CHAPITRE 10

RÉGRESSION MULTIPLE

EXERCICES

Dans ce qui suit, nous nous permettrons certains raccourcis afin d'alléger le langage.

- Nous dirons « x est significative » lorsque l'hypothèse que le coefficient de la variable exogène *x* est nul est rejetée à un certain niveau α.
- Par défaut, le niveau exigé d'un test d'hypothèse est présumé égal à 5 % (et le niveau d'un intervalle de confiance égal à 95 %).

10.1 [*Tableau* 10.4] Le tableau 10.4 présente des données sur un échantillon de 34 logements recueillies afin de déterminer une formule de prédiction du montant de la facture d'électricité. Les variables sont le montant de la facture (facture); le revenu du ménage (revenu); le nombre de personnes (personnes); et de la superficie (surface) du plancher du logement.

a) Déterminer la matrice de corrélation des variables

b) Analyser le modèle E(facture) = $\beta_o + \beta_1$(revenu) + β_2(personnes) + β_3(surface). Vérifier que la variable revenu n'est pas significative.

c) Analyser le modèle E(facture) = $\gamma_o + \gamma_1$(personnes) + γ_2(surface). Tester les hypothèses usuelles $\gamma_1 = 0$ et $\gamma_2 = 0$.

d) Vérifier que la variable revenu comme seule variable exogène est néanmoins significative.

e) Vérifier que la variable revenu demeure significative en présence de personnes.

f) Vérifier que la variable revenu en présence de surface est à nouveau non significative

g) Il semblerait que revenu est significative dans tout modèle, à condition que surface n'en fasse pas partie. Peut-on expliquer ceci?

10.2 [*Tableau* A03] Dans le modèle E(irm) = $\beta_o + \beta_1$g + β_2v + β_3p, le coefficient de corrélation multiple est $R = 0{,}5366586$. Vérifier numériquement que R est le coefficient de corrélation entre y = irm et $\hat{y} = \hat{\beta}_o + \hat{\beta}_1 g + \hat{\beta}_2 v + \hat{\beta}_3 p$.

10.3 [Tableau A02] Les données du tableau A02 portent sur un échantillon de maisons vendues. On tente de déterminer lesquelles des variables disponibles (le nombre de salles de bains, le nombre de chambres à coucher, l'âge) pourraient servir à estimer le prix d'une maison. On supposera que le montant de l'offre n'est pas connu et ne peut être utilisé comme variable exogène.

a) Déterminer la matrice de corrélation des variables.

b) S'il fallait utiliser une seule des variables exogènes pour prédire le prix, laquelle choisirait-on? Pourquoi?

c) Introduire les variables exogènes successivement, dans l'ordre décroissant de leur corrélation avec le prix. Examiner à chaque étape

i) l'estimation de l'écart-type conditionnel $\hat{\sigma}$ (il devrait décroître à chaque ajout);
ii) le coefficient R^2 *ajusté*. Commentez les gains apportés par l'ajout de chaque nouvelle variable.

d) À chaque étape, déterminer si la variable exogène ajoutée au modèle est significative.

On compare les coefficients dans les trois modèles suivants: E(prix) = $\beta_o + \beta_1$(bains); E(prix) = $\gamma_o + \gamma_1$(cac); et E(prix) = $\delta_o + \delta_1$(bains)+ δ_2(cac). Expliquer les inégalités suivantes (examiner les signes des corrélations entre les variables concernées): i) $\hat{\beta}_1 > \hat{\delta}_1$; ii) $\hat{\gamma}_1 > \hat{\delta}_2$.

f) On compare les coefficients dans les trois modèles suivants: E(prix) = $\beta_o + \beta_1$(bains); E(prix) = $\gamma_o + \gamma_1$(age); et E(prix) = $\delta_o + \delta_1$(bains)+ δ_2(age). Expliquer les inégalités suivantes (examiner les signes des corrélations entre les variables concernées): i) $\hat{\beta}_1 > \hat{\delta}_1$; ii) $\hat{\gamma}_1 > \hat{\delta}_2$.

g) Revenons au modèle qui ne comprend que le nombre de salles de bains comme variable exogène. Estimer ce qu'une salle de bains ajoute en moyenne au prix d'une maison et déterminer un

intervalle de confiance pour ce paramètre.

