CHAPITRE 10

RÉGRESSION MULTIPLE

SOLUTIONS

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

Matrice des corrélations

	facture	revenu	personnes	surface
facture	1.000	0.837	0.494	0.905
revenu	0.837	1.000	0.143	0.961
personnes	0.494	0.143	1.000	0.366
surface	0.905	0.961	0.366	1.000

b) Analyser le modèle E(facture) = β_o + β_1(revenu) + β_2(personnes) + β_3(surface). Vérifier que la variable revenu n'est pas significative.

Voici l'analyse fournie par le logiciel **R**

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-358.44157	198.73583	-1.804	0.0813	.
revenu	0.07514	0.13609	0.552	0.5850	
personnes	55.08763	29.04515	1.897	0.0675	.
surface	0.28110	0.22611	1.243	0.2234	

La dépendance est globalement significative (F = 57,28 sur 3 et 30 degrés de liberté, p = 1,58e-12); et assez forte (R^2 = 0,8514).

c) Analyser le modèle E(facture) = γ_o + γ_1(personnes) + γ_2(surface). Tester les hypothèses usuelles $\gamma_1 = 0$ et $\gamma_2 = 0$.

Voici les calculs fournis par le logiciel **R** :

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-255.94923	70.14230	-3.649	0.000959	***
personnes	42.03394	16.67947	2.520	0.017096	*
surface	0.40429	0.03615	11.183	2.08e-12	***

On remarque que la réduction du coefficient de détermination R^2 est négligeable : elle passe de 0,8514 à R^2 = 0,8499. Le fait d'éliminer la variable revenu ne fait pas perdre grand-chose à la précision d'une prédiction de facture.

Les deux variables exogènes retenues (personnes et surface) sont toutes deux hautement significatives, avec des valeurs p de 0,017 et 2,08e-12, respectivement.

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

Calculs fournis par le logiciel R

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-425.16251	124.92953	-3.403	0.00181	**
revenu	0.25899	0.02995	8.649	6.97e-10	***

Contrairement à ce qui a été conclu en b), la variable revenu est fortement significative, bien que le coefficient de détermination soit plus faible (R^2 = 0,7004) qu'avec les deux autres variables exogènes, personnes et surface (0,8499).

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

Calculs par le logiciel **R** :

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-575.4106	95.9010	-6.000	1.23e-06	***
revenu	0.2421	0.0222	10.905	3.89e-12	***
personnes	85.3352	16.0031	5.332	8.28e-06	***

Les deux variables sont hautement significatives, et le coefficient de détermination est à peine moins élevé que si on y ajoutait la surface comme troisième variable exogène. (R^2 = 0,8437 comparé à R^2 = 0,8514 avec les trois variables exogènes).

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

Calculs par le logiciel **R** :

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -52.23835  120.64872  -0.433    0.668
revenu       -0.13498    0.08229  -1.640    0.111
surface       0.64033    0.12857   4.981 2.27e-05 ***
```

On constate en effet que revenu n'est plus significatif (du moins pas à 5 % ou même 10 %).

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

Les variables revenu et surface sont très fortement corrélées (r = 0,961) : les maisons des gens à revenu élevé sont grandes; elles contiennent la même information utile à la prédiction de la facture. Elles sont donc redondantes. C'est la même chose avec la variable revenu : elle est utile toute seule, ainsi qu'en présence de personnes. On éliminera donc l'une des variables revenu ou surface. Les données ne favorisent pas un choix plutôt qu'un autre (le coefficient de détermination est de 0,8499 dans une régression comprenant personnes et surface; et de 0,8437 dans une régression comprenant personnes et revenu—une différence minuscule. Nous choisirons surface plutôt que revenu pour des raisons contextuelles : on s'attend à ce qu'une grande maison habitée par une personne à faible revenu consomme plus d'électricité qu'une petite maison habitée par un riche.)

10.2 [*Tableau* A03] Dans le modèle $E(\text{irm}) = \beta_o + \beta_1 g + \beta_2 v + \beta_3 p$, le coefficient de corrélation multiple est $R = 0{,}5366586$. Vérifier numériquement que R est le coefficient de corrélation entre y = irm et $\hat{y} = \hat{\beta}_o + \hat{\beta}_1 g + \hat{\beta}_2 v + \hat{\beta}_3 p$.

Il suffit de calculer les estimations $\hat{y}$ que voici :

