STT1000

CHAPITRE 8 RÉGRESSION LINÉAIRE SIMPLE

EXERCICES

8.1 Les données dans le tableau 8.3 ci-dessous portent sur un certain nombre d'indiens qui ont immigré des montagnes vers les plaines. Le but de cette collecte de données était de savoir si le pouls est affecté par le nombre d'années passées dans les plaines.

a) Faites un nuage de points avec le nombre d'années comme variable exogène (x) et le pouls (y) comme variable endogène.
b) Estimez les paramètres.
d) Retrouver les quantités que vous avez calculées dans le tableau suivant :
e) Dressez une table d'analyse de variance.

8.2 Une usine fabrique des toiles métalliques pour des usines de pâtes et papier. Afin de mieux répartir son personnel, le gérant aimerait prévoir le temps, T, requis pour la finition des toiles. Ce temps pourrait être lié, entre autres variables, à la surface de la toile, S. On a obtenu les données du tableau 8.3:

a) Quelle variable doit-on utiliser comme variable dépendante? (Justifier ce choix).
b) Déterminer l'équation de régression correspondante.
c) La régression est-elle utilisable pour des fins de prévision? (α = 5%).
d) Quel est le temps moyen de finition pour une toile de 20 m^2? Déterminer un intervalle de confiance à 80 % pour cette moyenne. 7,26; intervalle de confiance : [7,12832 ; 7,382748]
e) Prédire le temps de finition d'une toile de 20 m^2? Déterminer des limites de prédiction à 70 %. Prédiction : 7,26 ; limites de prédiction : [6,99 ; 7,52].
f) Quel est le pourcentage de variation expliquée par la droite de régression? Le pourcentage expliqué est le carré du coefficient de corrélation, r^2 = 87,82 %.
g) Faire un graphique des données. Tracer la droite de régression. Le modèle est-il raisonnable?

8.3 Un professeur de secondaire est responsable de l'enseignement de l'algèbre. Au début de l'année, il fait passer à 20 de ses étudiants un petit test mesurant les habiletés arithmétiques (H) de ses étudiants. À la fin du premier semestre, il examine les résultats (F) de ses étudiants à l'examen d'algèbre. Les résultats sont présentés au tableau 8.5.

a) Quelle variable doit-on utiliser comme variable dépendante? (Justifier ce choix).
b) Déterminer l'équation de régression correspondante.
c) La régression est-elle utilisable pour des fins de prévision? (Faire un test à 5%),
d) À quelle note d'algèbre devrait-on s'attendre d'un étudiant dont le score au test d'habileté est 25? Déterminer un intervalle de confiance à 80 % pour la moyenne en algèbre des étudiants dont la le score au test d'habileté est 25 Estimation : 60,4761; intervalle de confiance : [57,48 ; 63,47].
e) Déterminer des limites de prédiction, à 70 %, de la note en algèbre pour un étudiant dont le score en habileté est 25. Prédiction : 60,4761 ; limites de prédiction : [49,47 ; 71,48].
f) Quel est le pourcentage de variation expliquée par la droite de régression? 82,44 %
g) Faire un graphique des données. Tracer la droite de régression. Le modèle est-il raisonnable?

8.4 [*Données du tableau* A.3] On s'intéresse à la relation entre la taille du cerveau (*IRM*) et les différentes mesures d'aptitudes mentales.

a) Déterminez s'il y a une relation significative entre la taille du cerveau et (*i*) le score *G*; (*ii*) le score *V*; (*iii*) le score *P*.
b) Reprenez l'exercice ci-dessus en prenant pour variable dépendante la mesure relative *IRM*/*taille*.
c) Examinez la dépendance entre *IRM* et la mesure générale d'aptitude mentale $y = (G + V + P)/3$.

