Une exploration économétrique des arrestations en Californie

Hafsa Mousalli et Imane Lemkhayer

2024-03-21

#Partie A

##Question 1: Charger la base de donnees et vérifier qu'elle contient les variables indiquées. Vérifier que toutes les variables ont le bon format.

```
require(data.table)
require(stargazer)
require(ggplot2)
library(haven)
library(sandwich)
library(psych)

data <- read_dta("Data/crime1_simplified.dta")
```

```
#Pour vérifier que les variables ont le bon format:
data$narr86=as.numeric(sub(",", ".", data$narr86, fixed = TRUE))
data$pcnv=as.numeric(sub(",", ".", data$pcnv, fixed = TRUE))
data$avgsen=as.numeric(sub(",", ".", data$avgsen, fixed = TRUE))
data$ptime86=as.numeric(sub(",", ".", data$ptime86, fixed = TRUE))
data$qemp86=as.numeric(sub(",", ".", data$qemp86, fixed = TRUE))
data$inc86=as.numeric(sub(",", ".", data$inc86, fixed = TRUE))



summary(data)
```

```
##      narr86             pcnv            avgsen           ptime86
##  Min.   : 0.0000   Min.   :0.0000   Min.   : 0.0000   Min.   : 0.0000
##  1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000
##  Median : 0.0000   Median :0.2500   Median : 0.0000   Median : 0.0000
##  Mean   : 0.4044   Mean   :0.3578   Mean   : 0.6323   Mean   : 0.3872
##  3rd Qu.: 1.0000   3rd Qu.:0.6700   3rd Qu.: 0.0000   3rd Qu.: 0.0000
##  Max.   :12.0000   Max.   :1.0000   Max.   :59.2000   Max.   :12.0000
##      qemp86          inc86
##  Min.   :0.000   Min.   :  0.00
##  1st Qu.:1.000   1st Qu.:  0.40
##  Median :3.000   Median : 29.00
##  Mean   :2.309   Mean   : 54.97
##  3rd Qu.:4.000   3rd Qu.: 90.10
##  Max.   :4.000   Max.   :541.00
```

```
colnames(data)
```

```
## [1] "narr86"  "pcnv"    "avgsen"  "ptime86" "qemp86"  "inc86"
```

```r
head(data)
```

```
## # A tibble: 6 x 6
##   narr86  pcnv avgsen ptime86 qemp86  inc86
##    <dbl> <dbl>  <dbl>   <dbl>  <dbl>  <dbl>
## 1      0 0.380   17.6      12      0  0
## 2      2 0.440    0         0      1  0.800
## 3      1 0.330   22.8       0      0  0
## 4      2 0.25     0         5      2  8.80
## 5      1 0        0         0      2  8.10
## 6      0 1        0         0      4 97.6
```

```r
tail(data)
```

```
## # A tibble: 6 x 6
##   narr86  pcnv avgsen ptime86 qemp86  inc86
##    <dbl> <dbl>  <dbl>   <dbl>  <dbl>  <dbl>
## 1      0     0      0       0      3 119.
## 2      1     0      0       0      0   0
## 3      0     0      0       0      3  11.5
## 4      0     0      0       0      1   1.90
## 5      1     0      0       0      0   0
## 6      0     0      0       0      4 191.
```

```r
str(data)
```

```
## tibble [2,725 x 6] (S3: tbl_df/tbl/data.frame)
##  $ narr86 : num [1:2725] 0 2 1 2 1 0 2 5 0 0 ...
##  $ pcnv   : num [1:2725] 0.38 0.44 0.33 0.25 0 ...
##  $ avgsen : num [1:2725] 17.6 0 22.8 0 0 ...
##  $ ptime86: num [1:2725] 12 0 0 5 0 0 0 0 9 0 ...
##  $ qemp86 : num [1:2725] 0 1 0 2 2 4 0 0 0 3 ...
##  $ inc86  : num [1:2725] 0 0.8 0 8.8 8.1 ...
```

Calculer le carré de la variable inc86:

```r
data$inc86_squared <- data$inc86^2
```

##Question 2 : Combien y a-t-il d'observations dans les données ? Combien y a t-il de variables ? De quel type sont-elles ?

```r
str(data)
```

```
## tibble [2,725 x 7] (S3: tbl_df/tbl/data.frame)
##  $ narr86       : num [1:2725] 0 2 1 2 1 0 2 5 0 0 ...
##  $ pcnv         : num [1:2725] 0.38 0.44 0.33 0.25 0 ...
##  $ avgsen       : num [1:2725] 17.6 0 22.8 0 0 ...
##  $ ptime86      : num [1:2725] 12 0 0 5 0 0 0 0 9 0 ...
##  $ qemp86       : num [1:2725] 0 1 0 2 2 4 0 0 0 3 ...
##  $ inc86        : num [1:2725] 0 0.8 0 8.8 8.1 ...
##  $ inc86_squared: num [1:2725] 0 0.64 0 77.44 65.61 ...
```

```r
sapply(data,class)
```

```
##        narr86          pcnv        avgsen       ptime86        qemp86
##     "numeric"     "numeric"     "numeric"     "numeric"     "numeric"
##         inc86 inc86_squared
##     "numeric"     "numeric"
```

```
lignes <- nrow(data) # Nombre d'observations
colonnes <- ncol(data) # Nombre de variables
```

Il y a 2725 observations et 7 variables lorsque la variable inc86^2 est comptabilisée/incluse, sinon on en compte 6.