```
905975.7 975048.0 916172.2 937529.7 913242.2 930348.1 906688.2 850705.6
958306.6 895011.4 978690.5 852315.0 909074.6 862961.9 857503.2 884814.9
878617.0 839497.6 862110.1 854432.0 940797.5 965160.6 922718.0 963241.1
882437.6 906706.6 831323.9 898279.2
```

On calcule ensuite le coefficient de corrélation entre $\hat{y}$ irm. On trouvera $r = 0{,}5366586$, qui est en effet la racine carrée du coefficient de détermination R^2, tel que déterminé par le logiciel **R**.

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.

Voici la matrice des coefficients de corrélation :

```
          prix  Offre  bains    cac    age sousol garage
prix     1.000  0.959  0.650  0.358 -0.535  0.228  0.274
Offre    0.959  1.000  0.551  0.366 -0.434  0.086  0.138
bains    0.650  0.551  1.000  0.348 -0.463  0.215  0.362
cac      0.358  0.366  0.348  1.000  0.101 -0.243  0.260
age     -0.535 -0.434 -0.463  0.101  1.000 -0.636  0.048
sousol   0.228  0.086  0.215 -0.243 -0.636  1.000  0.213
garage   0.274  0.138  0.362  0.260  0.048  0.213  1.000
```

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

La variable bains, puisqu'elle est la plus fortement corrélée avec prix.

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.

Le tableau suivant montre que le R^2 non ajusté croît à mesure qu'on ajoute des variables exogènes, ce qui reflète le fait que l'ajustement ne peut que bénéficier de l'ajout d'information.

Normalement, σ devrait décroître à mesure qu'on ajoute des variables exogènes. Par exemple, dans une régression comprenant les variables bains, age et cac, σ représente la dispersion des prix des maisons ayant le même nombre de salles de bains, le même âge et le même nombre de chambres à coucher. Lorsque, en plus, on se limite à des maisons ayant le même nombre de garages, σ devrait diminuer. Mais le hasard fait que les estimations de σ ne se conforment pas nécessairement à cette logique. Ainsi donc l'ajout de la variable garage fait croître $\hat{\sigma}$; et l'ajout subséquent de sousol le fait croître encore plus.

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

	$\hat{\sigma}$	R^2 ajusté	R^2 non ajusté
bains	15605	0,4088	0,4229
bains + age	14810	0,4675	0,4928
bains + age + cac	14154	0,5136	0,5484
bains + age + cac + garage	14190	0,5112	0,5577
bains + age + cac + garage + sousol	14219	0,5091	0,5675

	Valeur *p* pour la dernière variable	Valeur *p* globale
bains	0,0000006	0,0000023606
bains+age	0,01929	0,0000009788
bains+age+cac	0,0357	0,0000004513
bains+age+cac+garage	0,3776	0,0000005218
bains+age+cac+garage+sousol	0,3648	0,0000005992

e) 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) $\hat{\gamma}_1 > \hat{\delta}_2$.

E(prix) = 18750 + 26439 (bains)
E(prix) = 28099 + 10417 (cac)
E(prix) = 8541 + 24322 (bains) + 4350 (cac)
bains et cac sont positivement corrélées: r = 0,348.

$\hat{\beta}_1$ = 26439 signifie que le prix augmente de 26 439 $ pour chaque salle de bains supplémentaire; $\hat{\delta}_1$ = 24322 signifie la même chose, mais avec une différence importante. Une maison avec un grand nombre de salles de bains a tendance à avoir en plus un grand nombre de chambres à coucher. Donc le coefficient $\hat{\beta}_1$ = 26439 représente l'effet non seulement d'une salle de bains de plus, mais aussi l'effet du plus grand nombre de chambres à coucher qui va avec. Alors que, dans la régression multiple, le coefficient $\hat{\delta}_1$ = 24322 est le taux d'accroissement du prix par rapport à bains lorsque cac *reste fixe*.

