

SÉANCE 4

M1 ACTUARIAT

Exercices

Exercice 1 : Lutte contre les infections nosocomiales

Tous les établissements de santé sont concernés par la lutte contre les infections nosocomiales. L'indicateur de consommation de solutions hydro-alcooliques pour l'hygiène des mains au sein de chaque établissement de santé (*icsha2*) est un marqueur indirect de la mise en oeuvre effective de l'hygiène des mains, une mesure-clé de prévention de nombreuses infections nosocomiales. A partir des données de 2011 d'un échantillon représentatif de 2790 établissements de santé français, on vous demande d'examiner le niveau d'engagement des établissements dans ce domaine et en particulier les disparités potentielles entre régions et type d'établissements.

Partie 1 : Pour ce faire, on dispose des résultats de la régression linéaire par moindres carrés ordinaires suivante :

. reg icsha2 i.Ambulatoire i.region						
Source	SS	df	MS	Number of obs	=	2,790
Model	91038.9262	4	22759.7315	F(4, 2785)	=	3.14
Residual	20209086.8	2,785	7256.40458	Prob > F	=	0.0139
Total	20300125.7	2,789	7278.63954	R-squared	=	0.0045
				Adj R-squared	=	0.0031
				Root MSE	=	85.185

icsha2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
1.Ambulatoire	41.45095	13.93261	2.98	0.003	14.13167 68.77023
region					
1	-3.416809	4.720231	-0.72	0.469	-12.67231 5.838696
2	-11.08198	5.591765	-1.98	0.048	-22.0464 -0.1175561
3	-2.683385	5.359603	-0.50	0.617	-13.19258 7.825811
_cons	88.19361	1.993815	44.23	0.000	84.28411 92.10312

où *icsha2*, exprimé en pourcentage, est le rapport entre le volume de produits hydro-alcooliques consommé réellement par l'établissement et son objectif personnalisé de consommation vers lequel il doit tendre (cet objectif est déterminé à partir d'un référentiel national). *Ambulatoire* est une variable indicatrice qui vaut 1 si l'établissement propose des soins en ambulatoire dans le cadre d'une hospitalisation (c'est-à-dire une prise en charge médicale du patient pour certaines chirurgies, dialyses, chimiothérapies, etc. sans hospitalisation prolongée, d'une durée de quelques heures, permettant aux patients de rentrer chez eux le jour même de l'intervention) et 0 sinon. La variable *region* a été codée comme suit : 1 pour Ile-de-France, 2 pour Rhône-Alpes, 3 pour Provence-Alpes-Côte d'Azur, 0 Autres.

1. Pourquoi dans la régression précédente, il n'y a pas de coefficient représentant la catégorie *region* = 0 (c'est-à-dire *Autre région*) ?
2. Au vu des résultats de la régression, que peut-on dire sur les niveaux d'engagement des établissements dans la lutte contre les infections nosocomiales ?
3. Quelles auraient été les valeurs des estimateurs MCO des paramètres $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ si on avait estimé le modèle : $icsha2 = \beta_0 + \beta_1 Ambulatoire + \beta_2 IledeFrance + \beta_3 PACA + \beta_4 Autres + u$?

Partie 2 : En complément, une deuxième régression a été effectuée :

. reg icsha2 i.Ambulatoire i.region i.region#i.Ambulatoire						
Source	SS	df	MS	Number of obs	=	2,790
Model	253812.18	7	36258.8829	F(7, 2782)	=	5.03
Residual	20046313.5	2,782	7205.72017	Prob > F	=	0.0000
Total	20300125.7	2,789	7278.63954	R-squared	=	0.0125
				Adj R-squared	=	0.0100
				Root MSE	=	84.887

icsha2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
1.Ambulatoire	108.8288	20.10686	5.41	0.000	69.40292 148.2547
region					
1	-.7475693	4.752614	-0.16	0.875	-10.06658 8.571437
2	-9.209631	5.628696	-1.64	0.102	-20.24647 1.827211
3	-1.645127	5.375983	-0.31	0.760	-12.18645 8.896193
region#Ambulatoire					
1 1	-147.0935	33.81534	-4.35	0.000	-213.3992 -80.78785
2 1	-121.0681	40.40982	-3.00	0.003	-200.3044 -41.83189
3 1	-94.99098	47.22975	-2.01	0.044	-187.5999 -2.38209
_cons	87.53232	1.991961	43.94	0.000	83.62645 91.43819