Ces variables sont quantitatives.

##Question 3: Y a t-il des observations manquantes ?

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

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

```
(new_DF <- data[rowSums(is.na(data)) > 0,])
```

```
## # A tibble: 0 x 7
## # i 7 variables: narr86 <dbl>, pcnv <dbl>, avgsen <dbl>, ptime86 <dbl>,
## #   qemp86 <dbl>, inc86 <dbl>, inc86_squared <dbl>
```

```
data=na.omit(data)
knitr::kable(new_DF, caption = "représentation")
```

Table 1: représentation

narr86	pcnv	avgsen	ptime86	qemp86	inc86	inc86_squared

Il n'y a pas d'observations manquantes. On observe également que le tableau est vide.

##Question 4 : Tableau de statistiques descriptives synthétiques pour les variables quantitatives.

```
descriptives <- describe(data)
descriptives2 <- descriptives[,c("min", "max", "mean", "median", "sd")]

colnames(descriptives2) <- c("Minimum", "Maximum", "Moyenne", "Médiane", "Ecart-type")

print(descriptives)
```

```
##               vars    n    mean       sd median trimmed     mad min      max
## narr86           1 2725    0.40     0.86   0.00    0.22    0.00   0     12.0
## pcnv             2 2725    0.36     0.40   0.25    0.32    0.37   0      1.0
## avgsen           3 2725    0.63     3.51   0.00    0.00    0.00   0     59.2
## ptime86          4 2725    0.39     1.95   0.00    0.00    0.00   0     12.0
## qemp86           5 2725    2.31     1.61   3.00    2.39    1.48   0      4.0
## inc86            6 2725   54.97    66.63  29.00   43.29   43.00   0    541.0
## inc86_squared    7 2725 7458.93 16361.24 841.00 3788.82 1246.87   0 292681.0
##                 range  skew kurtosis     se
## narr86           12.0  4.11    29.64   0.02
## pcnv              1.0  0.60    -1.19   0.01
## avgsen           59.2  7.72    76.57   0.07
## ptime86          12.0  5.31    27.35   0.04
## qemp86            4.0 -0.32    -1.50   0.03
## inc86           541.0  1.62     3.51   1.28
## inc86_squared 292681.0  6.11    68.42 313.42
```

```
knitr::kable(descriptives2, caption ="Tableau de statistiques synthétiques")
```

Table 2: Tableau de statistiques synthétiques

	Minimum	Maximum	Moyenne	Médiane	Ecart-type
narr86	0	12.0	0.4044037	0.00	8.590768e-01
pcnv	0	1.0	0.3577872	0.25	3.951920e-01
avgsen	0	59.2	0.6322936	0.00	3.508031e+00
ptime86	0	12.0	0.3871560	0.00	1.950051e+00
qemp86	0	4.0	2.3090275	3.00	1.610428e+00
inc86	0	541.0	54.9670459	29.00	6.662721e+01
inc86_squared	0	292681.0	7458.9326215	841.00	1.636124e+04

- La variable narr86 montre que la majorité des individus n'ont pas été arrêtés en 1986 (médiane de 0), mais il y a eu jusqu'à 12 arrestations pour certains, indiquant des cas extrêmes ou des individus avec un comportement nettement différent.
- Pour pcnv, la médiane inférieure à la moyenne suggère une distribution asymétrique, avec une concentration d'individus ayant une faible proportion d'arrestations (médiane de 0.29 et moyenne de 0.3656) menant à une condamnation.
- La variable avgsen a une médiane et un premier quartile de 0, indiquant que beaucoup d'individus n'ont pas reçu de peine ou que les peines étaient courtes, mais la présence d'une valeur maximale élevée (59,2) montre que certains ont eu des peines beaucoup plus longues. La moyenne est de 0.7962.
- ptime86 a des statistiques similaires à avgsen, avec la plupart des valeurs concentrées autour de 0, suggérant que beaucoup n'ont pas passé de temps en prison en 1986 et une valeur maximal de 12 mois.
- qemp86 indique que la plupart des individus ont travaillé au moins deux trimestres en 1986, avec une médiane à 3 et un maximum à 4, ce qui signifie que beaucoup ont travaillé toute l'année.
- Le revenu inc86 montre une large gamme, avec une médiane (44.7) beaucoup plus basse que la moyenne(66.63), ce qui pourrait indiquer la présence de revenus très élevés pour certains individus, influençant la moyenne vers le haut.
- Pour inc86_squared, la distribution est fortement inclinée vers la droite, avec une moyenne (9103.0) beaucoup plus élevée que la médiane (1998.1), suggérant la présence de revenus très élevés pour certains individus. Le maximum extrêmement élevé (292681.0) confirme la présence de valeurs extrêmes, indiquant une inégalité significative dans la distribution du revenu au sein de l'échantillon.

