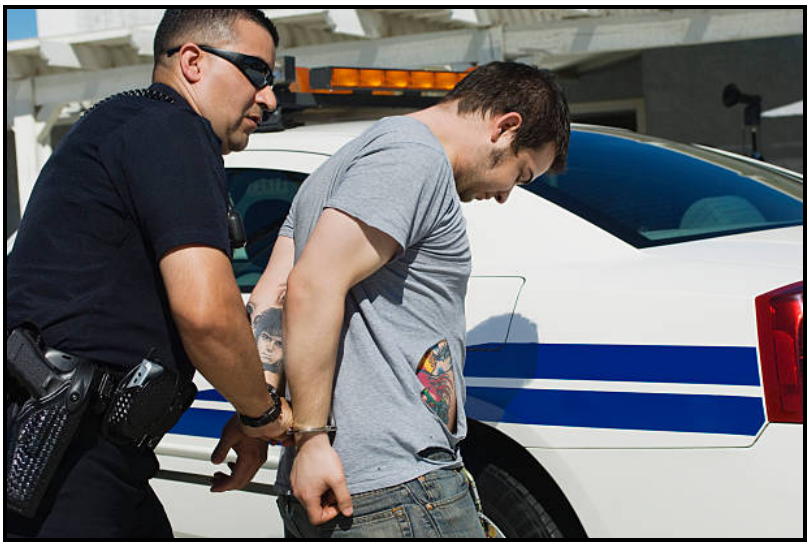
Projet sur R

Étude économétrique : probabilité d'être arrêté par la police



Réalisé par Quentin Ruel et Zaccharie Ennya

27 mars 2024

Table des matières

Exercice :

Nous allons, au cours de cet article, nous intéresser à la probabibilité de se faire arrêter par la police. Pour ce faire, nous allons nous appuyer sur une base de données portant sur des hommes nés en Californie entre 1960 et 1961 et regroupant le nombre d'arrestations pour chaque homme en 1986, sachant qu'ils ont tous déjà été arrêtés une fois avant 1986.

Nous disposons des variables suivantes :

- Le nombre d'arrestations par la police avant 1986 ($narr86$)
- La proportion des arrestations avant 1986 qui ont mené à une condamnation ($pcnv$)
- La durée moyenne de la peine pour les condamnations avant 1986 (en mois) ($avgsen$)
- Le nombre de mois passés en prison en 1986 ($ptime86$)
- Le nombre de trimestres en emploi en 1986 ($qemp86$)
- Le revenu disponible en 1986 (en centaines de dollars) ($inc86$)

Notre document sera organisé en 4 parties, à l'intérieur desquelles nous répondrons à un ensemble de questions.

Partie A

Question 1 : Charger la base de données et vérifier qu'elle contient les variables indiquées. Vérifier que toutes les variables ont le bon format. Calculer le carré de la variable inc86.

La base de données étant en .dta, nous devons utiliser le package **haven** afin d'importer nos données dans R. Nous avons donc utilisé les commandes suivantes :

```
library(haven)
df <- read_dta("crime1_simplified.dta")
```

La base de données étant chargée, nous vérifions que nous disposons bien des 6 variables indiquées précédemment :

```
colnames(df)
```

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

On constate donc qu'aucune variable n'est manquante. Afin de voir leur format, nous utilisons le code suivant :

```
str(df)
```

```
## 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 ...
##   ..- attr(*, "label")= chr "# times arrested, 1986"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ pcnv   : num [1:2725] 0.38 0.44 0.33 0.25 0 ...
##   ..- attr(*, "label")= chr "proportion of prior convictions"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ avgsen : num [1:2725] 17.6 0 22.8 0 0 ...
##   ..- attr(*, "label")= chr "avg sentence length, mos."
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ ptime86: num [1:2725] 12 0 0 5 0 0 0 0 9 0 ...
```

```
##   ..- attr(*, "label")= chr "mos. in prison during 1986"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ qemp86 : num [1:2725] 0 1 0 2 2 4 0 0 0 3 ...
##   ..- attr(*, "label")= chr "# quarters employed, 1986"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ inc86  : num [1:2725] 0 0.8 0 8.8 8.1 ...
##   ..- attr(*, "label")= chr "legal income, 1986, $100s"
##   ..- attr(*, "format.stata")= chr "%9.0g"
```

Nos 6 variables sont, comme on s'y attendait, exprimées en données numériques. Elles ont donc le bon format afin de les manipuler.

Enfin, nous calculons le carré de la variable $inc86$ que nous allons affecter à une nouvelle variable $inc86^2$ et que nous allons ajouter à notre base de données (sous le nom de "inc86_2"). Nous avons réalisé cela avec la commande suivante :

```
df$inc86_2 <- df$inc86^2
head(df$inc86_2)
```

```
## [1]    0.00000    0.64000    0.00000   77.44000   65.61001 9525.75970
```

La commande **head** nous permet d'avoir un aperçu de notre variable $inc86^2$, en nous renvoyant les 6 premières lignes parmi les 2725 qui la composent.

