Marianne Tenand - TD d'introduction à STATA®

Département d'économie de l'ENS (2016 – 2017)

Séance 2 : Travailler sur les données

CORRECTION

Exercice 1

La base de données **tabac.dta** (libre d'accès) fournit des informations au niveau individuel sur la consommation de tabac, le nombre d'années d'étude et le revenu ainsi que le prix local des cigarettes.

1. Quelle est la consommation moyenne de cigarettes dans l'échantillon ? Et le prix moyen des cigarettes ? Quelle est la consommation maximale de cigarettes observée ?

   Moyenne de 8,67 cigarettes
   Min de 0 cigarettes
   Max de 80 cigarettes
   Prix moyen : 60 (centimes par cigarette ?)

2. Quel est le pourcentage de fumeurs dans l'échantillon ?

   38,4 %

3. Estimer le modèle de régression simple suivant :

$$cigs = \alpha_0 + \alpha_1 educ + \alpha_2 pcig + \alpha_3 age + \alpha_4 rev + \alpha_5 nres + u \qquad (1)$$

   Donner une interprétation de chaque coefficient.
   Les signes sont-ils en accord avec vos attentes ?
   Le prix des cigarettes et l'interdiction de fumer dans les restaurants ont-ils un impact statistiquement significatif sur la consommation de cigarette ? Cet impact vous paraît-il substantiel ?

Source	SS	df	MS	Number of obs =	807
Model	2888.63535	5	577.72707	F(5, 801) =	3.11
Residual	148865.047	801	185.848998	Prob > F =	0.0087
				R-squared =	0.0190
Total	151753.683	806	188.280003	Adj R-squared =	0.0129
				Root MSE =	13.633

cigs	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	-.3680468	.169172	-2.18	0.030	-.7001196	-.0359739
pcig	.0046741	.1024819	0.05	0.964	-.1964908	.2058389
age	-.0438986	.028707	-1.53	0.127	-.1002484	.0124512
rev	.0001306	.000056	2.33	0.020	.0000207	.0002405
nres	-2.982192	1.130995	-2.64	0.009	-5.202256	-.7621274
_cons	13.01946	6.551227	1.99	0.047	.1598616	25.87906

Interprétation des coefficients :

- α_0 : donne la consommation de cigarettes pour un individu pour lequel les variables explicatives du modèle prennent des valeurs nulles (donc pour un individu fictif) ;
- α_1 : traduit l'impact d'une année d'étude supplémentaire sur la consommation de cigarettes
    - Négatif : les individus plus aisés ont moins de comportements à risques. Une augmentation d'une année d'étude diminue la consommation journalière de cigarettes d'un tiers d'une cigarette.
- α_2 : effet d'une augmentation de 1 centime du prix d'une cigarette sur la consommation de cigarettes
    - Négatif (élasticité-prix négative : attendu) ; magnitude faible et statistiquement non significatif au seuil de 10 % (cf. p-value)
- α_3 : mesure l'effet d'une année supplémentaire sur la consommation de cigarettes
    - Négatif : les individus plus âgés sont plus averses au risque… Mais coefficient non significatif au seuil de 10 %
- α_4 : effet de passer d'une catégorie de revenu à celle en dessus sur le nombre de cigarettes consommées
    - Positif : un peu surprenant ! Statistiquement significatif au seuil de 5 % (pas au seuil de1 %). Magnitude difficile à apprécier (car la variable de revenu est libellée en tranches)
- α_5 : mesure de combien augmente la consommation moyenne de cigarettes quand l'individu passe d'une zone où fumer est autorisé dans les restaurants à une zone où cela ne l'est pas
    - Fort effet positif, statistiquement significatif au seuil de 1 % : le fait d'interdire la cigarette dans les restaurants diminue la consommation en moyenne de 3 cigarettes par jour.

4. Estimer à nouveau le modèle mais seulement sur les fumeurs. Quelles différences avec les résultats du modèle précédent constatez-vous ?

