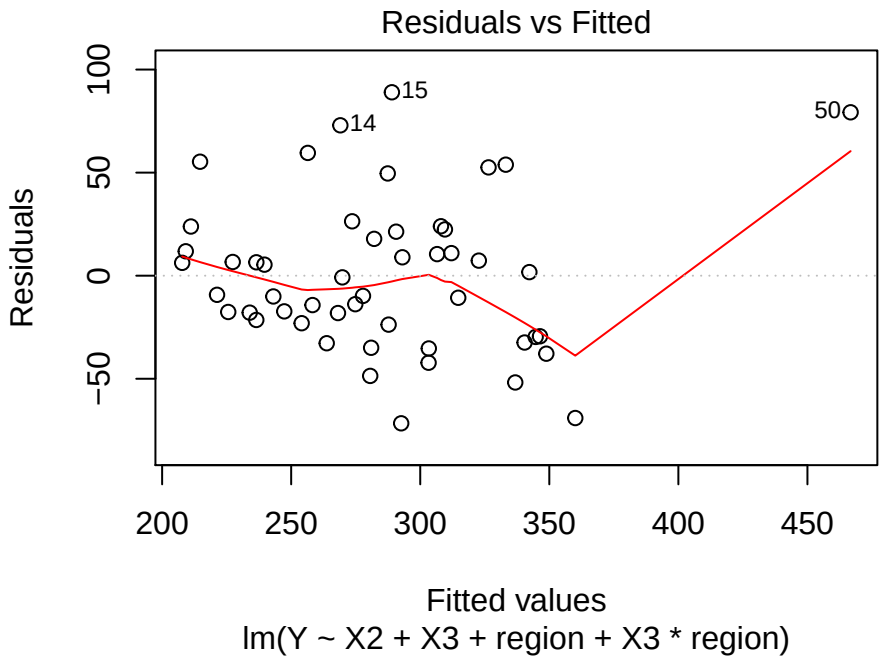
```
x<-seq(min(mod6$residuals),max(mod6$residuals),0.01)
y<-dnorm(x,mean(mod6$residuals),sd(mod6$residuals))
hist(mod6$residuals,proba=T,border=grey(0.7),col=grey(0.9),main="Courbe de densité des résidus")
lines(x,y,col="red",lwd=2)
lines(density(mod6$residuals),col="blue",lwd=2,lty=2)
```



On peut vérifier que les résidus suivent la loi normale avec un histogramme classique. Celui-ci s'approche de la courbe de densité d'une loi normale.

5.3 Test de Rainbow :

```
plot(mod6, which = 1)
```



D'après le graphique on voit que la relation entre Y et les variables explicatives existe mais la linéarité n'est pas parfaite. Pour qu'il y ait relation linéaire il faut que la ligne rouge soit approximativement horizontale.

On vérifie la linéarité par un test de Rainbow.

```
raintest(mod6) # test de rainbow
```

```
##
##  Rainbow test
##
## data:  mod6
## Rain = 1.6777, df1 = 25, df2 = 20, p-value = 0.1205
```

On obtient ici $p-value = 0.09672 > 0.05$, donc on peut conclure qu'il y a adéquation du modèle de régression.

5.4 Indépendance des résidus (test de Durbin-Watson)

Graphique l'indépendence des résidus :

```
acf(residuals(mod1), main="Regression du modèle 1")
```

Remarque :

Premier bâtonnet : très élevé, il y a une auto-corrélation des résidus avec eux-même.

Les bâtonnets suivants : indique l'auto-corrélation entre les résidus et les résidus n+1, on peut conclure qu'il y a une problème de auto-corrélation lorsque le bâtonnet dépasse les pointillées.

On vérifie l'indépendance avec un test de Durbin-Watson.

Le test Durbin - Watson

Pour ce test, on effectue des hypothèses :

H_0 : il existe une phénomène d'auto-corrélation

H_1 : absence de la phénomène d'auto-corrélation

```
durbinWatsonTest(mod6)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.07368159      2.043763   0.802
##  Alternative hypothesis: rho != 0
```

```
dwtest(mod6)
```

```
##
##  Durbin-Watson test
##
## data:  mod6
## DW = 2.0438, p-value = 0.4259
## alternative hypothesis: true autocorrelation is greater than 0
```

D'après le test de Durbin-Watson on conclut que les résidus sont indépendants car p-value>0.05.

