Supports à lire avant de commencer les exercices

Site du cours : http://eric.univ-lyon2.fr/~ricco/cours/cours_econometrie.html

A lire en particulier pour cette séance :

- Etudes des résidus (slides) : http://eric.univ-lyon2.fr/~ricco/cours/slides/Reg_Multiple_Etude_Des_Residus.pdf
- Points atypiques et influents (slides) : http://eric.univ-lyon2.fr/~ricco/cours/slides/Reg_Multiple_Points_Atypiques.pdf
- Colinéarité et sélection de variables (slides) : http://eric.univ-lyon2.fr/~ricco/cours/slides/Reg_Multiple_Colinearite_Selection_Variables.pdf
- Exogènes qualitatives (slides) : http://eric.univ-lyon2.fr/~ricco/cours/slides/Reg_Multiple_Exogenes_Qualitatives.pdf
- Pratique de la régression linéaire multiple (fascicule de cours) : http://eric.univ-lyon2.fr/~ricco/cours/cours/La_regression_dans_la_pratique.pdf

Exercice 1

On cherche à expliquer le montant de loyer d'appartements étudiants à partir des caractéristiques des logements. Voici les variables disponibles.

LOC	le loyer (€)
SURF	la surface (m^2)
CHAMB	le nombre de chambres
SCHAMB	la surface totale des chambres
SDB	le nombre de salles de bain
SSDB	la surface totale des salles de bain
PARK	le nombre de places de parking
PLAGE	la distance à la plage (km)
UNIV	la distance à l'université

LOC est la variable endogène. Nous disposons de **n = 25** observations. La régression fournit les résultats suivants R^2 = 0.995807 ; $\widehat{\sigma}_{\varepsilon}$ =16.089269 ;

Variable	Coef.	Ecart-type	t(16)	p-value
Intercept	273.9636	47.1889	5.8057	0.0000
SURF	7.4012	1.6991	4.3559	0.0005
CHAMB	-6.1686	19.6172	-0.3144	0.7572
SCHAMB	12.6325	2.4384	5.1808	0.0001
SDB	32.3051	31.0274	1.0412	0.3133
SSDB	1.8162	6.5814	0.2760	0.7861
PARK	30.2811	7.3729	4.1071	0.0008
PLAGE	-12.1928	1.7911	-6.8075	0.0000
UNIV	-3.3691	1.3994	-2.4075	0.0285

1. Avec les informations disponibles à ce stade, reconstituez le tableau d'analyse de variance. La régression est-elle globalement significative à 5% ?
2. On vous dit maintenant que SCT = 987694. Le modèle est-il pertinent au sens du critère AIC ? Au sens du critère BIC ?
3. Quelles sont les coefficients significatifs à 5% ? Avec les informations disponibles à ce stade, peut-on supprimer en bloc les variables qui ne sont pas pertinentes ?

On vous fournit ci-dessous la matrice des corrélations des exogènes.

CORR	SURF	CHAMB	SCHAMB	SDB	SSDB	PARK	PLAGE	UNIV
SURF	1	0.65	0.87	0.89	0.80	0.48	0.20	-0.17
CHAMB	0.65	1	0.88	0.49	0.55	0.59	0.00	0.33
SCHAMB	0.87	0.88	1	0.81	0.82	0.61	0.02	0.10
SDB	0.89	0.49	0.81	1	0.93	0.55	0.24	-0.24
SSDB	0.80	0.55	0.82	0.93	1	0.58	0.26	-0.13
PARK	0.48	0.59	0.61	0.55	0.58	1	0.11	0.11
PLAGE	0.20	0.00	0.02	0.24	0.26	0.11	1	-0.52
UNIV	-0.17	0.33	0.10	-0.24	-0.13	0.11	-0.52	1

4. Sur quelles couples de variables suspectez vous une corrélation qui peut s'avérer gênante pour la régression : appuyez-vous sur 3 règles différentes (règle simple, règle de Klein, cohérence des signes).

Voici maintenant l'inverse de la matrice des corrélations.