10.4 [Tableau A02] Considérer maintenant un modèle qui comprend les trois variables exogènes énumérées, soit E(prix) = β_o + β_1(bains) + β_2(cac) + β_3(age)

a) Estimer ce que vaut sur le marché une salle de bains supplémentaire (ce qu'elle ajoute en moyenne au prix d'une maison). Expliquer en quoi le sens de ce paramètre diffère de celui dans la question 10.3-g).

b) Déterminer un intervalle de confiance pour β_1, β_2 et β_3.

c) Prédire la valeur d'une maison vieille de 25 ans, ayant deux salles de bains et 3 chambres à coucher.

d) Déterminer les limites de la prédiction que vous avez faite en c) à 95 % et puis à 60 %.

10.5 [Tableau 10.2] Le tableau 10.2 présente des données sur un groupe de professeurs recueillies afin d'identifier les facteurs qui contribuent au salaire en 2012 (sal12). Les variables exogènes possibles sont l'ancienneté (anc), le salaire à l'entrée (sal0), le sexe (sexe) et l'expérience préalable à l'engagement (exp).

a) S'il fallait utiliser une seule variable comme variable exogène, laquelle devrait-on choisir? Pourquoi?

b) Peut-on expliquer pourquoi le coefficient de corrélation entre sal12 et sal0 est négatif?

c) Revenons à anc comme première variable exogène, et considérons l'ajout de sal0 comme deuxième variable exogène. On a donc deux modèles:

Modèle A: E(sal12) = β_o + β_1(anc); et Modèle B: E(sal12) = γ_o + γ_1(anc) + γ_2(sal0)

i) Comparer le coefficient de corrélation du modèle A au coefficient de corrélation multiple R du modèle B. Que peut-on conclure du fait que la différence est minuscule?
ii) Vérifier que si R^2 augmente à l'ajout de sal0, R^2 ajusté baisse. Est-ce normal?
iii) Considérer le modèle suivant` : Modèle C : : E(sal12) = δ_o + δ_1(sal0). Vérifier que *sal0* est hautement significative.
iv) Pouvez-vous expliquer pourquoi dans le modèle B le coefficient de sal0 n'est pas significatif alors qu'il est hautement significatif dans le modèle C?

d) Considérer deux modèles pour prédire le salaire à l'entrée sal0:

Modèle A: E(sal0) = β_o + β_1(exp) et Modèle B: E(sal0) = γ_o + γ_1(exp) + γ_2(anc)

Vérifier que dans le modèle A, on ne peut pas affirmer que $\beta_1 \neq 0$ alors que dans le modèle B, on peut conclure avec confiance que $\gamma_1 > 0$.

e) Comment peut-on expliquer le paradoxe révélé en d)? Imaginer un graphique du style des figures 10.4 ou 10.5 pour illustrer ce phénomène.

e) Définir les variables indicatrices danc2, danc3, danc4, danc5 qui identifient les quintiles de la variable anc, c'est-à-dire, danc2 prend la valeur 1 si anc se situe au 2e quintile, de même pour danc3, danc4, et danc5. Le tableau suivant présente l'analyse d'une régression dans laquelle ces variables, ainsi que la variable exp sont les variables exogènes et sal0 est la variable endogène.

	Estimate	Std._Error	t_value	Pr(>|t|)
(Intercept)	31099,69	2082,63	14,933	<2e-16
exp	158,47	66,14	2,396	0,0185
danc2	-16041,10	1359,05	-11,803	<2e-16
danc3	-22939,39	1154,61	-19,868	<2e-16
danc4	-26453,41	1290,39	-20,500	<2e-16
danc5	-28114,24	2193,14	-12,819	<2e-16

i) Que signifient les valeurs des coefficients de danc2, danc3, danc4, et danc5?
ii) Que signifie le fait que les coefficients des variables indicatrices décroissent en allant de danc2 à danc5?