#PARTIE B

##Question 1 : Quel est le signe attendu des différents paramètres B ?

Bo la constante est à priori positive.

Beta1: On pourrait s'attendre à ce que le coefficient soit positif. Une proportion plus élevée d'arrestations avant 1986 menant à des condamnations pourrait être associée à un nombre plus élevé d'arrestations en 1986. En effet, les individus ayant un historique d'arrestations menant à des condamnations sont plus susceptibles d'être arrêtés en 1986.

Beta2: On pourrait s'attendre à ce que ce coefficient soit négatif. Une durée moyenne de peine plus longue pour les condamnations avant 1986 pourrait dissuader les hommes de commettre à nouveau des infractions, ce qui pourrait se traduire par un nombre moins élevé d'arrestations en 1986. Cepdendant, dans le cas où la durée moyenne d'un individu est assez longue, il serait susceptible de commettre des infractions au sein de la prison et à sa sortie, donc le coeffiecient pourrait être positif et donc se traduire par un nombre plus élevé d'arrestations.

Beta3 : Le signe attendu de ce coefficient est négatif. En effet, passer plus de temps en prison peut dissuader les hommes à commettre à nouveau des infractions et c'est donc associé à moins d'arrestations.

Beta4 : On pourrait s'attendre à ce que ce coefficient soit négatif. Plus un individu travaille, moins il est susceptible de faire des activités qui pourrait l'emmener à être arrétés. L'emploi peut contribuer à réduire le comportement criminel.

Beta5 : Le signe attendu de ce coefficient est négatif. En effet, un homme avec un revenu plus élevé a davantage accès à des opportunités éducatives et professionnelles, pouvant réduire la probabilité de faire des actes répréhensibles et donc il sera moins sujet à se faire arrêter.

##Question 2: Estimation du modèle par la méthode des MCO

```
reg1 <- lm(narr86~pcnv+avgsen+ptime86+qemp86+inc86, data = data)
summary(reg1)
```

```
##
## Call:
## lm(formula = narr86 ~ pcnv + avgsen + ptime86 + qemp86 + inc86,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8839 -0.4736 -0.2609  0.3704 11.4226
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.6844436  0.0333366  20.531  < 2e-16 ***
## pcnv        -0.1529514  0.0406935  -3.759 0.000174 ***
## avgsen       0.0065626  0.0047180   1.391 0.164346
## ptime86     -0.0352246  0.0087695  -4.017 6.06e-05 ***
## qemp86      -0.0539058  0.0145494  -3.705 0.000216 ***
## inc86       -0.0016620  0.0003437  -4.836 1.40e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8379 on 2719 degrees of freedom
## Multiple R-squared:  0.05036,    Adjusted R-squared:  0.04862
## F-statistic: 28.84 on 5 and 2719 DF,  p-value: < 2.2e-16
```

##Question 3: Interprétation des coefficients

Pour savoir si un coefficient est significatif, il faut observer si la p-value est inférieure ou supérieur à Alpha. En effet, si la p-value est inférieur à alpha, alors le coefficient est significatif au seuil de alpha.

- B0 : Sa p-value est inférieure à 2x10^-16 donc il est significatif à 1%,5%, et 10%. Il est estimé à environ 0.7. Cela indique le nombre moyen d'arrestations en 1986 pour une personne qui a des valeurs de zéro pour toutes les autres variables dans le modèle.
- B1 (pcnv): Sa p-value est de 0.000174 donc il est significatif à 1%,5%, et 10%. En effet, les individus avec un historique plus important d'arrestations menant à des condamnations sont moins susceptibles d'être arrêtés en 1986. Une augmentation d'une unité dans la proportion des arrestations avant 1986 menant à une condamnation est associée, en moyenne, à une baisse d'environ 0.153 (B1 estimate) dU nombre d'arrestations en 1986, toutes choses égales par ailleurs.
- B2 (avgsen): Sa p-value est d'environ 0.164, elle est donc supérieur au seuil 5%,1% et 10%. B2 n'est donc pas significatif ni à 1%,5%, et 10%. Une augmentation de la durée moyenne de la peine pour les condamnations avant 1986 est associée, en moyenne, à une hausse d'environ 0.0065 du nombre de fois où l'homme a été arrêté en 1986, toutes choses égales par ailleurs. L'homme qui a passé un certain temps en prison, sera plus susceptible d'être arrêté après sa condamnation.