Le même phénomène explique l'inégalité $\hat{\gamma}_1 > \hat{\delta}_2$.

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) $\hat{\gamma}_1 > \hat{\delta}_2$.

Les variables bains et age sont négativement corrélées (r = -0,463): plus une maison est vieille, moins elle a de salles de bains.
E(prix) = 18750 + 26439 (bains)
E(prix) = 69481 - 398 (age)
E(prix) = 32595 + 20828 (bains) – 222 (age)
Les maisons avec une salle de bains de plus coûte en moyenne 26 439 $ de plus. Mais une maison avec une salle de bains de plus a tendance à être moins vieille, ce qui ajoute à son prix. Le taux de 26 439 $ par salle de bains reflète, en plus, l'effet de l'âge. Dans la régression multiple, le taux de 20 828 $ estime le taux d'accroissement par salle de bains *lorsque l'âge est fixé*.

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.

$\hat{\beta}_1$ = 26439,111, $\hat{\sigma}_{\hat{\beta}_1}$ = 4823,782. Point critique (Student, 41 degrés de liberté) = 2,019541.

Intervalle de confiance :
[26439,111-2,019541(4823,782); 26439,111+2,019541(4823,782)] = [16 697 $; 36 181 $].

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).

Droite estimée :
E(prix) = 18600 + 15409,1(bains) + 7714,1(cac) - 287,7(age).
Une salle de bains supplémentaire ajoute en moyenne 15 409 $ au prix d'une maison, pour un âge et un nombre donné de chambre à coucher.
Ce montant est inférieur au montant de 26 439 $ trouvé au numéro 10.3-g), car une salle de bains de plus est généralement accompagnée d'un plus grand nombre de salles de bains.
À titre d'exemples voici des estimations des prix moyens selon le modèle et le nombre de salles de bains.
Le Modèle 1 n'a que bains pour variable exogène alors que le Modèle 2 comprend bains, cac et age :

Modèle 1 1 salle de bains : E(prix) = 45 189 $.
2 salles de bains : E(prix) = 71 628 $.
Modèle 2 1 salle de bains, 3 chambres à coucher, 22 ans d'âge : E(prix) = 51 602 $.
2 salles de bains, 3 chambres à coucher, 22 ans d'âge : E(prix) = 67 011 $.

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

	Limite inférieure	Limite supérieure
β_1	4243	26575
β_2	590	14838
β_3	-480	-95

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

Estimation du prix moyen: 65 368 $; écart-type de l'estimateur : 3427,519; nombre de degrés de liberté : 39; intervalle de confiance : [58435 ; 72301].

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

95 % : [58 435 $; 72 301 $]
60 % : [62 451 $; 68 284 $]

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?

Le tableau des corrélations montre que l'ancienneté qui est la plus fortement corrélée avec le salaire. C'est donc l'ancienneté qu'on choisirait pour prédire le salaire.

	sal12	anc	sal0	sexe	exp
sal12	1.000	0.900	-0.844	0.359	-0.034
anc	0.900	1.000	-0.942	0.261	0.039
sal0	-0.844	-0.942	1.000	-0.229	0.102
sexe	0.359	0.261	-0.229	1.000	-0.142
exp	-0.034	0.039	0.102	-0.142	1.000

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

Ceux dont le salaire en 2012 est élevé ont généralement beaucoup d'ancienneté. Leur engagement date donc depuis plus longtemps, d'où un salaire inférieur à l'entrée.

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?
Modèle A : $r^2 = 0{,}8095$; Modèle B : $R^2 = 0{,}8096$, ce qui veut dire que l'ajout de sal0 n'apporte aucune amélioration dans la prédiction de sal12.
Analyse du modèle A :

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 50095.30    1595.61   31.40   <2e-16 ***
anc          1529.66      73.12   20.92   <2e-16 ***
```