1. Ecrire la spécification du modèle correspondant à cette estimation en mentionnant comment les variables ont été construites.
2. Interpréter les résultats de la régression en mettant en perspectives les différences entre régions et types de soins proposés.
3. Calculer la moyenne du score icsha2 pour les établissements proposant des soins ambulatoires en Ile-de-France.
4. Au vu de la régression, peut-on affirmer que les établissements de la région Rhône-Alpes remplissent moins bien leurs objectifs en termes de lutte contre les maladies nosocomiales ? Justifier votre réponse.
5. Proposer un classement des établissements (en fonction de leur région et de leur type) des plus performants en termes de prévention jusqu'aux moins performants ? Justifier votre réponse.
6. Proposer une spécification économétrique et expliquer en détails le(s) test(s) statistique(s) (formule de statistique de test, règle de décision, etc) qui permettraient de savoir si l'écart entre établissements avec et sans soins ambulatoires dans les scores icsha2 est croissant avec le nombre de patients accueillis dans les établissements ?

Exercice 2 : Analyse du salaire Homme/femmes

Un économiste spécialisé en économie du travail est intéressé par l'analyse des déterminants du salaire. En particulier, il cherche à étudier la relation liant la rémunération horaire des individus et différentes caractéristiques telles que leur niveau d'éducation, leur âge, le nombre d'heures travaillées, etc. Pour ce faire, il dispose d'un échantillon de 2813 hommes et de 2240 femmes pour lesquels il a les informations suivantes :

- Sexe = 1 si homme et 0 sinon
- Heures annuelles de travail
- Revenu hors travail
- Salaire : salaire horaire
- Age : âge en années
- Education : indice pour le niveau d'éducation variant de 1 à 5.

- Ville 1 : Variable dichotomique - Ville > 1 000 000 habitants
 - Ville 2 : Variable dichotomique - 500 000 < Ville ≤ 1 000 000 habitants
 - Ville 3 : Variable dichotomique - 100 000 < Ville ≤ 500 000 habitants
 - Ville 4 : Variable dichotomique - Ville < 100 000 habitants
 - Enfant6 : Nombre d'enfants de moins de 6 ans.
- Le modèle économétrique estimé s'écrit :

$$\begin{aligned} \text{Salaire} = & \beta_1 + \beta_2 \text{Revenu} + \beta_3 \text{Age} + \beta_4 \text{Education} + \beta_5 \text{Ville1} + \beta_6 \text{Ville3} + \\ & \beta_7 \text{Ville4} + \beta_8 \text{Enfant6} + \beta_9 \text{Heure} + \beta_{10} \text{Sexe} + u \end{aligned}$$

L'estimation par les moindres carrés ordinaires (MCO), effectuée sous le logiciel Stata, conduit aux résultats suivants (les ??? correspondent à des résultats qui ont été volontairement enlevés) :

. regress salaire revenu age education ville1 ville3 ville4 enfant6 heure sexe						
Source	SS	df	MS	Number of obs = 5053		
Model	104358.693	9	11595.4103	F(9, 5043) = ??????		
Residual	267925.354	5043	53.1281686	Prob > F = 0.0000		
Total	372284.047	5052	73.690429	R-squared = ??????		
				Adj R-squared = ??????		
				Root MSE = 7.2889		

salaire	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
-----	-----	-----	-----	-----	-----	-----
revenu	.0000523	.0000118	4.42	0.000	.0000291	.0000755
age	.1169023	.0133854	?????	0.000	.0906612	.1431434
education	1.675906	.0701702	23.88	0.000	???????	???????
ville1	3.26355	.2193638	14.88	0.000	2.833502	3.693598
ville3	-.6040621	.4140037	-1.46	0.145	-1.415689	.207565
ville4	3.179761	.2457474	12.94	0.000	2.697989	3.661532
enfant6	.1388817	.1541526	?????	0.368	-.1633244	.4410878
heure	-.0006906	.0001516	-4.55	0.000	-.0009879	-.0003934
sexe	4.88224	.2331337	20.94	0.000	4.425197	5.339284
_cons	-6.535071	.6784582	-9.63	0.000	-7.865144	-5.204998
-----	-----	-----	-----	-----	-----	-----