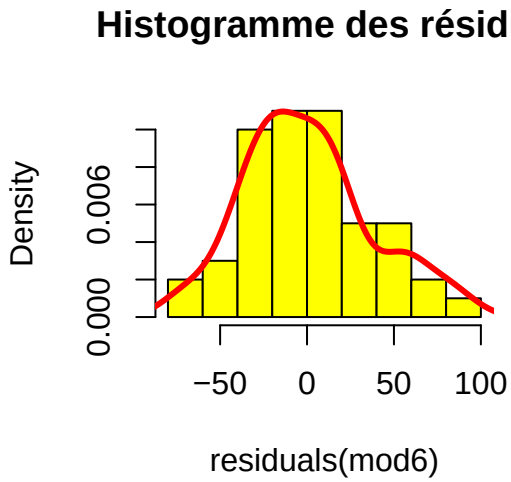
5.5 Distribution des résidus :

Observons l'histogramme et le diagramme Quantile-Quantile :

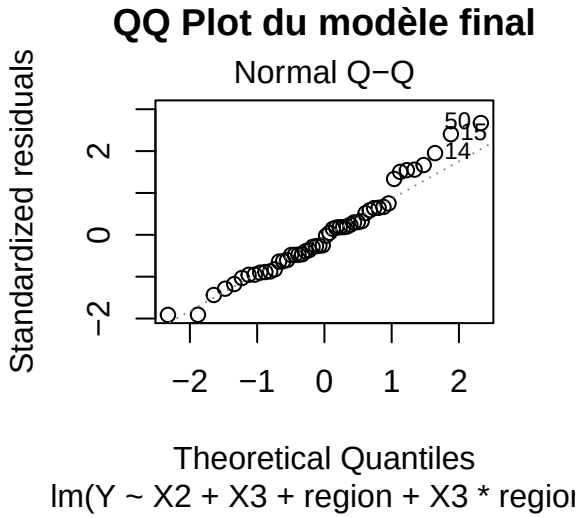
```
hist(residuals(mod6),col="yellow",freq=F,main="Histogramme des résidus")

densite <- density(residuals(mod6)) # estimer la densité que représente ces différentes valeurs

lines(densite, col = "red",lwd=3) # Superposer une ligne de densité à l'histogramme
```




```
plot(mod6, which = 2, main="QQ Plot du modèle final")
```

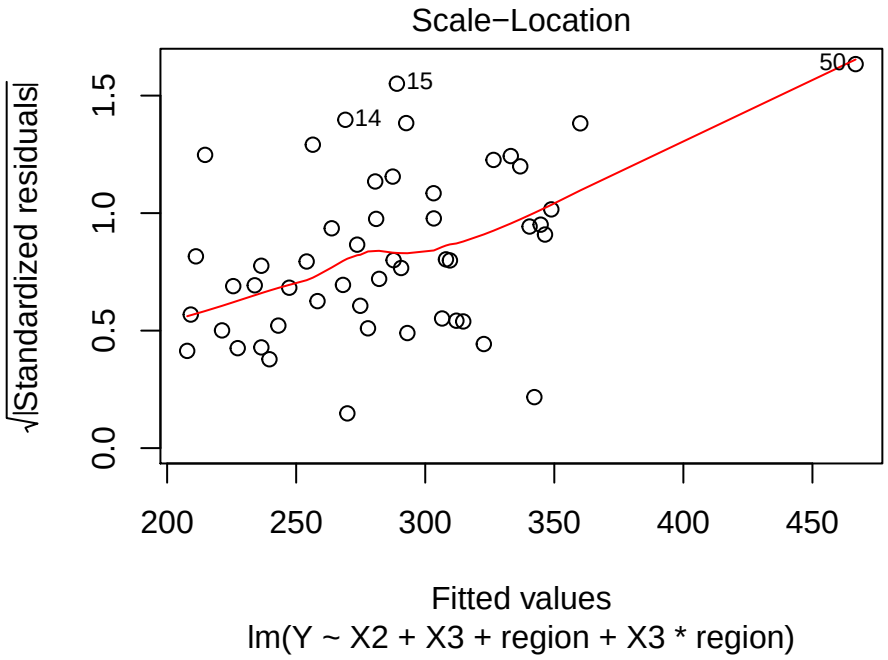


Avec ces deux graphes ainsi qu'avec notre vérification pars le test de Shapiro-Wilk on conclut que la distribution des résidus suit une loi normale