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

Afin de voir le nombre d'observations dans les données et le nombre de variables, nous utilisons les commandes suivantes :

```
nrow(df)
```

```
## [1] 2725
```

```
ncol(df)
```

```
## [1] 7
```

Le nombre de lignes obtenu grâce à la commande **nrow(df)** correspond au nombre d'observations, ici nous en disposons donc de 2725. Le nombre de colonnes indiqué avec la commande **ncol(df)** nous renvoie, elle, le nombre de variables observées. Nous en avons donc 7 à partir de notre base de données. Afin de voir précisément de quel type sont nos 7 variables, nous pouvons utiliser la commande **class()** pour chacune d'entre elles :

```
class(df$narr86)
```

```
## [1] "numeric"
```

```
class(df$pcnv)
```

```
## [1] "numeric"
```

```
class(df$avgsen)
```

```
## [1] "numeric"
```

```
class(df$ptime86)
```

```
## [1] "numeric"
```

```r
class(df$qemp86)
```

```
## [1] "numeric"
```

```r
class(df$inc86)
```

```
## [1] "numeric"
```

```r
class(df$inc86_2)
```

```
## [1] "numeric"
```

Nos 7 variables sont donc toutes numériques.

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

Afin de savoir s'il existe des valeurs manquantes au sein de notre base de donnée, nous utilisons la commande qui suit :

```r
any(is.na(df))
```

```
## [1] FALSE
```

Elle nous renvoie le message "FALSE", ce qui signifie qu'il n'y a aucune observation manquante au sein de notre base de données. C'est une bonne chose, parce que cela indique que nous pouvons garder la totalité de nos 2725 observations sans avoir à en supprimer ; nous ne perderons pas en précision d'estimation des paramètres de nos modèles.

Question 4 : Réaliser un tableau de statistiques descriptives synthétiques pour les variables quantitatives. Le sommaire inclut minimum, maximum, moyenne, médiane et écart-type des variables. Commenter ces statistiques.

Nous avons réalisé un tableau de statistiques descriptives de nos variables grâce aux lignes de code qui suivent :

```r
library(knitr)
library(kableExtra)

min <- apply(df,2,min)
max <- apply(df,2,max)
moy <- apply(df,2,mean)
med <- apply(df,2,median)
ectype <- apply(df,2,sd)

tab <- data.frame(
  Minimum = min,
  Maximum = max,
  Moyenne = moy,
  Médiane = med,
  "Ecart-type" = ectype)

kable_styling(kable(tab, align="c",
                    caption="Tableau d'aperçu de la distribution de nos variables"),
```

Table 1: Tableau d'aperçu de la distribution de nos variables

	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_2	0	292681.0	7458.9326215	841.00	1.636124e+04

```
              bootstrap_options = c("striped", "bordered"),
              full_width = FALSE)
```

Nous allons commenter les valeurs de ce tableau pour chaque variable :

- *narr*86 : Tout d'abord, nous remarquons que la moyenne du nombre d'arrestation en 1986 est très basse (≈ 0.40), ce qui signifie qu'en moyenne, les individus qui composent notre échantillon se sont fait arrêter 0.4 fois par la police en 1986. La valeur de l'écart-type est faible ce qui indique que la dispersion des valeurs de notre variable est bien regroupée autour de la moyenne. De plus, la médiane nous informe qu'au moins 50% des individus de notre échantillon ne s'est pas fait arrêter durant l'année 1986. Cependant, nous remarquons que le maximum est très éloigné de la moyenne, ce qui signifie qu'il y a une distribution asymétrique des données.
- *pcnv* : Cette variable prend des valeurs entre 0 et 1, il est donc logique que l'écart-type soit faible. La moyenne nous indique que sur l'ensemble de notre échantillon, 35.78% des arrestations avant 1986 ont mené à une condamnation. La médiane, elle, nous montre qu'au moins 50% des individus de notre échantillon ont une proportion d'arrestations qui ont mené à une condamnation inférieure ou égale à 25%.
- *avgsen* : Nous remarquons que la moyenne de la durée moyenne des condamnations avant 1986 est très basse (≈ 0.63). Cependant, l'étendue des données est considérablement élevée : elles vont de 0 à 59.2, ce qui signifie qu'il y a une distribution asymétrique des données et justifie le fait que l'écart-type est relativement élevé. En ce qui concerne la médiane, on peut affirmer qu'au moins 50% de notre échantillon n'ont pas purgé de peine pour leurs condamnation.
- *ptime*86 : Cette variable prend des valeurs entre 0 et 12, qui correspondent au nombre de mois passés en prison en 1986. La moyenne est d'environ 0.39 et nous avons un écart-type relativement élevé pour une moyenne très basse. Cela signifie qu'il y a une dispersion des valeurs non-négligeable autour de la moyenne. La médiane, elle, nous permet de dire qu'au moins 50% de notre échantillon n'a pas été en prison lors de l'année 1986.
- *qemp*86 : Cette variable prend des valeurs entre 0 et 4, qui correspondent au nombre de trimestres travaillés en 1986. La moyenne est d'environ 2.31, ce qui signifie qu'en moyenne, les individus de notre échantillon ont travaillés 2.31 trimestres cette année là. L'écart-type est non-négligeable et indique qu'il existe une certaine dispersion des valeurs autour de la moyenne. La médiane nous permet d'affirmer qu'au moins 50% de notre échantillon à travaillé 3 semestres ou moins en 1986.
- *inc*86 : L'étendue des données va de 0 à 541, ce qui indique qu'il existe une forte asymétrie dans la distribution. La moyenne est d'environ 54.97, ce qui signifie qu'en moyenne, notre échantillon a 5497$ de revenu disponible en 1986. L'écart-type est élevé, ce qui signifie qu'il y a une forte dispersion des valeurs autour de la moyenne. La médiane, elle, nous permet de dire qu'au moins 50% de notre échantillon a entre 0$ et 2900$ de revenu disponible en 1986.
- $inc86^2$: Cette variable est le carré de la variable *inc*86, les interprétations sont donc les mêmes (les valeurs sont juste mises au carré).

