

Partie IV

Méthodes de régression

17. Inférence classique pour la régression simple
18. Régression multiple: introduction
19. Ajustement du modèle de régression multiple
20. Inférence classique pour la régression multiple
21. Inférence par bootstrap pour la régression
22. Introduction à la régression logistique
23. Introduction à l'analyse de survie

Chapitre 17

Inférence classique pour la régression simple

Dans le Chapitre 3, le modèle de régression simple a été introduit. Ce modèle décrit la relation entre deux variables X et Y à l'aide d'une droite. X est la variable explicative et Y la réponse. Les coefficients a (intercept) et b (pente) de la droite sont déterminés à l'aide d'un échantillon $(x_1, y_1), \dots, (x_n, y_n)$: le critère des moindres carrés fournit les estimations \hat{a} et \hat{b} . Les méthodes d'inférence permettent de tester des hypothèses telles que " b (ou a) est égal à une valeur spécifiée" et de déterminer des intervalles de confiance pour a et b . Dans ce chapitre, nous utilisons les notations introduites au Chapitre 3.

17.1 Modèle classique pour l'inférence

Selon l'approche introduite au Chapitre 8, Section 8.3, nous décrivons les réponses à l'aide de variables aléatoires Y_1, \dots, Y_n . Il n'est pas nécessaire de supposer que les x_i sont obtenus de façon aléatoire. Les x_i pourraient être, par exemple, les doses d'un médicament, fixées arbitrairement lors d'une expérience où les Y_i représentent une mesure d'amélioration; plusieurs individus pourraient être soumis à la même dose et manifester des niveaux différents d'amélioration. L'approche classique à l'inférence, se fonde sur un ensemble de conditions connues comme le *modèle de Gauss*.

1. $Y_i = a + bx_i + U_i, i = 1, \dots, n$ où a et b sont des paramètres.

2. Les erreurs U_i sont i.i.d. et indépendantes de X_i .

La moyenne des erreurs est nulle et la variance est un paramètre noté σ^2 .

3. $U_i \sim \mathcal{N}(0, \sigma^2)$.

La distribution de Y_i en fonction de x_i est esquissée dans la Figure 1.

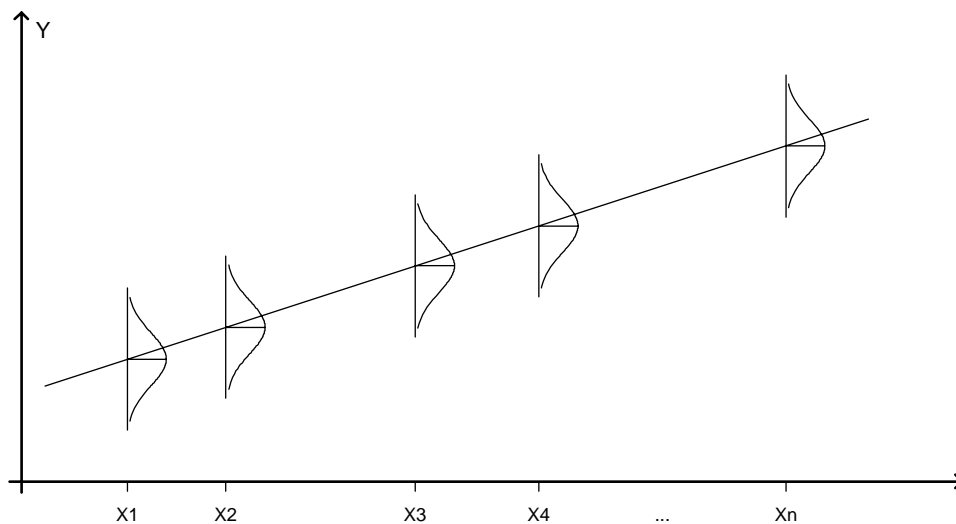


Figure 1. Distribution de Y_i en fonction de x_i

Remarques

1. Souvent les X_i sont obtenus de façon aléatoire simultanément aux Y_i . Dans ce cas, il faudra interpréter les résultats concernant la distribution des estimateurs de façon conditionnelle, les valeurs observées des X_i étant données.

2. Les équations $Y_i = a + bx_i + U_i$ pour les variables aléatoires correspondent à n équations pour les réponses observées:

$$y_i = a + bx_i + u_i, \quad i = 1, \dots, n.$$

Notez que les erreurs u_i ne sont pas observables (car a et b sont inconnus).