5.6 Homogénéité de la distribution :

Pour cela on trace le graphique suivant qui met en relation les racines carrées des résidus (résidus standardisés) en fonction des valeurs théoriques (fitted-values) de Y prédites par l'équation de la régression.

```
plot(mod6, which = 3)
```



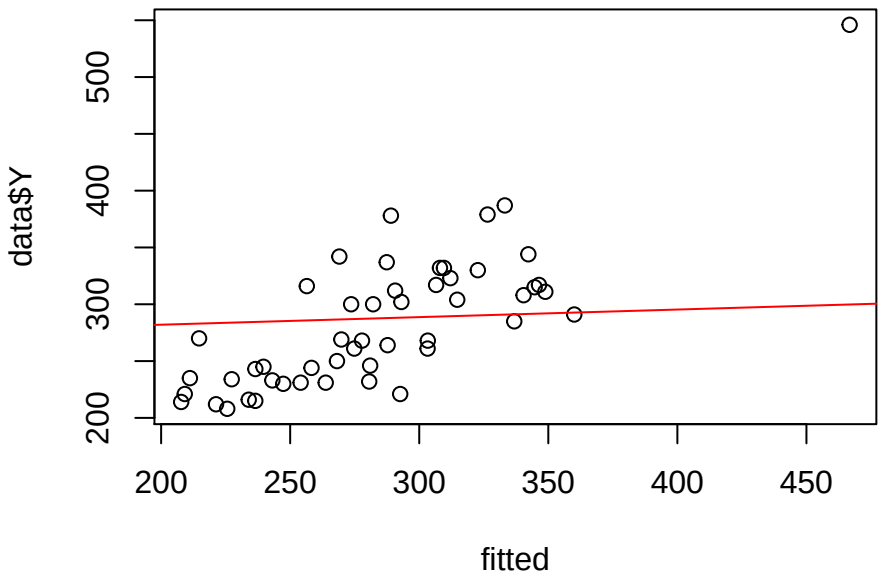
On cherche ici une courbe rouge plane. L'homogénéité est à rejeter si celle-ci n'est pas horizontale.

De plus, ce graphique est l'occasion de vérifier si certains points se regroupent, ce qui indiquerait un défaut d'indépendance.

D'après le graphique il y a un problème d'hétéroscédasticité car la droite rouge n'est pas horizontale. On vérifie cela avec un test de Breush-Pagan et le test de White. Quand à l'indépendance on l'a déjà confirmé avec le test de Durbin-Watson.

Graphique des résidus

```
fitted=predict(mod6, data)
plot(data$Y~fitted)
abline(mod6,  col = "red")
```



Les résidus du modèle semble avoir une distribution normale

5.7 Test de Breush-Pagan et test de White ou problème d'hétéroscédasticité

Hétéroscédasticité

- Variance des résidus non constante
- Exogène en abscisse pour détecter (traiter) dépendance

5.7.1 Test de Breush-Pagan

Pour faire ce test, on test un sous-modèle :

$$\epsilon_i^2 = \lambda_0 + \lambda_1 X2 + \lambda_2 X3 + \lambda_3 region + \lambda_4 region * X3$$

Si l'un des coefficients λ_i, $\forall i = 1, ...n$, est différents à 0, on peut conclure que l'hétéroscédasticité est présente dans ce modèle

Donc notre hypothèse dans ce cas est :

$$H_0 : \lambda_0^2 + \lambda_1^2 + \lambda_2^2 + \lambda_3^2 + \lambda_4^2 = 0$$

ou H_0 :(homoscédasticité)

$$H_1 : \lambda_0^2 + \lambda_1^2 + \lambda_2^2 + \lambda_3^2 + \lambda_4^2 \neq 0$$

ou H_1 :(hétéroscédasticité)

On calcule maintenant la statistique de Breusch-Pagan : $BP = nR^2$ qui suit une loi du Khi-deux de paramètre $(0.95, 4)$

```
bptest(mod6)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod6
## BP = 13.273, df = 4, p-value = 0.01002
```

On obtient un p-value inférieur à 0.05. On peut conclure qu'il y a une problème d'hétérosdasticité.

