Projet

Corre, Lenoir

Table des matières

1 Importation des données

Table 1 – Nombre de valeurs manquantes

	NA
Data SANTE	0

Nous allons travailler sur une base de données de 200 individus sur 5 années ce qui nous fait 1000 observations. Nous avons également 6 variables et aucunes données manquantes. Tout d'abord nous allons determiner la structure de panel :

```
## Balanced Panel: n = 200, T = 5, N = 1000
```

On a bien un panel cylindré de 200 individus observés sur 5 périodes. Nous avons donc un T faible et un nombre d'individus assez important ce qui signifie que l'on est en présence d'un panel court et cylindré.

Ensuite, voici une sommaire des vairables quantitatives.

Table 2 – Sommaire des variables quantitatives

	DEPSANTE	REV	AGE
Min	0.00	22.00	21.00
Max	13.00	146.00	70.00
Moyenne	4.86	73.57	46.42
Mediane	5.00	74.00	47.00
Ecart.type	2.84	18.60	13.76

On voit que pour chaque variables, la médiane et la moyenne sont assez proche :

- Pour les dépenses de santé, on peut observer une moyenne autour de 500$ par an.
- Pour le revenu, elle est de l'ordre de 74 000$ par an
- Et enfin, pour l'âge, elle se situe vers 47 ans.

Enfin, le modèle que nous allons étudié est le suivant :

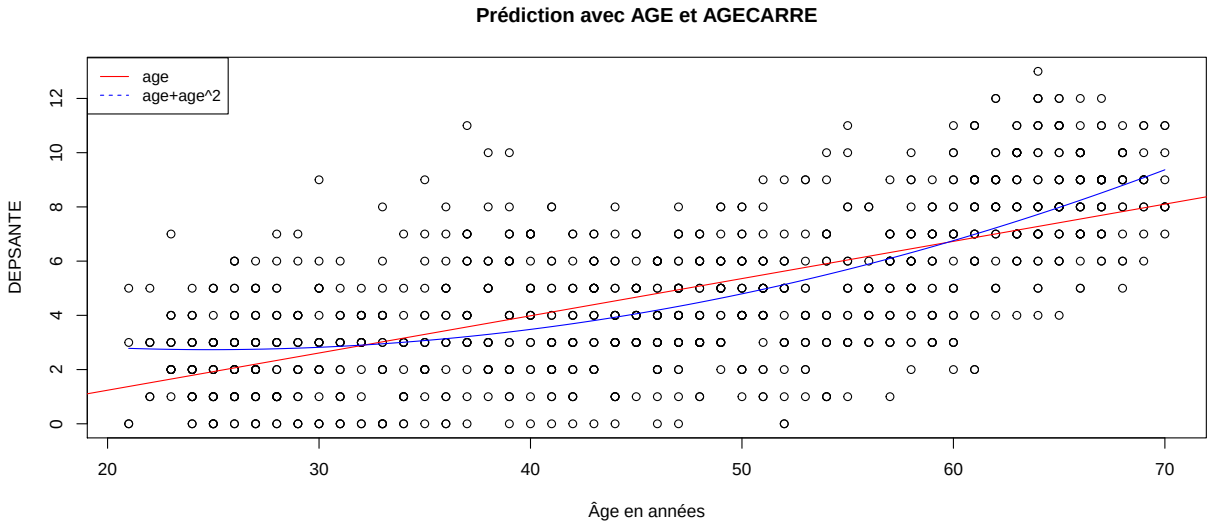
$$DEPSANTE_{it} = \beta_{1i} + \beta_2 ln(REV_{it}) + \beta_3 AGE_{it} + \beta_4 AGE_{it}^2 + \beta_5 ASSU_{it} + \varepsilon_{it}$$

2 Question 1

Il est plus intéressant d'utiliser le logarithme du revenu car le revenu est donné en milliers de dollars et les dépenses de santé en centaines de dollars. En utilisant le log on corrige cela, en quelque sorte on "lisse" les données. Cela permet que les valeurs extrêmes n'impactent pas trop la régression.

Nous aurons donc un modèle Niveau-log ce qui veut dire que, toutes choses égales par ailleurs, si le revenu augmente de 1%, une approximation de l'effet marginal du revenu sur les dépenses liées à la santé est de $\frac{\beta_2}{100}$%.