3. On dit que les équations $Y_i = a + bx_i + U_i$ caractérisent la “structure du modèle”, tandis que les conditions 2 et 3 caractérisent la “partie aléatoire du modèle”. Une autre expression de la structure du modèle est

$$E(Y|X = x) = a + bx.$$

Ici, $E(Y|X = x)$ est l’espérance conditionnelle de Y pour $X = x$ (c’est-à-dire, l’espérance de la distribution conditionnelle de Y pour $X = x$ donné).

17.2 Distributions des estimateurs

Les résultats suivants s’obtiennent sous le modèle de Gauss.

– les estimateurs \hat{a} et \hat{b} suivent des distributions de Gauss:

$$\hat{a} \sim \mathcal{N}(a, \sigma^2(\hat{a})), \quad \hat{b} \sim \mathcal{N}(b, \sigma^2(\hat{b})),$$

où

$$\sigma^2(\hat{a}) = \left[\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}^2} \right] \sigma^2, \quad \sigma^2(\hat{b}) = \frac{1}{s_{xx}^2} \cdot \sigma^2, \quad s_{xx}^2 = \sum_{i=1}^n (x_i - \bar{x})^2.$$

En outre, si $\hat{y}_x = \hat{a} + \hat{b}x$ indique la réponse calculée en fonction d’une valeur x donnée, alors \hat{y}_x suit une distribution de Gauss de moyenne $y_x = a + bx$ et de variance

$$\sigma^2(\hat{y}_x) = \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{s_{xx}^2} \right] \sigma^2.$$

Ces résultats pourraient permettre de réaliser des inférences si σ^2 était connu. Mais en pratique, σ^2 est presque toujours inconnu et il faut l’estimer. Dans ce but, on utilise l’estimateur

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2,$$

où $e_i = y_i - (\hat{a} + \hat{b}x_i)$. (Noter que les résidus observés e_i ne sont pas les erreurs aléatoires U_i et que $\hat{\sigma}^2$ est noté s_E^2 au Chapitre 3.) Des estimations $\hat{\sigma}^2(\hat{a})$, $\hat{\sigma}^2(\hat{b})$ et $\hat{\sigma}^2(\hat{y}_x)$ des variances de \hat{a} , \hat{b} et \hat{y}_x sont alors obtenues des expressions de $\sigma^2(\hat{a})$, $\sigma^2(\hat{b})$ et $\sigma^2(\hat{y}_x)$, en remplaçant σ^2 par $\hat{\sigma}^2$. On démontre alors que:

- La variable aléatoire $(n-2)\hat{\sigma}^2/\sigma^2$ suit une distribution χ^2 à $n-2$ degrés de liberté.
- Les estimateurs standardisés

$$(\hat{a} - a)/\hat{\sigma}(\hat{a}), \quad (\hat{b} - b)/\hat{\sigma}(\hat{b}), \quad (\hat{y}_x - y_x)/\hat{\sigma}(\hat{y}_x)$$

suivent une distribution t à $n-2$ degrés de liberté.

17.3 Intervalles de confiance usuels

Le résultats précédents permettent d'obtenir les intervalles de confiance pour a , b et $y_x = a + bx$ pour un x fixé. Soit α une probabilité préfixée (par exemple, $\alpha = 2.5\%$). Alors, des intervalles de confiance bilatéraux avec coefficient de couverture $1 - 2\alpha$ sont:

$$\begin{aligned} & [\hat{a} - \hat{\sigma}(\hat{a}) t_{1-\alpha, n-2}, \quad \hat{a} + \hat{\sigma}(\hat{a}) t_{1-\alpha, n-2}], \\ & [\hat{b} - \hat{\sigma}(\hat{b}) t_{1-\alpha, n-2}, \quad \hat{b} + \hat{\sigma}(\hat{b}) t_{1-\alpha, n-2}], \\ & [\hat{y}_x - \hat{\sigma}(\hat{y}_x) t_{1-\alpha, n-2}, \quad \hat{y}_x + \hat{\sigma}(\hat{y}_x) t_{1-\alpha, n-2}], \end{aligned}$$

où $t_{1-\alpha, n-2}$ indique le percentile $1 - \alpha$ de la distribution t à $n - 2$ degrés de liberté.