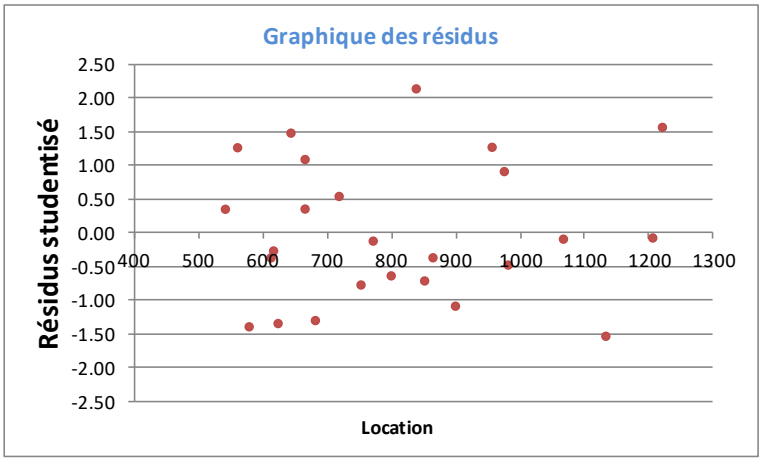
$CORR^{-1}$	SURF	CHAMB	SCHAMB	SDB	SSDB	PARK	PLAGE	UNIV
SURF	11	1.63	-9.83	-8.59	5.97	1.08	-1.63	0.12
CHAMB	1.63	27	-37.79	11.94	6.39	-2.72	-5.77	-3.69
SCHAMB	-9.83	-37.79	63	-12.90	-14.06	2.40	9.43	4.09
SDB	-8.59	11.94	-12.90	23	-9.34	-2.41	-0.87	0.14
SSDB	5.97	6.39	-14.06	-9.34	14	-0.37	-3.11	-1.68
PARK	1.08	-2.72	2.40	-2.41	-0.37	2	0.22	0.08
PLAGE	-1.63	-5.77	9.43	-0.87	-3.11	0.22	3	1.57
UNIV	0.12	-3.69	4.09	0.14	-1.68	0.08	1.57	2

5. Selon le critère VIF, quelles sont les variables redondantes ici ?

6. Nous effectuons une sélection « backward » avec le critère d'arrêt $\alpha = 5\%$. Les variables retenues sont (entre parenthèses le coefficient estimé) : SURF (7.124), SCHAMB (12.105), SDB (45.983), PARK (29.283), PLAGE (-12.349), UNIV (-3.487) ; la constante est égale à 273.971. Quelles sont les variables avec des coefficients significatifs à 5% dans ce modèle ? Décelez-vous des choses étonnantes par rapport au premier modèle ? Comment expliquez-vous cela ?
7. $R^2 = 0.995738$ pour ce nouveau modèle. C'est décevant par rapport au premier ?
8. Reconstituez le tableau d'analyse de variance de ce second modèle. Calculez les critères AIC et BIC. Que constatez-vous par rapport au premier modèle ? Quel modèle choisiriez-vous finalement ?
9. Quel serait le loyer d'un appartement avec les caractéristiques suivantes ?

	SURF	CHAMB	SCHAMB	SDB	SSDB	PARK	PLAGE	UNIV
N°26	39	1	13	1.5	6	2	6	8

10. Les sommes des résidus ont été calculées (somme simple, somme des carrés, sommes des cubes, sommes des puissances 4), nous obtenons respectivement : 0.000 ; 4209.645 ; 11531.618 ; 1394127.532. La normalité des résidus est-elle assurée au sens du test de Jarque-Bera ?
11. On réalise le graphique des résidus suivant avec, en abscisse la variable location, en ordonnée les résidus studentisés. A 5%, quelle est la valeur seuil à partir de laquelle on pourrait suspecter un point ? Y a-t-il des logements mal modélisés dans notre régression ?



Exercice 2

Suite à une revendication des syndicats, on souhaite comparer les salaires des hommes et de femmes dans une multinationale.

Voici les données du problème :

homme	Moyenne de SAL	3110.80
	Écartype de SAL2	1517.33
	Nombre de SEXE	20
femme	Moyenne de SAL	1947.25
	Écartype de SAL2	1021.59
	Nombre de SEXE	20

1. En passant par un test de comparaison, est-ce que les salaires sont identiques en moyenne à 5% ?
2. On décide de passer par une régression pour répondre à la question. Nous codons 1 : femme ; 0 : homme. Voici les sorties de DROITEREG

	a1	a0
coef.	-1163.55	3110.8
ec.type	409.019035	289.220133
	0.17557098	1293.43176
	8.09250659	38
	13538486	63572697

   Que constatez-vous au niveau des coefficients ? Testez la significativité de la pente de la droite de régression. Que constatez-vous par rapport à la comparaison de moyenne ci-dessus ?
3. Le patronat s'insurge et argue que les femmes font moins d'études que les hommes, c'est ce qui expliquerait le différentiel de salaire. Pour vérifier cela, vous introduisez les années d'études dans la régression. Vous obtenez les résultats suivants avec DROITEREG. Que faut-il en penser ?