Partie B :

On dispose du modèle suivant, que nous appellerons régression 1 :

$narr86_i = \beta_0 + \beta_1 pcnv_i + \beta_2 avgsen_i + \beta_3 ptime86_i + \beta_4 qemp86_i + \beta_5 inc86_i + u_i$

Question 1 : Quel est le signe attendu des differents paramètres ?

Au premier abord de cette régression linéaire, on pourrait s'attendre à ce que :

- $\beta_0 > 0$; En effet, on peut s'attendre à un coefficient positif car le nombre d'arrestation par la police en 1986 ne peut pas être négatif, cela est trivial.
- $\beta_1 < 0$; En effet, on pourrait penser que plus un individu a une proportion importante d'arrestations qui ont mené à une condamnation, le nombre de fois où cette même personne a été arrêtée par la police en 1986 baisse car elle serait moins tentée à récidiver qu'une personne qui n'a pas été condamnée.
- $\beta_2 < 0$; En effet, on pourrait penser que plus un individu a passé de temps derrière les barreaux, moins elle sera encline à recommettre des délits.
- $\beta_3 > 0$; En effet, le fait qu'un individu ait passé un ou plusieurs mois en prison en 1986 implique fortement que cette personne ait été arrêtée par la police en 1986.
- $\beta_4 < 0$; En effet, on pourrait penser que plus une personne a travaillé en 1986, moins elle sera tentée à commettre des délits et donc à se faire arrêter par la police en 1986.
- $\beta_5 < 0$; En effet, on pourrait penser que plus une personne a un revenu disponible élevé en 1986 (c'est-à-dire plus la personne est riche), moins elle sera tentée de faire des délits et donc de se faire arrêter par la police en 1986. En revanche, une personne avec un très faible revenu disponible sera plus encline à commettre des délits (vols, braquages...) afin de satisfaire ses besoins.

Question 2 : Estimer le modèle proposé par la méthode des MCO.

Nous formulons notre première régression linéaire, que nous appellerons régression 1, grâce à la fonction **lm()** de la façon suivante :

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

La fonction **stargazer()** permet d'obtenir une table récapitulative de notre modèle. On y trouve l'estimation de la constante et des coefficients associés aux variables du modèle, leurs écarts-types, leur significativité, le nombre d'observations, la valeur du Rš et du Rš ajusté, l'erreur standard résiduelle ainsi que la F-statistique. Cette fonction nous permet d'obtenir la table suivante :

```
library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
stargazer(reg1,title="Régression 1")
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: mer., mars 27, 2024 - 19:43:58

Table 2: Régression 1

	Dependent variable:
	narr86
pcnv	-0.153^{***}
	(0.041)
avgsen	0.007
	(0.005)
ptime86	-0.035^{***}
	(0.009)
qemp86	-0.054^{***}
	(0.015)
inc86	-0.002^{***}
	(0.0003)
Constant	0.684^{***}
	(0.033)
Observations	2,725
R^2	0.050
Adjusted R^2	0.049
Residual Std. Error	0.838 (df = 2719)
F Statistic	28.840^{***} (df = 5; 2719)
Note:	*p<0.1; **p<0.05; ***p<0.01

Question 3 : Déterminer les coefficients qui sont statistiquement significatifs en rappelant la règle de décision à utiliser. Interpréter les coefficients.

La significativité statistique des coefficients est indiquée par les astérisques (*) qui se trouvent sur la droite du tableau récapitulatif des différents coefficients. Les astérisques fonctionnent de la manière suivante :

- * : Le coefficient est significatif à 10%
- ** : Le coefficient est significatif à 5%
- *** : Le coefficient est significatif à 1%

D'après les résultats de la régression linéaire, on constate que les coefficients associés aux variables *pcnv*, *ptime*86, *qemp*86, et *inc*86 sont extrêmement significatifs (à 1%). En revanche, le coefficient de la variable *avgsen* n'a pas de symbole à sa droite ; cela signifie que β_2 n'est pas significatif dans ce modèle.

Interprétation des coefficients :

- $\beta_1 \approx 0.684 \rightarrow$ Si les valeurs de toutes les variables indépendantes sont nulles, le nombre d'arrestation d'un homme en 1986 est en moyenne de 0.684.
- $\beta_1 \approx$ -0.153 $\rightarrow$ Si la proportion des arrestations avant 1986 qui ont mené à une condamnation d'un individu augmente de 1%, le nombre de fois où cette personne a été arrêtée par la police en 1986 baisse en moyenne de 0.153, toutes choses égales par ailleurs.
- $\beta_2 \approx 0.007 \rightarrow$ Si la durée moyenne de la peine pour les condamnations d'un individu avant 1986 augmente d'1 mois, le nombre de fois où cette personne a été arrêtée par la police en 1986 augmente en moyenne de 0.007, toutes choses égales par ailleurs.
- $\beta_3 \approx$ -0.035 $\rightarrow$ Si le nombre de mois passés en prison d'un individu en 1986 augmente d'1 mois, le nombre de fois où cette personne a été arrêtée par la police en 1986 baisse en moyenne de 0.035, toutes choses égales par ailleurs.
- $\beta_4 \approx$ -0.054 $\rightarrow$ Si le nombre de trimestres travaillés d'un individu en 1986 augmente d'1 trimestre, le nombre de fois où cette personne a été arrêtée par la police en 1986 baisse en moyenne de 0.054, toutes choses égales par ailleurs.
- $\beta_5 \approx$ -0.002 $\rightarrow$ Si le revenu disponible d'un individu en 1986 augmente de 100€, le nombre de fois où cette personne a été arrêtée par la police en 1986 baisse en moyenne de 0.002, toutes choses égales par ailleurs.

Question 4 : Nous ajoutons le carré du revenu disponible en 1986 à notre modèle. Cet ajoût dans la régression est-il nécessaire ? Pourquoi ? Commenter la relation entre le revenu de la personne en 1986 et le nombre de fois où la personne a été arrêtée.

Nous avons désormais le modèle suivant, que nous appellerons régression 2 :

$narr86_i = \beta_0 + \beta_1 pcnv_i + \beta_2 avgsen_i + \beta_3 ptime86_i + \beta_4 qemp86_i + \beta_5 inc86_i + \beta_6 inc86_i^2 + u_i$

Nous la formulons dans R de la façon suivante :