Analyse du modèle B :

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.805e+04  7.546e+03   6.367 5.58e-09 ***
anc         1.587e+03  2.190e+02   7.246 8.43e-11 ***
sal0        4.913e-02  1.771e-01   0.277    0.782
```

ii) Vérifier que si R^2 augmente à l'ajout de sal0, R^2 ajusté baisse. Est-ce normal?
Les carrés des coefficients de corrélation ajustés dans les modèles A et B sont, respectivement,0,9076 et 0,8059. C'est normal car le coefficient de corrélation ajusté fait payer un prix pour l'ajout de variables exogènes.

iii) Considérer le modèle suivant` : Modèle C : : $E(sal12) = \delta_o + \delta_1(sal0)$. Vérifier que *sal0* est hautement significative.
Analyse du modèle C :

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.019e+05  1.580e+03   64.52   <2e-16 ***
sal0        -1.160e+00  7.273e-02  -15.94   <2e-16 ***
```

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?
Sal0 et anc sont fortement (et négativement) corrélés. Le fait qu'ils sont fortement corrélés signifie que l'un ou l'autre pourrait servir à prédire sal12, mais que l'un en présence de l'autre n'ajoute pas grand-chose à la capacité de prédiction.

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

Modèle A: $E(sal0) = \beta_o + \beta_1(exp)$ et Modèle B: $E(sal0) = \gamma_o + \gamma_1(exp) + \gamma_2(anc)$

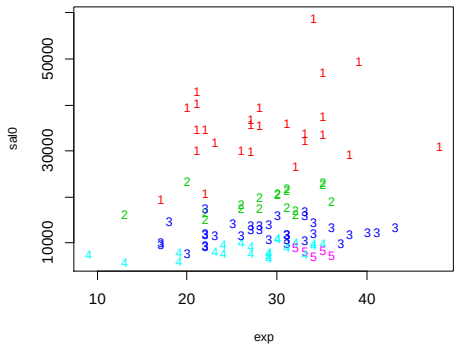
Vérifier que dans le modèle A, on ne peut pas affirmer que $\beta_1 \neq 0$ alors que dans le modèle B, on peut conclure avec confiance que $\gamma_1 > 0$.
Analyse du modèle A

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  13663.4     4826.8   2.831  0.00559 **
exp            171.7      165.1   1.040  0.30072
```

Analyse du modèle B

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 35118.09    1638.69  21.431  < 2e-16 ***
exp           233.93      50.96   4.591 1.26e-05 ***
anc         -1171.85      37.42 -31.319  < 2e-16 ***
```

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.
Le graphique suivant donne une idée de ce qui se passe. Il montre la relation entre sal0 et exp.
On voit assez bien que la relation entre sal0 et exp est faible.
Le rôle de l'ancienneté (anc) dans cette relation est indiqué par les chiffres 1 à 5, qui sont les quintiles de la variable anc. On constate que si l'on s'en tient aux individus appartenant au même quantile (1er quantile, par exemple), la relation est positive et assez forte. De même si l'on s'en tient aux valeurs du 2e quantile de anc; et possiblement des 3e et 4e quantiles également. Ce qui veut dire, que la relation existe quand on se limite à des individus de même ancienneté. Mais étant donné la grande dispersion des niveaux d'ancienneté pour une même expérience, sal0 reste très variable malgré un même niveau d'expérience.



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?
Pour exemple, le coefficient -16041,10 de danc2 signifie que le salaire à l'entrée est en moyenne inférieure de 16041,10 lorsque l'ancienneté se situe au 2e quintile que lorsqu'il se situe au premier quintile. Les coefficients de danc3, danc4 et danc5 comparent les salaires des 2e au 5e quantiles au premier quantile.

ii) Que signifie le fait que les coefficients des variables indicatrices décroissent en allant de danc2 à danc5?
Cette croissance reflète le fait que le salaire décroît avec l'ancienneté.

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.
Les échantillons des femmes et des hommes sont de tailles $n_1 = 40$ et $n_2 = 65$, respectivement; les moyennes sont 73310,12 et 84782,02; la différence (la moyenne des hommes moins la moyenne des femmes) est 11471.89; les écarts-type des échantillons sont $S_1 = 13800,71$ et $S_2 = 15118,03$. En supposant l'égalité des variances on estime l'écart-type commun par $S = 14633,19$. La statistique t est $T = 3,901108$, ce qui, à 103 degrés de libertgé donne la valeur $p\ vp = 0,00017095$. On rejette donc l'hypothès que la moyenne des femmes est égale à celle des hommes : on peut conclure qu'elle lui est inférieure.

b) Vérifier, cependant, que les hommes ont plus d'ancienneté que les femmes.
Les femmes ont en moyenne 16,775 années d'ancienneté alors que les hommes en ont 21,692 années.

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).

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	48638.92	1603.39	30.335	< 2e-16	***
sexe	4239.94	1368.68	3.098	0.00252	**
anc	1470.71	72.78	20.208	< 2e-16	***

On estime donc le salaire moyenne des femmes à 48638,92 + 1470,71(anc), et celui des hommes de même ancienneté par 48638,92 + 4239,94 + 1470,71(anc). Donc pour un même niveau d'ancienneté, le salaire moyen des hommes est de 4339,94 $ supérieur à celui des femmes.

d) La différence est maintenant réduite par rapport à celle calculée en a). Est-elle significativement différente de 0?
Elle est significativement différente de 0, avec une valeur p de 0,00252.

e) Résumer les conclusions en termes concrets.
Le salaire moyen des hommes est supérieur (de 11471,89 $) à celui des femmes, et ce n'est pas seulement parce que les femmes ont moins d'ancienneté. Si elles avaient eu le même niveau d'ancienneté, la différence aurait été moindre (de 4239,94 $) mais pas nulle.

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 $\beta_o = \gamma_o$.

Ces paramètres ont été estimés en c). $\hat{\beta}_1 = 1470{,}75$; $\hat{\beta}_o = 48638{,}92$ et $\hat{\gamma}_o = 48638{,}92 + 4239{,}94 = 52\,878{,}86$.