En outre,

$$[(n - 2)\hat{\sigma}^2/\chi_{1-\alpha, n-2}^2, \quad (n - 2)\hat{\sigma}^2/\chi_{\alpha, n-2}^2],$$

où $\chi_{\alpha, n-2}^2$ est le percentile α de la distribution χ^2 à $n - 2$ degrés de liberté, est un intervalle de confiance avec coefficient de couverture $1 - 2\alpha$ pour σ^2

17.4 Tests usuels

L'hypothèse

$$H_0 : b = b_0,$$

où b_0 est une valeur donnée, peut être rejetée au niveau α , en faveur de l'alternative $H_1 : b \neq b_0$, si la statistique de test

$$T = \frac{(\hat{b} - b_0)}{\hat{\sigma}(\hat{b})}$$

n'appartient pas à l'intervalle $[t_{\alpha/2, n-2}, t_{1-\alpha/2, n-2}]$. Un exemple fréquent est $b_0 = 0$, auquel cas H_0 signifie que la covariable n'explique pas la réponse. De façon équivalente, on peut rejeter H_0 en faveur de H_1 au niveau α si l'intervalle de confiance avec coefficient de couverture $1 - \alpha$ pour b ne contient pas b_0 . L'hypothèse $H_0 : a = a_0$ contre l'alternative $H_1 : a \neq a_0$, où a_0 est une valeur donnée, est traitée de la même manière.

Remarques

1. Il est possible d'ajuster aux données une droite qui passe par l'origine, c'est-à-dire, d'imposer la condition $a = 0$ au modèle. On peut alors étudier les distributions de \hat{b} , $\hat{y}_x = \hat{b}x$ et $\hat{\sigma}$ et établir de nouvelles formules pour les intervalles de confiance et les tests. Voir Chapitre 18 pour une approche générale à la régression qui inclue le modèle $Y_i = bx_i$.
2. Les logiciels de statistique courants fournissent dans leurs outputs standards les valeurs de $\hat{\sigma}(\hat{a})$ et de $\hat{\sigma}(\hat{b})$, ainsi que celles des statistiques $\hat{a}/\hat{\sigma}(\hat{a})$ et $\hat{b}/\hat{\sigma}(\hat{b})$ et les P-values correspondantes. Par exemple, R et S-plus calculent

$$P(|t_{n-2}| > |\hat{a}/\hat{\sigma}(\hat{a})|) \quad \text{et} \quad P(|t_{n-2}| > |\hat{b}/\hat{\sigma}(\hat{b})|),$$

où t_{n-2} indique une variable aléatoire qui suit une distribution t à $n - 2$ degrés de liberté et $\hat{a}/\hat{\sigma}(\hat{a})$ et $\hat{b}/\hat{\sigma}(\hat{b})$ désignent les valeurs observées des statistiques correspondantes.

17.5 Analyse des résidus

Si le modèle de Gauss est approprié, les résidus ont approximativement une distribution de Gauss. Il faut donc examiner cette condition à l'aide d'un qq-plot. En outre, la variance des résidus ne doit pas dépendre de la variable explicative. Il est donc opportun de représenter graphiquement les résidus en fonction des valeurs observées de X . Aucune

relation (relation non linéaire, variance non homogène) ne doit apparaître. Si une relation apparaît le modèle de Gauss et les inférences obtenues avec son appui doivent être mis en doute.

17.6 Exemple

La Table 1 donne les temps t [s] de chute d'une bille lâchée de différentes hauteurs h [m]. Les mesures ont été prises par une étudiante du gymnase aux travaux pratiques de physique, dans le but de vérifier la relation $h = (1/2)\gamma t^2$ avec $\gamma = 9.81$ [m/s²].

Table 1. Mesures des hauteurs h et des temps t

h [m]	t [s]	h [m]	t
0.15	0.173	0.15	0.179
0.15	0.177	0.15	0.184
0.20	0.199	0.20	0.201
0.20	0.218	0.20	0.202
0.25	0.244	0.25	0.225
0.25	0.227	0.25	0.226
0.30	0.244	0.30	0.253
0.30	0.244	0.30	0.248
0.35	0.275	0.35	0.270
0.35	0.268	0.35	0.264
0.40	0.289	0.40	0.284
0.40	0.288	0.40	0.283
0.45	0.308	0.45	0.298
0.45	0.305	0.45	0.302
0.50	0.331	0.50	0.318
0.50	0.319	0.50	0.319
0.55	0.332	0.55	0.333
0.55	0.355	0.55	0.331
0.60	0.360	0.60	0.350
0.60	0.347	0.60	0.349