Source	SS	df	MS		
Model	3297.95012	5	659.590024	Number of obs	310
Residual	50831.5983	304	167.209205	F(5, 304)	3.94
				Prob > F	0.0018
Total	54129.5484	309	175.176532	R-squared	0.0609
				Adj R-squared	0.0455
				Root MSE	12.931

cigs	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.7923911	.2940706	2.69	0.007	.2137196	1.371063
pcig	.0593401	.1528494	0.39	0.698	-.2414367	.360117
age	.1204724	.05022	2.40	0.017	.0216496	.2192952
rev	.0001679	.0000835	2.01	0.045	3.53e-06	.0003322
nres	-1.874924	1.879242	-1.00	0.319	-5.572893	1.823044
_cons	1.962497	9.885841	0.20	0.843	-17.49084	21.41584

L'estimation est faite maintenant avec seulement 310 observations (= le nombre de fumeurs).
La part de la variance expliquée par le modèle augmente sensiblement, ainsi que les signes de certains coefficients :

- Le niveau d'éducation a un effet positif sur le nombre moyen de cigarettes fumées
- L'âge a lui aussi un effet statistiquement significatif et positif
- L'effet de l'interdiction de fumer n'est plus significatif

5. Créer une variable binaire, qui vaut 1 pour les fumeurs et 0 pour les non-fumeurs. Introduisez cette variable comme variable explicative dans le modèle (1). Comparez les résultats obtenus à ceux obtenus aux questions 3. et 4.

Source	SS	df	MS		
Model	98692.2319	6	16448.7053	Number of obs	= 807
Residual	53061.4509	800	66.3268136	F(6, 800)	= 247.99
				Prob > F	= 0.0000
Total	151753.683	806	188.280003	R-squared	= 0.6503
				Adj R-squared	= 0.6477
				Root MSE	= 8.1441

cigs	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.2014175	.1021679	1.97	0.049	.0008687	.4019663
pcig	.0240783	.0612247	0.39	0.694	-.0961017	.1442584
age	.0390769	.0172879	2.26	0.024	.0051418	.073012
rev	.0000697	.0000335	2.08	0.038	3.96e-06	.0001354
nres	-.6367175	.678468	-0.94	0.348	-1.968505	.6950702
fum	22.86311	.6015738	38.01	0.000	21.68226	24.04396
_cons	-6.859479	3.948494	-1.74	0.083	-14.61011	.8911529

L'effet d'être fumeur a évidemment un fort impact sur la consommation moyenne de cigarettes ! L'interdiction de fumer dans les restaurants n'a plus d'effet significatif.

6. Calculer le nombre moyen de cigarettes fumées selon le nombre d'années d'études

On observe que l'évolution de la consommation moyenne de cigarettes n'est pas une fonction linéaire du nombre d'années d'étude : les consommations moyennes les plus faibles s'observent chez les moins éduqués, mais aussi chez ceux ayant 15 ans d'études.

7. Créer la variable *educ2* qui correspond au nombre d'années d'études au carré. Estimer le modèle de régression suivant :

$$cigs = \alpha_0 + \beta_1 educ + \beta_2 educ2 + \alpha_2 pcig + \alpha_3 age + \alpha_4 rev + \alpha_5 nres + u \qquad (2)$$

Que pouvez-vous en conclure à propos de l'impact du niveau d'éducation sur la consommation de cigarettes ?

Le coefficient devant les années d'études est positif, égal à 2,75. Mais le coefficient devant le nombre d'années d'études au carré est égal à -0.12 (les deux estimateurs sont significativement différents de 0 au seuil de 1 %). L'effet total d'une année supplémentaire d'étude doit se calculer en prenant en compte ses deux coefficients. Comme l'un est positif et l'autre est négatif, il existe un âge pour lequel le sens de l'effet s'inverse. On cherche *educ** tel que :

$\beta_1 + 2\beta_2 educ^* > 0$ ➔ $2\beta_2 educ^* > -\beta_1$

⇨ $educ^* < -\beta_1 / 2\beta_2$ (attention, β_2 est négatif !).

Au-delà de 11 ans d'études environ (2,75/(-2*0,12)), une année de plus d'éducation conduit à une diminution du nombre de cigarettes consommées.