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
On définit une nouvelle variable, sexanc, qui prend la valeur anc si le sujet est un homme et 0 si le sujet est une femme. Voici l'analyse du modèle :

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 49158.47    2271.95  21.637   <2e-16 ***
sexe         3327.01    3133.56   1.062    0.291
anc          1439.74     120.29  11.969   <2e-16 ***
sexanc         49.11     151.47   0.324    0.746
```

Donc les droites estimées sont :
Femmes : E(sal12) = 49158,47 + 1439,74(anc).
Hommes : E(sal12) = (49158,47 + 3327,01) + (1439.74 + 49,11)(anc).
Le taux d'accroissement du salaire par année d'ancienneté des hommes serait donc supérieur à celui des femmes de 49,11. Cette différence, cependant, est loin d'être significative. Un modèle réaliste et parcimonieux éliminerait ce facteur, ce qui nous ramènerait au modèle en f).

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).

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    54009     224081   0.241 0.811426
Taille         12454       3283   3.794 0.000799 *
```

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

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 645244.5    87530.3   7.372 7.92e-08 ***
Poids         1741.0      584.8   2.977  0.00622 **
```

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?

```
      Estimate Std. Error t value Pr(>|t|)
(Intercept) 124595.6   264940.1   0.470   0.6422
Taille       10474.2     5065.1   2.068   0.0491 *
Poids          434.7      838.3   0.519   0.6086
```

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

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    857432      15016   57.10  < 2e-16 ***
sexe           105967      22938    4.62 9.17e-05 ***
```

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.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  13009.6      226.5  57.439   <2e-16
sexe           555.8      346.0   1.607     0.12
```

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

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   6369.7      201.1  31.668   <2e-16
sexe          -386.4      307.2  -1.258     0.22
```

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 :

Femmes: $E(\text{poids}) = \beta_o + \beta_1(\text{taille})$
Hommes: $E(\text{poids}) = \gamma_o + \gamma_1(\text{taille})$

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

On définit une nouvelle variable sextail = sexe*taille.
Ensuite détermine une régression avec sexe, taille et sextail pour variables exogènes. On obtient ceci` :

E(poids) = -143,2724 + 40,1774(sexe) + 4,2399(taille) - 0,4915(sextail)

Donc les deux droites sont :
Femmes : E(poids) = -143,2724 + 4,2399(taille)
Hommes : E(poids) = -143,2724 + (4,2399 - 0,4915)(taille)

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

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -143.2724   112.9266  -1.269   0.2167
sexe          40.1774   151.2235   0.266   0.7928
taille         4.2399     1.7111   2.478   0.0206 *
sextail       -0.4915     2.2192  -0.221   0.8266
```

$\hat{\gamma}_o - \hat{\beta}_o$ est le coefficient de la variable sexe. Il est estimé par 40,1774 et son écart-type est estimé par 151,2235. Le point critique à 95 % est $t_{24;0,025} = 2,06$. L'intervalle de confiance est [-271,9 ; 352,3]. Du fait que l'intervalle comprend 0, on ne peut rejeter que la différence est nulle.

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

$\hat{\gamma}_1 - \hat{\beta}_1$ est le coefficient de la variable sextail. Il est estimé par -4915 et son écart-type est estimé 2,2192.

Le point critique à 95 % demeure à $t_{24;0,025} = 2,06$. L'intervalle de confiance est [-05,07 ; 4,089]. Ici non plus on ne peut rejeter que l'hypothèse que la différence entre hommes et femmes est nulle.

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.)

On détermine deux modèles de régression, le modèle complet et le modèle réduit. Le modèle complet est celui déterminé en a). Le modèle réduit n'a que la taille comme variable exogène. Les sommes des carrés résiduelles sont :

Modèle complet : SCR = 6012,1, à 24 degrés de liberté; MCR = $\frac{6012,1}{24}$ = 3288,287.

Modèle réduit : SCRo = 6191,7 à 26 degrés de liberté; SCRo – SCR = 2903,366, à 26 - 24 degrés de liberté; $\frac{SCR_o - SCR}{26-24}$ = 1451,683. La statistique F est $F = \frac{1451,683}{3288,287}$ = 0,4415, à 2 et 24 degrés de liberté. La valeur p est 0,648. On ne peut rejeter H_o.

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.

Régression de B1 sur A1 (calculs par le logiciel R)

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.8524     1.1853   1.563    0.123
A1            0.6358     0.1158   5.491 7.36e-07 ***
```

R^2 = 0,3202

Régression de B1 sur A2 (calculs par le logiciel R)

