

SÉANCE 3

M1 ACTUARIAT

Exercices

Exercice 1 : Effet du temps de travail sur les salaires

On s'intéresse à l'impact des caractéristiques individuelles sur le salaire horaire. On dispose pour cela d'un échantillon de 34 209 travailleurs, pour qui on observe les informations suivantes :

- le sexe (*sex* = 1 pour les femmes, 0 pour les hommes) ;
- l'âge de fin d'étude (*adfe*) ;
- l'expérience (*exp*) ;
- le nombre d'heures travaillées (*hh*).

On estime dans un premier temps le modèle suivant :

$$\ln_{sal} = \beta_0 + \beta_1 sexe_i + \beta_2 adfe_i + \beta_3 exp_i + \beta_4 ln_{hh} + \varepsilon_t$$

avec \ln_{sal} qui correspond à la variable de salaire exprimée en logarithme, et \ln_{hh} la variable du nombre d'heures travaillées, elle aussi exprimée en logarithme.

L'estimation du modèle par la méthode des MCO nous donne les résultats suivants :

```
. regress ln_sal sexe adfe exp ln_hh
```

Source	SS	df	MS	Number of obs	=	34209
Model	4231.61527	4	1057.90382	F()	=
Residual	4619.3454	34204	.135052783	Prob > F	=	
Total	8850.96067	34208	.258739496	R-squared	=	

ln_sal	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
sex	-.1862118	.0041758			
adfe	.0708604	.0007403	95.72	0.000	.0694094 .0723114
exp	.0140278	.0002902	48.34	0.000	.013459 .0145967
ln_hh	.9361701	.0077067	121.47	0.000	.9210647 .9512756
_cons	2.267233	.0328055	69.11	0.000	2.202933 2.331533

1. La variable sexe est-elle significative à un seuil de 5% ? Justifier votre réponse et interpréter.
2. Déterminer l'intervalle de confiance (à 95%) pour le coefficient associé à la variable sexe et donner en une interprétation.

3. Le modèle est-il globalement significatif ?
4. Calculer et interpréter le coefficient de détermination. Que pouvez-vous en conclure sur la qualité de l'ajustement du modèle ?
5. On décide maintenant d'intégrer les formes quadratiques de l'âge de fin d'étude (*adfe* * *adfe*, soit *adfe2*) et de l'expérience (*exp* * *exp*, soit *exp2*). On obtient les résultats suivants :

```
. regress ln_sal sexe adfe adfe2 exp exp2 enfcc90 ln_hh
```

Source	SS	df	MS	Number of obs = 34209
Model	4285.98152	7	612.283074	F(7, 34201) = 4587.25
Residual	4564.97915	34201	.13347502	Prob > F = 0.0000
Total	8850.96067	34208	.258739496	R-squared = 0.4842 Adj R-squared = 0.4841 Root MSE = .36534

ln_sal	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
sexe	-.1919385	.0041687	-46.04	0.000	-.2001092 -.1837678
adfe	.1385335	.0047085	29.42	0.000	.1293046 .1477623
adfe2	-.0016162	.0001136	-14.23	0.000	-.0018388 -.0013936
exp	.0241956	.001287	18.80	0.000	.021673 .0267182
exp2	-.0002025	.0000275	-7.35	0.000	-.0002565 -.0001485
enfc90	-.0002968	.0018809	-0.16	0.875	-.0039834 .0033898
ln_hh	.9297629	.0076888	120.92	0.000	.9146925 .9448333
_cons	1.499996	.0541422	27.70	0.000	1.393875 1.606116

- (a) Quel est l'intérêt d'intégrer ces deux nouvelles variables ?
- (b) Ces deux nouvelles variables permettent-elles d'améliorer la qualité d'ajustement du modèle ? Quelle statistique doit-on regarder ?

Sous Stata, la commande *test* nous renvoie le résultat suivant :

```
( 1) adfe2 = 0
( 2) exp2 = 0

F( 2, 34201) = 201.49
Prob > F = 0.0000
```

- (c) De quel test s'agit-il ? Comment l'interpréter ?
- (d) L'ajout des formes quadratiques de l'âge à la fin des études et de l'expérience est-il pertinent ?
6. On souhaite également contrôler de l'effet du type de ménage (personne seule, plusieurs personnes, famille monoparentale, couple sans enfant ou couple avec enfant), soit la variable *typmen* avec 5 modalités.
 - (a) Quelle stratégie adopter pour prendre en compte l'effet du type de ménage ?

Stata nous renvoie les résultats suivants :

Source	SS	df	MS	Number of obs = 34209 F(11, 34197) = 2927.04 Prob > F = 0.0000 R-squared = 0.4849 Adj R-squared = 0.4848 Root MSE = .36511			
Model	4292.20217	11	390.200198				
Residual	4558.7585	34197	.133308726				
Total	8850.96067	34208	.258739496				

ln_sal	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
sexe	-.1900816	.0042381	-44.85	0.000	-.1983885 -.1817748
adfe	.1374707	.0047089	29.19	0.000	.1282411 .1467004
adfe2	-.0015908	.0001136	-14.00	0.000	-.0018134 -.0013681
exp	.0245085	.0012923	18.96	0.000	.0219755 .0270416
exp2	-.0002091	.0000277	-7.55	0.000	-.0002634 -.0001548
enfc90	-.0058527	.0027901	-2.10	0.036	-.0113214 -.000384
ln_hh	.9297178	.0076869	120.95	0.000	.9146511 .9447844
pers_seul	.0024539	.0078631	0.31	0.755	-.012958 .0178658
pls	-.0888115	.0177238	-5.01	0.000	-.1235508 -.0540722
monop	-.0150647	.00995	-1.51	0.130	-.0345669 .0044376
couple_ac_enf	.0158079	.0076594	2.06	0.039	.0007953 .0308205
_cons	1.505973	.0542357	27.77	0.000	1.399669 1.612277