- B3 (ptime86): Sa p-value est de 6.06*10^-5, elle est donc inférieur aux 3 seuils. B3 est donc signficative à 1%,5%, et 10%. Pour chaque mois supplémentaire passé en prison en 1986, le nombre d'arrestations diminue en moyenne d'environ 0.035, toutes choses égales par ailleurs. Ceci est conforme à l'attente que le temps passé en prison pourrait réduire la probabilité d'engager dans des comportements criminels.
- B4 (qemp86) : Sa p-value est 0.000216. Il est donc significatif à 1%,5% et 10%. Chaque trimestre supplémentaire de travail en 1986 est associé à une diminution moyenne de 0.0539058 dans le nombre d'arrestations, à toutes autres facteurs fixés. Ceci suggère que l'emploi peut jouer un rôle dans la réduction du comportement criminel.
- B5 (inc86) : Sa p-value est de 1.40e-06. Il est donc significatif à 1%,5% et 10%. Cela indique qu'une augmentation de 100 dollars du revenu est associée à une diminution d'environ 0.0017 du nombre d'arrestations, toutes choses égales par ailleurs. Bien que l'effet soit relativement petit, il est statistiquement significatif, suggérant que le revenu a un impact négatif sur la probabilité d'être arrêté.

Donc : B0,B1,B3,B4,et B5 sont signficatifs.

##Question 4:

```
reg2 <- lm(narr86~pcnv+avgsen+ptime86+qemp86+inc86+inc86_squared, data = data)
summary(reg2)
```

```
##
## Call:
## lm(formula = narr86 ~ pcnv + avgsen + ptime86 + qemp86 + inc86 +
##     inc86_squared, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8816 -0.4955 -0.2385  0.3554 11.4085
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.790e-01  3.332e-02  20.379  < 2e-16 ***
## pcnv          -1.568e-01  4.064e-02  -3.859 0.000116 ***
## avgsen         6.031e-03  4.712e-03   1.280 0.200715
## ptime86       -3.449e-02  8.757e-03  -3.938 8.41e-05 ***
## qemp86        -2.054e-02  1.776e-02  -1.156 0.247588
## inc86         -4.099e-03  8.217e-04  -4.989 6.47e-07 ***
## inc86_squared  8.549e-06  2.619e-06   3.264 0.001113 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8365 on 2718 degrees of freedom
## Multiple R-squared:  0.05407,    Adjusted R-squared:  0.05198
## F-statistic: 25.89 on 6 and 2718 DF,  p-value: < 2.2e-16
```

On observe que la p-value du revenu au carré (inc86_squared) est de 0.001113; B6 est donc significatif aux seuils 1%, 5% et 10%. Cette significativité permet d'indiquer qu'il y a une relation non-linéaire entre le revenu et le nombre d'arrestations. Le signe positif du coefficient pour inc86_squared suggère que, tout en tenant compte de l'effet du revenu (qui a un coefficient négatif), l'augmentation du revenu a un effet décroissant sur le nombre d'arrestations jusqu'à un certain point, après lequel de plus grands revenus peuvent être associés à une augmentation du nombre d'arrestations. Cela pourrait représenter un effet de seuil où, après avoir atteint un certain niveau de revenu, des augmentations supplémentaires sont associées à des comportements qui peuvent augmenter le risque d'arrestation (par exemple, une plus grande consommation d'alcool, de drogues, implications dans des meurtres...).

De plus, l'ajout de cette variable améliore également le R carré ajusté du modèle (on passe de 0.04862 à

0.05198), ce qui suggère que le modèle avec le terme quadratique explique une plus grande partie de la variance dans le nombre d'arrestations que le modèle sans ce terme.

Donc, cette hausse du R carré ajusté suggère que l'ajout du revenu au carré dans la régression est nécessaire.

##Question 5: Interprétation de chaque coefficients

-> Ici les interprétations sont quasi similaires à la question 3 (à part quelques exceptions). Les valeurs pour B0,B1,B2,B3 sont presque identiques ainsi que leur p-value.

