Exercices d'économétrie

TD 3 : Le modèle de régression multiple.

Janvier, 2020

Exercice 3.1

Dans le domaine de la santé publique, un problème important pour les décideurs politiques est de déterminer les effets de la consommation de tabac par la mère durant la grossesse (`cigs`) sur la santé de son enfant. Le poids à la naissance (`bwght`) est une mesure de la santé infantile. Un poids à la naissance trop faible peut augmenter le risque de contracter différentes maladies. Il existe d'autres facteurs qui affectent le poids du bébé à la naissance. Comme ces facteurs sont susceptibles d'être corrélés avec la consommation de cigarettes, nous devons les prendre en compte. Par exemple, un revenu familial (`faminc`) plus élevé facilite l'accés aux soins avant la naissance et assure également une meilleure alimentation de la mère. Considérons l'équation suivante, qui tient compte du revenu comme variable de contrôle :

$$\texttt{bwght}_i = \beta_0 + \beta_1 \texttt{cigs}_i + \beta_2 \texttt{faminc}_i + u_i$$

i Quel est le signe le plus probable pour β_2 ?

ii Pensez-vous que `cigs` et `faminc` soient susceptibles d'être corrélés ? Expliquez pourquoi la corrélation pourrait être positive ou négative.

iii En utilisant la base de données `bwght`, estimez l'équation par les MCO avec et sans `faminc`. L'ajout de `faminc` change-t-il de manière substantielle l'effet estimé de `cigs` sur `bwght` ?

Exercice 3.2

La base de données `ceosal2` contient un jeu de données portant sur 177 PDG aux États-Unis. Ces données peuvent être utilisées pour examiner les effets de la performance des entreprises sur le salaire des PDG.

i Estimez un modèle expliquant le salaire annuel du PDG (`salary`) par le chiffre d'affaires (`sales`) et la valeur de marché (`mktval`) de son entreprise. Faites en sorte que ce modèle soit un modèle à élasticité constante pour les deux variables indépendantes.

ii Ajoutez la variable `profits` au modèle utilisé dans la question précédente. Pourquoi ne doit-on pas inclure cette variable sous forme de logarithmique ? Considérez-vous que ces variables de performance des entreprises expliquent l'essentiel de la variation des salaires des PDG ?

iii Ajoutez l'expérience du PDG (`ceoten`) comme variable explicative au modèle utilisé dans la question précédente. Toutes choses étant égales par ailleurs, quel est le « rendement » estimé d'une année suplémentaire d'expérience ?

iv Calculez le coefficient de corrélation entre les variables log(`mktval`) et `profits` dans l'échantillon. Ces variables sont-elles fortement corrélées ? Qu'est-ce-que cela implique pour les estimateurs des MCO ?

Exercice 3.3

On utilise la base de données `meap93`.

i Estimez un modèle :

$$\texttt{math10}_i = \beta_0 + \beta_1 \log(\texttt{expend}_i) + \beta_2 \texttt{lnchprg}_i + u_i$$

où `math10` est le pourcentage d'élèves réussissant le test de mathématiques, `expend` est la dépense par élève ($), et `lnchprg` est le pourcentage d'élèves participant au « *shool lunch program*» utilisé comme une proxy de la pauvreté des élèves. Les signes des estimateurs sont-ils ceux que vous anticipiez ?

ii Pouvez-vous interpréter l'estimateur de l'ordonnée à l'origine, $\hat{\beta}_0$?

iii Reprenez l'estimation du modèle en éliminant la variable `lnchprg`. Comparez les résultats avec la première estimation. L'estimation de l'effet des dépenses est-elle plus grande ou plus petite que dans le premier modèle ?

iv Calculez la corrélation entre log(`expend`) et `lnchprg`. Pouvez-vous justifier son signe ?

v Utilisez le résultat sur le calcul de la corrélation pour justifier la différence entre les deux estimations.