On peut retrouver la forme en cloche dessinée par l'effet du niveau d'éducation sur la consommation de cigarettes

L'effet du niveau d'éducation est donc non-linéaire (qui est en réalité plus complexe qu'un effet quadratique)

Pour information : un article sur l'effet des interdictions de fumer dans les lieux publics sur la santé des fumeurs et des non-fumeurs vient d'être publié dans le journal *Health Economics* (très bonne revue en économie de la santé). Voici les références et le résumé :

Kuehnle, D*., and* Wunder, C. *(*2017*)* The Effects of Smoking Bans on Self-Assessed Health: Evidence from Germany. Health Econ*.,* 26*:* 321–337. http://onlinelibrary.wiley.com/doi/10.1002/hec.3310/full

We examine the effects of smoking bans on self-assessed health in Germany taking into account heterogeneities by smoking status, gender and age. We exploit regional variation in the dates of enactment and dates of enforcement across German federal states. Using data from the German Socio-Economic Panel, our difference-in-differences estimates show that non-smokers' health improves, whereas smokers report no or even adverse health effects in response to bans. We find statistically significant health improvements especially for non-smokers living in households with at least one smoker. Non-smokers' health improvements materialise largely with the enactment of smoking bans.

Ce n'est évidemment pas le seul article à avoir traité ce sujet ! Beaucoup de travaux antérieurs ont utilisé les différences inter-régionales (ou entre états dans les pays fédéraux) dans la sévérité de la législation anti-tabac pour évaluer l'effet des interdictions de fumer ou celui de la hausse du prix du tabac. L'hypothèse d'identification cruciale dans cette approche est que les différences de législation sont dues à des facteurs qui n'ont pas d'effet direct sur la santé (ou autre variable d'outcome).

Exercice 2

La base de données **busind** fournit des informations sur le revenu national brut par habitant (GNI per capita), le nombre de jours nécessaires pour créer une entreprise et faire appliquer un type de contrat donné, pour un échantillon de 135 pays. Cette base est extraite de la base plus riche « Doing Business » constituée par la Banque Mondiale. Elle vise à recueillir des indicateurs sur la plus ou moins grande facilité à mener à bien une entreprise économique, et notamment sur les lourdeurs administratives.

NB : La base de données busind ne vous est pas fournie d'emblée. Deux bases distinctes vous sont fournies :

- GNIpc.xls

- business.xls

Chacune de ces bases (en format Excel et non en format .dta) vous fournit des informations différentes. Il faut donc procéder en 2 étapes :

1. Convertir les données en format .dta. Vous pouvez le faire en copiant-collant les données depuis le tableur Excel dans l'éditeur Stata.

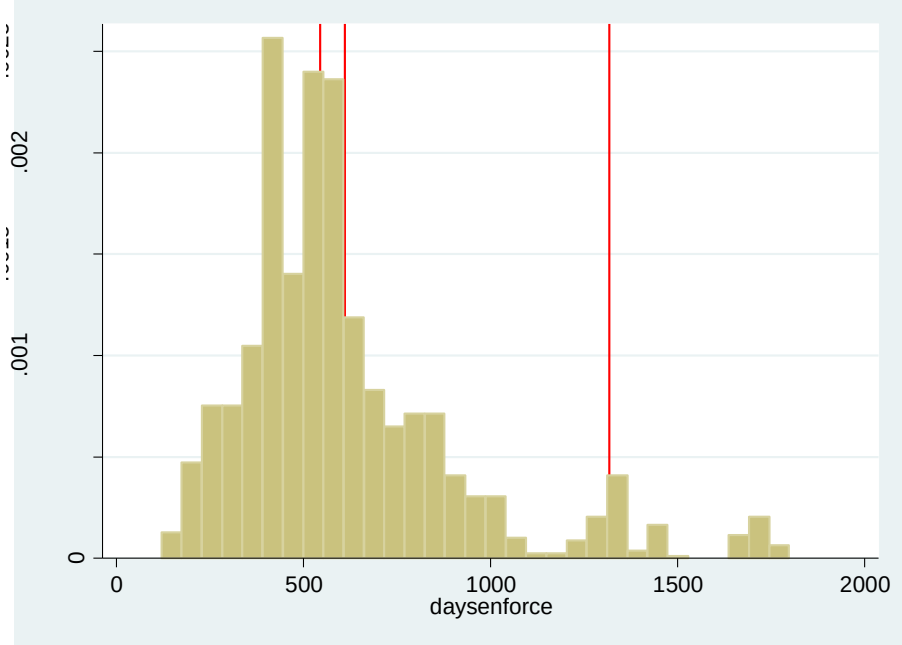
2. Apparier les deux bases de données sur les deux variables en commun, le pays et l'année. Attention à bien harmoniser au préalable les noms des variables.