Ou regardons sur l'indice BP :

Notre $\chi^2_{(0.95,4)}$

```
qchisq(.95,4)
```

```
## [1] 9.487729
```

Notre indice $BP = 13.273 > \chi^2_{(0.95,4)} = 9.487729$

On peut conclure l'existence du problème d'hétéroscédasticité aussi en regardant l'indice BP.

5.7.2 Test de White

Notre sous-modèle du test de White a une petite modification :

$$\epsilon_i^2 = \lambda_0 + \lambda_1 X2_i + \lambda_2 X2_i^2 + \lambda_3 X3_i + \lambda_4 X3_i^2 + \lambda_5 region_i + \lambda_6 region_i^2 + \lambda_7 X3_i * region_i + \lambda_8 X2_i * X3_i + \lambda_9 X2_i * region_i$$

Donc notre hypothèse dans ce cas est :

$$H_0 : \lambda_0^2 + \lambda_1^2 + \lambda_2^2 + \lambda_3^2 + \lambda_4^2 + \lambda_5^2 + \lambda_6^2 + \lambda_7^2 + \lambda_8^2 + \lambda_9^2 = 0$$
$$H_1 : \lambda_0^2 + \lambda_1^2 + \lambda_2^2 + \lambda_3^2 + \lambda_4^2 + \lambda_5^2 + \lambda_6^2 + \lambda_7^2 + \lambda_8^2 + \lambda_9^2 \neq 0$$

Testons notre sous-modèle

```
data$resi <- mod6$residuals
ressq=data$resi^2
white<-lm(ressq~X2+I(X2^2)+X3+I(X3^2)+region+I(region^2)+X3*region+X2*X3+X2*region, data=data)
summary(white)
```

```
##
## Call:
## lm(formula = ressq ~ X2 + I(X2^2) + X3 + I(X3^2) + region + I(region^2) +
##     X3 * region + X2 * X3 + X2 * region, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2039.7  -980.5  -309.0   468.8  5426.5
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.365e+04  5.750e+04   1.281    0.208
## X2          -4.866e+00  9.797e+00  -0.497    0.622
## I(X2^2)     -3.946e-04  6.366e-04  -0.620    0.539
## X3          -5.164e+02  3.403e+02  -1.517    0.137
## I(X3^2)      7.726e-01  5.678e-01   1.361    0.181
```

```
## region       1.024e+04  8.730e+03   1.173    0.248
## I(region^2)  2.295e+02  3.197e+02   0.718    0.477
## X3:region   -3.369e+01  2.467e+01  -1.365    0.180
## X2:X3        3.190e-02  2.249e-02   1.418    0.164
## X2:region   -2.380e-01  5.558e-01  -0.428    0.671
##
## Residual standard error: 1624 on 40 degrees of freedom
## Multiple R-squared:  0.3569, Adjusted R-squared:  0.2122
## F-statistic: 2.467 on 9 and 40 DF,  p-value: 0.02422
```