10.6 [Tableau 10.2]

a) Comparer les salaires moyens en 2012 des femmes et des hommes: estimer la différence de salaire et montrer qu'elle est significative.

b) Vérifier, cependant, que les hommes ont plus d'ancienneté que les femmes.

c) Est-ce possible que la différence de salaires ne soit due qu'au fait que les hommes ont plus d'ancienneté? Estimer la différence des salaires en tenant compte de l'ancienneté (déterminer une régression multiple avec l'ancienneté et le sexe comme variables exogènes).

d) La différence est maintenant réduite par rapport à celle calculée en a). Est-elle significativement différente de 0?

e) Résumer les conclusions en termes concrets.

f) Le modèle développé en c) postule que la relation entre sal12 et anc s'exprime par deux droites parallèles, l'une pour les femmes l'autre pour les hommes. Formellement, cela équivaut au modèle suivant:

$$\text{Femmes: } E(sal12) = \beta_o + \beta_1(anc); \quad \text{Hommes: } E(sal12) = \gamma_0 + \beta_1(anc)$$

Remarquez que β_1 est le même pour les femmes et les hommes, ce qui assure que les droites sont parallèles. Estimer β_o, β_1, et γ_o et tester l'hypothèse que $\beta_o = \gamma_o$.

g) Maintenant analyser un modèle qui n'impose pas de parallélisme entre les deux droites, donc un modèle qui postulerait l'existence de deux droites :

$$\text{Femmes: } E(sal12) = \beta_o + \beta_1(anc); \quad \text{Hommes: } E(sal12) = \gamma_0 + \gamma_1(anc)$$

Estimer les paramètres β_o, β_1, γ_o et γ_1 et tester l'hypothèse que chaque année d'ancienneté rapporte (en salaire) le même gain aux femmes qu'aux hommes

10.7 [Tableau A03] Le tableau A03 présente des données visant à déterminer si un lien peut être établi entre la grosseur du cerveau (irm) et certains traits physiques et psychologiques. Dans ce numéro nous comparons la grosseur du cerveau des femmes et des hommes en tentant d'éliminer l'effet de la grandeur du corps.

a) Montrer que la grosseur du cerveau (irm) est liée à la taille (tester pour montrer que la dépendance est significative).

b) Montrer également que la grosseur du cerveau est liée au poids

c) Montrer que dans le modèle $E(irm) = \beta_o + \beta_1(taille) + \beta_2(poids)$ on ne peut pas conclure que $\beta_2 \neq 0$, ce qui semble contredire la conclusion en b). Qu'est-ce qui pourrait expliquer la contradiction?

d) Tester (indépendamment de toute autre variable) l'hypothèse que le cerveau des femmes est en moyenne égal à celui des hommes.

e) La conclusion en d) (que les hommes ont un plus gros cerveau) serait-elle due uniquement au fait que les hommes sont plus grands? Il faudrait, pour que la comparaison soit juste, que la grosseur du cerveau soit relativisée par rapport à la taille (ou au poids). Une façon de le faire est de mesurer la grosseur du cerveau par irmT = irm/taille. Montrer qu'avec cette mesure, on ne peut pas conclure à une différence entre hommes et femmes quant à la grosseur du cerveau.

f) Refaire l'analyse en e) en prenant pour mesure de la grosseur du cerveau le rapport irmp = irm/poids.

10.8 [Tableau A03; *voir l'exercice* 10.6] Le tableau A03 présente entre autres des données sur le poids (poids) et la taille (taille) d'un groupe d'hommes et de femmes. Considérer un modèle liant la taille et le poids par deux droites (femmes et hommes) de même ordonnée à l'origine mais de pentes différentes :

$$\text{Femmes: } E(poids) = \beta_o + \beta_1(taille)$$

Hommes: E(poids) = $\gamma_o + \gamma_1$(taille)

a) Déterminer $\hat{\beta}_o$, $\hat{\beta}_1$, $\hat{\gamma}_o$ et $\hat{\gamma}_1$

b) Déterminer un intervalle de confiance pour $\gamma_o - \beta_o$.

c) Déterminer un intervalle de confiance pour $\gamma_1 - \beta_1$.

d) En b) et c) on a testé séparément les hypothèses $\gamma_o = \beta_o$ et $\gamma_1 = \beta_1$. Tester maintenant ces deux hypothèses simultanément, c'est-à-dire, tester l'hypothèse $H_o : \gamma_o = \beta_o$ *et* $\gamma_1 = \beta_1$ (au moyen d'une comparaison de sommes de carrés résiduelles.)