- B0 : Sa p-value est inférieure à 2x10^-16 donc il est significatif à 1%,5%, et 10%. Il est estimé à environ 0.679. Cela indique le nombre moyen d'arrestations en 1986 pour une personne qui a des valeurs de zéro pour toutes les autres variables dans le modèle.
- B1 (pcnv): Sa p-value est de 0.000116 donc il est significatif à 1%,5%, et 10%. Une augmentation d'une unité dans la proportion des arrestations avant 1986 menant à une condamnation est associée, en moyenne, à une baisse d'environ 0.153 (B1 estimate) du nombre d'arrestations en 1986, toutes choses égales par ailleurs.
- B2 (avgsen): Sa p-value est de 0.2. B2 n'est donc pas significatif ni à 1%,5%, et 10%. Une augmentation de la durée moyenne de la peine pour les condamnations avant 1986 est associée, en moyenne, à une hausse d'environ 0.006 du nombre de fois où l'homme a été arrêté en 1986, toutes choses égales par ailleurs.
- B3 (ptime86): Sa p-value est de 8.41e-05. B3 est donc signficative à 1%,5%, et 10%. Pour chaque mois supplémentaire passé en prison en 1986, le nombre d'arrestations diminue en moyenne d'environ 0.0344, toutes choses égales par ailleurs.
- B4 (qemp86) : Sa p-value est d'environ 0.25. Il n'est donc pas significatif (ni à 1%,5% et 10%). Chaque trimestre supplémentaire de travail en 1986 est associé à une diminution moyenne de 0.02 dans le nombre d'arrestations, à toutes autres facteurs fixés.
- B5 (inc86) : Sa p-value est de 6.47e-07. Il est donc significatif à 1%,5% et 10%. Cela indique qu'une augmentation de 100 dollars du revenu est associée à une diminution d'environ 0.004 du nombre d'arrestations, toutes choses égales par ailleurs.
- B6 (inc86_squared): Sa p-value est de 0.001113. Il est donc signficifatif à 1% et 5%. Pour chaque augmentation d'une unité dans le revenu au carré (c'est-à-dire pour chaque 100^2), le nombre d'arrestations augmente en moyenne de 8.549e-06, toutes choses égales par ailleurs.

##Question 6: Existe-t-il un problème d'hétéroscédasticité dans les données ? Mettre en oeuvre un test que vous connaissez pour tester cela.

Pour tester si la variance des résidus est constante ou non, on peut réaliser le test de Breusch-Pagan afin de savoir s'il existe un problème d'hétéroscédasticitié dans les données.

Hypothèse nulle: L'hypothèse nulle du test de Breusch-Pagan stipule qu'il n'y a pas d'hétéroscédasticité dans le modèle de régression, ce qui signifie que la variance des erreurs est constante pour toutes les valeurs des variables explicatives.

Hypothèse alternative: L'hypothèse alternative du test de Breusch-Pagan est que l'hétéroscédasticité est présente dans le modèle de régression, ce qui signifie que la variance des erreurs n'est pas constante et dépend des valeurs des variables explicatives.

```
require(lmtest)
```

```
## Le chargement a nécessité le package : lmtest
```

```
## Warning: le package 'lmtest' a été compilé avec la version R 4.2.3
```

```
## Le chargement a nécessité le package : zoo
```

```
## Warning: le package 'zoo' a été compilé avec la version R 4.2.3
```

```
##
## Attachement du package : 'zoo'
```

```
## Les objets suivants sont masqués depuis 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(reg2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  reg2
## BP = 40.013, df = 6, p-value = 4.529e-07
```

La p-value est égale à 4.529e-07 et inférieure à 0.05. On peut rejeter H0. Le test confirme qu'on peut rejeter l'hypothèse nulle d'homoscédasticité. Il existe donc un problème d'hétéroscédasticité dans les données.

##Question 7 : Corriger les erreurs-types d'un éventuel problème d'hétéroscédasticité, quelle que soit la réponse à la question précédente. Est-ce que cela crée des modifications majeures ?

En cas d'hétéroscedasticité, on peut calculer les écarts-types robustes:

```
library("lmtest")
coeftest(reg2, vcov = vcovHC(reg2, type = "HC1"))
```

```
##
## t test of coefficients:
##
##                   Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)    6.7902e-01  4.1409e-02 16.3978 < 2.2e-16 ***
## pcnv          -1.5684e-01  3.3684e-02 -4.6561 3.377e-06 ***
## avgsen         6.0311e-03  5.1654e-03  1.1676    0.2431
## ptime86       -3.4488e-02  6.1026e-03 -5.6514 1.757e-08 ***
## qemp86        -2.0540e-02  1.7550e-02 -1.1704    0.2420
## inc86         -4.0990e-03  6.5797e-04 -6.2298 5.395e-10 ***
## inc86_squared  8.5495e-06  2.0166e-06  4.2396 2.314e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#HC1 permet de faire une correction
summary(reg2)
```

```
##
## Call:
## lm(formula = narr86 ~ pcnv + avgsen + ptime86 + qemp86 + inc86 +
##     inc86_squared, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8816 -0.4955 -0.2385  0.3554 11.4085
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.790e-01  3.332e-02  20.379  < 2e-16 ***
## pcnv          -1.568e-01  4.064e-02  -3.859 0.000116 ***
## avgsen         6.031e-03  4.712e-03   1.280 0.200715
## ptime86       -3.449e-02  8.757e-03  -3.938 8.41e-05 ***
```

```
## qemp86        -2.054e-02  1.776e-02  -1.156 0.247588
## inc86         -4.099e-03  8.217e-04  -4.989 6.47e-07 ***
## inc86_squared  8.549e-06  2.619e-06   3.264 0.001113 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8365 on 2718 degrees of freedom
## Multiple R-squared:  0.05407,    Adjusted R-squared:  0.05198
## F-statistic: 25.89 on 6 and 2718 DF,  p-value: < 2.2e-16
```