```
reg2 <- lm(narr86 ~ pcnv + avgsen + ptime86 + qemp86 + inc86 + inc86_2, df)
```

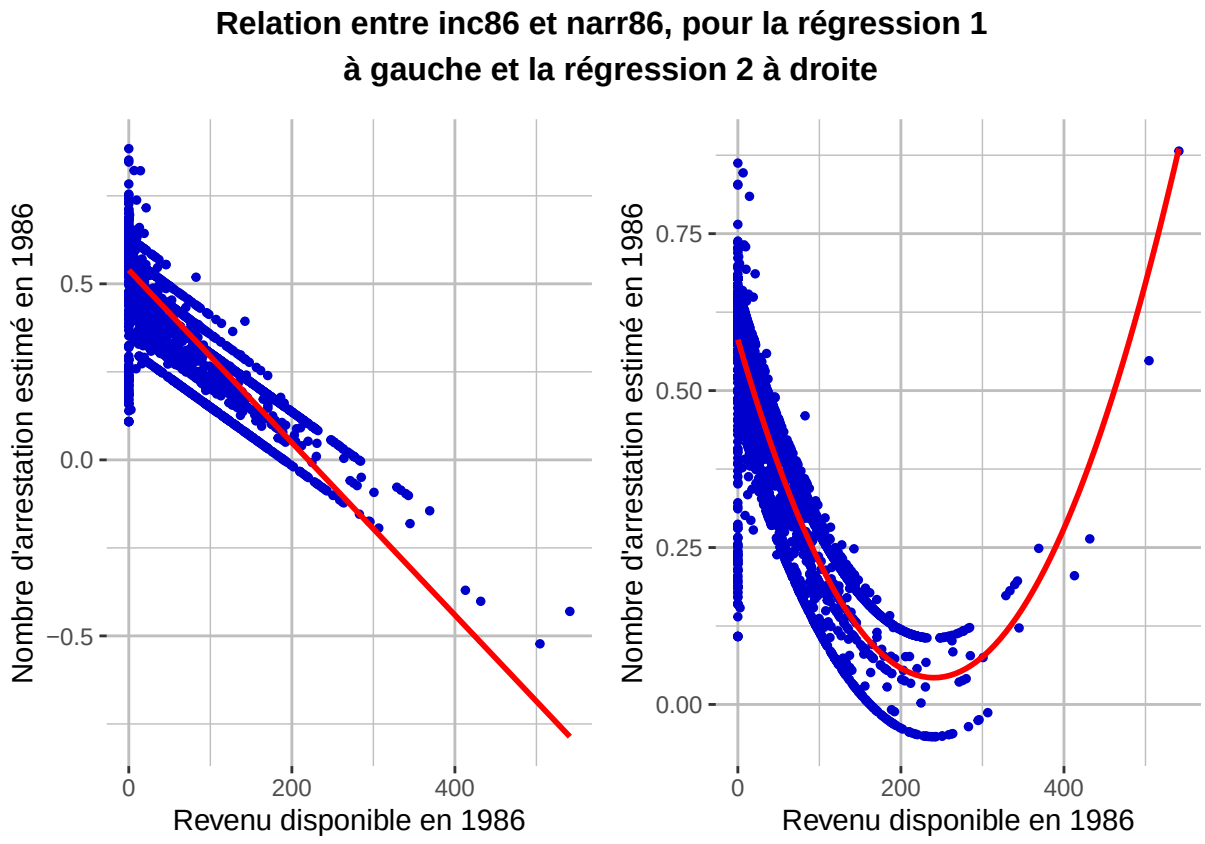
L'ajout du carré d'une variable au sein d'une régression linéaire multiple permet de modéliser des relations non-linéaires entre cette variable et la variable dépendante. Ici, cela permet un effet de courbure entre la variable *inc*86 (le revenu disponible en 1986) et la variable *narr*86 (le nombre d'arrestations en 1986). En effet, au vu du graphique ci-dessous, la relation entre le revenu disponible en 1986 et le nombre d'arrestations en 1986 n'est pas linéaire :