1. Retrouver le coefficient de détermination R^2 et \bar{R}^2 (R-squared et Adj R-squared). Quelles sont les particularités de ces deux coefficients ? Quelle conclusion peut être tirée sur la qualité de l'ajustement ?
2. Donner une estimation de la variance résiduelle.
3. A l'aide d'un test de Fisher, tester l'hypothèse $\beta_2 = \beta_3 = \dots = \beta_{10} = 0$ (retrouver la valeur de $F(9, 5043)$ et conclure sur la significativité de la régression).
4. Tester la significativité des coefficients et interpréter les résultats obtenus. Calculer les statistiques "t" manquantes.
5. Déterminer l'intervalle de confiance (à 95%) pour le coefficient associé à la variable "éducation" (cf. valeurs manquantes dans les résultats).
6. Donner une prévision du salaire horaire d'une femme âgée de 35 ans, n'ayant pas de revenus hors travail, pas d'enfants, travaillant 1900 heures par an, habitant une ville de 600 000 habitants et ayant un niveau d'éducation égal à 5.
7. *Tests de contraintes linéaires* : L'économiste veut tester désormais les restrictions conjointes : $\beta_4 = 1.5$ et $\beta_5 = \beta_7$.
 - (a) Dans le cours, nous avons vu que ce test pouvait être effectué grâce à l'estimation de deux modèles : un modèle général et un modèle constraint. Donner l'expression de ces deux modèles et suggérer la méthode (statistique de test, règle de décision) qui permettrait de réaliser le test voulu.

- (b) Sous Stata, la commande “test” effectue le test des contraintes linéaires à partir de la seule estimation du modèle non contraint. Les résultats sont présentés ci-dessous :

```
. test (education=1.5) (ville1=ville4)

( 1) education = 1.5
( 2) ville1 - ville4 = 0

F(  2,  5043) =     3.21
Prob > F =    0.0406
```

Donner la formule qui a permis de calculer $F(2,5043)$ et conclure sur la pertinence des hypothèses imposées.

8. *Test de stabilité des paramètres* : Notre économiste spécialisé en économie du travail s'intéresse particulièrement à l'influence du sexe sur la rémunération.

- (a) L'introduction d'une variable qualitative ($\text{sexe}=0$ ou 1) est-elle totalement pertinente pour étudier l'effet du sexe sur les salaires ?
- (b) Trois estimations sont effectuées : la première sur l'échantillon total, la seconde sur uniquement les hommes et la dernière sur uniquement les femmes. Est-ce qu'à l'aide de ces régressions, on peut préciser l'analyse ?

Estimation sur l'échantillon total :

```
. regress salaire revenu age education ville1 ville3 ville4 enfant6 heure
```

Source	SS	df	MS	Number of obs = 5053		
Model	81058.8577	8	10132.3572	F(8, 5044) = 175.49		
Residual	291225.19	5044	57.7369527	Prob > F = 0.0000		
Total	372284.047	5052	73.690429	R-squared = 0.2177		
				Adj R-squared = 0.2165		
				Root MSE = 7.5985		
<hr/>						
salaire	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
revenu	.0000513	.0000123	4.16	0.000	.0000271	.0000755
age	.1706277	.0136952	12.46	0.000	.1437791	.1974762
education	1.650021	.0731391	22.56	0.000	1.506636	1.793405
ville1	3.267379	.2286806	14.29	0.000	2.819066	3.715693
ville3	-.7054227	.4315578	-1.63	0.102	-1.551464	.1406181
ville4	2.83367	.2556048	11.09	0.000	2.332573	3.334766
enfant6	.5684684	.1592706	3.57	0.000	.2562288	.880708
heure	.0006846	.0001425	4.81	0.000	.0004053	.0009639
_cons	-8.318762	.7016783	-11.86	0.000	-9.694356	-6.943168
<hr/>						

Estimation sur l'échantillon des hommes :

```
. regress salaire revenu age education ville1 ville3 ville4 enfant6 heure
if sexe==1
```