Exercice 3.4

On utilise la base de données `charity`.

i Estimez le modèle suivant par les MCO :

$$\texttt{gift}_i = \beta_0 + \beta_1 \texttt{mailsyear}_i + \beta_2 \texttt{giftlast}_i + \beta_3 \texttt{propresp}_i + u_i$$

où `gift` est le montant des dons (en florin néerlandais), `mailsyear` est le nombre moyen de sollicitations envoyées par la poste sur l'année, `giftlast` le montant moyen du dernier don, et `propresp` le taux de réponse aux envois postaux. Comparez le R^2 de cette régression avec celui que nous obtiendrions en excluant les deux dernières variables (`giftlast` et `propresp`).

ii Interprétez le coefficient associé à `mailsyear`. Est-il plus grand ou plus petit que le coefficient estimé à partir de la régression simple ?

iii Interprétez le coefficient associé à `propresp`.

iv Rééstimez le modèle en ajoutant la variable `avggift` (montant moyen des dons effectués les années précédentes). Que devient l'effet estimé de `mailsyear` ?

v Dans l'équation de la dernière estimation, qu'est-il arrivé au coefficient de `gitftlast` ? Que se passe-t-il ?

Solution 3.1

(i) Le paramètre β_2 devrait être positif, une mère plus riche bénéficie d'un environnement plus favorable (par exemple elle ne souffre pas de carances alimentaires) et a accès à de meilleurs soins pré-nataux. (ii) La corrélation pourrait être positive car une personne plus riche n'est pas financièrement contrainte pour s'acheter des cigarettes, mais en même temps on sait que les personnes plus éduquées, ce qui généralement est le cas des personnes plus riches, fument moins. On peut donc penser que cette corrélation devrait être négative. Dans l'échantillon la corrélation estimée est faible mais négative : -0,173. (iii) L'estimation avec les deux variables explicatives est résumée dans le tableau suivant :

OLS Regression Results

Dep. Variable:	bwght	R-squared:	0.030
Model:	OLS	Adj. R-squared:	0.028
Method:	Least Squares	F-statistic:	21.27
Date:	Sun, 23 Feb 2020	Prob (F-statistic):	7.94e-10
Time:	17:21:31	Log-Likelihood:	-6130.4
No. Observations:	1388	AIC:	1.227e+04

```
Df Residuals:                     1385   BIC:                         1.228e+04
Df Model:                            2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    116.9741      1.049    111.512      0.000     114.916     119.032
cigs          -0.4634      0.092     -5.060      0.000      -0.643      -0.284
faminc         0.0928      0.029      3.178      0.002       0.036       0.150
==============================================================================
Omnibus:                      116.751   Durbin-Watson:                   1.922
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              619.781
Skew:                          -0.154   Prob(JB):                    2.61e-135
Kurtosis:                       6.259   Cond. No.                         67.4
==============================================================================
```

On note que $\hat{\beta}_1 < 0$ (significativement) et $\hat{\beta}_2 > 0$ (significativement) comme attendu. Si on élimine la dernière variable, en ne considérant que la régression de `bwght` sur `cigs` (et une constante), on obtient :

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  bwght   R-squared:                       0.023
Model:                            OLS   Adj. R-squared:                  0.022
Method:                 Least Squares   F-statistic:                     32.24
Date:                Sun, 23 Feb 2020   Prob (F-statistic):           1.66e-08
Time:                        17:21:31   Log-Likelihood:                -6135.5
No. Observations:                1388   AIC:                         1.227e+04
Df Residuals:                    1386   BIC:                         1.229e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    119.7719      0.572    209.267      0.000     118.649     120.895
cigs          -0.5138      0.090     -5.678      0.000      -0.691      -0.336
==============================================================================
Omnibus:                      118.187   Durbin-Watson:                   1.924
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              635.742
Skew:                          -0.156   Prob(JB):                    8.92e-139
Kurtosis:                       6.301   Cond. No.                         6.72
==============================================================================
```

La consommation de cigarettes a toujours un effet négatif sur la santé des enfants, même s'il est un peu plus faible. La différence est faible car `cigs` et `faminc` sont peu corrélés et $\hat{\beta}_2$ est proche de zéro.

Solution 3.2

(i) Afin que les paramètres puissent s'interpréter comme des élasticités (constantes puisqu'il s'agit de paramètres) nous allons considérer un modèle linéaire avec les variables en logarithme :

$$\log(\texttt{salary}_i) = \beta_0 + \beta_1 \log(\texttt{sales}_i) + \beta_2 \log(\texttt{mktval}_i) + u_i$$

L'estimation par les MCO donne :

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          logged_salary   R-squared:                       0.299
```

```
Model:                            OLS   Adj. R-squared:                  0.291
Method:                 Least Squares   F-statistic:                     37.13
Date:                Sun, 23 Feb 2020   Prob (F-statistic):           3.73e-14
Time:                        18:52:55   Log-Likelihood:                -130.56
No. Observations:                 177   AIC:                             267.1
Df Residuals:                     174   BIC:                             276.6
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      4.6209      0.254     18.163      0.000       4.119       5.123
logged_sales   0.1621      0.040      4.087      0.000       0.084       0.240
logged_mktval  0.1067      0.050      2.129      0.035       0.008       0.206
==============================================================================
Omnibus:                       17.241   Durbin-Watson:                   2.092
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               63.383
Skew:                          -0.038   Prob(JB):                     1.72e-14
Kurtosis:                       5.931   Cond. No.                         70.4
==============================================================================
```