```
pred_reg1=predict(reg1,df)
pred_reg2=predict(reg2,df)
```

```r
library(patchwork)
library(ggplot2)
library(grid)
library(gridExtra)
library(ggpubr)

graphreg1 <- ggplot(df, aes(x=inc86, y=pred_reg1))+
  geom_point(size=1, color="blue3")+
  theme(panel.background = element_rect(fill="transparent"),
        panel.grid.major = element_line(color="grey"),
        panel.grid.minor = element_line(color="grey"))+
  xlab("Revenu disponible en 1986")+
  ylab("Nombre d'arrestation estimé en 1986")+
  geom_smooth(method=lm, color="red", se=FALSE)

graphreg2 <- ggplot(df, aes(x=inc86, y=pred_reg2))+
  geom_point(size=1, color="blue3")+
  theme(panel.background = element_rect(fill="transparent"),
        panel.grid.major = element_line(color="grey"),
        panel.grid.minor = element_line(color="grey"))+
  xlab("Revenu disponible en 1986")+
  ylab("Nombre d'arrestation estimé en 1986")+
  geom_smooth(method=lm, color="red", se=FALSE, formula=y~poly(x,2))

g1g2 <- ggarrange(graphreg1+graphreg2)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```r
titre <- "Relation entre inc86 et narr86, pour la régression 1 \n à gauche et la régression 2 à droite"
Titre <- textGrob(titre, gp = gpar(fontface = "bold"))
grid.arrange(g1g2, top=Titre)
```



On remarque qu'aux extrêmes du revenu disponible (c'est-à-dire si la personne a un très faible ou un très élevé revenu disponible), le nombre d'arrestation a tendance à augmenter. On peut penser que ce n'est pas pour les mêmes délits : une personne à faible revenu sera concernée par des vols ou des braquages, alors qu'une personne avec un revenu élevé sera plutôt arrêtée pour fraude fiscale.

Les personnes avec un revenu moyen, elles, ont largement moins tendance à se faire arrêter, et donc à commettre des délits.

Question 5 : Commenter les résultats après avoir interprété chaque coefficient.

La table récapitulative du modèle de la régression 2 est la suivante :

```
stargazer(reg2,title="Régression 2")
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: mer., mars 27, 2024 - 19:44:01

Table 3: Régression 2

	Dependent variable:
	narr86
pcnv	-0.157^{***}
	(0.041)
avgsen	0.006
	(0.005)
ptime86	-0.034^{***}
	(0.009)
qemp86	-0.021
	(0.018)
inc86	-0.004^{***}
	(0.001)
inc86_2	0.00001^{***}
	(0.00000)
Constant	0.679^{***}
	(0.033)
Observations	2,725
R^2	0.054
Adjusted R^2	0.052
Residual Std. Error	0.836 (df = 2718)
F Statistic	25.894^{***} (df = 6; 2718)
Note:	*p<0.1; **p<0.05; ***p<0.01

Interprétation des coefficients :

- $\beta_1 \approx$ -0.157 → Si la proportion des arrestations avant 1986 qui ont mené à une condamnation d'un individu augmente de 1%, le nombre de fois où cette personne a été arrêtée par la police en 1986 baisse en moyenne de 0.157, toutes choses égales par ailleurs.
- $\beta_2 \approx 0.006$ → Si la durée moyenne de la peine pour les condamnations d'un individu avant 1986 augmente d'1 mois, le nombre de fois où cette personne a été arrêtée par la police en 1986 augmente en moyenne de 0.006, toutes choses égales par ailleurs.
- $\beta_3 \approx$ -0.034 → Si le nombre de mois passés en prison d'un individu en 1986 augmente d'1 mois, le nombre de

fois où cette personne a été arrêtée par la police en 1986 baisse en moyenne de 0.034, toutes choses égales par ailleurs.
- $\beta_4 \approx$ -0.021 $\rightarrow$ Si le nombre de trimestres travaillés d'un individu en 1986 augmente d'1 trimestre, le nombre de fois où cette personne a été arrêtée par la police en 1986 baisse en moyenne de 0.021, toutes choses égales par ailleurs.
- $\beta_5 \approx$ -0.004 $\rightarrow$ Si le revenu disponible d'un individu en 1986 augmente de 100€, le nombre de fois où cette personne a été arrêtée par la police en 1986 baisse en moyenne de 0.004, toutes choses égales par ailleurs.
- $\beta_6 \approx 0.00001 \rightarrow$ Ce coefficient indique la relation quadratique entre $inc86$ et $narr86$. $\beta_6 > 0$, cela signifie que le nombre d'arrestation en 1986 augmente à un taux de plus en plus rapide à mesure que le revenu disponible augmente. Cela justifie bien la relation en "U" vue précédemment.

On constate que les coefficients des variables $pcnv$, $inc86$ et de $inc86^2$ sont significatifs au seuil d'1%. En revanche, les coefficients associés à $avgsen$ et à $qemp86$ ne sont pas significatifs dans ce modèle. On peut noter que l'ajout de la variable $inc86^2$ a diminué l'effet du nombre de trimestre à travailler en 1986 sur le nombre d'arrestation par la police ; son coefficient est passé de -0.054 à -0.021.

La F-statistique a légèrement baissée, passant de 28.840 à 25.894, mais le modèle reste globalement significatif au seuil d'1%.

Du côté du R^2 et du R^2 ajusté, ils ont légèrement augmenté par rapport à la régression 1, mais la qualité de la régression reste cependant très faible (R^2 ajusté = 0.052) ; le modèle de régression ne parvient pas à expliquer une grande part de la variance de $narr86$.

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

En économétrie, l'hétéroscédasticité se réfère à une situation où la variance des résidus d'un modèle de régression n'est pas constante pour toutes les valeurs des variables indépendantes. Afin de vérifier cela au sein de notre modèle, nous allons réaliser un test de Breusch-Pagan dont les hypothèses sont les suivantes :

$$\begin{cases} H_0 : \text{La variance des } u_i \text{ est constante} \\ H_1 : \text{La variance des } u_i \text{ n'est pas constante} \end{cases}$$

```
require(lmtest)
```

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

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

```
##
## 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 du test de Breusch-Pagan est extrêmement faible. Cela indique une forte évidence contre l'hypothèse nulle selon laquelle il n'y a pas d'hétéroscédasticité dans ce modèle. Après ce test, nous pouvons donc conclure qu'il y a une forte hétéroscédasticité au sein de la régression 2.

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 ?

Afin de corriger l'hétéroscédasticité au sein de notre modèle, nous pouvons calculer les écarts-types robustes.