8.5 [*Données du tableau* 8.1] Il existe une prédiction naturelle du prix y d'une maison à partir de sa surface de plancher x, soit $\hat{y}_R = \hat{R}x$, où $\hat{R} = \sum_{i=1}^{n} y_i / \sum_{i=1}^{n} x_i$ est une estimation du quotient du prix par mètre carré. Faites une prédiction $\hat{y}_{iR} = \hat{R}\, x_i$ pour chaque i, puis calculez la somme des carrés des erreurs $\sum_{i=1}^{n} (y_i - \hat{y}_{iR})^2$. Comparez cette somme avec $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$, où $\hat{y}_i$ est la prédiction de y_i basée sur la droite de régression. Dites pourquoi l'inégalité $\sum_{i=1}^{n} (y_i - \hat{y}_{iR})^2 \geq \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ est nécessairement vérifiée.

8.6 Dans l'exemple du tableau 8.1, déterminez la cote Z d'une maison dont le prix est de 800 000 $. Ensuite déterminer une cote Z qui tient compte du fait que la superficie de la maison est de 5000 m^2.

8.7 [Crowder, M. and Hand, D. (1990), *Analysis of Repeated Measures*, Chapman and Hall] Le tableau 8.6 présente les poids de 45 poussins à différents moments : à 6 semaines de la naissance, à 10 semaines et à 21 semaines. Faites une analyse pour déterminer laquelle des trois variables suivantes prédit le mieux le poids à l'âge de 21 semaines a) x_1 : le poids à la sixième semaine; b) x_2 : le poids à la dixième semaine; ou c) x_3 : la moyenne $(x_1+x_2)/2$.

8.8 Considérer un modèle de régression sans la constante β_o : $y_i = \beta x_i + \varepsilon_i$, $i = 1, \ldots, n$, où les ε_i sont indépendantes, de moyenne 0 et de variance σ^2.

a) Montrer que l'estimateur des moindres carrés de β est $\hat{\beta} = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$.

b) Montrer que $\mathrm{Var}(\hat{\beta}) = \dfrac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$ [*Rappelez-vous que les x sont des constantes et non des variables aléatoires. Ceci devrait faciliter la démonstration*].

c) Considérons un autre critère pour la sélections d'un estimateur de β. On commence par limiter le choix à l'ensemble des estimateurs linéaires, c'est-à-dire des estimateurs de la forme $\tilde{\beta} = \sum_{i=1}^{n} \ell_i y_i$, puis on les limite encore aux estimateurs *sans biais*. Dans cet ensemble réduit, on choisit celui qui a la plus petite variance.

i) Montrer que $\tilde{\beta}$ est sans biais si et seulement si $\sum_{i=1}^{n} \ell_i = 1$.

ii) Montrer que pour tout estimateur sans biais $\tilde{\beta} = \sum_{i=1}^{n} \ell_i y_i$, $\mathrm{Var}(\tilde{\beta}) \geq \sigma^2 \dfrac{\left(\sum_{i=1}^{n} \ell_i x_i\right)^2}{\sum_{i=1}^{n} x_i^2}$ [Utiliser l'inégalité de Cauchy-Schwartz, $\left(\sum_{i=1}^{n} \ell_i x_i\right)^2 \geq \dfrac{\sum_{i=1}^{n} \ell_i^2}{\sum_{i=1}^{n} x_i^2}$]

iii) Montrer $\mathrm{Var}(\tilde{\beta}) = \sigma^2 \dfrac{\left(\sum_{i=1}^{n} \ell_i x_i\right)^2}{\sum_{i=1}^{n} x_i^2}$ si et seulement si $\ell_i = a\, x_i$ (cela découle directement du théorème de Cauchy-Schwart), auquel cas $\mathrm{Var}(\tilde{\beta}) = \sigma^2 \dfrac{\left(\sum_{i=1}^{n} \ell_i x_i\right)^2}{\sum_{i=1}^{n} x_i^2}$

8.9 [*Données du tableau* A.3] On considère la relation entre la taille et le poids.

