Exercices d'économétrie

TD 2 : Le modèle de régression linéaire simple.

Janvier, 2020

Exercice 2.1

On utilise la base de données `401k` pour cet exercice. Ces données cherchent à rendre compte de la participation des travailleurs au plan d'épargne-pension « 401k», un système de retraite par capitalisation très largement utilisé aux États-Unis. La variable `prate` représente le taux de participation des travailleurs. La variable `mrate` mesure la proportion de la contribution venant de l'employeur relativement à celle du salarié (si `mrate` est égal à $1/2$ alors une contribution de 1$ du travailleur est complétée par une contribution de 0,5$ de l'employeur).

i Calculez le taux de participation moyen ainsi que la contribution relative moyenne de l'employeur dans l'échantillon.

ii Estimer le modèle de régression simple pour obtenir :

$$\widehat{\texttt{prate}}_i = \hat{\beta}_0 + \hat{\beta}_1 \texttt{mrate}_i$$

iii Interprétez les coefficients estimés.

iv Prédisez `prate` lorsque `mrate` est égal à 3,5.

v Quel pourcentage de la variation de `prate` est expliqué par `mrate` ?

Exercice 2.2

On utilise la base de données `sleep75` pour cet exercice. Ces données cherchent à rendre compte de l'arbitrage entre le temps passé à dormir (par semaine) et le temps passé à travailler (on ne compte que les activités rémunérées). Le sens de la causalité n'est pas évident, mais dans la suite nous considérerons la quantité de sommeil comme la variable dépendante, c'est-à-dire le modèle suivant :

$$\texttt{sleep}_i = \beta_0 + \beta_1 \texttt{totwrk}_i + u_i$$

`sleep` est le nombre de minutes de sommeil par semaine, `totwrk` le nombre de minutes travaillées par semaine.

i Estimez les paramètres et reportez le R^2. Interprétez la constante de la régression.

ii Calculez la variation du temps passé à dormir si la quantité de travail augmente de deux heures.

Exercice 2.3

On s'intéresse à l'investissement en Recherche et Développement des firmes de l'industrie chimique. La variable `rd` représente le montant des dépenses en R&D (en millions de dollars), la variable `sales` représente le montant des dépenses des firmes (toujours en million de dollar).

i Écrivez un modèle impliquant une élasticité constante entre les deux variables.

ii Estimez le modèle en utilisant la base de données `rdchem`. Quelle est l'élasticité estimée entre `rd` et `sales` ?

Exercice 2.4

Dans cet exercice nous n'utiliserons pas de données mais des données engendrées aléatoirement, afin d'illustrer les propriétés de l'estimateur des Moindres Carrés Ordinaires.

i Créez cinq cents observations pour une variable explicative, x_i, en tirant dans une loi uniforme entre 0 et 10. Utilisez la fonction textttrand dans le module Python `numpy.random`.

ii Générez cinq cents erreurs, u_i, en tirant dans une loi normale d'espérance nulle et de variance 36. Vous utiliserez la fonction `randn` du même module `numpy.random`. Calculez la moyenne et l'écart type des erreurs u_i. Que remarquez-vous ?

iii Créez la variable y selon le modèle :

$$y_i = b_0 + b_1 x_i + u_i$$

avec $\beta_0 = 1$ et $\beta_1 = 2$. Représentez graphiquement l'échantillon (utilisez le module `matplotlib`).

iv En utilisant les données engendrées dans les questions précédentes, estimez le modèle :

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

par les Moindres Carrés Ordinaires. Comparez $\hat{\beta}_0$ et $\hat{\beta}_1$ avec les vraies valeurs du modèles b_0 et b_1 (c'est-à-dire les valeurs de la constante et de la pente utilisées pour engendrer la variable expliquée y).

v Calculez les résidus estimés $\hat{e}$ et vérifiez que l'on a bien (aux erreurs numériques près) :

$$\sum_{i=1}^{500} \hat{e}_i = 0 \quad \text{et} \quad \sum_{i=1}^{500} x_i \hat{e}_i = 0$$

vi Calculez les deux mêmes sommes en utilisant les « vraies » erreurs plutôt que les résidus estimés. Commentez.

vii En utilisant une boucle engendrez 10 échantillons $(y_i, x_i)_{i=1}^{500}$ en tirant des erreurs différentes dans la même loi normale. Pour chaque échantillon estimez le modèle par les MCO et stockez les valeurs de $\hat{\beta}_1$ dans un vecteur. À la sortie de la boucle calculez la moyenne des $\hat{\beta}_1$.

viii Reprenez la question précédente avec 100 échantillons simulés, puis 1000 échantillons simulés. Comparez les moyennes de $\hat{\beta}_1$ calculées sur la base de 10, 100 et 1000 échantillons. Que remarquez vous ? À quoi correspond cette moyenne ?