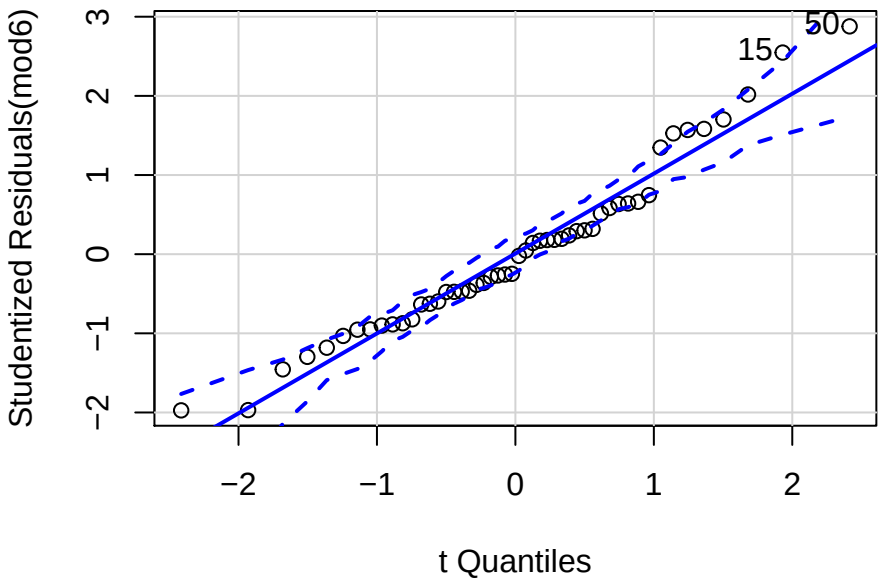
```
bptest(mod6,~X2+I(X2^2)+X3+I(X3^2)+region+I(region^2)+X3*region+X2*X3+X2*region, data=data)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod6
## BP = 17.846, df = 9, p-value = 0.037
```

On obtient un p-value de $0.037 < 0.05$. Donc il existe un problème d'homoscédasticité.

5.8 Graphique

```
qqPlot(mod6, id.n=4)
```




```
## [1] 15 50
```

Donc il existe un problème d'hétéroscédasticité. En plus, d'après le graphique, il y a deux observations aberrantes qui peuvent influencer à l'homogénéité du modèle.

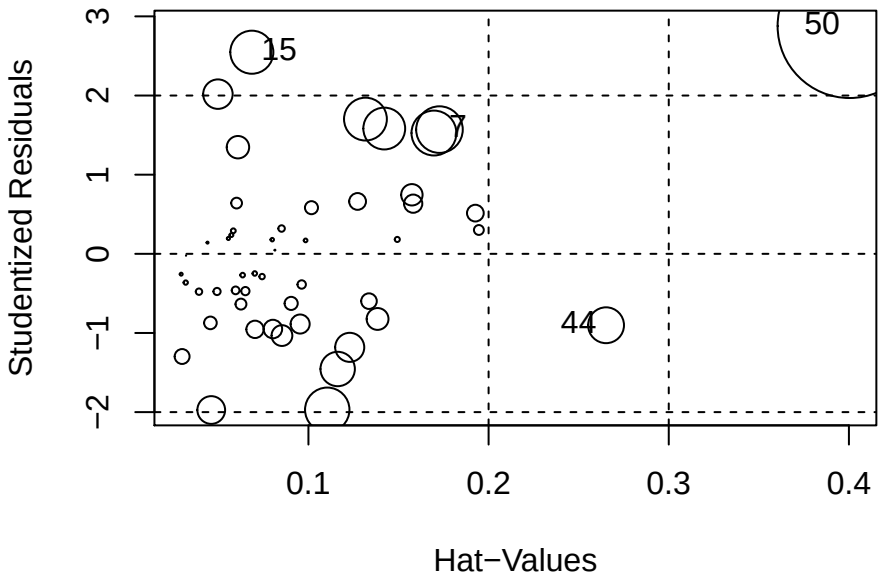
6 Résoudre problème d'hétérosdasticité

6.1 Avec distance de Cook

La distance de Cook permet d'évaluer les points qui auront une (trop) grande influence sur le modèle de régression.

Le graphique suivant permet de voir les points normaux et trop influents.

```
influencePlot(mod6) %>% kable(format = "latex",booktabs=T, caption = "Sommaire de la distance de Cook")
```



Tab. 16 : Sommaire de la distance de Cook

	StudRes	Hat	CookD
7	1.5694793	0.1727330	0.0996262
15	2.5473009	0.0685851	0.0851716
44	-0.9022171	0.2652042	0.0590017
50	2.8781573	0.4003971	0.9522079

L'observations 50 présente la distance de Cook la plus élevée, il pourrait s'agir d'une observation aberrante. On peut l'éliminer pour voir l'effet sur les résultats des estimations.