Les deux élasticités sont positives, la performance de l'entreprise a bien un effet positif sur le salaire du PDG. (ii) Nous ne pouvons pas inclure la variable `profits` sous un logarithme, car quelques observations sont négatives. Nous considérons donc le modèle suivant :

$$\log(\texttt{salary}_i) = \beta_0 + \beta_1 \log(\texttt{sales}_i) + \beta_2 \log(\texttt{mktval}_i) + \beta_3 \texttt{profits}_i + u_i$$

Une alternative aurait pu être d'éliminer les observations pour lesquelles les profits sont négatifs. L'estimation par les MCO donne :

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          logged_salary   R-squared:                       0.299
Model:                            OLS   Adj. R-squared:                  0.287
Method:                 Least Squares   F-statistic:                     24.64
Date:                Sun, 23 Feb 2020   Prob (F-statistic):           2.53e-13
Time:                        18:52:55   Log-Likelihood:                -130.53
No. Observations:                 177   AIC:                             269.1
Df Residuals:                     173   BIC:                             281.8
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      4.6869      0.380     12.343      0.000       3.937       5.436
logged_sales   0.1614      0.040      4.043      0.000       0.083       0.240
logged_mktval  0.0975      0.064      1.531      0.128      -0.028       0.223
profits        3.566e-05   0.000      0.235      0.815      -0.000       0.000
==============================================================================
Omnibus:                       17.054   Durbin-Watson:                   2.097
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               62.074
Skew:                          -0.029   Prob(JB):                     3.32e-14
Kurtosis:                       5.901   Cond. No.                     4.52e+03
==============================================================================
```

On note que les estimateurs de β_1 et β_2 sont pratiquement inchangés. On remarque aussi que le R^2 ne change pas, on voit même une légère réduction du R^2 ajusté (qui pénalise le nombre de variables dans le modèle),

autrement dit la variable `profits` n'apporte pas grand chose ici. Dans ces deux modèles le R^2 n'excède pas 30%, autrement dit ces modèles expliquent moins d'un tiers de l'hétérogénéité des salaires des PDG. Enfin notons que le paramètre associé à la variable `profits` n'apparaît pas significativement différent de zéro (la statistique de Student est largement inférieure à 1,96), et qu'avec l'introduction de cette variable $\hat{\beta}_2$ n'est plus significativement différent de zéro. (iii) Nous reprenons le dernier modèle en rajoutant une variable rendant compte de l'expérience du PDG :

$$\log(\texttt{salary}_i) = \beta_0 + \beta_1 \log(\texttt{sales}_i) + \beta_2 \log(\texttt{mktval}_i) + \beta_3 \texttt{profits}_i + \beta_4 \texttt{ceoten}_i + u_i$$

L'estimation par les MCO donne :

OLS Regression Results

Dep. Variable:	logged_salary	R-squared:	0.318
Model:	OLS	Adj. R-squared:	0.302
Method:	Least Squares	F-statistic:	20.08
Date:	Sun, 23 Feb 2020	Prob (F-statistic):	1.39e-13
Time:	18:52:55	Log-Likelihood:	-128.10
No. Observations:	177	AIC:	266.2
Df Residuals:	172	BIC:	282.1
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.5578	0.380	11.986	0.000	3.807	5.308
logged_sales	0.1622	0.039	4.109	0.000	0.084	0.240
logged_mktval	0.1018	0.063	1.614	0.108	-0.023	0.226
profits	2.905e-05	0.000	0.193	0.847	-0.000	0.000
ceoten	0.0117	0.005	2.187	0.030	0.001	0.022

Omnibus:	25.236	Durbin-Watson:	2.047
Prob(Omnibus):	0.000	Jarque-Bera (JB):	121.011
Skew:	-0.281	Prob(JB):	5.28e-27
Kurtosis:	7.012	Cond. No.	4.57e+03

L'introduction de la variable d'expérience du PDG améliore marginalement le pouvoir explicatif du modèle (le R^2). Toute chose égale par ailleurs, une augmentation d'une année de l'expérience du PDG augmente de 1,2% son salaire. (iv) La corrélation entre les variables `profits` et `mktval` est positive et relativement élevée. En moyenne une firme qui a un profit élevé est aussi une firme avec une valeur de marché importante. Ce n'est guère surprenant puisque la valeur de marché est (en partie) déterminée par le profit courrant et les profits anticipés. Malgré cette forte corrélation, l'absence de la variable `profits` dans la première estimation ne biaise pas les estimations des résultats car le paramètre associés (son estimateur) est proche de zéro. Tout au plus, la forte corrélation contribue à augmenter la variance de l'estimateur des MCO (on peut clairement le voir en comparant les deux premières estimations) ; rappelons que la variance de l'estimateur est proportionnelle à $(X'X)^{-1}$, où X est la matrice $N \times k$ des régresseurs, si deux colonnes sont proches d'être colinéaires (une valeur propre de $X'X$ est proche de zéro) alors la variance doit être importante.

Solution 3.3

(i) L'estimation par les MCO donne :

OLS Regression Results

Dep. Variable:	math10	R-squared:	0.180

