T. P. n° 6
Régression linéaire multiple

Les exercices $1, 2, 5, 6$ *sont issus du livre « Probabilités, Statistique et technique de régression » de Gérald Baillargeon, Les éditions SMG et les exercices* $3, 4$ *sont issus du livre « Analyse de régression appliquée » de Yadolah Dodge, Édition Dunod.*

Exercice 1 Un bureau de conseil en ressources humaines a effectué une étude sur le niveau d'anxiété Y mesuré sur une échelle de 1 à 50 de cadres d'entreprises au cours d'une période de deux semaines. Nous voulons examiner si les facteurs suivants peuvent influencer sur le niveau d'anxiété des cadres :
- X_1 : pression artérielle systolique
- X_2 : test évaluant les capacités managériales
- X_3 : niveau de satisfaction du poste occupé.

Le tableau d'analyse de la variance indique l'apport de chaque variable introduite dans l'ordre indiqué et ceci pour 22 cadres.

Source de variation	Somme des carrés	*ddl*
Régression due à X_1	$981,326$	1
Régression due à X_2	$190,232$	1
Régression due à X_3	$129,431$	1
Résiduelle	$442,292$	18
Totale	$1743,281$	21

1. Quelle est la somme des carrés due à la régression pour l'ensemble des trois variables explicatives ?
2. Quelle proportion de la variation dans le niveau d'anxiété est expliquée par les trois variables explicatives ?
3. Pouvons-nous conclure que dans l'ensemble les trois variables explicatives ont un effet significatif sur le niveau d'anxiété ? Utiliser un seuil de signification $\alpha = 5\%$. Préciser les hypothèses que nous voulons tester.
4. Si nous ne tenons compte que de la variable explicative X_1, quel serait alors le tableau d'analyse de la variance correspondant ?

Source de variation	Somme des carrés	*ddl*
Régression due à X_1	$981,326$	
Résiduelle		
Totale		

5. Tester les hypothèses nulles suivantes, au seuil de signification $\alpha = 5\%$, en utilisant un rapport F approprié :

a) $\mathcal{H}_0 : \beta_1 = 0$ dans le modèle $Y = \beta_0 + \beta_1 X_1 + \varepsilon$;
b) $\mathcal{H}_0 : \beta_2 = 0$ dans le modèle $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$;
c) $\mathcal{H}_0 : \beta_3 = 0$ dans le modèle $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$.

6. Quelle est la valeur du coefficient de détermination R^2 associée à l'estimation de chaque modèle spécifié à la question 5. ?
7. Lequel des trois modèles semble le mieux approprié pour expliquer les fluctuations du niveau d'anxiété des cadres d'entreprises ?

Exercice 2 L'entreprise CITRON fabrique un matériau en matière plastique qui est utilisé dans la fabrication de jouets. Le département de contrôle de qualité de l'entreprise a effectué une étude qui a pour but d'établir dans quelle mesure la résistance à la rupture (en kg/cm^2) de cette matière plastique pouvait être affectée par l'épaisseur du matériau ainsi que la densité de ce matériau. Douze essais ont été effectués et les résultats sont présentés dans le tableau ci-dessous.

Essai numéro	Résistance à la rupture Y_i	Épaisseur du matériau X_{i_1}	Densité X_{i_2}
1	$37,8$	4	$4,0$
2	$22,5$	4	$3,6$
3	$17,1$	3	$3,1$
4	$10,8$	2	$3,2$
5	$7,2$	1	$3,0$
6	$42,3$	6	$3,8$
7	$30,2$	4	$3,8$
8	$19,4$	4	$2,9$
9	$14,8$	1	$3,8$
10	$9,5$	1	$2,8$
11	$32,4$	3	$3,4$
12	$21,6$	4	$2,8$

Diverses analyses ont été effectuées sur ordinateur et nous les résumons dans les trois tableaux suivants :

Régression de la résistance à la rupture Y en fonction de l'épaisseur X_1

Coefficients $\widehat{\beta}_j$	Erreurs-types $s(\widehat{\beta}_j)$
$\widehat{\beta}_0 = 3,523$	
$\widehat{\beta}_1 = 6,036$	$s(\widehat{\beta}_1) = 1,279$

Source de variation	Somme des carrés	*ddl*
Régression X_1	$980,64$	1
Résiduelle	$440,03$	10

Régression de la résistance à la rupture Y en fonction de la densité X_2

Coefficients $\widehat{\beta}_j$	Erreurs-types $s(\widehat{\beta}_j)$
$\widehat{\beta}_0 = -36,373$	
$\widehat{\beta}_2 = 17,464$	$s(\widehat{\beta}_2) = 6,069$