a) Déterminez une droite de régression du poids sur la taille, et testez l'hypothèse que ces deux variables sont indépendantes.

b) La droite de régression n'est probablement pas la même pour les hommes et les femmes. Déterminez deux droites de régression, disons $\hat{\mu}_{of} = \hat{\beta}_{of} + \hat{\beta}_{1f} x_f$ et $\hat{\mu}_{om} = \hat{\beta}_{om} + \hat{\beta}_{1m} x_m$, pour les femmes et pour les hommes, respectivement. Considérez la variable

$$Z = \frac{\hat{\beta}_{1f} - \hat{\beta}_{1m} - \mathrm{E}(\hat{\beta}_{1f} - \hat{\beta}_{1m})}{\sqrt{\hat{\sigma}_f^2 + \hat{\sigma}_m^2}},$$

où $\hat{\sigma}^2_{b_{1f}}$ et $\hat{\sigma}^2_{b_{1m}}$ sont les variances estimées de $\hat{\beta}_{1f}$ et de $\hat{\beta}_{1m}$, respectivement.

Admettons que Z suit à peu près une loi $\mathcal{N}(0 ; 1)$. Utilisez ce résultat pour tester l'hypothèse que les coefficients β_{1f} et β_{1m} sont égaux.

8.10 [*Données du tableau* A.5] On s'intéresse à la relation entre la température (y) et le nombre de battements du cœur (x).

a) Déterminez s'il y a une relation significative entre y et x.

b) Reprenez l'exercice *a*) en considérant séparément les femmes et les hommes.

c) Utilisez le test approximatif décrit en 8.9 b) pour tester l'hypothèse que le taux de croissance de y par rapport à x est le même chez les hommes et chez les femmes.

de la variance $\sigma^2_{\hat{\mu}_{x_o}}$ de cet estimateur en utilisant le fait (non démontré dans ce cours) que $\bar{y}$ et $\hat{\beta}_1$ sont des variables aléatoires indépendantes et que $\hat{\mu}_{x_o} = \bar{y} + \hat{\beta}_1(x_o - \bar{x})$. Ensuite déterminez une expression pour un estimateur sans biais $\hat{\sigma}^2_{\hat{y}_o}$ de $\sigma^2_{\hat{y}_o}$.

e) Soit $\beta_o+\beta_1 x$ et de $\gamma_o+\gamma_1 x$ les droites de régression de y sur x pour les femmes et les hommes respectivement et soit $\hat{\nu}_o = \hat{\beta}_o + \hat{\beta}_1 x_o$ et $\hat{\mu}_o = \hat{\gamma}_o + \hat{\gamma}_1 x_o$ les estimateurs $\beta_o+\beta_1 x_o$ et de $\gamma_o+\gamma_1 x_o$, respectivement, où x_o = 73,8, la valeur moyenne de x pour l'ensemble de l'échantillon. Soit $\hat{\sigma}^2_{\hat{\nu}_o}$ et $\hat{\sigma}^2_{\hat{\mu}_o}$ les variances estimées des prédictions. Admettons que la variable $Z = \dfrac{\hat{\mu}_o - \hat{\nu}_o - \mathrm{E}(\hat{\mu}_o - \hat{\nu}_o)}{\sqrt{\hat{\sigma}^2_{\hat{\mu}_o} + \hat{\sigma}^2_{\hat{\nu}_o}}}$ suit approximativement une loi $\mathcal{N}(0 ; 1)$. Utilisez ce fait pour tester l'hypothèse que $\beta_o+\beta_1 x_o = \gamma_o+\gamma_1 x_o$. Expliquez le sens concret de cette hypothèse.