Une fois que vous avez réussi à créer la base busind.dta, vous pouvez répondre aux questions suivantes :

/!\ Pour réaliser ce corrigé, j'ai utilisé la base comprenant 1 446 observations. Il manque donc les observations correspondant aux pays qui ont un accent circonflexe ou un autre caractère spécial dans leur nom. Ce n'est pas très embêtant pour cet exercice, mais dans une « vraie » application il faut être extrêmement vigilant avec les importations de données : il y souvent des bugs, mais ceux-ci sont plus ou moins apparents (disparition d'observations, problème de format des données, premières lignes qui correspondent non pas à des observations mais à des commentaires, etc.).

1. Déterminer le revenu national brut moyen par habitant, le nombre moyen de jours nécessaires pour créer une entreprise et le nombre de jours moyens pour appliquer un type de contrat donné.

   Revenu national brut moyen : 14 810$ (courants, PPP)
   Il faut en moyenne 39 jours pour ouvrir une entreprise et 611 jours pour faire appliquer un contrat.
   La moyenne du délai pour faire appliquer un contrat est très élevée → nous invite à regarder la distribution de cette variable. La médiane est à 545 jours, et pour 5 % des observations (couple pays*année) il faut en moyenne plus de 1 318 jours pour faire appliquer un contrat !



2. Dans combien de pays faut-il en moyenne moins de 5 jours pour ouvrir une entreprise ? Quel est le nombre de jours maximum pour créer son entreprise dans la base de données ? A quel pays ce record correspond ?

   /!\ Petite subtilité : il faut d'abord créer une variable qui calcule pour chaque pays la moyenne sur les différentes années du nombre de jours nécessaires pour créer une entreprise.
   Dans 4 pays ce nombre est en moyenne inférieur à 5.
   Le maximum est de 690,6 jours.
   Ce maximum est atteint au Suriname.

3. Quels est le nombre de jours minimum sur lesquels il faut compter pour créer une entreprise dans les 10 % des pays pour lesquels il est le plus long de réaliser cette opération

- Indice : utiliser la commande `sum variable, d`

Pour répondre à cette question, on est obligé de prendre une année donnée (pour que tous les pays ne soient comptés qu'une fois).

Par exemple, le 90ème percentile (valeur au-delà de laquelle se situe 10 % des pays) était de 62 en 2010, et il fallait même au minimum 93 jours pour ouvrir une entreprise dans les 5 % des pays les plus lents au regard de ce critère.

4. Estimer le modèle de régression linéaire suivant :

$$gnipc = \alpha_0 + \alpha_1 daysopen + u \qquad (a)$$

Donner une interprétation des deux coefficients. Les signes sont-ils cohérents avec ce à quoi vous vous attendiez ?