On constate qu'il n'y a pas de modifications majeures. En comparant les résultats standard avec les résultats corrigés pour l'hétéroscédasticité, il semble que les significations statistiques des coefficients restent largement inchangées.

#Partie C :

##Question 1: Nous introduisons maintenant une indicatrice "condamne" valant 1 pour les personnes qui ont déjà été condamnées avant 1986 (pcnv > 0) et 0 sinon. Construire cette indicatrice.

L'équation du modèle est:

$$narr86_i = \beta_0 + \beta_1 condamne_i + \beta_2 avgsen_i + \beta_3 ptime86_i + \beta_4 qemp86_i + \beta_5 inc86_i + \beta_6 inc86_squared_i + \varepsilon_i \quad \forall i \in [1, N] \tag{1}$$

où $pcnv_i$ est une variable indicatrice égale à 1 si les personnes qui ont déjà été condamnées avant 1986 et zéro sinon.

```
data$condamne <- 0
data$condamne[data$pcnv > 0] <- 1 #remplace la valeur 0 par 1 pour toutes les observations ou la condit

str(data)
```

```
## tibble [2,725 x 8] (S3: tbl_df/tbl/data.frame)
##  $ narr86       : num [1:2725] 0 2 1 2 1 0 2 5 0 0 ...
##  $ pcnv         : num [1:2725] 0.38 0.44 0.33 0.25 0 ...
##  $ avgsen       : num [1:2725] 17.6 0 22.8 0 0 ...
##  $ ptime86      : num [1:2725] 12 0 0 5 0 0 0 0 9 0 ...
##  $ qemp86       : num [1:2725] 0 1 0 2 2 4 0 0 0 3 ...
##  $ inc86        : num [1:2725] 0 0.8 0 8.8 8.1 ...
##  $ inc86_squared: num [1:2725] 0 0.64 0 77.44 65.61 ...
##  $ condamne     : num [1:2725] 1 1 1 1 0 1 1 1 1 1 ...
```

```
head(data)
```

```
## # A tibble: 6 x 8
##   narr86  pcnv avgsen ptime86 qemp86  inc86 inc86_squared condamne
##    <dbl> <dbl>  <dbl>   <dbl>  <dbl>  <dbl>         <dbl>    <dbl>
## 1      0 0.380   17.6      12      0  0              0            1
## 2      2 0.440    0         0      1  0.800          0.640        1
## 3      1 0.330   22.8       0      0  0              0            1
## 4      2 0.25     0         5      2  8.80          77.4          1
## 5      1 0        0         0      2  8.10          65.6          0
## 6      0 1        0         0      4 97.6         9526.           1
```

```
attach(data)
```

##Question 2 : Estimer la régression suivante

```
reg3 <- lm(narr86~condamne+avgsen+ptime86+qemp86+inc86+inc86_squared, data = data)
summary(reg3)
```

```
##
## Call:
## lm(formula = narr86 ~ condamne + avgsen + ptime86 + qemp86 +
##     inc86 + inc86_squared, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8065 -0.5016 -0.2362  0.3766 11.4201
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.276e-01  3.562e-02  17.621  < 2e-16 ***
## condamne      -4.179e-03  3.296e-02  -0.127  0.89910
## avgsen         5.850e-03  4.741e-03   1.234  0.21728
## ptime86       -3.610e-02  8.833e-03  -4.087 4.49e-05 ***
## qemp86        -2.257e-02  1.780e-02  -1.268  0.20487
## inc86         -4.007e-03  8.250e-04  -4.857 1.26e-06 ***
## inc86_squared  8.268e-06  2.628e-06   3.146  0.00167 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8387 on 2718 degrees of freedom
## Multiple R-squared:  0.04889,    Adjusted R-squared:  0.04679
## F-statistic: 23.29 on 6 and 2718 DF,  p-value: < 2.2e-16
```

##Question 3: Toutes choses égales par ailleurs, existe-t-il une différence dans le nombre d'arrestations moyen en 1986 pour les personnes déjà condamnées et celles qui ne l'ont jamais été ? Interpréter.