```
Model:                            OLS   Adj. R-squared:                  0.176
Method:                 Least Squares   F-statistic:                     44.43
Date:                Sat, 07 Mar 2020   Prob (F-statistic):           3.59e-18
Time:                        10:32:46   Log-Likelihood:                -1497.1
No. Observations:                 408   AIC:                             3000.
Df Residuals:                     405   BIC:                             3012.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -20.3608     25.073     -0.812      0.417     -69.650      28.928
logged_expend  6.2297      2.973      2.096      0.037       0.386      12.073
lnchprg       -0.3046      0.035     -8.614      0.000      -0.374      -0.235
==============================================================================
Omnibus:                       52.915   Durbin-Watson:                   1.903
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               83.720
Skew:                           0.816   Prob(JB):                     6.61e-19
Kurtosis:                       4.504   Cond. No.                     1.58e+03
==============================================================================
```

On trouve les signes attendus pour les pentes, la réussite des élèves en mathématiques dépend bien positivement des dépenses par élève et négativement du taux de participation au « *school lunch program* » qui mesure indirectement la pauvreté des élèves dans chaque école. (ii) De façon générale, la constante s'interprête comme le niveau moyen de la variable expliquée lorsque les variables explicatives sont nulles. Ici cela ne fait pas beaucoup de sens, puisque la constante estimée est négative (alors que la variable expliquée, le pourcentage de réussite à un test de matématiques, doit être positive). Notons que la première variable explicative est le logarithme de dépense par élève. Pour que cette variable soit nulle il faudrait que la dépense par élève soit exactement de un dollar. Un tel scénario est totalement hors des valeurs possibles de cette variable, dès lors il n'est pas étonnant que la prédiction du modèle soit difficile à interpréter. (iii) L'estimation par les MCO avec seulement la première variable explicative donne :

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 math10   R-squared:                       0.030
Model:                            OLS   Adj. R-squared:                  0.027
Method:                 Least Squares   F-statistic:                     12.41
Date:                Sat, 07 Mar 2020   Prob (F-statistic):           0.000475
Time:                        10:32:46   Log-Likelihood:                -1531.4
No. Observations:                 408   AIC:                             3067.
Df Residuals:                     406   BIC:                             3075.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -69.3412     26.530     -2.614      0.009    -121.495     -17.188
logged_expend  11.1644      3.169      3.523      0.000       4.935      17.394
==============================================================================
Omnibus:                       28.397   Durbin-Watson:                   1.615
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               34.609
Skew:                           0.591   Prob(JB):                     3.05e-08
Kurtosis:                       3.800   Cond. No.                         440.
==============================================================================
```