```
library("lmtest")
library("sandwich")
stargazer(coeftest(reg2, vcov = vcovHC(reg2, type="HC0")),
          title="Régression 2 avec écarts-types robustes")
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: mer., mars 27, 2024 - 19:44:01

Table 4: Régression 2 avec écarts-types robustes

	Dependent variable:
pcnv	-0.157^{***}
	(0.034)
avgsen	0.006
	(0.005)
ptime86	-0.034^{***}
	(0.006)
qemp86	-0.021
	(0.018)
inc86	-0.004^{***}
	(0.001)
inc86_2	0.00001^{***}
	(0.00000)
Constant	0.679^{***}
	(0.041)
Note:	*p<0.1; **p<0.05; ***p<0.01

On constate de légers changements dans les écarts-types de certaines variables :

- Celui de *pncv* passe de 0.041 à 0.034
- Celui de *ptime*86 passe de 0.009 à 0.006
- Celui de la constante β_0 passe de 0.033 à 0.041

Cependant, ces changements sont mineurs. On ne peut pas dire que l'on a réglé le problème d'hétéroscédasticité au sein de notre modèle.

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.

Afin de construire cette indicatrice, nous entrons les commandes suivantes :

```
df$pcnv <- ifelse(df$pcnv > 0, 1, 0)
names(df)[names(df) == "pcnv"] <- "condamne"
```

Voici un aperçu de l'indicatrice *condamne* :

```
str(df$condamne)
```

```
##  num [1:2725] 1 1 1 1 0 1 1 1 1 1 ...
```

Question 2 : Estimer une nouvelle régression, en remplaçant la variable *pcnv* par l'indicatrice *condamne*.

Nous cherchons donc à estimer le modèle suivant, que nous appellerons régression 3 :

$narr86_i = \beta_0 + \beta_1 condamne_i + \beta_2 avgsen_i + \beta_3 ptime86_i + \beta_4 qemp86_i + \beta_5 inc86_i + \beta_6 inc86_i^2 + u_i$

Nous formulons la régression 3 de la manière suivante :

```
attach(df)
reg3 <- lm(narr86 ~ condamne + avgsen + ptime86 + qemp86 + inc86 + inc86_2, df)
```

Nous obtenons la table récapitulative qui suit :

```
stargazer(reg3,title="Régression 3")
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: mer., mars 27, 2024 - 19:44:01

Table 5: Régression 3

	Dependent variable:
	narr86
condamne	-0.004
	(0.033)
avgsen	0.006
	(0.005)
ptime86	-0.036^{***}
	(0.009)
qemp86	-0.023
	(0.018)
inc86	-0.004^{***}
	(0.001)
inc86_2	0.00001^{***}
	(0.00000)
Constant	0.628^{***}
	(0.036)
Observations	2,725
R^2	0.049
Adjusted R^2	0.047
Residual Std. Error	0.839 (df = 2718)
F Statistic	23.287^{***} (df = 6; 2718)
Note:	*p<0.1; **p<0.05; ***p<0.01

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.

Regardons tout d'abord la distribution du nombre d'arrestations en 1986 de notre échantillon, en comparant les personnes qui ont déjà été condamnées avant 1986 et celles qui ne l'ont pas été. Pour ce faire, nous avons réalisé des diagrammes en boîte ainsi qu'un tableau d'effectif.

```
par(mfrow=c(1,2))

boxplot(df$narr86[df$condamne==0],
        range=0, lty=1, col="green2",
        main="Personnes non condamnées \navant 1986",
        ylab= "Nombre d'arrestations en 1986")
legend("topright", legend = c("Effectif: 1260","Moyenne: 0.389",
                              "Médiane: 0","3è Quartile: 1",
                              "Min: 0","Max: 6"),cex=0.6,bty="n")