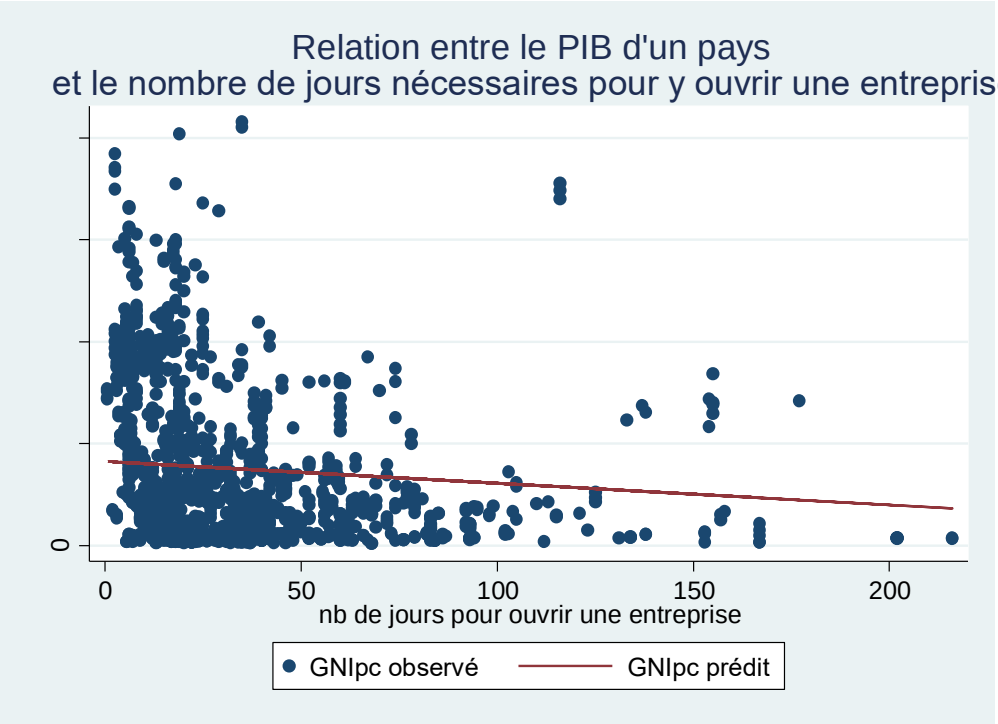
Source	SS	df	MS	Number of obs =	1446
				F(1, 1444) =	31.89
Model	9.3789e+09	1	9.3789e+09	Prob > F =	0.0000
Residual	4.2467e+11	1444	294090472	R-squared =	0.0216
				Adj R-squared =	0.0209
Total	4.3405e+11	1445	300377526	Root MSE =	17149

gnipc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
daysopen	-42.81502	7.581603	-5.65	0.000	-57.68716	-27.94289
_cons	16484.09	539.818	30.54	0.000	15425.17	17543

L'effet du délai d'ouverture d'une entreprise sur le revenu national par habitant est à négatif (ce à quoi on pourrait s'attendre). La p-value associée étant très faible (0.000), cet effet est statistiquement significatif aux seuils conventionnels (1%, et a fortiori 5% et 10%). Avec ce modèle, on prédit qu'un jour de plus nécessaire pour ouvrir une entreprise est associée à une baisse de 43 $ du PIB par tête.

5. Quels sont les facteurs regroupés dans *u* ? Sont-ils susceptibles d'être corrélés avec le nombre de jours nécessaires pour créer sa propre entreprise ? Qu'en déduisez-vous pour l'interprétation du modèle ?

On peut penser à énormément de facteurs liés à la fois au revenu national et aux délais pour ouvrir une entreprise. Si le pays connait une guerre ou une crise politique, il est probable qu'ouvrir une entreprise soit long et que le pays ait un revenu par habitant faible, sans que les difficultés à ouvrir une entreprise soit la cause (principale) du faible niveau de revenu dans le pays. Il faut donc se garder de toute interprétation causale d'un modèle pour lequel on attend des biais de variables omises importants.



NB : pour réaliser le graphique, les observations correspondant à des pays dans les 1% les plus riches ou dans les 1% les plus lents à l'ouverture d'une entreprise ont été enlevées.

6. Quel est, selon ce modèle, le revenu national brut moyen prédit pour un pays où il faut 5 jours pour créer son entreprise ? Et le revenu prédit pour un pays où il faut 200 jours ? Montrer comment vous pouvez calculer vos réponses à la main, après avoir obtenu les réponses en utilisant STATA. Est-ce que les valeurs obtenues vous semblent vraisemblables ? Expliquer.

   Les coefficients estimés du modèle conduisent à prédire que dans un pays où il faut 5 jours en moyenne pour créer son entreprise le revenu national brut par habitant devrait être de 16 270 $ (courant, en PPP).
   Dans un pays où le nombre de jours nécessaire s'élèvent à 200 jours, le revenu prédit est de 7 291 $ par tête.

7. Estimer le modèle suivant :

$$gnipc = \alpha_0 + \alpha_1 daysenforce + u \quad (b)$$

   Comment interprétez-vous α_1 ?