On constate que l'effet estimé des dépenses par élève sur la réussite en mathématiques est beaucoup plus élevé

(presque deux fois). (iv) Notons que ces deux variables sont négativement corrélées ($\simeq -0,2$), en moyenne quand la population est moins riche (participation au « *school lunch program* plus élevée) la dépense par élève est moins élevée. (v) En omettant la variable `lnchprg`, proxy de la pauvreté, dans la seconde régression on tend à surestimer l'effet des dépenses sur la réussite en mathématiques. Ce résultat se comprend dans la mesure où la variable omise, dont on sait qu'elle affecte négativement la réussite en mathématiques via la première régression, est négativement corrélée avec la log(`expend`). Le second modèle estimé est de la forme :

$$\texttt{math10}_i = \beta_0 + \beta_1 \log(\texttt{expend}_i) + v_i$$

Nous savons que l'esimateur de MCO de la pente est :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} \left(\log(\texttt{expend}_i) - \overline{\log(\texttt{expend})}\right) \left(\texttt{math10}_i - \overline{\texttt{math10}}\right)}{\sum_{i=1}^{N} \left(\log(\texttt{expend}_i) - \overline{\log(\texttt{expend})}\right)^2}$$

En supposant que le vrai modèle, c'est-à-dire celui utilisé par la nature pour générer les données, est de la forme :

$$\texttt{math10}_i = \beta_0 + \beta_1 \log(\texttt{expend}_i) + \beta_2 \texttt{lnchprg}_i + \varepsilon_i$$

comme dans la première régression, avec ε_i une variable aléatoire iid d'espérance nulle indépendante des variables explicatives, il s'ensuit que :

$$\overline{\texttt{math10}} = \beta_0 + \beta_1 \overline{\log(\texttt{expend})} + \beta_2 \overline{\texttt{lnchprg}} + \bar{\varepsilon}$$

par linéarité de la moyenne, et donc :

$$\texttt{math10}_i - \overline{\texttt{math10}} = \beta_1 \left(\log(\texttt{expend}_i) - \overline{\log(\texttt{expend})}\right) + \beta_2 \left(\texttt{lnchprg}_i - \overline{\texttt{lnchprg}}\right) + \varepsilon_i - \bar{\varepsilon}$$

En substituant dans l'expression de $\hat{\beta}_1$ et en simplifiant, il vient :

$$\hat{\beta}_1 = \beta_1 + \beta_2 \frac{\sum_{i=1}^{N} \left(\log(\texttt{expend}_i) - \overline{\log(\texttt{expend})}\right) \left(\texttt{lnchprg}_i - \overline{\texttt{lnchprg}}\right)}{\sum_{i=1}^{N} \left(\log(\texttt{expend}_i) - \overline{\log(\texttt{expend})}\right)^2} + \frac{\sum_{i=1}^{N} \left(\log(\texttt{expend}_i) - \overline{\log(\texttt{expend})}\right) (\varepsilon_i - \bar{\varepsilon})}{\sum_{i=1}^{N} \left(\log(\texttt{expend}_i) - \overline{\log(\texttt{expend})}\right)^2}$$