Correction du modèle

```
data2=data1[-50,]
mod6.new<- update(mod6, data=data2)
```

```
stargazer(mod6,mod6.new, type= "latex",title="Comparer les deux modèles après la supression d'une obser
```

Tab. 17 : Comparer les deux modèles après la supression d'une observation

	Dependent variable :	
	Y	
	(1)	(2)
X2	0.067***	0.058***
	(0.009)	(0.009)
X3	−1.019	−0.655
	(1.133)	(1.059)
region	−221.830**	−137.627
	(102.392)	(99.395)
X3 :region	0.715**	0.451
	(0.320)	(0.311)
Constant	268.661	195.958
	(376.775)	(350.458)
Observations	50	49
R^2	0.641	0.514
Adjusted R^2	0.609	0.470
Residual Std. Error	38.338 (df = 45)	35.568 (df = 44)
F Statistic	20.109*** (df = 4 ; 45)	11.656*** (df = 4 ; 44)
Note :		*p<0.1 ; **p<0.05 ; ***p<0.01

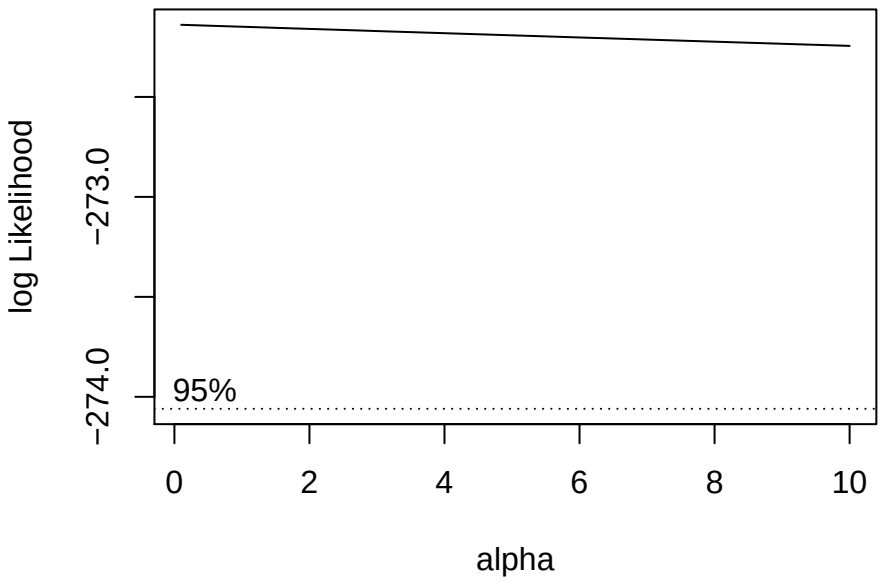
La suppression d'une observation entraîne une diminution du coefficient de détermination R^2. Donc cette méthode ne peut pas être utilisée dans ce cas.

6.2 Par transformation logarithmique

Idée est : Trouvez et éventuellement tracez la probabilité marginale (profil) pour alpha pour un modèle de transformation de la forme $log(y + \alpha) = x1 + x2 + ...$

Le "meilleur" réel α peut être estimé.

```
library(MASS)
graph = logtrans(mod6, alpha = seq(0.1, 10, length = 50))
```




```
hat_alpha = graph$x[which.max(graph$y)]
mod6.log = lm(log(Y + hat_alpha) ~ X2 + X3 + region + X3*region)
summary(mod6.log)
```

```
##
## Call:
## lm(formula = log(Y + hat_alpha) ~ X2 + X3 + region + X3 * region)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.258531 -0.076315 -0.002579  0.058610  0.272200
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.941e+00  1.446e+00   3.417  0.00144 **
## X2           1.995e-04  3.629e-05   5.498 2.23e-06 ***
## X3          -9.122e-04  4.384e-03  -0.208  0.83621
## region2     -1.579e+00  1.841e+00  -0.857  0.39617
## region3     -9.807e-01  1.453e+00  -0.675  0.50356
## region4     -1.728e+00  1.419e+00  -1.218  0.23021
## X3:region2   4.907e-03  5.812e-03   0.844  0.40335
## X3:region3   3.045e-03  4.610e-03   0.660  0.51265
## X3:region4   5.671e-03  4.516e-03   1.256  0.21633
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1252 on 41 degrees of freedom
## Multiple R-squared:  0.6626, Adjusted R-squared:  0.5967
## F-statistic: 10.06 on 8 and 41 DF,  p-value: 1.259e-07
```