8.11 [*Données du tableau* A.4] Soit x = (A1+A2)/2 et y = (B1+B2)/2. Considérez la régression de y sur x pour le groupe 1 et pour le groupe 2 séparément. Soit x_o = 7,4 et testez l'hypothèse que $\beta_o + \beta_1 x_o = \gamma_o + \gamma_1 x_o$, $\beta_1 + \beta_1 x$ et $\gamma_o + \gamma_1 x_o$ étant, respectivement, les droites de régression pour les groupes 1 et 2. Expliquez le sens concret de cette hypothèse *Voir l'exercice* 8.10].

8.12 Les poussins dont il est question au numéro 8.7 ont été répartis en 4 groupes, et chaque groupe a été soumis à un régime alimentaire différent. Soit y_2 et y_3 les poids à la 21e semaine des poussins des groupes 2 et 3, respectivement; et x_2 et x_3 leurs poids à la 10e semaine. Supposons deux modèles différents de régression, soit $y_2 = \beta_o+\beta_1 x_2$ et $y_3 = \gamma_o+\gamma_1 x_3$.

a) Supposons que vous voulez tester l'hypothèse que $\beta_1 = \gamma_1$. Les techniques pour le faire n'ont pas été présentées dans ce chapitre, mais vous pouvez vous en faire une idée à l'aide d'un test grossièrement approximatif. Sachant que la statistique $Z = \dfrac{\hat{\beta}_1 - \hat{\gamma}_1}{\sigma_{\hat{\beta}_1 - \hat{\gamma}_1}}$ est de loi $\mathcal{N}(0 ; 1)$ sous cette hypothèse, remplacez l'écart-type au dénominateur par une estimation $\hat{\sigma}_{\hat{\beta}_1 - \hat{\gamma}_1}$ et dites ce que vous en concluriez si $Z = \dfrac{\hat{\beta}_1 - \hat{\gamma}_1}{\hat{\sigma}_{\hat{\beta}_1 - \hat{\gamma}_1}}$ était en fait de loi $\mathcal{N}(0 ; 1)$.

$\hat{\beta}_1$ = 2,751153; $\hat{\gamma}_1$ = 2,712720; $\hat{\sigma}_{\hat{\beta}_1}$ = 0,5889297; $\hat{\sigma}_{\hat{\gamma}_1}$ = 0,8081447; Z = 0,0384.

b) De la même manière, testez l'hypothèse que $\mu_o = \beta_o + \beta_1(110) = \nu_o = \gamma_o+\gamma_1(110)$.
$\hat{\mu}_o$ = 218,8267; $\hat{\nu}_o$ = 251,0397; $\hat{\sigma}_{\hat{\mu}_o}$ = 13,60286; $\hat{\sigma}_{\hat{\nu}_o}$ = 16,50872; Z = -1,51.

8.13 a) Montrer que (8.2.3) peut s'écrire comme $\hat{\sigma}^2 = \dfrac{n-1}{n-2} S_y^2(1-r^2)$, où r est le coefficient de corrélation.

b) Montrer que (8.3.5) peut s'écrire comme $\hat{\sigma}^2_{\hat{\beta}_1} = \dfrac{S_y^2(1-r^2)}{(n-2)}$

8.14 Soit T_1 la statistique (8.4.1) pour tester l'hypothèse $\mathrm{H_o}$: $\beta_1 = 0$ dans le modèle (8.1.7). Montrer que $T_1^2 = \dfrac{(n-2)r^2}{1-r^2}$, où r est le coefficient de corrélation.

Sous $\mathrm{H_o}$, $T_1^2 \sim \mathcal{F}_{1;n-2}$. Ceci découle du fait que $T_1 \sim t_{n-2}$. Donc la valeur p pour tester $\mathrm{H_o}$ (contre une alternative bilatérale) est donnée par $vp = \mathrm{P}\left(\mathcal{F}_{1;n-2} > \dfrac{(n-2)r_o^2}{1-r_o^2}\right)$, où $\mathcal{F}_{1;n-2}$ désigne une variable de loi de loi de Fisher à 1 et $n-2$ degrés de liberté et r_o est le coefficient de corrélation observé.