Les points $(h_i, t2_i)$, avec $t2_i = t_i^2$ sont représentés dans la Figure 2. L'allure est celle d'une relation linéaire; la relation entre h et $t2$ peut donc être décrite par le modèle $h = a + b \cdot (t2)$. Un programme de régression simple donne les résultats suivants:

Coefficients:

	Value	Std.Error	t value	Pr(> t)
Intercept a	0.0001	0.0076	0.0176	0.9861
Pente b	4.8320	0.0917	52.6938	0.0000

Residual standard error: 0.01712 on 38 degrees of freedom
Multiple R-Squared: 0.9865

Correlation of Coefficients:

Intercept
b -0.9346

Dans les notations des sections précédentes, nous avons donc:

$$\begin{aligned}\hat{a} &= 0.0001, & \hat{b} &= 4.8320 \\ \hat{\sigma}(\hat{a}) &= 0.0076, & \hat{\sigma}(\hat{b}) &= 0.0917.\end{aligned}$$

La proportion de variance expliquée par le modèle est $R^2 = 0.9865$ et l'erreur standard des résidus est $\hat{\sigma} = 0.01712$. (Le programme nous donne aussi le coefficient de corrélation entre \hat{a} et \hat{b} : ce coefficient vaut -0.9346 .) En outre,

$$\begin{aligned}\frac{\hat{a}}{\hat{\sigma}(\hat{a})} &= 0.0176, & P(|t_{38}| > 0.0176) &= 0.9861, \\ \frac{\hat{b}}{\hat{\sigma}(\hat{b})} &= 52.6938, & P(|t_{38}| > 52.6938) &= 0.0000,\end{aligned}$$

où t_{38} indique une variable aléatoire qui suit une distribution t à 38 degrés de liberté. Il faut donc retenir l'hypothèse $a = 0$ et rejeter l'hypothèse $b = 0$. En supprimant l'intercept on obtient:

Coefficients:

	Value	Std.Error	t value	Pr(> t)
Pente b	4.8335	0.0322	150.1507	0.0000

Residual standard error: 0.0169 on 39 degrees of freedom

Multiple R-Squared: 0.9983

La pente de la droite est maintenant $\hat{b} = 4.8335$. On remarquera que $2\hat{b} = 9.6670$ est une estimation de l'accélération de gravité γ . Pour construire un intervalle de confiance pour γ calculons le percentile 97.5% de la distribution t à 39 degrés de liberté. On trouve $t_{97.5\%,39} = 2.0226$, et donc

$$[9.6670 - 2 \cdot 0.0322 \cdot 2.0226, 9.6670 + 2 \cdot 0.0322 \cdot 2.0226] = [9.537, 9.797]$$

est un intervalle de confiance avec coefficient de couverture 95% pour γ . Selon ce calcul il faut alors rejeter l'hypothèse que l'accélération est $9.81 \text{ [m/s}^2\text{]}$ (et ceci, au niveau 5%). Toutefois, l'analyse des résidus des Figures 3 et 4 indique que la condition de normalité des erreurs n'est pas bien satisfaite. L'inférence basée sur cette condition est alors douteuse. Voir la remarque ci-dessous.

Dans la Figure 5 plusieurs intervalles de confiance pour les hauteurs $h = a + b(t2)$ sont représentés par les lignes traitillées. Pour leur calcul, la valeur de $t_{97.5\%,38} = 2.024$ a été utilisée. La ligne continue est obtenue selon la règle décrite dans le Complément 2 (avec $F_{95\%,2,38} = 3.245$).

Remarque. Nous avons ajusté le modèle $h = b \cdot (t2) + \text{erreur}$ car il fournit directement une estimation et un intervalle de confiance pour $\gamma = 2b$ selon les formules des sections précédentes. Toutefois, dans l'expérience, les temps de chute ont été mesurés en fonction d'hauteurs préfixées. Il est donc préférable d'ajuster le modèle $t2 = c + d \cdot h + \text{erreur}$.

L'hypothèse $c = 0$ peut être retenue et on obtient

Coefficients:

	Value	Std.Error	t value	Pr(> t)
Pente d	0.2065	0.0014	150.1507	0.0000

Residual standard error: 0.003493 on 39 degrees of freedom

Multiple R-Squared: 0.9983

L'estimation de γ est alors $2/\hat{d} = 9.6837$. Pour construire un intervalle de confiance pour $2/d$ nous utilisons le procédé, décrit dans le Complément 3, qui sert à construire un intervalle de confiance pour le rapport entre deux paramètres. On obtient l'intervalle $[9.5560, 9.8159]$ avec un coefficient de couverture de 95%.