Source de variation	Somme des carrés	*ddl*
Régression X_2	$643,57$	1
Résiduelle	$777,10$	10

Régression de la résistance à la rupture Y en fonction de l'épaisseur X_1 et de la densité X_2

Coefficients $\widehat{\beta}_j$	Erreurs-types $s(\widehat{\beta}_j)$
$\widehat{\beta}_0 = -30,081$	
$\widehat{\beta}_1 = 4,905$	$s(\widehat{\beta}_1) = 1,014$
$\widehat{\beta}_2 = 11,072$	$s(\widehat{\beta}_2) = 3,621$

Source de variation	Somme des carrés	*ddl*
Régression (X_1, X_2)	$1204,86$	2
Résiduelle	$215,81$	9

1. Quel pourcentage de variation dans la résistance à la rupture est expliquée par chacune des régressions ?
2. Pour chaque régression linéaire, compléter le tableau suivant :

	Carré moyen résiduel	Écart-type des résidus
Régression due à X_1		
Régression due à X_2		
Régression due à (X_1, X_2)		

3. Compléter le tableau d'analyse de la variance suivant pour la régression comportant les deux variables explicatives.

Source de variation	Somme des carrés	*ddl*	Carrés moyens	F_{obs}
Régression due à (X_1, X_2)				
Résiduelle				
Totale				

4. Tester au seuil de signification $\alpha = 5\%$, l'hypothèse nulle

$$\mathcal{H}_0 : \beta_1 = \beta_2 = 0$$

contre l'hypothèse alternative

$$\mathcal{H}_1 : \text{ au moins un des } \beta \neq 0.$$

Quelle est votre conclusion ?

5. Dans le cas du modèle de régression linéaire ne comportant que l'épaisseur du matériau comme variable explicative, déterminer un intervalle de confiance à 95% pour β_1.
6. Avec l'intervalle de confiance calculé à la question 5., pouvons-nous affirmer, au seuil de signification $\alpha = 5\%$, que la régression linéaire est significative entre la résistance à la rupture et l'épaisseur du matériau ? Justifier votre conclusion.
7. Quel est l'apport marginal de la variable X_2 lorsqu'elle est introduite à la suite de la variable X_1 ?

8. Est-ce que la contribution marginale de la variable « densité du matériau », lorsqu'elle est introduite à la suite de la variable « épaisseur du matériau » est significative au seuil de signification $\alpha = 5\%$? Utiliser les deux façons équivalentes d'effectuer ce test.

« F partiel »	F_c	t_{obs}	t_c

9. Nous voulons obtenir diverses estimations et prévisions de la résistance à la rupture. Quelle est, en moyenne, la résistance à la rupture de jouets dont l'épaisseur du matériau utilisé et la densité du matériau sont celles indiquées dans le tableau suivant ?

Épaisseur X_1	Densité X_2
4	3,8
3	3,4
4	2,9

Estimation de la résistance moyenne	Écart-type de l'estimation
	2,10
	1,43
	2,57

10. Entre quelles valeurs peut se situer la résistance moyenne à la rupture, pour des jouets dont l'épaisseur du matériau est $X_1 = 4$ et la densité du matériau est $X_2 = 3,8$, si l'entreprise utilise un niveau de confiance de 95% ?
11. Quelle est la marge d'erreur dans l'estimation effectuée à la question 10. ?
12. Nous désirons un intervalle de prédiction de la résistance à la rupture pour un jouet ayant comme épaisseur du matériau et densité celles précisées à la question 10. Quel est cet intervalle au niveau de confiance de 95% ?

Exercice 3 Nous utilisons le modèle de régression linéaire multiple :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

1. Compléter le tableau d'analyse de variance suivant :

Source de variation	Somme des carrés	*ddl*	Carrés moyens	F_{obs}
Régression	1 504,4			
Résiduelle			19,6	
Totale	1 680,8			

2. Tester l'hypothèse nulle
$$\mathcal{H}_0 : \beta_1 = \beta_2 = 0.$$
contre l'hypothèse alternative
$$\mathcal{H}_1 : \text{ au moins un des } \beta \neq 0.$$
3. Quel est le coefficient de détermination R^2 du modèle ?
4. Donner une estimation de la variance de ε.

Exercice 4 Les données proposées dans cet exercice présentent le taux de décès par attaque cardiaque ches les hommes de 55 à 59 ans dans différents pays.
Les variables sont les suivantes :
- Y : 100 [log(nombre de décès par crise cardiaque pour 100 000 hommes de 55 à 59 ans) -2] ;
- X_1 : téléphones par millier d'habitants ;
- X_2 : calories grasses en pourcentage du total des calories ;
- X_3 : calories provenant de protéines animales en pourcentage du total des calories.