- (b) Que dire de la qualité d'ajustement du modèle ?
- (c) Que peut-on dire de l'effet du type de ménage sur les salaires ?
- (d) Peut-on affirmer que les personnes en situation de couple avec au moins un enfant (*couple_ac_enf*) gagnent en moyenne plus que les personnes qui vivent à plusieurs mais sans enfant (*pls*) ?
- (e) Quel test doit-on utiliser pour différencier l'effet de chaque type de ménage sur le salaire ?
- (f) Le test Stata nous renvoie le résultat ci-dessous, que pouvez-vous conclure ?

```
test (pls==couple_ac_enf)

( 1)  pls - couple_ac_enf = 0

      F(  1, 34197) =   34.13
      Prob > F = 0.0000
```

Exercice 2 : Coût d'un véhicule

Afin d'étudier comment varie le coût de maintenance d'un véhicule utilitaire en fonction de l'âge de celui-ci, une entreprise a collecté les données suivantes :

Age (x_1) (en mois)	Coût annuel (y) (en centaine d'euros)
15	48
8	43
36	77
41	89
16	50
8	40
21	56
21	62
53	100
10	47
32	71
17	58
58	102
6	35
20	60

Les valeurs suivantes ont été calculées :

$$\begin{aligned} \sum_{i=1}^{15} x_{1i} &= 362 & \sum_{i=1}^{15} x_{1i}^2 &= 12490 & \sum_{i=1}^{15} x_{1i}y_i &= 27437 \\ \sum_{i=1}^{15} y_i &= 938 & \sum_{i=1}^{15} y_i^2 &= 64926 \end{aligned}$$

Nombre d'observations : $N = 15$.

Partie 1

On cherche à estimer les coefficients d'une régression linéaire entre les variables x_1 et y , de la forme :

$$y_i = a_0 + a_1 x_{1i} + u_i \quad (1)$$

On suppose que $u \sim \mathcal{N}(0, \sigma^2 I_N)$.

- Démontrer la formule permettant de calculer \hat{a}_0 et \hat{a}_1 , les estimateurs des MCO des paramètres a_0 et a_1 .

Montrer que ces estimateurs sont sans biais.

Calculer la valeur de ces estimateurs à partir des données et interpréter les résultats.

2. Rappeler l'équation de décomposition de la variance.

Sachant que la valeur de la somme des carrés expliqués (SCE) par le modèle est :

$$\sum_{i=1}^{15} (\hat{y}_i - \bar{\hat{y}})^2 = 6137.71889, \text{ calculer } \sum_{i=1}^{15} e_i^2, \text{ la somme des carrés des résidus (SCR).}$$

3. Calculer le coefficient de détermination R^2 .

Interpréter littérairement la valeur obtenue.

4. Donner la formule d'un estimateur sans biais de σ^2 (démonstration non demandée).

Calculer sa valeur.

En déduire une estimation des variances de \hat{a}_0 et \hat{a}_1 .

5. Expliquer la construction d'un intervalle de confiance au seuil α pour \hat{a}_1 .

Le calculer pour $\alpha = 5\%$.

Interpréter littérairement.

6. Tester si les coefficients a_0 et a_1 sont significativement différents de 0 au seuil $\alpha = 5\%$.

Interpréter littérairement les résultats des tests.

7. Déterminer une prévision du coût de maintenance pour un véhicule de 4 ans.

Calculer son intervalle de confiance au seuil de 5 %.

Interpréter littérairement.

Partie 2

On souhaite maintenant comparer les résultats de l'estimation du modèle (1) avec ceux de l'estimation du modèle suivant :

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + u_i \quad (2)$$

où :

y_i et x_{1i} sont les mêmes variables que celles utilisées dans le modèle (1).

x_{2i} est une variable dichotomique prenant la valeur 1 si le véhicule est de couleur claire, 0 si la couleur est foncée.

x_{3i} est une variable dichotomique prenant la valeur 1 si le véhicule a un moteur diesel et 0 s'il a un moteur essence.

Les résultats de l'estimation de ce modèle sur le même échantillon sont les suivants :

$$\begin{aligned} \hat{y}_i &= 31.748 &+& 1.152 &x_{1i} &-& 0.025 &x_{2i} &+& 5.600 &x_{3i} \\ &(1.253) && (0.061) &&& (1.549) &&& (2.022) \end{aligned}$$

(.) écart-types estimés

$$SCE' = \sum_{i=1}^{15} (\hat{y}_i - \bar{\hat{y}})^2 = 6194.34598$$

1. Tester la significativité des variables x_{1i} , x_{2i} et x_{3i} .

Interpréter leurs coefficients estimés (\hat{b}_1, \hat{b}_2 et \hat{b}_3) et commenter.

2. Tester la significativité globale du modèle (2) au seuil $\alpha = 5\%$.

3. Comparer les résultats de l'estimation des modèles (1) et (2).

Quel modèle privilégié ? Argumenter la réponse.

Exercice 1:

$$t_{\text{sex}}^* = \frac{-0.1862118}{0.0041738} = -44,59$$

$$|t_{32,204}^{\alpha/2}| = 1,96 \quad \Rightarrow |t^*| > |t_{7-k}^{\alpha/2}|$$

\Rightarrow Rejet de H_0

D'après le test de significativité la va est statistiquement significative au seuil de 5%

Cela signifie que le sexe est une va explicative du salaire