Solution 2.1

(i) Le taux de participation moyen est 87,36%, la contribution moyenne des employeurs 0,73. (ii) Pour estimer le modèle on utilise le module `statsmodels` (voir le code plus bas), et on trouve :

OLS Regression Results

Dep. Variable:	prate	R-squared:	0.075
Model:	OLS	Adj. R-squared:	0.074
Method:	Least Squares	F-statistic:	123.7
Date:	Sun, 12 Jan 2020	Prob (F-statistic):	1.10e-27
Time:	15:17:59	Log-Likelihood:	-6437.0
No. Observations:	1534	AIC:	1.288e+04
Df Residuals:	1532	BIC:	1.289e+04
Df Model:	1		
Covariance Type:	nonrobust		

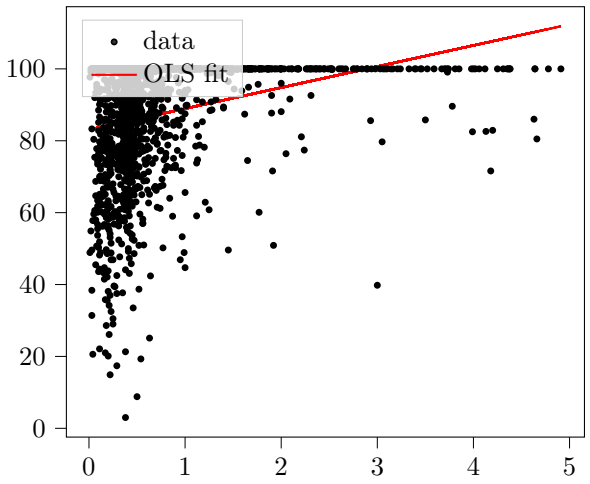
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	83.0755	0.563	147.484	0.000	81.971	84.180
mrate	5.8611	0.527	11.121	0.000	4.827	6.895

Omnibus:	394.767	Durbin-Watson:	1.908
Prob(Omnibus):	0.000	Jarque-Bera (JB):	870.172
Skew:	-1.444	Prob(JB):	1.11e-189
Kurtosis:	5.296	Cond. No.	2.32

(iii) Le coefficient associé à la variable `mrate` ($\hat{\beta}_1 = 5,86$) est positif et significativement différent de zéro. En moyenne, quand la contribution des employeurs est plus forte le taux de participation des salariés est plus élevé (l'incitation est plus importante). (iv) On calcule la prédiction du taux de participation des salariés de la façon suivante :

$$\widehat{prate} = \hat{\beta}_0 + \hat{\beta}_1 \times 3,5$$

et on trouve 103,59, ce qui n'a pas beaucoup de sens étant donné que le taux de participation des salariés ne saurait être supérieur à 100%. On comprend facilement ce qui se passe en représentant graphiquement les variables et la droite de régression :



le modèle linéaire n'est probablement pas adapté à la question qui nous intéresse ici. (v) Le pourcentage de la variation de `prate` expliqué par `mrate` est donné par le coefficient d'ajustement, c'est-à-dire le R^2, de la régression. Il est très faible, on a $R^2 = 0,075$, ainsi 7,5% de la volatilité de `prate` est expliquée par `mrate`. La contribution relative des employeurs explique peu le taux de participation des salariés. Ce résultat est cohérent avec ce que nous voyons sur le graphique : les points sont très éloignés de la droite de régression.