3 Question 2



L'interêt de mettre également la variable AGE au carré est de prendre en compte la non linéarité. Si l'on compare les deux courbes du graphique précèdent, on peut voir que si l'on prend en compte l'âge au carré, la régression semble beaucoup mieux représenter les données entre la variable explicative et la variable expliqué.

Mettre l'âge au carré va nous permettre d'augmenter le R^2 et donc la qualité de notre régression.La variable "ageˆ2" nous montre comment évolue les dépenses de santé par rapport à la convexité de la variable âge. Pour pouvoir interpréter le coefficient associé à cette variable, il faut que celui-ci soit significatif ($p<0.05$). Si celui-ci est positif, l'âge au carré décrit une courbe convexe, si il est négatif la courbe sera concave. Pour déterminer le minimum ou le maximum de la fonction age^2, il faut faire le calcul suivant:

$$\frac{-\beta_1}{2\beta_2}$$

où β_1 est le coefficient associé à l'âge et β_2 le coefficient associé à l'âgeˆ2.

4 Question 3

La variable ASSU est mise en variable dummy. Donc l'écriture de ce modèle varie selon cette variable :

Si $ASSU = 0$ alors :

$$DEPSANTE_{it} = \beta_{1i} + \beta_2 ln(REV_{it}) + \beta_3 AGE_{it} + \beta_4 AGE_{it}^2 + \varepsilon_{it}$$

Si $ASSU = 1$ alors :

$$DEPSANTE_{it} = \beta_{1i} + \beta_{5i} + \beta_2 ln(REV_{it}) + \beta_3 AGE_{it} + \beta_4 AGE_{it}^2 + +\varepsilon_{it}$$

4.1 MCO

Voici les résultats pour l'éstimation par les MCO :

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                DEPSANTE
## ------------------------------------------------
## lrev                            0.392*
##                                 (0.220)
##
## AGE                            -0.208***
##                                 (0.033)
##
## agecarre                       0.004***
##                                (0.0004)
##
## ASSU1                          1.517***
##                                 (0.123)
##
## Constant                       3.439***
##                                 (1.221)
##
## ------------------------------------------------
## Observations                     1,000
## R2                               0.550
## Adjusted R2                      0.548
## Residual Std. Error       1.910 (df = 995)
## F Statistic           303.859*** (df = 4; 995)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Ici, on peut voir que le logarithme du revenu n'est pas significatif au seuil 5%. Le coefficient est positif donc si le revenu augmente de 1%, les dépenses de santé augmentent de $\frac{0.392}{100} = 0,04\%$

On peut constater que le coefficient associé à l'âge est négatif et celui de l'âge au carré positif donc nous constatons que l'âge au carré est convexe. Donc les dépenses de santé diminuent jusqu'à son minimum situé en $\frac{-\beta_3}{2\beta_4} = \frac{0.208}{0.008} = 26$. A partir de cet âge, les dépenses de santé augmentent plus une personne vieillit.

On constate que le coefficient associé à l'assurance maladie est positif, ce qui signifie que si l'individu est assuré ses dépenses de santé vont augmenter. En effet, la constante est de $\beta_1 + \beta_5 = 4,956$ si l'individu est assuré contre $\beta_1 = 3,439$ s'il ne l'est pas.

4.2 Effets fixes individuels

Le modèle à estimer est :

$$DEPSANTE_{it} - \overline{DEPSANTE}_i = \beta_2(ln(Rev_{it}) - \overline{ln(Rev_i)}) + \beta_3(AGE_{it} - \overline{AGE}_i) + \beta_4(AGE_{it}^2 - \overline{AGE^2}_i) +$$

$$\beta_5(ASSU_{it} - \overline{ASSU}_i) + (\varepsilon_{it} - \bar{\varepsilon}_i) \quad i = 1, ..., N; \quad t = 1, ..., T$$

Testons si il y a présence d'autocorrélation et d'héteroscédasticité dans le modèle :

```
##
##  Wooldridge's test for serial correlation in FE panels
##
```

```
## data:  within
## F = 2.3665, df1 = 1, df2 = 798, p-value = 0.1244
## alternative hypothesis: serial correlation

##
##  studentized Breusch-Pagan test
##
## data:  within
## BP = 12.978, df = 4, p-value = 0.01138
```

Pour l'autocorrélation, la p-value est > 0.05 donc on rejette l'hypothèse alternative d' autocorrélation. Pour l'hétéroscédasticité, la p-value est < 0.05 donc on conserve l'hypothèse de présence d'hétéroscédaticité.