10.9 [Tableau A09]
Le tableau A09 présente, entre autres, les scores B1, B2 et B3 de trois posttests composés par 66 sujets à la fin d'une période de formation. On se concentre ici sur B1, le posttest de compréhension et les scores A_1 et A_2 à deux pré-tests composés avant la période de formation.

a) Montrer que chacune des variables A1 et A2 est séparément utile à la prédiction de B1.
b) Montrer cependant que A2 n'est plus significative (à 5 %) en présence de A1. Comment explique-t-on ce phénomène?

Les données du tableau ont été prélevées afin de comparer trois méthodes d'enseignement (trois *traitements*). L'échantillon est constitué de trois groupes, 1, 2 et 3, identifiés dans le tableau par la variable T (Traitement). On veut donc savoir si B1 dépend du traitement. À cette fin on définit les variables dichotomiques t1, t2 et t3 qui indiquent l'appartenance aux groupes 1, 2 et 3, respectivement.

c) Montrer par une régression multiple qu'on peut conclure que B1 dépend en effet du traitement (sans tenir compte des pré-tests).

d) Il serait normalement à craindre que les différences entre les moyennes de B1 ne soient pas attribuables aux traitements mais plutôt aux différences initiales A1.

i) Montrer pour commencer que ce n'est probablement pas le cas en comparant les moyennes de A1 dans les 3 groupes.

ii) Le fait que les différences entre les groupes (par rapport à A1) ne sont pas significatives pourrait justifier qu'on ne tienne pas compte de A1 dans la comparaison des traitements. Élaborer.

iii) Les différences dans les moyennes initiales A1 ne sont pas significatives, mais elles ne sont pas nulles. Il y aurait donc lieu d'éliminer l'effet de A1 dans la comparaison des traitements. Considérer le modèle suivant: E(B1) = $\gamma_j + \beta$(A1), où $\gamma_j = \gamma_1$, γ_2 ou γ_3, selon le traitement. Estimer les paramètres γ_j et β.

e) Tester séparément (mais dans le cadre du même modèle) les trois hypothèses $\gamma_1 = \gamma_2$; $\gamma_1 = \gamma_3$; et $\gamma_2 = \gamma_3$.

f) La variable A2 est une autre indication des capacités initiales des sujets; on aurait intérêt à en tenir compte (dans le sens d'éliminer sont effet dans la comparaison des traitements). Considérer le modèle suivant: E(B1) = $\gamma_j + \beta_1$(A1) + β_2(A2), où $\gamma_j = \gamma_1$, γ_2 ou γ_3, selon le traitement. Estimer les paramètres γ_j , β_1 et β_2.

10.10 [Tableau 10.4] Le tableau 10.4 présente les données suivantes sur un échantillon de 34 logements:

a) Déterminer la matrice de corrélation des variables

b) Déterminer une régression multiple pour estimer le montant de la facture à partir du revenu du ménage, du nombre de personnes et de la superficie du plancher du logement.

c) Vérifier que la variable revenu n'est pas significative dans le modèle en b), et refaire la régression sans elle.

d) Le tableau suivant présente des estimations dans quatre modèles pour la prédiction de facture, toutes comprenant la variable revenu comme variable exogène.

Il semblerait que le *revenu* est significatif dans tout modèle, tant que *surface* n'en fasse pas partie. Comment expliquer cela? Résumer en une phrase et en termes concrets la conclusion que suggèrent ces résultats.