Source	SS	df	MS	Number of obs = 2813		
Model	60437.6161	8	7554.70201	F(8, 2804) = 111.55		
Residual	189897.525	2804	67.7237963	Prob > F = 0.0000		
Total	250335.141	2812	89.0238766	R-squared = 0.2414		
				Adj R-squared = 0.2393		
				Root MSE = 8.2294		
<hr/>						
salaire	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
revenu	.0000463	.0000149	3.11	0.002	.0000171	.0000755
age	.1584387	.0196248	8.07	0.000	.1199581	.1969193
education	1.854258	.1046637	17.72	0.000	1.649033	2.059484
ville1	3.758862	.3314454	11.34	0.000	3.10896	4.408763
ville3	-.3553427	.6297836	-0.56	0.573	-1.590229	.8795436
ville4	4.088821	.3721086	10.99	0.000	3.359186	4.818455
enfant6	.2316143	.2261213	1.02	0.306	-.2117668	.6749954
heure	-.001408	.0002338	-6.02	0.000	-.0018664	-.0009496
_cons	-3.465295	1.094066	-3.17	0.002	-5.610551	-1.32004
<hr/>						

Estimation sur l'échantillon des femmes :

```
. regress salaire revenu age education ville1 ville3 ville4 enfant6 heure
if sexe==0
```

Source	SS	df	MS	Number of obs = 2240		
Model	21015.7863	8	2626.97328	F(8, 2231) = 79.83		
				Prob > F = 0.0000		

Residual	73412.7901	2231	32.9057777	R-squared	=	0.2226
Total	94428.5763	2239	42.1744423	Adj R-squared	=	0.2198
				Root MSE	=	5.7364
<hr/>						
salaire	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
revenu	.0000779	.0000213	3.66	0.000	.0000362	.0001197
age	.0592558	.0165694	3.58	0.000	.0267628	.0917488
education	1.447657	.084729	17.09	0.000	1.281501	1.613813
ville1	2.610704	.2599715	10.04	0.000	2.100893	3.120516
ville3	-.8614145	.4865582	-1.77	0.077	-1.815569	.0927398
ville4	2.10793	.2908743	7.25	0.000	1.537517	2.678342
enfant6	.1992775	.1922788	1.04	0.300	-.1777867	.5763416
heure	.0001354	.0001779	0.76	0.447	-.0002134	.0004843
_cons	-3.621694	.8166567	-4.43	0.000	-5.22318	-2.020207

Exercice 3 : Rendement salarial de l'éducation

Sur un échantillon de 500 individus, on obtient les résultats suivants concernant une régression du salaire horaire sur le niveau d'éducation :

$$\hat{W}_i = 12.68 + 2.78BAC2_i + 6.72BACSUP2_i \quad (1)$$

où W est le salaire horaire en euros, $BAC2$ est une variable prenant la valeur 1 si l'individu a comme diplôme le plus élevé un diplôme de niveau bac+2, 0 sinon et $BACSUP2$ est une variable prenant la valeur 1 si l'individu a comme diplôme le plus élevé un diplôme de niveau supérieur à bac+2, 0 sinon.

1. Soit $Bac2inf$ une variable prenant la valeur 1 si l'individu a un diplôme de niveau inférieur à bac+2, 0 sinon. Pourquoi ne pas avoir introduit $Bac2inf$ comme variable explicative supplémentaire dans le modèle (1) ?
2. Quel est le salaire horaire estimé d'un individu ayant un diplôme de niveau bac+2 ?
3. Quelle est la différence de salaire estimée entre un individu d'un niveau bac+2 et un autre d'un niveau bac+5 ?
4. Quelles sont les valeurs des estimateurs des MCO des paramètres β_0 , β_1 , β_2 pour le modèle : $W_i = \beta_0 + \beta_1 BAC2inf_i + \beta_2 BACSUP2_i + u_i$?
5. Proposer une spécification économétrique et expliciter en détails le test statistique (formule de statistique de test, règle de décision, etc) qui permettraient de savoir si les individus ayant un diplôme de niveau bac+3 ont des salaires plus faibles que les individus ayant un niveau de diplôme bac+5.
6. Proposer une spécification économétrique et expliquer en détails le(s) test(s) statistique(s) (formule de statistique de test, règle de décision, etc) qui permettraient de savoir si l'écart de salaire entre hommes et femmes est plus marqué parmi les non-diplômés.