1. Régresser Y sur X_1 et tester la signification de cette régression linéaire simple.
2. Trouver l'équation de la régression linéaire multiple de Y sur X_1 et X_2.
3. Effectuer un test conjoint d'hypothèse nulle $\mathcal{H}_0 : \beta_1 = \beta_2 = 0$.
4. Tester si l'adjonction de la variable X_2 à l'équation trouvée à la question 1. a significativement amélioré l'estimation.
5. Construire la régression linéaire multiple de Y sur X_1, X_2 et X_3.
6. Donner les limites de l'intervalle de confiance à 95% pour β_3 dans cette équation.
7. Donner les limites de l'intervalle de confiance à 95% pour $\widehat{Y}$ au point $X_1 = 221$, $X_2 = 39$ et $X_3 = 7$.
8. Tester si X_2 et X_3 ensemble apportent quelque chose à la régression linéaire simple de Y sur X_1.
9. Régresser X_1 sur X_2 et X_3.
10. Donner les limites de l'intervalle de confiance à 95% pour les coefficients de la régression linéaire de X_1 sur X_3.

Observation i	Pays	X_1 x_{1i}	X_2 x_{2i}	X_3 x_{3i}	Y x_{4i}
1	Australie	124	33	8	81
2	Autriche	49	31	6	55
3	Canada	181	38	8	80
4	Ceylan	4	17	2	24
5	Chili	22	20	4	78
6	Danemark	152	39	6	52
7	Finlande	75	30	7	88
8	France	54	29	7	45
9	Allemagne	43	35	6	50
10	Irlande	41	31	5	69
11	Israel	17	23	4	66
12	Italie	22	21	3	45
13	Japon	16	8	3	24
14	Mexique	10	23	3	43
15	Pays-Bas	63	37	6	38
16	Nouvelle-Zélande	170	40	8	72
17	Norvège	125	38	6	41
18	Portugal	12	25	4	38
19	Suède	221	39	7	52
20	Suisse	171	33	7	52
21	Grande-Bretagne	97	38	6	66
22	États-Unis	254	39	8	89

Exercice 5 Dans une étude de régression linéaire multiple comportant quatre variables explicatives X_1, X_2, X_3, X_4, nous avons obtenu le tableau d'analyse de la variance suivant, et ceci pour 20 observations.

Source de variation	Somme des carrés	*ddl*	Carrés moyens
Régression (X_1, X_2, X_3, X_4)	85 570	4	21 392,50
Résiduelle	1 426	15	95,07
Totale	86 996	19	

1. Est-ce que la régression est significative dans son ensemble ? Utiliser $\alpha = 5\%$.
2. Une de vos collègues mentionne que les variables X_3 et X_4 sont inutiles dans le modèle de régression linéaire. Une autre analyse de régression linéaire ne comportant cette fois que les variables explicatives X_1 et X_2 conduit au tableau d'analyse de la variance suivant.

Source de variation	Somme des carrés	*ddl*	Carrés moyens
Régression (X_1, X_2)	62 983	2	31 491,50
Résiduelle	24 013	17	1 412,53
Totale	86 996	19	

Est-ce que l'affirmation de votre collègue est vraisemblable au seuil de signification $\alpha = 5\%$? Effectuer le test approprié.

Exercice 6 Dans une étude de régression multiple comportant quatre variables explicatives et 20 observations, nous avons introduit, dans l'ordre, les variables X_1, X_2, X_3, X_4.

Variables explicatives dans l'équation	Variable additionnelle	Écart-type des résidus	Proportion de la variation expliquée
Aucune		$s_Y = 2,7456$	
X_1	X_1	1,8968	0,5480
X_1, X_2	X_2	1,6352	0,6830
X_1, X_2, X_3	X_3	0,7349	0,9400
X_1, X_2, X_3, X_4	X_4	0,6281	0,9590

1. Dans quelle proportion la variation non expliquée par X_1 est réduite avec l'ajout de X_2 dans l'équation de régression ?
2. Quelle est la corrélation partielle entre Y et X_3 après s'être débarrassé de l'influence des variables explicatives X_1 et X_2 sur Y et X_3 ? Nous considérons que le coefficient de régression $\widehat{\beta}_3$ est négatif.
3. Déterminer la somme de carrés résiduelle lorsque les variables explicatives X_1 et X_2 sont dans l'équation de régression.
4. Quelle est la somme de carrés de régression attribuable à X_3 lorsque nous ajoutons cette variable à la suite de X_1 et X_2 ?
5. Quelle est la somme de carrés de régression attribuable à X_4 lorsque nous ajoutons cette variable à la suite de X_1, X_2 et X_3 ?