10.11 [Tableau 10.5] Le tableau 10.5 présente des données (fictives) sur l'âge (age), le score en un test de vocabulaire (voc) et le niveau de scolarité (scol).

a) Considérer les deux modèles suivants:

$$E(voc) = \beta_o + \beta_1(age) \quad \text{et } E(voc) = \gamma_o + \gamma_1(age) + \gamma_2(scol)$$

Qu'est-ce qui explique la différence entre $\hat{\beta}_1 = 0{,}0037$ et $\hat{\gamma}_1 = 0{,}2415$ (pourquoi $\hat{\beta}_1$ est-il tellement plus petit que $\hat{\gamma}_1$)?

b) Considérer maintenant la scolarité comme une variable catégorielle : analyser le modèle

$$E(voc) = \gamma_j + \beta_1(age)\ ,\ j = 1, \ldots, 5\ ,$$

où j est le niveau de scolarité. Déterminer les 5 équations liant voc à age.

c) À partir du modèle en b) tester l'hypothèse que $\gamma_1 = \gamma_2$.

d) À partir du modèle en b) tester l'hypothèse que $\gamma_2 = \gamma_3$.

e) Déterminer le coefficient de corrélation entre voc et age; et le coefficient de corrélation partiel entre voc et age étant donné scol. Interpréter.

10.12 [Tableau 10.6; suite de l'exemple 10.7.2]

a) Qu'est-ce qui explique la différence entre $\hat{\beta}_1 = 0{,}9961$ et $\hat{\gamma}_1 = -0{,}02389$? Pourquoi $\hat{\beta}_1$ est-il significativement différent de 0 alors que $\hat{\gamma}_1$ ne l'est pas?

b) Considérer maintenant la scolarité comme une variable catégorielle : analyser le modèle

$$E(dext) = \gamma_j + \beta(voc)\ ,\ j = 1, \ldots, 9\ ,\ \text{où } j \text{ est le groupe d'âge.}$$

[Voici les estimations des γ: $\hat{\gamma}_1 = 32{,}618$; $\hat{\gamma}_2 = 38{,}701$; $\hat{\gamma}_3 = 45{,}052$; $\hat{\gamma}_4 = 50{,}251$; $\hat{\gamma}_5 = 55{,}334$; $\hat{\gamma}_6 = 60{,}722$; $\hat{\gamma}_7 = 66{,}552$; $\hat{\gamma}_8 = 71{,}329$; $\hat{\gamma}_9 = 77{,}580$ et $\hat{\beta} = -0{,}1060$.]

10.13 [Tableau 10.7] Le tableau 10.7 présente des données médicales sur un échantillon de 332 sujets d'origine indienne Pima d'Arizona. L'objectif dans ce numéro est de tenter d'identifier les facteurs qui contribuent à la tension artérielle.

a) Analyser le modèle $E(tension) = \beta_o + \beta_1(imc) + \beta_2(age) + \beta_3(peau) + \beta_4(glu) + \beta_4(gros)$.

b) À partir du modèle en a), éliminer les variables exogènes non significative s'il y lieu. Procéder par étapes: refaire une régression après avoir éliminé la variable exogène la moins significative (dont la *valeur p* est la plus grande). Recommencer avec le modèle réduit, ainsi de suite jusqu'à ce que toutes les variables exogènes soient significatives. Comparer le R du modèle final au R du modèle initial, question de s'assurer que l'élimination des variables exogènes ne cause pas d'importantes pertes de précision.

c) Vérifier la relation entre tension et imc, ainsi que la relation entre tension et peau sont toutes deux significatives.

d) Vérifier, cependant, que dans le modèle $E(tension) = \beta_o + \beta_1(imc) + \beta_2(peau)$, on ne peut pas conclure que $\beta_2 \neq 0$. Comment s'expliquerait cette apparente contradiction?

e) Vérifier que dans le modèle E(tension) = $\gamma_o + \gamma_1$(age) + γ_2(peau), on peut conclure avec confiance que $\gamma_2 > 0$. Expliquer pourquoi cette conclusion ne contredit pas nécessairement celles énoncées en a) et en b).

f) Le tableau suivant présente quelques résultats d'analyse de quatre modèles. La variable endogène est tension et une des variables exogènes est gros dans tous les cas. Chaque modèle comprend une autre variable exogène. Comment se fait-il que gros est signicatif dans trois cas et non significatif dans le quatrième?