Dans les résultats de régression, le coefficient pour condamne est de -0.1579 et sa p-value est de 0.89190, ce qui est bien au-dessus du seuil de significativité. Cela signifie qu'il n'y a pas de différence statistiquement significative du nombre moyen d'arrestations en 1986 entre les personnes ayant été condamnées et celles qui ne l'ont pas été, toutes choses égales par ailleurs.

##Question 4: Nous souhaitons maintenant savoir si l'effet du nombre de trimestres où la personne a été employé sur le nombre d'arrestations en 1986 est différent selon que la personne a déjà été condamnée ou pas. Ecrire la spécification du modèle.

Nouveau modèle :

$$narr86_i = \beta_0 + \beta_1 condamne_i + \beta_2 avgsen_i + \beta_3 ptime86_i + \beta_4 qemp86_i + \beta_5 inc86_i + \beta_6 inc86_squared_i + \beta_7 (qemp86_i * comdamne_i) \varepsilon_i \quad (2)$$

condamne_i * qemp86_i est l'interaction entre les variables dummy condamne et qemp86.

##Question 5 : Estimer le modèle

```
reg4 <- lm(narr86~condamne+avgsen+ptime86+qemp86+inc86+inc86_squared+(condamne*qemp86), data = data)
summary(reg4)
```

```
##
## Call:
## lm(formula = narr86 ~ condamne + avgsen + ptime86 + qemp86 +
##     inc86 + inc86_squared + (condamne * qemp86), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8889 -0.4958 -0.2405  0.3505 11.3687
##
```

```
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.108e-01  4.511e-02  11.325  < 2e-16 ***
## condamne         2.028e-01  5.924e-02   3.423  0.00063 ***
## avgsen           5.601e-03  4.727e-03   1.185  0.23615
## ptime86         -4.333e-02  8.973e-03  -4.829 1.45e-06 ***
## qemp86           2.582e-02  2.116e-02   1.220  0.22260
## inc86           -4.104e-03  8.228e-04  -4.988 6.50e-07 ***
## inc86_squared    8.422e-06  2.620e-06   3.214  0.00132 **
## condamne:qemp86 -8.659e-02  2.063e-02  -4.198 2.78e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8362 on 2717 degrees of freedom
## Multiple R-squared:  0.05502,    Adjusted R-squared:  0.05259
## F-statistic:  22.6 on 7 and 2717 DF,  p-value: < 2.2e-16
```

##Question 6: Quelle est l'effet moyen sur le nombre d'arrestations en 1986 d'une augmentation du nombre de trimestre travaillés pour une personne déjà condamnée ? Et pour une personne qui l'a jamais été ? La différence est-elle significative ?

a) Effet moyen avec nombre de trimestre et condamné

```
Espérance_condamne <- predict(reg4, newdata = data.frame(condamne =1, avgsen = 1, ptime86 = 0, qemp86 =
Espérance_condamne
```

```
##        1
## 0.658406
```

L'effet moyen est de 0.658406. Lorsqu'une personne a été condamnée, l'augmentation d'un trimestre de travail est associée à une diminution prévue de 0.658406 dans le nombre d'arrestations en 1986, toutes chose égales par ailleurs.

b) Effet moyen avec nombre de trimestre et pas condamné

```
Espérance_pascondamne <- predict(reg4, newdata = data.frame(condamne =0, avgsen = 1, ptime86 = 0, qemp8
Espérance_pascondamne
```

```
##        1
## 0.542242
```

L'effet moyen lorsque la personne n'est pas condamnée est de 0.542242.

Pour une personne n'ayant pas été condamnée, l'augmentation d'un trimestre de travail est associée à une diminution prévue de 0.542242 dans le nombre d'arrestations en 1986, toutes choses égales par ailleurs.

c) Faisons la différence :

```
diff <- Espérance_condamne - Espérance_pascondamne
diff
```

```
##        1
## 0.116164
```

Pour savoir si la différence est signficative on regarde la p-value de la variable (condamne*qemp86). Elle est égale à 2.78e-05. Donc la différence est bien significative (à 1%, 5%, 10%).

La différence dans la diminution prévue du nombre d'arrestations entre une personne condamnée et une non condamnée pour un trimestre supplémentaire de travail est de 0.116164.

#Partie D

##Question 1: Faire un resumé de votre analyse. Indiquer les faiblesses selon vous de l'analyse à laquelle vous avez abouti (vous pouvez par exemple parler de l'echantillon, du modéle, d'éventuels problémes d'endogènéité ...)