Solution 2.2

(i) Les résultats de la régression sont obtenus avec le module `statsmodels` et résumés dans le tableau suivant :

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sleep   R-squared:                       0.103
Model:                            OLS   Adj. R-squared:                  0.102
Method:                 Least Squares   F-statistic:                     81.09
Date:                Sun, 12 Jan 2020   Prob (F-statistic):           1.99e-18
Time:                        19:26:12   Log-Likelihood:                -5267.1
No. Observations:                 706   AIC:                         1.054e+04
Df Residuals:                     704   BIC:                         1.055e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   3586.3770     38.912     92.165      0.000    3509.979    3662.775
totwrk        -0.1507      0.017     -9.005      0.000      -0.184      -0.118
==============================================================================
Omnibus:                       68.651   Durbin-Watson:                   1.955
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              192.044
Skew:                          -0.483   Prob(JB):                     1.99e-42
Kurtosis:                       5.365   Cond. No.                     5.71e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
↪  specified.
[2] The condition number is large, 5.71e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

nous avons donc :

$$\widehat{\texttt{sleep}}_i = 3586,38 + -0,15\texttt{totwrk}_i$$

et la quantité de travail explique 10,3% de l'hétérogénéité du temps de sommeil. La constante estimée (3586 minutes, soit environ 60 heures) s'interprète comme le temps de sommeil moyen d'un individu qui ne travaillerait pas. (ii) Notons $\Delta\texttt{totwrk}$ la variation de deux heures du temps de travail, on a donc $\Delta\texttt{totwrk} = 120$ (minutes). Cette variation ne changera

pas la valeur de la constante β_0 (c'est une constante) ; puisque le modèle est linéaire, nous devons avoir :

$$\Delta \texttt{sleep} = -0,15 \Delta \texttt{totwrk} \simeq -18 \texttt{ minutes}$$

soit moins de trois minutes par nuit.

Solution 2.3

(i) Pour que l'élasticité entre `rd` et `sales` se réduise à un paramètre estimable, il faut que celle-ci soit constante. C'est bien le cas si la forme fonctionnelle est de type Cobb-Douglas :

$$\texttt{rd} = b_0 \texttt{sales}^{b_1}$$

On vérifie facilement que la dérivée de `rd` par rapport à `sales` rapportée au ratio $^{\texttt{rd}}/_{\texttt{sales}}$ est égale à b_1. (ii) Pour estimer le modèle, on considère le logarithme de l'équation précédente, de sorte que le modèle devient linéaire par rapport aux paramètres, et on ajoute un terme d'erreur. Le modèle empirique est donc :

$$\log \texttt{rd}_i = \beta_0 + \beta_1 \log \texttt{sales}_i + u_i$$

avec $\beta_0 = \log b_0$ et $\beta_1 = b_1$. L'estimation par Moindres Carrés Ordinaires est résumée par :

OLS Regression Results

Dep. Variable:	logged_rd	R-squared:	0.910
Model:	OLS	Adj. R-squared:	0.907
Method:	Least Squares	F-statistic:	302.7
Date:	Tue, 14 Jan 2020	Prob (F-statistic):	3.20e-17
Time:	08:55:16	Log-Likelihood:	-24.021
No. Observations:	32	AIC:	52.04
Df Residuals:	30	BIC:	54.97
Df Model:	1		
Covariance Type:	nonrobust		

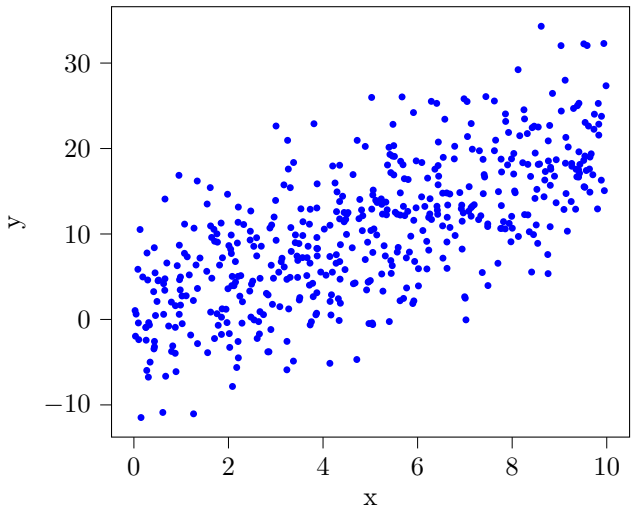
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-4.1047	0.453	-9.066	0.000	-5.029	-3.180
logged_sales	1.0757	0.062	17.399	0.000	0.949	1.202

Omnibus:	1.407	Durbin-Watson:	1.847
Prob(Omnibus):	0.495	Jarque-Bera (JB):	1.025
Skew:	0.139	Prob(JB):	0.599
Kurtosis:	2.168	Cond. No.	36.1

L'élasticité estimée est donc d'environ 1,08. Une augmentation de 1% des ventes engendre une augmentation de 1,08% des dépenses en R&D.