Etudes	Sexe	Constante
217.00754	-881.440198	181.198209
82.6088649	395.402307	1147.22271
0.30516282	1203.37046	#N/A
8.12494264	37	#N/A
23531466.2	53579716.7	#N/A

Exercice 3

Nous souhaitons expliquer la qualité de prothèses dentaires (qualité variable cible quantitative) en or à l'aide de la méthode de malaxage (qualitative : M1, M2 et M3) et le type de l'or (qualitative : T1 et T2). Nous disposons de n = 30 observations.

1. Nous effectuons la régression avec le type de l'or dans un premier temps, nous codons le type de l'or en une indicatrice Tor1 qui prend la valeur 1 lorsque type l'or est égal à T1, 0 sinon. Voici les résultats de DROITEREG

TOr1	const
-95.667	820.600
45.214	31.971
0.138	123.825
4.477	28
68640.833	429310.533

Quelle est la qualité moyenne des dents chez les T2 ? Chez les T1 ?

Quelle est la qualité moyenne des dents sachant qu'il y a autant de T1 que de T2 ?

Le type de l'or a-t-il un impact significatif sur la qualité à 5% ?

2. Nous décidons d'évaluer l'influence du mode de malaxage seul cette fois-ci. Voici les résultats de la régression :

Mix2	Mix1	const
161.400	130.400	675.500
51.007	51.007	36.068
0.295	114.056	#N/A
5.639	27	#N/A
146717.067	351234.300	#N/A

Quelle est la qualité moyenne des dents chez les M3 ? Chez les M1 ? Chez les M2 ?

Quelle est la qualité moyenne des dents sachant qu'il y a autant de M1, M2 et M3 ? Ce résultat concorde-t-il avec celui obtenu à la question précédente ?

Le mode de malaxage a-t-il un impact significatif à 5% sur la qualité des dents ?

3. La méthode M1 se démarque-t-elle significativement à 5% de M3 ? M2 se démarque-telle de M3 ? M1 et M2 se démarque-t-elle également ? A titre d'information, voici la matrice $(X'X)^{-1}$:

	const	mix1	mix2
const	0.10	-0.10	-0.10
mix1	-0.10	0.20	0.10
mix2	-0.10	0.10	0.20

4. Nous cherchons maintenant à évaluer l'action conjointe de 2 facteurs sur la qualité des dents. Nous réalisons la régression avec toutes les variables indicatrices.

TOr1	Mix2	Mix1	const
-95.667	161.400	130.400	723.333
38.068	46.624	46.624	38.068
0.432	104.254	#N/A	#N/A
6.605	26	#N/A	#N/A
215357.900	282593.467	#N/A	#N/A

La régression est-elle globalement significative à 5 % ?

La présence de TOR1 dans la régression amène-t-elle significativement de l'information dans l'explication de la qualité des dents à 5 % ?

Idem, la présence du mode de mixage a-t-elle un impact significatif à 5 % ?

5. Nous souhaitons introduire les termes d'interactions dans la régression :

Mix2*Tor1	Mix1*TOr1	TOr1	Mix2	Mix1	const
-76.4	-50.4	-53.4	199.6	155.6	702.2
95.75127501	95.75127501	67.70637587	67.70637587	67.70637587	47.87563751
0.447638427	107.0531799	#N/A	#N/A	#N/A	#N/A
3.889960051	24	#N/A	#N/A	#N/A	#N/A
222902.1667	275049.2	#N/A	#N/A	#N/A	#N/A

A partir des résultats obtenus dans les différentes régressions ci-dessus, essayez de reconstituer le tableau des moyennes de la qualité des dents en fonction du type de l'or et de la méthode de malaxage. La structure du tableau se présente comme suit :

Moyenne de	Méthode			
Type	M1	M2	M3	**Marge**
T1				
T2				
Marge				

Reste à le remplir avec les bonnes valeurs !