Avec ce méthode, on a obtenu une amélioration de R^2

```
bptest(mod6.log)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod6.log
## BP = 16.615, df = 8, p-value = 0.03438
```

Mais d'après le test de Breush-Pagan, même s'il y a une amélioration de R^2, il reste encore le problème hétérosdasticité

6.3 Par l'estimation des moindres carrés généralisés (MCG)

On observe ici deux modèles de régression linéaire :

- Modèle original : qui est avec le problème hétérosdascité
- Modèle moindres carrés généralisés (MCG) pour l'hétéroscédasticité avec erreurs corrélées.

Lorsque l'hétéroscédasticité est présente, le meilleur estimateur linéaire sans biais dépend du σ_i^2 inconnu. Introduisons l'estimateur de cette méthode de la forme :

$$var(e_i) = \sigma_i^2 = \sigma^2 * x_i^{\gamma}$$

où γ est le paramètre inconnu que nous devons estimer. Pour obtenons l'estimateur, commençons par prendre des logs de l'équation :

$$ln(\sigma_i^2) = ln(\sigma^2) + \gamma * ln(x_i)$$

d'où un anti-log donne

$$\sigma_i^2 = exp[ln(\sigma^2) + \gamma * ln(x_i)] = exp(\alpha_1 + \alpha_2 * z_i)$$

avec $\alpha_1 = ln(\sigma^2)$; $\alpha_2 = \gamma$; et $z_i = ln(x_i)$

Donc :

$$ln(\sigma_i^2) = \alpha_1 + \alpha_2 * z_i$$

Plus précis, dans nos cas, il y a plus d'une variable explicative. On peut écire sous la forme :

$$ln(\sigma_i^2) = \alpha_1 + \alpha_2 * z_1 i + \alpha_3 * z_2 i + ... + \alpha_{k+1} * z_k i$$

avec k : le nombre des variables explicatives.

Si on estime le modèle qu'on vient d'obtenir avec $\alpha_1, \alpha_2, ..., \alpha_{k+1}$ comme des variables inconue, notre estimateur $\hat{e}_i^2$

$$ln(\hat{e}_i^2) = ln(\sigma_i^2) + v_i = \alpha_1 + \alpha_2 * z_i + v_i$$

```
data$resi6 <- mod6$residuals
varfunc.ols <- lm(log(resi6^2) ~ log(X2)+log(X3)+log(X3*region)+region, data = data)
summary(varfunc.ols)
```

```
##
## Call:
## lm(formula = log(resi6^2) ~ log(X2) + log(X3) + log(X3 * region) +
##     region, data = data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9975 -0.7523  0.0685  0.8956  3.2527
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -88.120     35.913  -2.454   0.0181 *
## log(X2)             4.541      1.923   2.361   0.0226 *
## log(X3)            12.746      5.425   2.350   0.0232 *
## log(X3 * region)   -2.976      2.526  -1.178   0.2448
## region              0.721      1.150   0.627   0.5339
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.732 on 45 degrees of freedom
## Multiple R-squared:  0.2491, Adjusted R-squared:  0.1823
## F-statistic: 3.731 on 4 and 45 DF,  p-value: 0.01049
```

Notre fonction de variance estimée est

$$ln(\hat{\sigma}_i^2) = -88.120 + 4.541z_1 + 12.746z_2 - 2.976z3 + 0.721z_4$$

Transforme les observations de manière que notre nouveau modèle ait une variance d'erreur constante.