Nous allons donc corriger l'hétéroscédasticité et présenter les résultats en utilisant un estimateur à effets fixes individuels :

```
##
## ================================================
##                       Dependent variable:
##               ---------------------------------
##                      DEPSANTE
##                        panel          coefficient
##                       linear             test
##                         (1)               (2)
## ------------------------------------------------
## lrev                  -0.105            -0.105
##                       (0.754)           (0.741)
##
## AGE                    0.065             0.065
##                       (0.094)           (0.092)
##
## I(AGE2)               0.0003            0.0003
##                       (0.001)           (0.001)
##
## ASSU1                1.351***          1.351***
##                       (0.114)           (0.100)
##
## ------------------------------------------------
## Observations           1,000
## R2                     0.166
## Adjusted R2           -0.046
## F Statistic  39.678*** (df = 4; 796)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

En corrigeant le modèle, les coefficients ne changent pas, seuls les écarts-type se modifient.

Dans ce modèle, le logarithme du revenu n'est pas significatif et son coefficient est négatif. En ce qui concerne l'âge, les deux coefficients sont positifs donc si l'âge augmente, les dépenses augmentent de plus en plus. Toutefois ces variables ne sont pas significatives, même au seuil 10%.

On peut en conclure également que les personnes ayant une assurance maladie dépensent plus dans la santé qu'une personne qui n'ayant pas d'assurance maladie. De plus cette variable est très significative.

4.3 Effets aléatoires

Le modèle à estimer s'écrit :

$$DEPSANTE_{mcqgit} = \beta_1 + \beta_2 ln(Rev_{mcqgit}) + \beta_3 AGE_{mcqgit} + \beta_4 AGE^2_{mcqgit} + \beta_5 ASSU_{mcqgit} + v_{it} \quad i = 1, ..., N; \quad t = 1, ..., T$$

$DEPSANTE_{mcqgit} = DEPSANTE_{it} - \hat{\theta}\overline{DEPSANTE}_i$;

$AGE_{mcqgit} = AGE_{it} - \hat{\theta}\overline{AGE}_i$;

$AGE^2_{mcqgit} = AGE^2_{it} - \hat{\theta}\overline{AGE^2}_i$;

$ln(Rev_{mcqgit}) = ln(Rev_{it}) - \hat{\theta}\overline{ln(Rev}_{3i})$;

$ASSU_{mcqgit} = ASSU_{it} - \hat{\theta}\overline{ASSU}_i$

avec

$\hat{\theta} = 1 - \sqrt{\frac{\sigma^2_\varepsilon}{(\sigma^2_\varepsilon + T\sigma^2_\mu)}}$.

Nous testons tout d'abord la présence d'hétéroscédasticité :

```
##
##  studentized Breusch-Pagan test
##
## data:  random
## BP = 12.978, df = 4, p-value = 0.01138
```

Nous sommes en présence d'hétéroscédasticité (p-value < 0.05), nous allons donc corriger cela et présenter les résultats pour un estimateur à effets aléatoires :

```
##
## ========================================
##                  Dependent variable:
##              ---------------------------
##                     DEPSANTE
##                      panel       coefficient
##                      linear         test
##                        (1)            (2)
## ----------------------------------------
## lrev                  -0.149        -0.149
##                      (0.292)        (0.268)
##
## AGE                  -0.090*       -0.090**
##                      (0.051)        (0.044)
##
## I(AGE2)              0.002***      0.002***
##                      (0.001)       (0.0005)
##
## ASSU1                1.362***      1.362***
##                      (0.107)        (0.097)
##
## Constant             3.443**        3.443**
##                      (1.642)        (1.454)
##
## ----------------------------------------
## Observations          1,000
## R2                    0.308
## Adjusted R2           0.305
```