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.3729     1.0001   5.372 1.16e-06 ***
A2            0.5294     0.1799   2.942  0.00454 **
```

R^2 = 0,1191

Dans les deux cas, la valeur p est très petite, ce qui permet d'affirmer avec confiance que la dépendance existe bien dans les deux cas. Mais la relation est plutôt faible, en particulier dans le cas de A2. Donc toute prédiction à partir de l'une ou l'autre de ces deux variables exogènes sera peu précise.

b) Montrer cependant que A2 n'est plus significative (à 5 %) en présence de A1. Comment explique-t-on ce phénomène?

Régression de B1 sur A1 et A2 (calculs par le logiciel R)

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.1252     1.2533   0.898    0.373
A1            0.5699     0.1213   4.696 1.48e-05 ***
A2            0.2688     0.1656   1.623    0.110
```

R^2 = 0,3475.

La variable A2 se révèle encore moins utile ici, avec une valeur p supérieure à 10 %; et R^2 est à peine supérieur à celui d'une régression simple avec A1 comme seule variable exogène.

Si A2 n'est pas significative en présence de A1 alors qu'il l'était tout seul, c'est qu'elle est corrélée avec A1 (r = 0,335) et donc n'apporte pas grand-chose comme information qui ne soit contenue déjà dans A1.

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).

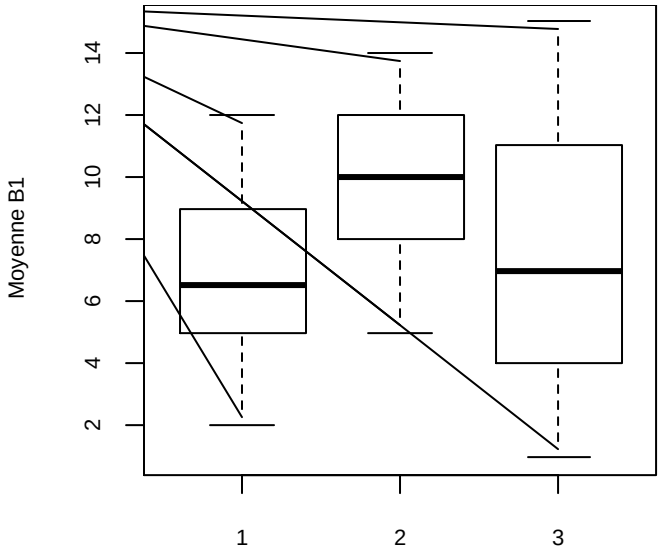
Il s'agit ici de voir si on peut affirmer qu'il y a des différences entre les moyennes de B1 selon le traitement suivi.

Les trois moyennes sont :

Traitement	1	2	3
Moyenne de B1	6,681818	9,772727	7,772727

Le graphique suivant semble révéler une différence entre le traitement 2 et les deux autres.

On définira les variables indicatrices t1, t2, et t3, ti prenant la valeur 1 si le sujet suit le traitement i. On détermine une régression de B1 sur 2 de ces variables. On choisira les variables



Il s'agit d'une analyse de variance à un facteur. Voici la table d'analyse de variance :

```
                    Df Sum Sq Mean Sq F value   Pr(>F)
Traitement           2 108.12  54.061  5.3174 0.007347 **
Residuals           63 640.50  10.167
```

Le facteur Traitement est nettement significatif.

La même analyse peut se faire à l'aide d'une régression multiple de B1 sur deux des variables indicatrices du traitement, t1, t2, et t3, ti prenant la valeur 1 si le sujet suit le traitement *i*. Puisque c'est le traitement 2 qui semble se distinguer des deux autres, on choisira t1 et t3 pour variables exogènes. C'est le choix qui permettra une comparaison immédiate des traitements 1 et 3 au traitement 2.

Voici l'analyse :

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.7727     0.6798  14.376  < 2e-16 ***
t1           -3.0909     0.9614  -3.215  0.00206 **
t3           -2.0000     0.9614  -2.080  0.04157 *
```

À un niveau de 5 %, on peut conclure que les moyennes des groups 2 et 3 sont différentes de la moyenne du groupe 2. On peut ensuite comparer les groupes 1 et 3 par différents moyens. L'un consiste à un test d'égalité de deux moyennes. Les moyennes du groupe 1 et du groupe 2 sont, respectivement,6,681818 et 7,772727, une différence de 1,090909. L'écart-type commun est estimé par $\hat{\sigma}$ = 3,188521 et la statistique *T* est égale à 1,134738, ce qui à 63 degrés de liberté donne *vp* = 0,26078.

Alternativement, on peut déterminer une régression de B1 sur t2 et t3, dont voici un résumé :

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.6818     0.6798   9.829 2.44e-14 ***
t2            3.0909     0.9614   3.215  0.00206 **
t3            1.0909     0.9614   1.135  0.26078
```

On voit là que la valeur *p* qui compare le groupe 3 au groupe 1 est 0,26078.