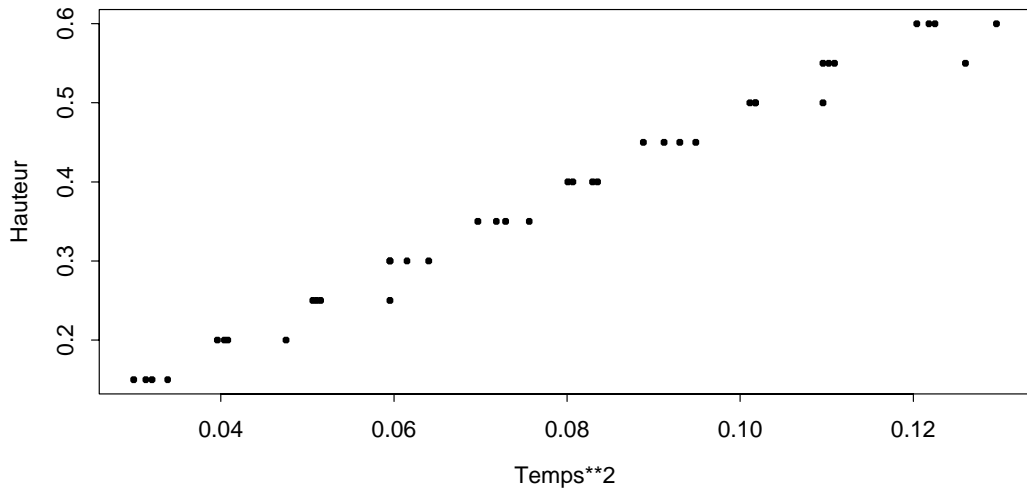


Figure 2. Diagramme de dispersion hauteur/(temps²)