```
## F Statistic    441.870***
## ========================================
## Note:         *p<0.1; **p<0.05; ***p<0.01
```

Le logarithme du revenu n'est pas significatif même au seuil 10%. On remarque que les coefficients sont du même signe que dans notre MCO, celui de l'âge est négatif, l'âge au carré positif. Donc les dépenses de santé diminuent jusqu'à son minimum situé en $\frac{-\beta_3}{2\beta_4} = \frac{0.090}{0.004} = 22,5$. A partir de cet âge, les dépenses de santé augmentent plus une personne vieillit.

De plus, le coefficient associé à l'assurance maladie est positif donc les dépenses augmentent si les individus possède cette assurance. Grâce à la correction de l'hétéroscédasticité, le coefficient de l'âge devient significatif à 5%.

5 Question 4

Pour tester l'hypothèse des termes constants, il faut utiliser le modèle MCO et le modèle à effets fixes individuels.

L'hypothèse H_0 est donc

$$H_0 : \alpha_i = \alpha \quad \forall_i \in [1; N]$$

Pour ceci, nous allons utiliser la statistique de Fisher :

$$F = \frac{(SCR_{c1} - SCR_{c2})/(N-1)}{SCR_{c2}/N(T-1) - K} \sim F[(N-1), N(T-1) - K]$$

où $N - 1 = 199$ et $N(T-1) - K = 796$

Calculons les sommes des carrées des résidus :

- SCRc1 : la somme des carrés des résidus du modèle contraint en utilisant MCO sur les données empilés
- SCRc2 : somme des carrés des résidus du modèle non contraint en utilisant l'estmiateur à effets individuels (within).

Table 3 – SCR en fonction du modèle

	MCO	Within
SCR	3630.26	845.4305

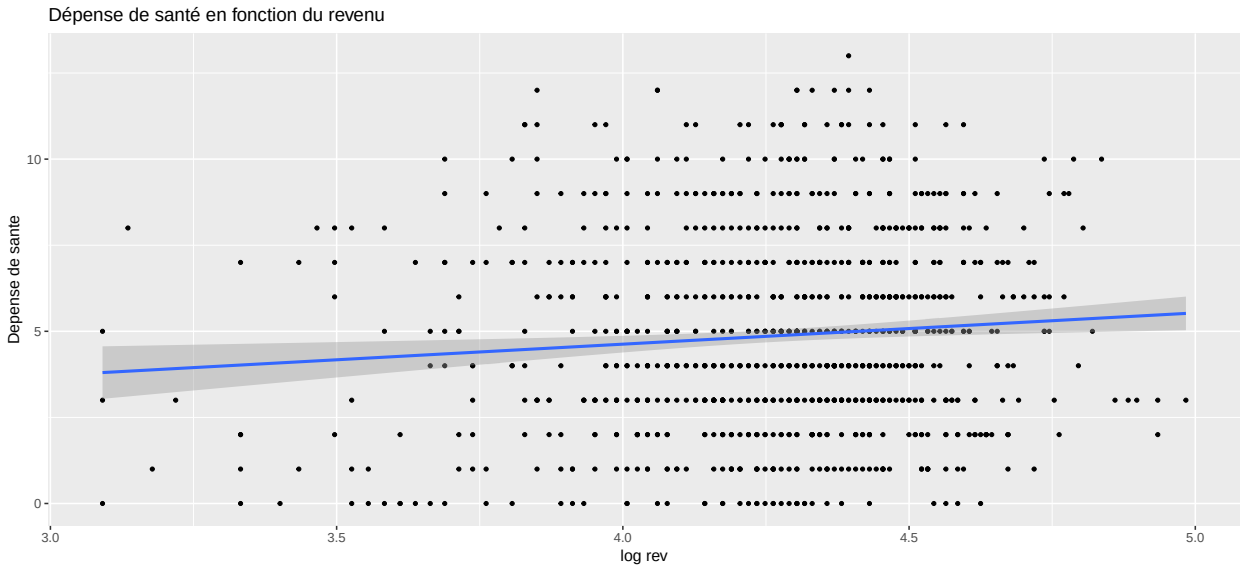
Calculons la statistique de Fisher et comparons avec celle de la table de loi :

Table 4 – Statistique de Fisher

	Fobs
Statistique	13.17591

$F_{tab} = F(199, 796) = 1, 2$. Donc $F_{obs} > F_{tab}$ nous pouvons rejetter H_0. Les constantes ne sont donc pas égales. On peut juste conclure que le modèle "pooled" estimé par les MCO n'est pas possible car $\alpha_i \neq \alpha$

6 Question 5



Sur le graphique, nous pouvons voir que plus le revenu augmente, plus les dépenses de santé semblent augmenter mais légèrement.

Pour les modèles à effets fixes individuels et effets aléatoires, le coefficient associé au logarithme du revenu est toujours négatif.

Pour le modèle à effets fixes, le coefficient β associé au log du revenu est de -0,105, ce qui signifie que l'augmentation de 1% du revenu va baisser de $\frac{-0,105}{100} = -0,01\%$ les dépenses de santé. Pour le modèle à effets aléatoires, le coefficient β associé au log du revenu est de -0,149, ce qui signifie que l'augmentation de 1% du revenu va baisser de $\frac{-0,149}{100} = -0,015\%$ les dépenses de santé.

Toutefois, le coefficient n'est jamais significatif (même au seuil 10%),ce qui rend l'interprétation impossible.

Pour les MCO, le coefficient est positif mais non significatif donc si le revenu augmente de 1%, les dépenses de santé augmentent de $\frac{0.392}{100} = 0,04\%$.

Nous avons donc des contradictions entre les différents modèles. Toutefois c'est dans le modèle des MCO que le revenu est le plus significatif avec un coefficient positif, donc une relation croissante entre le revenu et les dépenses de santé.

7 Question 6

Comparons les résultats entre le modèle à effets individuels et le modèle à effets aléatoires :