La bonne nouvelle est que l'estimateur des MCO, $\hat{\beta}_1$, dépend de la vraie valeur du paramètre associé aux dépenses par élève (β_1 le premier terme sur le membre de droite de la dernière équation). L'estimateur des MCO n'est pas exactement égal à la vraie valeur (cela n'arrive jamais) à cause de la présence des autres termes sur le membre de droite. Le dernier terme est lié à la variabilité de l'échantillon (l'estimateur est différent de la vraie valleur car la nature nous donne un échantillon mais elle aurait pu nous en donner un autre). L'avant dernier terme est lié à l'omission de la variable `lnchprg`. On note que les deux derniers termes ont le même dénominateur est que celui-ci nous est familier. Si nous divisons par N le dénominateur, nous obtenons l'expression de l'estimateur de la variance de log(`expend`) :

$$\frac{1}{N} \sum_{i=1}^{N} \left(\log(\texttt{expend}_i) - \overline{\log(\texttt{expend})}\right)^2$$

comme la moyenne des écarts à la moyenne au carré. Sous des conditions générales, quand la taille de l'échantillon (N) tend vers l'infini, cet estimateur tend [1] vers la variance (de la population) :

$$\frac{1}{N} \sum_{i=1}^{N} \left(\log(\texttt{expend}_i) - \overline{\log(\texttt{expend})}\right)^2 \xrightarrow[N \to \infty]{} \mathbb{V}\left[\log(\texttt{expend})\right]$$

1. On ne rentre pas dans les détails ici, mais il s'agit d'une convergence en probabilité, voir votre cours de probabilité.

De la même façon les numérateurs des deux derniers termes divisés par N tendent vers une covariance quand N tend vers l'infini :

$$\frac{1}{N}\sum_{i=1}^{N}\left(\log(\texttt{expend}_i) - \overline{\log(\texttt{expend})}\right)\left(\texttt{lnchprg}_i - \overline{\texttt{lnchprg}}\right) \xrightarrow[N\to\infty]{} \mathbb{C}ov(\log(\texttt{expend}), \texttt{lnchprg})$$

et

$$\frac{1}{N}\sum_{i=1}^{N}\left(\log(\texttt{expend}_i) - \overline{\log(\texttt{expend})}\right)(\varepsilon_i - \overline{\varepsilon}) \xrightarrow[N\to\infty]{} \mathbb{C}ov(\log(\texttt{expend}), \varepsilon)$$

Puisque la perturbation ε est supposée indépendante des variables explicatives, la seconde covariance doit être nulle. Ainsi, asymptotiquement nous devons avoir :

$$\hat{\beta}_1 \xrightarrow[N\to\infty]{} \beta_1 + \beta_2 \frac{\mathbb{C}ov(\log(\texttt{expend}), \texttt{lnchprg})}{\mathbb{V}\left[\log(\texttt{expend})\right]}$$

ou encore, par définition de la corrélation :

$$\hat{\beta}_1 \xrightarrow[N\to\infty]{} \beta_1 + \beta_2 \text{corr}(\log(\texttt{expend}), \texttt{lnchprg}) \sqrt{\frac{\mathbb{V}\left[\texttt{lnchprg}\right]}{\mathbb{V}\left[\texttt{expend}\right]}}$$

Ainsi, si la corrélation entre log(`expend`) et `lnchprg` est négative et si β_2 est négatif (comme le suggère l'estimateur $\hat{\beta}_2$), l'estimateur $\hat{\beta}_1$ tendra à surestimer β_1, c'est-à-dire l'effet des dépenses par élève sur la réussite en mathématiques.

Solution 3.4

L'estimation par les MCO donne :