boxplot(df$narr86[df$condamne==1],
        range=0, lty=1, col="orange",
        main="Personnes déjà condamnées \navant 1986",
        ylab= "Nombre d'arrestations en 1986")
legend("topright", legend = c("Effectif: 1465","Moyenne: 0.418",
                              "Médiane: 0","3è Quartile: 0",
                              "Min: 0","Max: 12"),cex=0.6,bty="n")
```

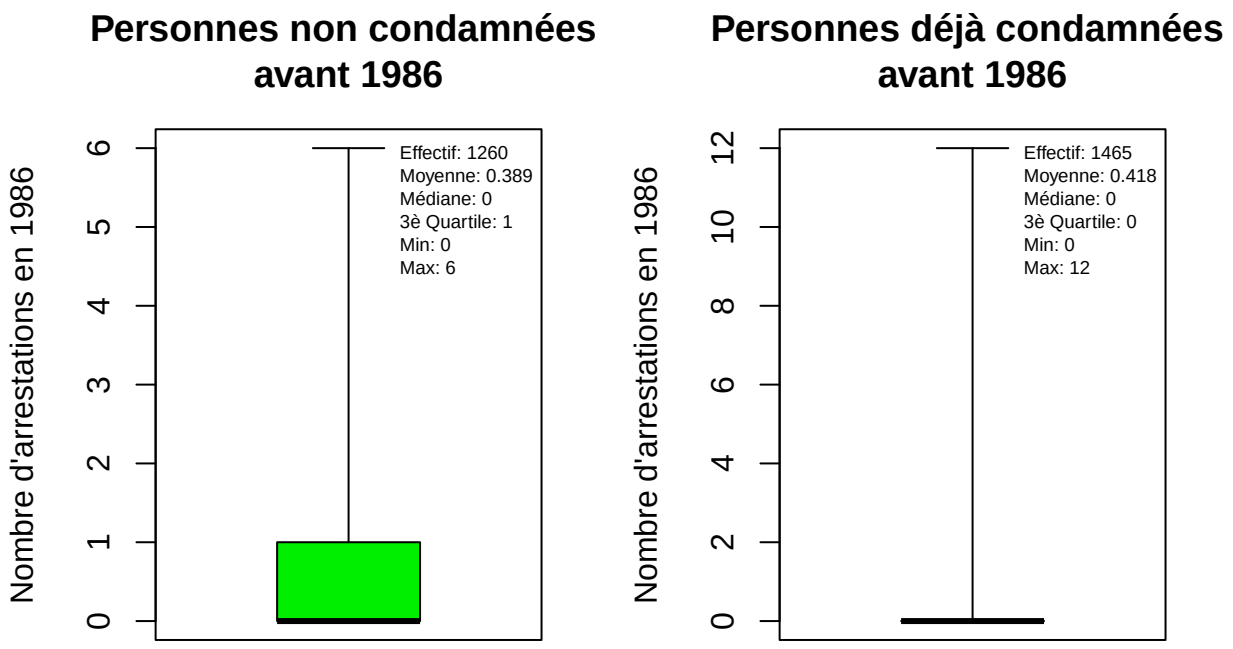


Table 6: Nombre d'arrestations en 1986 selon le passé judiciaire de notre échantillon

	0	1	2	3	4	5	6	7	9	10	12	Total
Jamais condamné	867	331	46	6	3	5	2	0	0	0	0	1260
Déjà condamné	1103	228	75	36	9	8	2	1	1	1	1	1465
Total	1970	559	121	42	12	13	4	1	1	1	1	2725

```
Table<-function(x,y){
  z<-table(x,y) %>%
        addmargins()
  return(z)
}
table_narr_condamne <- Table(df$condamne,df$narr86)
colnames(table_narr_condamne)[12] <- "Total"
row.names(table_narr_condamne)[1] <- "Jamais condamné"
row.names(table_narr_condamne)[2] <- "Déjà condamné"
row.names(table_narr_condamne)[3] <- "Total"

kable_styling(
  kable(
    table_narr_condamne,
    align = "c",
    caption = "Nombre d'arrestations en 1986 selon le passé judiciaire de notre échantillon"),
  full_width = FALSE,
  position = "center",
  bootstrap_options = c("striped", "bordered"),)
```

Au premier abord, nous ne décelons pas de tendance particulière, si ce n'est que pour nos deux groupe, la grande majorité ne s'est pas faite arrêtée par la police en 1986. De plus, la distribution est relativement similaire entre nos deux groupes pour un nombre d'arrestation en 1986 allant de 0 à 2, mais on remarque une certaine "cassure" à partir de 3 arrestations ou plus. En effet, il semble que le groupe ayant déjà été condamné avant 1986 a tendance à se faire arrêter plus de fois que le groupe n'ayant pas de casier judiciaire. Cependant, les effectifs sont extrêmement faible comparé à l'effectif total ; ces personnes peuvent être considérées comme "déviantes", et nous ne pouvons donc pas conclure grand chose.

Pour ce qui est de l'interprétation du coefficient de la variable *condamne* de notre modèle de régression 3, nous pouvons dire qu'une personne ayant déjà été condamnée avant 1986 a en moyenne 0.04 arrestation par la police de moins en 1986 qu'une personne n'ayant jamais été condamnée, toutes choses égales par ailleurs.

Cette différence est négligeable au sein de notre régression. Le coefficient associé à la variable *condamne* n'est pas significatif. En effet, le fait qu'une personne ait été ou non condamnée avant 1986 ne semble pas réellement influencer son nombre d'arrestation cette année là.

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. Écrire la spécification du modèle.

Nous allons rajouter une intéraction entre les variables *qemp*86 et *condamne* au sein de notre modèle. La régression 4 est un modèle de la forme suivante :

$narr86_i = \beta_0 + \beta_1 condamne_i + \beta_2 avgsen_i + \beta_3 ptime86_i + \beta_4 qemp86_i + \beta_5 inc86_i + \beta_6 inc86_i^2 + \beta_7 (qemp86_i * condamne_i) + u_i$

```
attach(df)
```

```
## Les objets suivants sont masqués depuis df (pos = 3):
##
##     avgsen, condamne, inc86, inc86_2, narr86, ptime86, qemp86
```

```
reg4 <- lm(narr86 ~ condamne + avgsen + ptime86 + qemp86 + inc86 + inc86_2 + qemp86*condamne, df)
```

Question 5 : Estimer le modèle.

Afin d'estimer le modèle, nous allons une nouvelle fois utiliser la fonction **stargazer()** afin de générer la table récapitulative de notre régression 4.