Modèle	Variables exogènes	Coefficient de *gros*	*Valeur p*
A	imc et gros	0,7207	0,0003
B	peau et gros	0,6362	0,0024
C	glu et gros	0,6329	0,0026
D	age et gros	-0,2624	0,3363

10.14 [Tableau A03] Le tableau A.3 présente des données visant à déterminer si un lien peut être établi entre la grosseur du cerveau et certains traits physiques et psychologiques. On désigne par irm la grosseur du cerveau, par x_1 et x_2 les résultats aux tests V et P de Wechsler. (Nous délaissons la variable *g* car les données la concernant semblent entachées d'erreurs).

a) i) Vérifier que la relation entre irm et x_2 est significative.

ii) Noter, cependant, que dans le modèle E(irm) = $\beta_o + \beta_1 x_1 + \beta_2 x_2$, on ne peut rejeter aucune des deux hypothèses $\beta_1 = 0$ et $\beta_2 = 0$.

b) Comment s'explique la contradiction entre les conclusions b-i) et b-ii)?

10.15 [Tableau A.3] On désigne par sexe une variable dichotomique désignant le sexe (1 = homme; 0 = femme).

a) Montrer que dans le modèle E(irm) = $\beta_o + \beta_1$(sexe) + $\beta_2 x_2$, sexe et x_2 sont tous deux significatifs.

b) Supposons qu'on prenne pour mesure de la grosseur du cerveau le rapport irmt = irm/taille. Montrer que dans le modèle E(irmt) = $\gamma_o + \gamma_1$(sexe) + $\gamma_2 x_2$ sexe n'est plus significatif mais x_2 l'est encore.

c) Résumer concrètement les résultats en a) et b).

10.16 [Tableau A.3] Le modèle de régression n'est pas uniquement un outil de prédiction. Parfois, le but est simplement d'établir qu'une corrélation existe. Dans ce cas, il n'y a pas lieu de distinguer les variables endogènes des variables exogènes. On vous demande ici de vérifier empiriquement que le choix qu'on fait n'a pas d'importance.

a) Considérer les deux modèles suivants: E(irm) = $\beta_o + \beta_1 x_2$ et E(x_2) = $\gamma_o + \gamma_1$(irm). Vérifier que la *valeur p* correspondant à l'hypothèse $\beta_1 = 0$ est identique à la *valeur p* correspondant à l'hypothèse $\gamma_1 = 0$.

b) Considérer les deux modèles suivants:

Modèle A: E(irm) = $\beta_o + \beta_1 x_2 + \beta_2$(sexe) et Modèle B: E($x_2$) = $\gamma_o + \gamma_1$(irm) + γ_2(sexe).

Soit p_A la *valeur p* obtenue dans le modèle A et p_B la *valeur p* obtenue dans le modèle B. Avant de calculer, dites quelles seraient vos interprétations sous les hypothèses suivantes? (α est le niveau du test): i) $p_A < \alpha$ et $p_B \geq \alpha$; ii) $p_A \geq \alpha$ et $p_B < \alpha$).

c) Maintenant vérifier que la *valeur p* correspondant à l'hypothèse $\beta_1 = 0$ dans le modèle A est identique à la *valeur p* correspondant à l'hypothèse $\gamma_1 = 0$ dans le modèle B.

d) Comment concilier les *valeurs p* correspondant à sexe dans les deux modèles? (Concrètement, qu'est-ce qu'on affirme dans un cas et qu'est-ce qu'on affirme dans l'autre?)

10.17 Voici une autre façon de traiter l'exercice 6.8. Rappelons qu'il s'agissait d'évaluer l'efficacité d'un régime alimentaire à partir de données appariées. Voici les données :

Individus	1	2	3	4	5	6	7	8	9	10
Avant (X)	148	179	125	149	147	151	145	169	138	120
Après (Y)	144	162	126	131	132	146	145	152	127	118
Différence ($Y - X$)	-4	-17	1	-18	-15	-5	0	-17	-11	-2

Une autre solution pourrait être basée sur la distribution *conditionnelle* de Y étant donné $X = x$. Admettons les hypothèses d'une régression linéaire simple, soit $E(Y_i \mid X = x_i) = \beta_o + \beta_1 x_i$ (ainsi que l'hypothèse d'homoscédasticité). Si le régime n'a aucun effet, le poids Y_i à la fin du régime devrait, en moyenne, égaler le poids avant, soit x_i. L'hypothèse nulle est donc $H_o : \beta_o + \beta_1 x_i = x_i$, une condition qui doit être satisfaite pour tout x_i, ce qui entraîne

$$H_o : \beta_o = 0 \textit{ et } \beta_1 = 1.$$

Avec un changement de notation, le modèle est $E(z_i \mid x_i) = \beta_o + \gamma x_i$, où $z_i = y_i - x_i$ et $\gamma = \beta_1 - 1$.