```
##
## ================================================
##                       Dependent variable:
##               ---------------------------------
##                             DEPSANTE
##                          (1)              (2)
## ------------------------------------------------
## lrev                   -0.105           -0.149
##                        (0.754)          (0.292)
##
## AGE                     0.065           -0.090*
##                        (0.094)          (0.051)
```

```
##
## I(AGE2)              0.0003           0.002***
##                      (0.001)          (0.001)
##
## ASSU1               1.351***          1.362***
##                      (0.114)          (0.107)
##
## Constant                              3.443**
##                                       (1.642)
##
## ------------------------------------------------
## Observations          1,000             1,000
## R2                    0.166             0.308
## Adjusted R2          -0.046             0.305
## F Statistic  39.678*** (df = 4; 796) 441.870***
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Le premier modèle (1) est le modèle "within" et le second (2) est "random".

Nous pouvons voir que :

- Le logarithme du revenu n'est pas significatif dans les deux modèles et le coefficient est négatif.
- L'âge n'est pas signicatif pour le modèle within et significatif au seuil de 10% pour le modèle random. Le coefficient est négatif pour le modèle random et positif pour le modèle within.
- L'âgeˆ2 n'est pas significatif pour le modèle à effets fixes mais significatif et positif pour le modèle à effets aléatoires, ce qui signifie pour les deux modèles que plus l'on vieillit, plus les dépenses de santé augmentent.
- Pour l'assurance maladie, les coefficients pour les deux modèles sont tous les deux signicatifs, positifs et presque identiques.

Faisons un test de Hausman pour déterminer quel modèle est la meilleure spécification.

```
##
##  Hausman Test
##
## data:  DEPSANTE ~ lrev + AGE + I(AGEˆ2) + ASSU
## chisq = 16.608, df = 4, p-value = 0.002303
## alternative hypothesis: one model is inconsistent
```

Les hypothèses de ce test sont les suivantes :

$$H_0 : \hat{\beta}_{inter} - \hat{\beta}_{MCQG} = 0 \rightarrow \text{Modèle aléatoire}$$

$$H_1 : \hat{\beta}_{inter} - \hat{\beta}_{MCQG} \neq 0 \rightarrow \text{Modèle à effets fixes}$$

On rejette l'hypothèse nulle (pvalue < 0.05) donc le modèle à effets aléatoires. Le modèle à effets fixes individuels est la meilleure spécification.

8 Question 7

8.1 Effet de l'âge

```
##
## ========================================
##                   Dependent variable:
##               ---------------------------
##                         DEPSANTE
## ----------------------------------------
## lrev                     -0.149
##                          (0.292)
##
## AGE                      -0.090*
##                          (0.051)
##
## I(AGE2)                 0.002***
##                          (0.001)
##
## ASSU1                   1.362***
##                          (0.107)
##
## Constant                 3.443**
##                          (1.642)
##
## ----------------------------------------
## Observations              1,000
## R2                        0.308
## Adjusted R2               0.305
## F Statistic           441.870***
## ========================================
## Note:         *p<0.1; **p<0.05; ***p<0.01
```

Le coefficient de l'âge est négatif ce qui montre que plus un individu est agé, moins ces dépenses de santé vont être élevé. De plus le coefficient de l'âgeˆ2 est positif, ce qui veut dire que plus l'âge augmente, plus les dépenses de santé diminue de plus en plus rapidement.