Notre modèle est de forme :

$$Y_i = \beta_0 + \beta_1 X2_i + \beta_2 X3_i + \beta_3 region_i + \beta_4 X3_i * region_i + \epsilon_i$$

On divise deux côtés par σ_i pour obtien un nouveau modèle de régression

$$\frac{Y_i}{\hat{\sigma}_i} = \beta_0 \frac{1}{\hat{\sigma}_i} + \beta_1 \frac{X2_i}{\hat{\sigma}_i} + \beta_2 \frac{X3_i}{\hat{\sigma}_i} + \beta_3 \frac{X3_i * region_i}{\hat{\sigma}_i} + \beta_4 \frac{region_i}{\hat{\sigma}_i} + \frac{\epsilon_i}{\hat{\sigma}_i}$$

Posons : $\frac{Y_i}{\hat{\sigma}_i} = Y^*$; $\frac{1}{\hat{\sigma}_i} = X_1^*$; ... ; $\frac{region_i}{\hat{\sigma}_i} = X_5^*$; $\epsilon_i \hat{\sigma}_i = \epsilon^*$

On obtient nouveau modèle de régression de la forme :

$$Y^* = X_1^* + X_2^* + X_3^* + X_4^* + X_5^* + \epsilon^*$$

Ici le variance du nouveau modèle est homodasticité car :

$$var(e_i^*) = var(\frac{\epsilon_i}{\sigma_i^2}) = (\frac{1}{\sigma_i^2}) * var(\epsilon_i) = \frac{1}{\sigma_i^2} * \sigma_i^2 = 1(constant)$$

```
data$varfunc <- exp(varfunc.ols$fitted.values)

mod6.gls <- lm(Y~X2+X3+region+X3*region, weights = 1/sqrt(varfunc), data = data)
```

```
stargazer(mod6,mod6.gls, type= "latex",title="Comparer les deux modèles après l'estimation MCG" ,table.
```

TAB. 18 : Comparer les deux modèles après l'estimation MCG

	Dependent variable :	
	Y	
	(1)	(2)
X2	0.067***	0.067***
	(0.009)	(0.008)
X3	−1.019	−1.007
	(1.133)	(1.122)
region	−221.830**	−186.382*
	(102.392)	(104.429)
X3 :region	0.715**	0.612*
	(0.320)	(0.326)
Constant	268.661	257.922
	(376.775)	(369.213)
Observations	50	50
R^2	0.641	0.654
Adjusted R^2	0.609	0.624
Residual Std. Error (df = 45)	38.338	7.559
F Statistic (df = 4 ; 45)	20.109***	21.294***
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01	

Nos deux modèles estimés sont :

Origine :

$$\widehat{Y} = 268.661 + 0.067X2 - 1.019X3 - 221.830region + 0.715X3 * region$$

Corrigé :

$$\widehat{Y} = 257.922 + 0.067X2 - 1.007X3 - 186.382region + 0.612X3 * region$$

7 Conclusion

Pour parvenir à la modélisation la plus appropriée on a premièrement procédé à l'étude des données avec des méthodes de statistiques descriptives. Ensuite on a cherché les modèles les plus intéressants sous les critères du plus grand R^2 et du plus petit AIC. Puis, ayant sélectionné le modèle pouvant le mieux expliquer notre variable dépendante Y, on a testé si celui-ci vérifiais les hypothèses.

On a donné les moyennes et écart-type de chaque variable par chaque région ainsi que les corrélations entre les variables. Ceci nous a permis d'avoir des a prioris sur le modèle le plus approprié.

Ensuite nous avons donné des modèles par la méthode stepwise et on a comparé le R^2 ainsi que les AIC. Nous avons retenu le modèle avec le plus grand R^2 et plus petit R^2.

Ayant trouvé que le modèle $Y = X2 + X3 + X3 * region + region$ correspondait à la meilleure régression on a cherché à savoir si celui-ci validait les hypothèses d'indépendance, distribution normale des résidus et homogénéité des résidus. Avec les tests de Breusch-Pagan et test de White on a réalisé que la variance des résidus n'était pas homogène. Ceci-dit le modèle n'était pas valable.

On a essayé de résoudre le problème d'hétéroscédasticité par trois façons différentes : premièrement en essayant d'exclure les outliers en calculant la distance de Cook, deuxièmement en faisant un transformation logarithmique, puis finalement en procédant par l'estimation des moindres carrés généralisés.

L'estimation des moindres carrés ordinaires a été la seule façon pour nous de corriger l'hétéroscédasticité dans notre modèle tout en gardant un modèle pertinent.