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   gift   R-squared:                       0.083
Model:                            OLS   Adj. R-squared:                  0.083
Method:                 Least Squares   F-statistic:                     129.3
Date:                Sat, 07 Mar 2020   Prob (F-statistic):           3.86e-80
Time:                        22:15:31   Log-Likelihood:                -17446.
No. Observations:                4268   AIC:                         3.490e+04
Df Residuals:                    4264   BIC:                         3.492e+04
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -4.5515      0.803     -5.668      0.000      -6.126      -2.977
mailsyear      2.1663      0.332      6.526      0.000       1.516       2.817
giftlast       0.0059      0.001      4.138      0.000       0.003       0.009
propresp      15.3586      0.875     17.562      0.000      13.644      17.073
==============================================================================
Omnibus:                     5103.126   Durbin-Watson:                   1.667
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           941920.719
Skew:                           6.166   Prob(JB):                         0.00
Kurtosis:                      74.726   Cond. No.                         721.
==============================================================================
```

Si nous réestimons le modèle en éliminant les deux dernières variables, on obtient :

OLS Regression Results

Dep. Variable:	gift	R-squared:	0.014
Model:	OLS	Adj. R-squared:	0.014
Method:	Least Squares	F-statistic:	59.65
Date:	Sat, 07 Mar 2020	Prob (F-statistic):	1.40e-14
Time:	22:15:31	Log-Likelihood:	-17602.
No. Observations:	4268	AIC:	3.521e+04
Df Residuals:	4266	BIC:	3.522e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.0141	0.739	2.724	0.006	0.564	3.464
mailsyear	2.6495	0.343	7.723	0.000	1.977	3.322

Omnibus:	4951.411	Durbin-Watson:	1.529
Prob(Omnibus):	0.000	Jarque-Bera (JB):	784986.187
Skew:	5.889	Prob(JB):	0.00
Kurtosis:	68.387	Cond. No.	8.34

Le R^2 de la première régression est relativement faible (environ 8%), le pouvoir explicatif est réduit par un facteur 6 si on ne retient que la variable `mailsyear` dans le modèle. Le pouvoir explicatif des deux variables omises est donc important relativement à celui de `mailsyear`. (ii) En contrôlant pour les variations de `giftlast` et `propresp`, la première régression nous dit qu'un courrier supplémentaire par an (une augmentation d'une unité de `mailsyear`) induit une augmentation de 2,17 florins en dons. Si on ne contrôle pas des variations de `giftlast` et `propresp`, l'augmentation induite des dons est plus importante (2,65 florins). (iii) La variable `propresp` représente le taux de réponse aux envois postaux (il s'agit d'un pourcentage). Si `propresp` augmente de 0.1 (soit 10 points de pourcentage), alors les dons augmentent de $15,36 \times 0,1 = 1,5$ florins. (iv) L'esimation du modèle augmenté avec `avggift` par les MCO :

OLS Regression Results

Dep. Variable:	gift	R-squared:	0.201
Model:	OLS	Adj. R-squared:	0.200
Method:	Least Squares	F-statistic:	267.3
Date:	Sat, 07 Mar 2020	Prob (F-statistic):	2.82e-205
Time:	22:15:31	Log-Likelihood:	-17154.
No. Observations:	4268	AIC:	3.432e+04
Df Residuals:	4263	BIC:	3.435e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-7.3278	0.758	-9.664	0.000	-8.814	-5.841
mailsyear	1.2012	0.312	3.845	0.000	0.589	1.814
giftlast	-0.2609	0.011	-24.251	0.000	-0.282	-0.240
propresp	16.2046	0.818	19.821	0.000	14.602	17.807
avggift	0.5269	0.021	24.996	0.000	0.486	0.568

Omnibus:	3922.109	Durbin-Watson:	1.787
Prob(Omnibus):	0.000	Jarque-Bera (JB):	698065.217
Skew:	3.806	Prob(JB):	0.00

```
Kurtosis:                       65.189   Cond. No.                         814.
==============================================================================
```

En contrôlant du montant moyen des dons effectués les années précédentes, l'effet induit d'une augmentation de `mailsyear` est très sensiblement réduit (1,2 florins). Notons que le R^2 de cette dernière régression est beaucoup plus important, le modèle explique maintenant 20% de la volatilité des dons. (v) On note que la contribution du dernier don (`giftlast`) devient négative lorsque l'on contrôle des dons effectués les dernières années. Notons que ces deux variables, `giftlast` et `avggift`, sont quasiment parfaitement corrélées (le coefficient de corrélation est 0,99, les personnes qui ont beaucoup contribué l'année précédente sont aussi celles qui contribuaient beaucoup par le passé). Une possible explication pour le renversement de signe suite à l'introduction de `avggift`, est que les personnes qui donnent beaucoup une année auront tendance à donner moins l'année suivante (même si elles donnent généralement beaucoup).