Il faudra donc tester l'hypothèse $\beta_o = 0$ *et* $\gamma = 0$. Une approche pour ce faire est décrite à la section 9.6; la statistique de test est (9.6.4).

a) Dans le modèle restreint ($\beta_o = 0$ et $\gamma = 0$) la somme des carrés résiduelle est $SCR_o = \sum_{i=1}^{10} z_i^2$; vérifier que SCR_o =1294;

b) Vérifier que dans le modèle complet, SCR = 271,0665;

c) La statistique $F = \dfrac{(SCR_o\text{-}SCR)/2}{MCR}$ suit, sous H_o, une loi $F(2 ; 8)$. Justifier.

d) Vérifier que F = 16,09 et que la valeur p est vp = 0,0019.

10.18 Au numéro 9.9, soit x le score avant et Y le score après. Fixons les valeurs de X et basons notre analyse sur la distribution conditionnelle de Y étant donné $X = x$. Considérons le modèle de régression, $E(Y_i) = \beta_o + \beta_1 x_i + \beta_2$(méthode) + β_3(sexe), où:

sexe : sexe = 1 pour un garçon, = 0 pour fille
méthode : méthode = 0 pour LOGO, = 1 pour DELTA

a) Considérer le modèle $E(Y) = \beta_o + \beta_1 x + \beta_2$ (méthode) + β_3 (sexe)

i) Interpréter la valeur p (vp = 0,02)

ii) Interpréter les valeurs des coefficients β_1, β_2 et β_3 (en considérant pour l'instant qu'elles sont toutes significativement différentes de 0).

iii) On ne rejette pas l'hypothèse que $\beta_3 = 0$. Concrètement, qu'est-ce qu'on conclut?

b) Considérer le modèle $E(Y) = \beta_o + \beta_1 x + \beta_2$ (*méthode*).

i) La valeur de R^2 a baissé très peu par rapport à R^2 dans le dernier modèle. Qu'est-ce que cela signifie?

ii) Interpréter le coefficient β_2 ainsi que la valeur p qui lui est associée.

c) Le but de l'expérience est de savoir si le traitement a un effet. La différence $Z = Y - x$ mesure cet effet. Considérer le modèle initial $E(Y_i - x_i) = \beta_o + \beta_1 x_i - x_i + \beta_2$ (méthode) + β_3 (sexe) ou $E(Z_i|x_i) = \beta_o + \gamma x_i + \beta_2$ (méthode) + β_3 (sexe), où $\gamma = \beta_1 - 1$. (Voir la section 9.6)

i) Montrer que la variable sexe n'est pas significative.

ii) Éliminer la variable sexe et considérer le modèle $E(Y_i - x_i) = \beta_o + \beta_1 x_i - x_i + \beta_2$ (méthode)

iii) Dans le modèle déterminé en 3), tester l'hypothèse que $E(Y_i \mid x_i) = x_i$

Deux groupes de 11 enfants de troisième année du cycle primaire ont complété le test psychologique IAR (*Intelligence Achievement Responsibility*) avant et après une période de quatre mois et demi d'expérimentation avec l'un ou l'autre de deux langages informatiques : LOGO et Delta Drawing. Contrairement au LOGO, le langage Delta Drawing n'attache pas une grande importance à la décomposition d'un problème complexe ou à l'apprentissage par la correction des erreurs. Le test IAR mesure la propension du sujet à se sentir maître de ses apprentissages et de son succès intellectuel. Les chercheurs ont voulu montrer que l'exercice du langage LOGO augmente cette propension. Voici les résultats obtenus :

Tableau 9.3

Comparaison des langages Logo et Delta Drawing

LOGO			DELTA		
	Score			Score	
sexe	Avant	Après	sexe	Avant	Après
F	16	29	F	15	21
F	20	24	M	18	22
M	21	23	F	21	21
M	22	21	F	21	19
M	22	26	F	22	20
F	23	30	F	22	20
F	24	26	F	23	23
F	24	23	F	23	30
F	25	32	M	26	21
M	27	34	M	27	25
M	28	29	M	30	27