   α_1 mesure l'effet d'un jour de plus pour faire appliquer un contrat sur le revenu national. Le coefficient estimé est de -8.7. Il y a un effet négatif du délai d'exécution des contrats, statistiquement significatif au seuil de 1%.

Source	SS	df	MS	Number of obs =	1446
Model	9.8677e+09	1	9.8677e+09	F(1, 1444) =	33.59
Residual	4.2418e+11	1444	293751975	Prob > F =	0.0000
				R-squared =	0.0227
				Adj R-squared =	0.0221
Total	4.3405e+11	1445	300377526	Root MSE =	17139

gnipc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
daysenforce	-8.678269	1.497325	-5.80	0.000	-11.61543	-5.741104
_cons	20116.67	1020.733	19.71	0.000	18114.39	22118.94

8. Comparer les résultats des modèles (a) et (b). Lequel explique une plus grande partie de la variation des revenus par habitant observée entre les pays ? Pouvez-vous en déduire si le nombre de jours pour créer son entreprise ou le nombre de jours pour faire appliquer un contrat est davantage corrélé avec le revenu par habitant ?

   On trouve que l'effet estimé d'un jour nécessaire en plus pour ouvrir son entreprise est plus élevé (5 fois plus fort) que l'effet estimé d'un jour nécessaire en plus pour faire appliquer un contrat. Cependant, la comparaison des coefficients de détermination (R^2, un peu plus élevé pour le second modèle) montre que la variation dans le nombre de jours nécessaires pour ouvrir une entreprise, d'une observation à l'autre, explique un peu moins de variation dans le revenu national par tête que la variation dans le nombre de jours nécessaires pour faire appliquer un contrat. Il faut bien avoir en tête que la magnitude de l'effet d'une variable X sur une autre variable Y ne permet pas de prédire le pouvoir explicatif de X en termes de variations de Y.

   On notera que, comme le montre la faiblesse des R^2, les deux modèles ont très peu de pouvoir explicatif, ce qui n'est pas surprenant du fait de leur parcimonie (beaucoup d'autres variables que les contraintes administratives expliquent le revenu par tête d'un pays !).

   Observez également que la corrélation est plus forte (en valeur absolue) entre revenu national et nombre de jours pour appliquer un contrat est plus forte que la corrélation entre revenu national et nombre de jours pour ouvrir une entreprise (*commande Stata « corr » suivi des noms des deux variables dont on veut calculer le coefficient de corrélation*). C'est logique puisque, dans un modèle de régression univariée, le R^2 est tout simplement égal au carré du coefficient de corrélation entre la variable dépendante et l'unique variable explicative.

9. Estimer le modèle suivant :

$$log(gnipc) = \alpha_0 + \alpha_1 daysopen + u \qquad (a')$$

   Quelle est maintenant l'interprétation de α_1 ?

   Un jour de plus nécessaire à l'ouverture d'une entreprise augmente en moyenne le revenu national de α_1 %.

On appelle cette spécification une spécification « log-level ».

10. D'après vos résultats, quelles conclusions pouvez-vous tirer à propos des politiques qui visent à réduire le nombre de jours pour créer son entreprise dans certains pays en développement (programmes de la Banque Mondiale notamment) ?

    Eh bien pas grand-chose ! On serait tentés, au vu des signes des estimateurs, de préconiser de réduire les lourdeurs administratives… Mais il faut se garder de faire des recommandations de politiques publiques : le manque de pouvoir explicatif du modèle suggère qu'il existe un grand nombre de variables à prendre en compte conjointement aux lourdeurs administratives avant de conclure quant à leurs effets causaux sur la croissance.

11. La base de données contient 156 pays, soit beaucoup moins que l'ensemble des pays existant à l'heure actuelle (et au moment de la collecte des données). Pensez-vous qu'il faille tenir compte de cette information pour interpréter les résultats ? Pourquoi ?

    Il faut tenir compte de cette information, même si a priori on ne sait pas bien « comment »… Pour certaines raisons, nous n'avons pas d'information sur les autres pays. Est-ce que ce sont des pays plutôt riches ? Plutôt « business friendly » ? Ou des pays qui n'étaient pas dans le « réseau » de la Banque Mondiale ? etc.