- Résumé PARTIE A

Dans cette partie, tout d'abord nous avons commencé notre exploration par une analyse descriptive des données, découvrant ainsi une répartition asymétrique pour certaines variables. Par exemple, le nombre d'arrestations en 1986 présente une concentration de valeurs à zéro, mais aussi des valeurs extrêmes, ce qui pourrait indiquer un groupe restreint d'individus avec une activité criminelle élevée. Cette analyse (cf tableau de statistiques) nous a révélé une variété de profils économiques et criminels dans notre échantillon.

- Résumé PARTIE B

Dans cette section, nous avons estimé un modèle de régression linéaire pour déterminer les facteurs influençant le nombre d'arrestations en 1986. Plusieurs variables étaient significatives, suggérant des effets variés sur le nombre d'arrestations. Par exemple, les individus avec un historique plus important d'arrestations menant à des condamnations étaient moins susceptibles d'être arrêtés, et une augmentation du revenu était associée à une diminution du nombre d'arrestations.

De plus, la significativité des coefficients a été testée, et le test de Breusch-Pagan a été appliqué pour aborder l'hétéroscédasticité, garantissant ainsi que les erreurs-types sont robustes et que les inferences sont fiables.

On a également au début, fais des hypothèses sur les signes attendus des coefficients Beta, suivi de leurs interprétations.

- Résumé PARTIE C

En ajoutant une variable indicatrice pour les personnes précédemment condamnées, nous avons pu examiner l'impact du passé judiciaire sur la probabilité d'arrestation en 1986. Cependant, le coefficient pour cette variable indicatrice n'était pas significatif. Donc il n'y avait pas de différence significative dans le nombre moyen d'arrestations entre ces deux groupes.

Ensuite, nous avons examiné l'effet de l'emploi sur le nombre d'arrestations, en trouvant une interaction significative : l'effet de l'emploi sur le nombre d'arrestations diffère selon le statut de condamnation d'un individu.

- LES FAIBLESSES

Notre analyse a mis en avant des interactions complexes entre les antécédents juridiques, l'emploi, et le revenu en relation avec le nombre d'arrestations. Nous avons en effet identifié des faiblesses potientielles.

1/ Premièrement, il existe un risque d'endogénéité, car des variables non observées pourraient influencer à la fois la probabilité d'une condamnation et celle d'une arrestation. C'est-à-dire que certaines variables explicatives sont corrélées avec l'erreur de terme, biaisant ainsi les estimations.

->Par exemple:

- l'accès à l'éducation : une bonne éducation peut offrir plusieurs bonnes opportunites (emploi, réseau, valeurs et principes) pour l'individu. A l'inverse cela peut augmenter le risque de comportement déliquants.
- la catégorie socio-professionnelle des parents : son interprétation rejoint celle de l'éducation.
- les expériences de violences : par exemple un enfant ayant vécu dans un foyer violent sera plus susceptible d'adopter des comportements violents aussi.
- la zone géographique : elle peut affecter les taux d'arrestations, une personne vivant en banlieu "dangereuse" est plus susceptible d'être arrétée qu'une personne vivant dans un quartier chic.

Ces variables omises, pourraient influencer les résultats et entrainer un biais dans l'analyse économétrique.

2/ Deuxièmement, bien que nous ayons corrigé l'hétéroscédascticité, il se peut que cela ne résolve pas entièrement la mauvaise spécification de notre modèle pour chaque observation. Il se pourrait aussi que notre modèle n'ait pas capté toutes les nuances dans la façon dont nos variables sont reliées entre elles.

-> Par exemple, il se pourrait qu'à un certain niveau de revenu très élevé, le nombre d'arrestations commence à augmenter à nouveau, peut-être en raison de crimes ou d'autres délits; Notre modèle actuel, s'il ne tient compte que d'une relation linéaire simple entre le revenu et les arrestations, pourrait manquer cette nuance, donc ne pas refléter précisément la réalité, et donc ne pas être vraiment juste.

3/ L'échantillon est très restreint, limité. Il pourrait ne pas être représentatif de l'ensemble de la population ou de ses sous-groupes, ce qui pose donc un problème sur le fait de généraliser nos résultats et nos interprétations.

Dans notre étude on se concentre seulement sur des hommes nés en Californie entre 1960 et 1961. Cela signifie que nos conclusions pourraient ne pas s'appliquer à des personnes d'autres âges, sexes, ou régions.

4/ La multicollinéarité pourrait être une autre faiblesse potentielle c'est à dire lorsque deux variables ou plus dans notre modèle de régression sont fortement corrélées entre elles. On aura du mal à distinguer l'effet individuel de chaque variable sur la variable dépendante (un changement dans une variable dépendra des changements dans une autre variable).

->Par exemple, si dans notre étude le revenu et le nombre de trimestres travaillés sont fortement corrélés (les personnes avec des revenus plus élevés pourraient également avoir tendance à travailler plus de trimestres), il pourrait être compliqué d'évaluer l'impact spécifique de chacun de ces facteurs sur le nombre d'arrestations.