10.19 Le test d'égalité de moyennes avec données appariées (section 6.6) peut être plongé dans un contexte plus riche permettant de formuler et de tester une hypothèse plus forte que la simple égalité de moyennes. On considère le cas d'un même sujet qui passe un test avant et un test après une certaine intervention (voir, par exemple, l'exercice 6.26). Soit X le score avant et Y le score après et μ_X, μ_Y leurs moyennes théoriques. Le but de l'expérience est de tester l'hypothèse n'a aucun effet, une hypothèse exprimée au chapitre 6 par $\mu_X = \mu_Y$ et réduite à $\mu_Z = 0$, où $\mu_Z = \mu_X - \mu_Y$, la moyenne de la variable $Z = Y - X$. Une approche plus nuancée considère la distribution *conditionnelle* de Y étant donné $X = x$. Soit $\mu_{Y.x} = E(Y \mid X = x)$. Sous l'hypothèse que l'intervention n'a pas d'effet, il est normal de s'attendre à ce que $\mu_{Y.x} = x$. On peut tester cette hypothèse dans le cadre de divers modèles de régression, dont voici trois ($Z = Y - X$):

Modèle	Expression du modèle	H_o
Modèle 1	$E(Y \mid x) = x + \beta_o$, ou $E(Z \mid x) = \beta_o$	$\beta_o = 0$
Modèle 2	$E(Y \mid x) = \beta_1 x$, ou $E(Z \mid x) = \gamma_1 x$, $\gamma_1 = \beta_1 - 1$	$\gamma_1 = 0$
Modèle 3	$E(Y \mid x) = \beta_o + \beta_1 x$, ou $E(Z \mid x) = \beta_o + \gamma_1 x$, $\gamma_1 = \beta_1 - 1$	$\beta_o = 0$ *et* $\gamma_1 = 0$

Considérons un échantillon de n paires $[y_i ; x_i]$, et soit $z_i = y_i - x_i$. Dans chacun de ces modèles, H_o peut être testé par les moyens décrits à la section 9.6, la statistique F étant définie par (9.6.4)

a) Modèle 1.

i) Montrer que $SCR_o = \sum_{i=1}^{n} z_i^2$

ii) Montrer que $SCR = (n\text{-}1)\, S_Z^2 = \sum_{i=1}^{n} (z_i - \overline{z})^2$.

ii) Montrer que la statistique F est identique la statistique de Student (6.6.1).

b) Modèle 2.

i) Montrer que $SCR_o = \sum_{i=1}^{n} z_i^2$

ii) Montrer que SCR $= \sum_{i=1}^{n} z_i^2 - \frac{[\sum_{i=1}^{n} x_i z_i]^2}{\sum_{i-1}^{n} x_i^2}$.

c) Modèle 3.

i) Montrer que $SCR_o = \sum_{i=1}^{n} z_i^2$

ii) Montrer que $SCR = (n-1)\left(S_Z^2 - \frac{S_{ZX}^2}{S_X^2}\right)$, où S_X^2 et S_Z^2 sont les variances échantillonnales de X et Z et S_{ZX} est la covariance échantillonnale entre Z et X.

d) Les données suivantes portent sur un échantillon de 40 individus, où X est un score avant et Y un score après une certaine intervention. Développer chacun des trois modèles ci-dessus et tester H_o dans chacun.

y	x	y	x	y	x	y	x
157,84	166	170,22	175	90,2	103	140,33	136
171,21	169	113,75	116	133,81	132	177,55	175
150,67	152	156,34	149	168,23	160	163,57	173
156,75	163	131,68	141	166,88	159	164,95	151
136,75	137	185,89	169	195,63	196	98,46	94
132,56	152	160,31	163	141,2	136	137,31	119
163,81	144	127,29	117	179,48	174	140,3	137
92,63	93	179,05	178	149,53	141	175,71	190
135,99	131	88,52	91	148	149	104,41	111
89,93	100	163,55	158	156,14	171	163,97	163