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.

Les trois moyennes sont :

Traitement	1	2	3
Moyenne de A1	10,5	9,7273	9,1364

Une analyse de variance ne permet pas de conclure que les différences sont significatives :

```
                        Df Sum Sq Mean Sq F value Pr(>F)
as.factor(Traitement)    2  20.58 10.2879  1.1322 0.3288
Residuals               63 572.45  9.0866
```

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.

Ce que l'on pourrait craindre, lorsqu'on détecte des différences dans B1, c'est de conclure erronément que ces différences sont dues aux traitements alors qu'elles ne pourraient être que des différences dans les aptitudes initiales des sujets. On se protège contre ce risque en assignant les sujets aux traitements *au hasard*. C'est l'avantage d'une expérience planifiée comme celle décrite ici, par opposition à une enquête auprès d'une population dont les groupements échappent au contrôle de l'expérimentateur.

Le fait que les différences dans A1 ne sont pas significatives nous rassure : il n'y a pas de raison de croire en un vice dans la procédure, et donc ce serait légitime de ne pas tenir compte de A1. Mais ce n'est pas une raison pour exclure A1 d'emblée. Significatives ou non, les différences initiales existent et méritent qu'on en tienne compte.

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.

Ce modère suppose que B1 dépend de A1. Ce sont les γ_i qui distinguent les traitements. Les différences entre elles sont des différences de moyennes *pour une valeur de* A1 *fixe*.

Le tableau suivant permet de montrer que le facteur A1 est significatif comme prédicteur de B1 et permet d'estimer les paramètres.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.5966     1.1845  -0.504  0.61625
t2            3.6266     0.7362   4.926 6.55e-06 ***
t3            2.0362     0.7450   2.733  0.00816 **
A1            0.6932     0.1015   6.831 4.21e-09 ***
```

On obtient les estimations suivantes :

$\hat{\gamma}_1 = -0{,}5966$; $\hat{\gamma}_2 = -0{,}5966 + 3{,}6266 = -0{,}5966$; $\hat{\gamma}_3 = -0{,}5966 + 2{,}0362 = 1{,}4395$; $\hat{\beta} = 0{,}6932$.

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

Le tableau en d-ii) permet de conclure que le groupe 1 diffère des groupes 2 et 3 quant à son effet sur B1. Afin de déterminer si la différence entre γ_3 et γ_2 est significative on refait la régression en remplaçant t2 et t3 par t1 et t3:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0299     1.1145   2.719  0.00849 **
t1           -3.6266     0.7362  -4.926 6.55e-06 ***
t3           -1.5904     0.7345  -2.165  0.03421 *
A1            0.6932     0.1015   6.831 4.21e-09 ***
```

La valeur *p* de 0,03421 montre que la différence entre γ_2 et γ_3 est significative.

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: $E(B1) = \gamma_j + \beta_1(A1) + \beta_2(A2)$, où $\gamma_j = \gamma_1$, γ_2 ou γ_3, selon le traitement. Estimer les paramètres γ_j , β_1 et β_2.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.3171     1.2216  -1.078  0.28517
t2            3.6244     0.7214   5.024 4.69e-06 ***
t3            2.0316     0.7300   2.783  0.00716 **
A1            0.6274     0.1053   5.957 1.38e-07 ***
A2            0.2676     0.1415   1.891  0.06337 .
```

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.

Modèle	Variables exogènes	Coefficient de revenu	*Valeur p*
A	revenu	0,25900	6,9735e-10
B	revenu, personnes	0,24210	3,8873e-12
C	revenu, surface	-0,13498	0,1110
D	revenu, personnes, surface	0,0751369 6	0,5850

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.
Le coefficient de corrélation entre voc et age est 0,0477; le coefficient de corrélation partiel est 0,6723. Ce qui veut dire que pour des personnes de même scolarité il y a bel et bien une relation (positive) entre l'âge et le vocabulaire : le vocabulaire s'accroît avec l'âge. Mais étant donné que les personnes plus âgées sont moins scolarisées, le gain dû à l'âge est compensé par la perte due à la scolarité réduite.

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) + β_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 $\mathcal{F}(2 ; 8)$. 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, $E(Y_i) = \beta_o + \beta_1 x_i + \beta_2(\text{méthode}) + \beta_3(\text{sexe})$, 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)