Solution 2.4

(i) et (ii) Voir le code. On remarque que la moyenne des erreurs simulées (-0,14) est différente de la moyenne théorique (0). De même l'écart-type calculé sur la base des erreurs simulées (5,88) est différent de l'écart-type théorique (6). Ces différences sont inhérentes à la nature « aléatoire » de l'expérience. Pour d'autres erreurs simulées, les moments simulés pourraient être plus proches des moments théoriques, mais pour d'autres échantillons, les différences pourraient être encore plus importantes. (iii) Voir le code.



(iv) Voir le code, les résultats sont résumés dans le tableau suivant :

OLS Regression Results

Dep. Variable:	y	R-squared:	0.543
Model:	OLS	Adj. R-squared:	0.542
Method:	Least Squares	F-statistic:	590.8
Date:	Mon, 20 Jan 2020	Prob (F-statistic):	1.25e-86
Time:	22:43:20	Log-Likelihood:	-1595.1
No. Observations:	500	AIC:	3194.
Df Residuals:	498	BIC:	3203.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.4243	0.501	0.847	0.397	-0.560	1.408
x	2.0870	0.086	24.306	0.000	1.918	2.256

Omnibus:	2.481	Durbin-Watson:	1.928
Prob(Omnibus):	0.289	Jarque-Bera (JB):	2.141
Skew:	-0.051	Prob(JB):	0.343
Kurtosis:	2.696	Cond. No.	11.3

On note que les estimateurs de la constante et de la pente (0,42 et 2,09) sont sensiblement différents des vraies valeurs des paramètres (1 et 2) utilisées pour construire l'échantillon, surtout la constante. (v) On trouve :

$$\sum_{i=1}^{500} \hat{e}_i = -7,34 \times 10^{-13} \quad \text{et} \quad \sum_{i=1}^{500} x_i \hat{e}_i = -9,61 \times 10^{-12}$$

Ces résultats sont attendus puisqu'il s'agit des conditions qui permettent d'identifier l'estimateur des MCO. En effet, celui-ci est défini par la minimisation de la somme des carrés des résidus :

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{\{\beta_0, \beta_1\}} \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2$$

où $e_i = y_i - \beta_0 - \beta_1 x_i$ est le résidu, et $\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ représentera le résidu estimé. La CNO par rapport à β_0 implique que la somme des résidus estimés doit être nulle :

$$-2 \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\Leftrightarrow \sum_{i=1}^{N} \hat{e}_i = 0$$

La CNO par rapport à β_1 implique que le produit scalaire des résidus estimés et de la variable explicative doit être nul :

$$-2\sum_{i=1}^{N} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\Leftrightarrow \sum_{i=1}^{N} x_i e_i = 0$$

(vi) Si on utilise les « vraies » erreurs, on trouve :

$$\sum_{i=1}^{500} \hat{u}_i = -72,25 \quad \text{et} \quad \sum_{i=1}^{500} x_i \hat{u}_i = 50,93$$

Il n'y a en effet aucune raison pour que les « vraies » erreurs vérifient les conditions d'identification de l'estimateur des MCO. (vii) et (viii) Voir le code. On observe que la distance entre la moyenne des estimateurs des MCO de β_1 et la vraie valeur de la pente (2) tend à décroître lorsque le nombre d'échantillons simulés s'accroît. La moyenne des $\hat{\beta}_1$, dans la coupe des échantillons simulés, est un estimateur de l'espérance de $\hat{\beta}_1$. La loi des grands nombres nous dit que cet estimateur converge vers l'espérance de $\hat{\beta}_1$ lorsque le nombre d'échantillons simulés tend vers l'infini. Ces simulations illustrent l'absence de biais de l'estimateur des MCO dans ce modèle. Même si pour une réalisation particulière de l'échantillon, $\hat{\beta}_1$ peut être éloigné de la vraie valeur de β_1, en moyenne $\hat{\beta}_1$ est égal à la vraie valeur de la pente.