Note : Le modèle de régression traite le vecteur $\boldsymbol{x}$ comme fixe. Si en fait $\boldsymbol{x}$ n'est pas fixe, le même modèle peut quand même s'appliquer, à condition d'interpréter l'espérance comme une espérance *conditionnelle* : $E(y_i \mid x_i) = \beta_o + \beta_1 x_i$, $i = 1, \ldots, n$. La valeur p est conditionnelle aussi : $P\left(\mathcal{F}_{1;n-2} > \frac{(n-2)r_o^2}{1-r_o^2} \middle| \boldsymbol{x} \right)$. Mais puisque cette probabilité ne dépend pas de $\boldsymbol{x}$, c'est aussi une probabilité non conditionnelle.

Tableau 8.2 *Pouls (y) et Nombre d'années dans les plaines (x)*

	Nombre d'années	***Pouls***
1	1	88
2	6	64
3	5	68
4	1	52
5	1	72
6	19	72
7	5	64
8	25	80
9	6	76
10	13	60
11	13	68
12	10	72
13	15	88
14	18	60
15	2	60
16	12	72
17	15	84
18	16	64
19	17	72
20	10	64
21	18	80
22	11	76
23	11	60
24	21	64
25	24	64
26	14	68
27	25	76
28	32	60
29	5	76
30	12	88
31	25	72
32	26	68
33	10	60
34	19	74
35	18	72
36	10	56
37	1	64
38	43	72
39	40	92

Tableau 8.3
Temps de finition d'une toile (T) et surface de la toile (S)

i	*T*	*S*
1	5,50	9,30
2	5,90	13,50
3	5,80	11,10
4	6,30	14,90
5	7,00	16,70
6	7,50	23,20
7	5,50	11,10
8	7,20	20,40
9	6,50	15,80
10	6,50	14,90
11	7,10	18,60
12	7,00	15,80
13	6,90	16,70
14	6,80	15,80
15	6,60	16,70
Totaux	98,10	234,50

Tableau 8.4
Habileté mathématique (H) et résultat à un examen d'algèbre (F)

i	*F*	*H*
1	36	9
2	23	10
3	22	13
4	36	15
5	49	16
6	32	18
7	44	20
8	52	22
9	51	23
10	83	24
11	59	26
12	58	28
13	72	30
14	87	31
15	86	32
16	79	33
17	74	34
18	78	36
19	99	38
20	85	40
Totaux	1 205	498

Tableau 8.5

Poids de 45 poussins aux âges de 6, 10, *et* 21 *semaine*

#	Groupe	Sixième semaine	Dixième semaine	21e semaine	#	Groupe	Sixième semaine	Dixième semaine	21e semaine
1	1	64	93	205	24	2	73	114	233
2	1	72	103	215	25	2	74	106	309
3	1	67	99	202	26	2	72	98	150
4	1	67	87	157	27	3	73	102	256
5	1	60	106	223	28	3	82	129	305
6	1	74	124	157	29	3	77	111	147
7	1	71	112	305	30	3	85	134	341
8	1	68	96	98	31	3	87	158	373
9	1	63	81	124	32	3	76	116	220
10	1	84	139	175	33	3	68	83	178
11	1	62	88	205	34	3	74	109	290
12	1	60	67	96	35	3	78	109	272
13	1	79	128	266	36	3	79	120	321
14	1	72	89	142	37	4	85	124	204
15	1	62	71	157	38	4	84	126	281
16	1	58	73	117	39	4	96	157	200
17	2	86	163	331	40	4	78	117	196
18	2	77	95	167	41	4	82	120	238
19	2	73	103	175	42	4	79	123	205
20	2	74	68	74	43	4	80	125	322
21	2	78	124	265	44	4	85	128	237
22	2	74	114	251	45	4	84	122	264
23	2	73	100	192					