```
stargazer(reg4, title = "Régression 4")
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: mer., mars 27, 2024 - 19:44:01

Table 7: Régression 4

	Dependent variable:
	narr86
condamne	0.203^{***}
	(0.059)
avgsen	0.006
	(0.005)
ptime86	-0.043^{***}
	(0.009)
qemp86	0.026
	(0.021)
inc86	-0.004^{***}
	(0.001)
inc86_2	0.00001^{***}
	(0.00000)
condamne:qemp86	-0.087^{***}
	(0.021)
Constant	0.511^{***}
	(0.045)
Observations	2,725
R^2	0.055
Adjusted R^2	0.053
Residual Std. Error	0.836 (df = 2717)
F Statistic	22.600^{***} (df = 7; 2717)
Note:	*p<0.1; **p<0.05; ***p<0.01

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 ne l'a jamais été ? Écrivez la formule et le calcul. La différence est-elle significative ?

Effet moyen d'une augmentation du nombre de trimestre travaillés pour une personne déjà condamnée sur le nombre d'arrestations en 1986, toutes choses égales par ailleurs :

Le calcul à effectuer est le suivant :

$$\begin{aligned}
&E(narr86|condamne = 1, \Delta qemp86 = +1) - E(narr86|condamne = 1) \\
&= \beta_0 + \beta_1 + \beta_2 avgsen + \beta_3 ptime86 + \beta_4(qemp86 + 1) + \beta_5 inc86 + \beta_6 inc86^2 + \beta_7(qemp86 + 1) - (\beta_0 + \beta_1 + \beta_2 avgsen + \\
&\beta_3 ptime86 + \beta_4 qemp86 + \beta_5 inc86 + \beta_6 inc86^2 + \beta_7 qemp86) \\
&= \beta_4 + \beta_7 \\
&= 0.026 - 0.087 \\
&= -0.061
\end{aligned}$$

Une augmentation d'un trimestre travaillé en plus pour une personne déjà condamné provoque en moyenne une baisse de 0.061 du nombre d'arrestation en 1986, toutes choses égales par ailleurs.

Effet moyen d'une augmentation du nombre de trimestre travaillés pour une personne n'ayant jamais été condamnée sur le nombre d'arrestations en 1986, toutes choses égales par ailleurs :

Le calcul à effectuer est le suivant :

$$\begin{aligned}
&E(narr86|condamne = 0, \Delta qemp86 = +1) - E(narr86|condamne = 0) \\
&= \beta_0 + \beta_2 avgsen + \beta_3 ptime86 + \beta_4(qemp86 + 1) + \beta_5 inc86 + \beta_6 inc86^2 - (\beta_0 + \beta_2 avgsen + \beta_3 ptime86 + \beta_4 qemp86 + \\
&\beta_5 inc86 + \beta_6 inc86^2) \\
&= \beta_4 \\
&= 0.026
\end{aligned}$$

Une augmentation d'un trimestre travaillé en plus pour une personne n'ayant jamais été condamnée provoque en moyenne une augmentation de 0.026 du nombre d'arrestation en 1986, toutes choses égales par ailleurs.

Le coefficient associé à l'intéraction entre *qemp*86 et *condamne* est significatif à 1%. Les effets moyens d'une augmentation du nombre de trimestre travaillés chez nos deux groupes sont très différents : pour le groupe ayant déjà été condamné, cette augmentation provoque une baisse du nombre d'arrestation en 1986, alors que pour le groupe n'ayant pas été condamné, cela provoque une augmentation du nombre d'arrestation. La différence est donc significative.

Partie D :

Question 1 : Faire un résumé de votre analyse. Indiquer les faiblesses selon vous de l'analyse à laquelle vous avez abouti (vous pouvez par exemple parler de l'échantillon, du modèle, d'éventuels problèmes d'endogénéité. . .)

Il est difficile de résumer notre analyse. Les quatre régressions que nous avons faites ne semblaient pas bien retranscrire la part de la variance de notre variable dépendante *narr*86 expliquée par la variance de nos variables explicatives (R^2 ajusté aux alentours de 5% pour chacune de nos régressions). Nous pouvons donc affirmer que le nombre d'arrestation par la police en 1986 ne semble pas dépendre des variables que nous avons utilisées, ou du moins il en existe d'autres qui expliqueraient bien mieux sa variation.

Nous pensons que l'échantillon du modèle était problématique, pour deux raisons :

- Sa taille : nous ne disposions seulement d'un effectif de 2725 individus. Pour pouvoir expliquer un sujet aussi complexe que la probabilité de se faire arrêter par la police, il nous faudrait un échantillon bien plus large.
- Sa spécificité : les 2725 observations issues de la base de données étaient sélectionnées de manière très précise : seulement des hommes, nés en Californie entre 1960 et 1961, et ayant tous déjà été arrêté au moins une fois avant 1986. Les individus sur lesquels nous avons estimé nos modèles ne sont pas représentatifs de la globalité de la population. De plus, le fait qu'ils aient tous déjà été arrêtés au moins une fois par la police avant 1986 peut causer une surestimation du nombre d'arrestation en 1986, du fait que ce sont des gens "à problème".

En plus de cela, les variables qui composent notre base de données ne semblent pas être adaptées pour expliquer le nombre d'arrestations en 1986. Nous pouvons noter qu'il y a sans doute un problème d'endogénéité entre les variables *ptime*86 et *qemp*86 : le temps passé en prison en 1986 est lié au nombre de trimestre travaillé cette même année, car une personne emprisonnée ne peut pas travailler.

Enfin, des variables pourtant essentielles pour expliquer notre problématique ne sont pas spécifiées dans la base de données et n'ont donc pas pu être inclues dans nos modèles. On peut par exemple citer l'ethnie, le quartier dans lequel l'individu réside, s'il y a des cas de délinquance dans sa famille ou dans son entourage, si la personne a grandit avec ses deux parents. . .