```
> LM1<-lm(y~x+tr+sexe)
> summary(LM1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.3235     5.1996   3.139  0.00567 **
x             0.4866     0.2343   2.077  0.05238 .
tr1          -4.2809     1.5263  -2.805  0.01171 *
sexe         -1.0362     1.7121  -0.605  0.55259

Residual standard error: 3.564 on 18 degrees of freedom
Multiple R-squared:  0.4127,    Adjusted R-squared:  0.3149
F-statistic: 4.217 on 3 and 18 DF,  p-value: 0.02007

> anova(LM1)
Analysis of Variance Table
Response: y
          Df  Sum Sq Mean Sq F value  Pr(>F)
x          1  58.852  58.852  4.6338 0.04516 *
```

```
tr          1  97.160  97.160  7.6501 0.01273 *
sexe        1   4.652   4.652  0.3663 0.55259
Residuals 18 228.609  12.700
```

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*).

```
LM2<-lm(y~x+tr)
summary(LM2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.2241     4.8984   3.516  0.00231 **
x             0.4267     0.2088   2.044  0.05508 .
tr1          -4.2085     1.4960  -2.813  0.01110 *


Residual standard error: 3.504 on 19 degrees of freedom
Multiple R-squared:  0.4008,    Adjusted R-squared:  0.3377
F-statistic: 6.354 on 2 and 19 DF,  p-value: 0.00771

> anova(LM2)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value  Pr(>F)
x          1  58.852  58.852  4.7937 0.04125 *
tr         1  97.160  97.160  7.9141 0.01110 *
Residuals 19 233.261  12.277
```

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) $+ \beta_3$ (sexe) ou $E(Z_i|x_i) = \beta_o + \gamma x_i + \beta_2$ (méthode) $+ \beta_3$ (sexe), où $\gamma = \beta_1 - 1$. (Voir la section 9.6)

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

```
LM3<-lm(z~x+tr+sexe)
> summary(LM3)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.3235     5.1996   3.139  0.00567 **
x            -0.5134     0.2343  -2.192  0.04181 *
tr1          -4.2809     1.5263  -2.805  0.01171 *
sexe         -1.0362     1.7121  -0.605  0.55259

Residual standard error: 3.564 on 18 degrees of freedom
Multiple R-squared:  0.4476,    Adjusted R-squared:  0.3555
F-statistic: 4.861 on 3 and 18 DF,  p-value: 0.01196

> anova(LM3)
Analysis of Variance Table

Response: z
          Df  Sum Sq Mean Sq F value  Pr(>F)
```

```
x          1  83.397  83.397  6.5665 0.01958 *
tr         1  97.160  97.160  7.6501 0.01273 *
sexe       1   4.652   4.652  0.3663 0.55259
Residuals 18 228.609  12.700
```

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

```
> summary(LM4)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.2241     4.8984   3.516  0.00231 **
x            -0.5733     0.2088  -2.746  0.01285 *
tr1          -4.2085     1.4960  -2.813  0.01110 *

Residual standard error: 3.504 on 19 degrees of freedom
Multiple R-squared:  0.4363,    Adjusted R-squared:  0.377
F-statistic: 7.354 on 2 and 19 DF,  p-value: 0.004313

> anova(LM4)
Analysis of Variance Table

Response: z
          Df  Sum Sq Mean Sq F value  Pr(>F)
x          1  83.397  83.397  6.7930 0.01735 *
tr         1  97.160  97.160  7.9141 0.01110 *
Residuals 19 233.261  12.277
---
> SCR0 =  510;   SCR =  233.2606;   SCR0-SCR =  276.7394;
dl0 =  22;   dl =  19;   (SCR0-SCR)/(dl0-dl) =  92.24645;
MCR =  12.27688;    f3.19<-(SCR0-SCR)/(dl0-dl)/MCR;    1-
pf(f3.19,3,19) =  0.001637719
```

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

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 $SCR = \sum_{i=1}^{n} z_i^2 - \dfrac{[\sum_{i=1}^{n} x_i z_i]^2}{\sum_{i-1}^{n} x_i^2}$.

c) Modèle 3.

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

ii) Montrer que $SCR = (n-1)\left(S_Z^2 - \dfrac{S_{ZX}^2}{S_X^2} \right)$, 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

scr0=3186,949 dans les trois modèles

Modèle	scr	scro-scr	F	d.l.	vp
Modèle 1	3168,949	17.424	0,2144	1 et 39	0,6459
Modèle 2	3174,172	12,20017	0,1499	1 et 39	0,7007
Modèle 3	3157.707	28,66545	0,17248	2 et 38	0,8422