Le coefficient de l'âge est négatif mais non significatif au seuil de 5% donc nous allons nous interessé au coefficient de l'âge au carré. Ce coefficient est positif donc cela décrit une courbe convexe qui atteint son minimum en $\frac{-\beta_3}{2\beta_4} = \frac{0.09}{0.004} = 22.5$.

Comme notre âge minimum est de 21 et le minimum de la fonction se situe à 22.5 ans, on peut en conclure que plus une personne vieillit, plus les dépenses de santé vont augmenter.

8.2 Effet de l'assurance maladie

Voici un graphique et un test de moyenne qui représentent l'effet de l'assurance maladie sur les dépenses de santé :

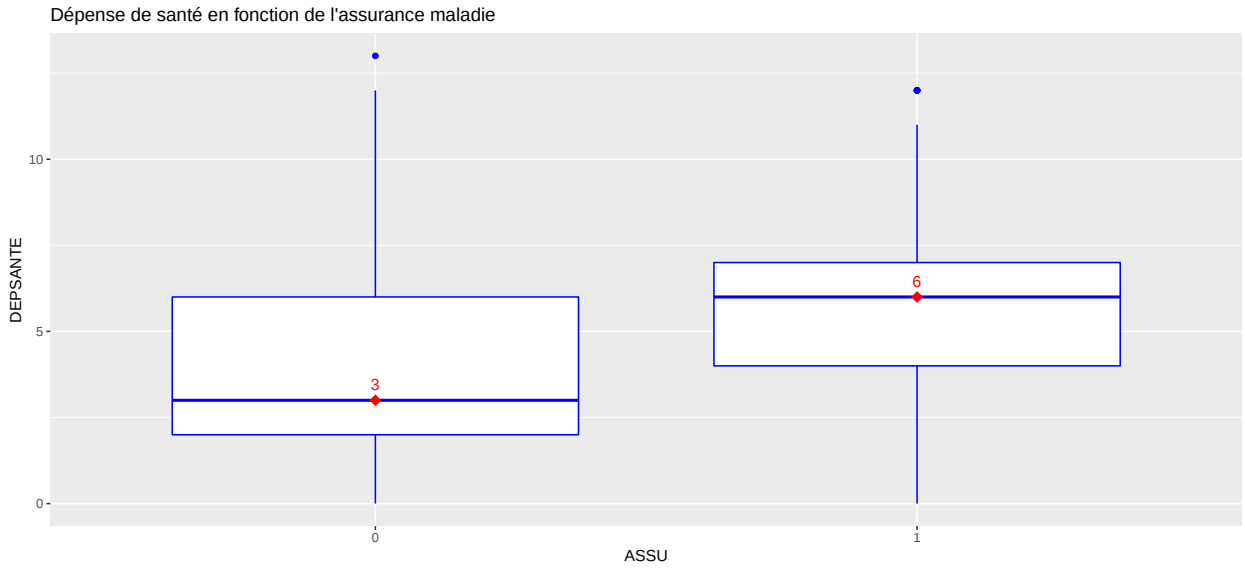


Table 5 – Test de moyenne sur l'assurance maladie

	DEPSANTE
Moyenne hors assurance maladie	4.093690
Moyenne assurance maladie	5.710692
Borne inf. de l'IC à 95% de la diff.	-1.954088
Borne sup. de l'IC à 95% de la diff.	-1.279915
p-value	0.000000

Nous pouvons constater que la moyenne des dépenses de santé est 410 euros si l'individu ne possède pas d'assurance maladie contre 570 euros s'il en possède une. Et grâce au test, on peut dire que la différence de moyenne est significative.

En ce qui conserne les résultats du modèle à effets aléatoires, le fait d'avoir une assurance maladie privée influe sur les dépenses de santé très significativement. Si un individu en possède une, les dépenses de santé sont 1,362 fois supérieure aux dépenses pour un individu qui n'en possède pas.

9 Question 8

Dans notre base de données, certains individus changent de statut d'assurance durant les cinq années, la variable est donc variante dans le temps et l'estimateur à effets fixes prend donc en compte cette variable dans le modèle.

Si les individus ne changent pas de statut d'assurance sur les cinq années alors cette variable est invariante dans le temps. Le modèle à effets fixes individuels (modèle within) n'estime pas ce genre de variable contrairement à un modèle à effets aléatoires. L'estimateur des effets fixes sera donc inefficace.