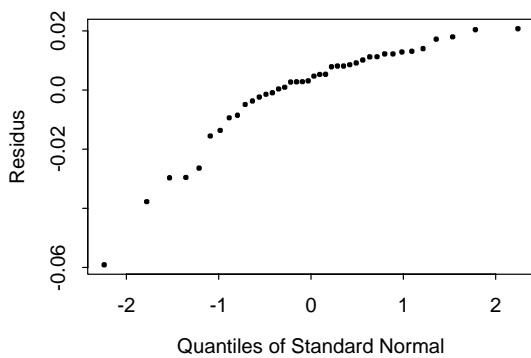


Figure 3. qq-plot des résidus

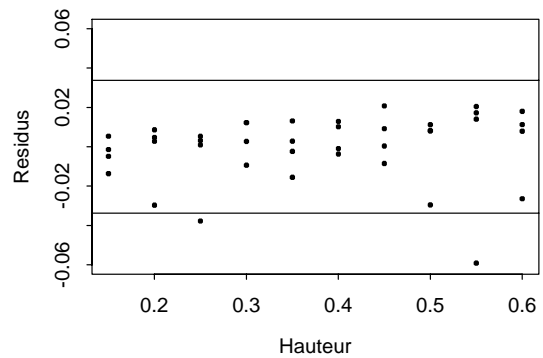


Figure 4. Diagramme résidus/hauteurs

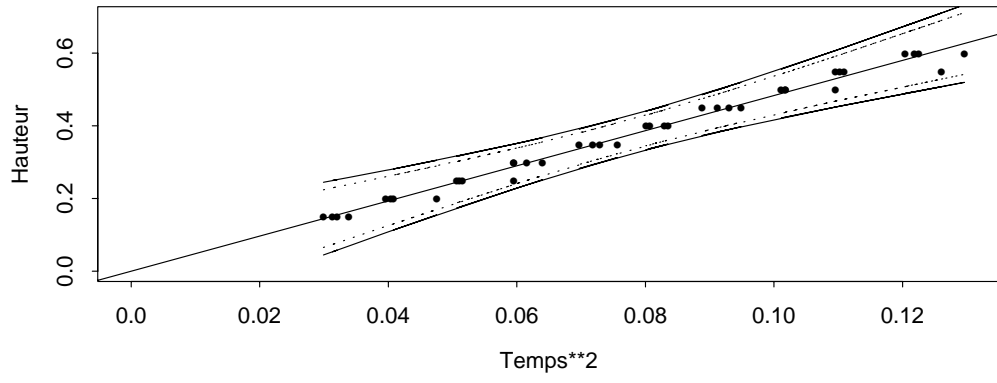


Figure 5. Intervalles de confiance pour les hauteurs h en fonction de t^2

Compléments

1. Conséquences théoriques du modèle de Gauss

- Les conditions 1 et 2 impliquent que les estimateurs des moindres carrés \hat{a} et \hat{b} ne sont pas biaisés pour a et b (Chapitre 9, Complément 1: $E(\hat{a}) = a$, $E(\hat{b}) = b$).
- Sous les conditions 1 et 2, $\hat{\sigma}^2$ est un estimateur sans biais de σ^2 .
- Les conditions 1 et 2 impliquent que les estimateurs \hat{a} et \hat{b} sont les estimateurs de variance minimale parmi tous les estimateurs linéaires en y_1, \dots, y_n et sans biais (théorème de Gauss-Markov).
- Les conditions 1, 2 et 3 impliquent que les estimateurs \hat{a} et \hat{b} sont les estimateurs de variance minimale parmi tous les estimateurs de a et b .

2. Bandes de confiance

Supposons de construire des intervalles de confiance avec coefficient de couverture $1 - 2\alpha$ pour $y_x = a + bx$ et pour différentes valeurs de x : $x = x_1$, $x = x_2$, etc. Supposons ensuite que nous joignons les extrémités supérieures et les extrémités inférieures, obtenant ainsi les deux courbes comme celles indiquées en traitillé dans la Figure 5. Il serait faux d'affirmer que la région entre les deux courbes couvre l'ensemble de toutes les valeurs de $a + bx$ avec probabilité $1 - 2\alpha$. (Si I_i est l'intervalle de confiance pour y_{x_i} et $P(y_{x_i} \in I_i) = 1 - 2\alpha$ pour $i = 1, \dots, n$, on ne peut pas conclure que $P(y_{x_1} \in I_1 \cap \dots \cap y_{x_n} \in I_n) = 1 - 2\alpha$.) Une région de confiance "simultanée" pour tous les y_x peut être obtenue (Miller R.G, 1966, p. 111) en joignant les extrémités supérieures et les extrémités inférieures des intervalles

$$[\hat{y}_x - \hat{\sigma}(\hat{y}_x)\sqrt{2F_{1-2\alpha,2,n-2}} \quad , \quad \hat{y}_x + \hat{\sigma}(\hat{y}_x)\sqrt{2F_{1-2\alpha,2,n-2}}],$$

où $F_{2\alpha,2,n-2}$ est le percentile $1 - 2\alpha$ de la distribution F à 2 et $n - 2$ degrés de liberté.

3. Intervalle de confiance pour un rapport

Soient \hat{a} et \hat{b} des estimateurs sans biais de deux paramètres a et b . Notre objectif est d'estimer le rapport $r = a/b$ et de construire un intervalle de confiance pour r . Supposons que \hat{a} et \hat{b} suivent approximativement une distribution de Gauss et que

$$V(\hat{a}) = v_{aa}\sigma^2, \quad V(\hat{b}) = v_{bb}\sigma^2, \quad V(\hat{a}, \hat{b}) = v_{ab}\sigma^2,$$

où v_{aa} , v_{ab} , v_{bb} et σ sont connues. Alors, $V(\hat{a} - r\hat{b}) = (v_{aa} - 2rv_{ab} + r^2v_{bb})\sigma^2$, et

$$P\left(\frac{(\hat{a} - r\hat{b})^2}{V(\hat{a} - r\hat{b})} \leq z_{1-\alpha}^2\right) \approx 1 - 2\alpha,$$

où $z_{1-\alpha}$ est le percentile $1 - \alpha$ de la distribution de Gauss standard. Pour trouver les limites r_l et r_u d'un intervalle de confiance avec coefficient de couverture $1 - 2\alpha$ pour r , il suffit donc de résoudre pour r l'équation quadratique $(\hat{a} - r\hat{b})^2 = z_{1-\alpha}^2 V(\hat{a} - r\hat{b})$. Les solutions sont

$$(r_l, r_u) = \left[\hat{r} - g \left(\frac{v_{ab}}{v_{bb}} \right) \pm \frac{z_{1-\alpha}\sigma}{|\hat{b}|} \left\{ v_{aa} - 2\hat{r}v_{ab} + \hat{r}^2v_{bb} - g \left(v_{aa} - \frac{v_{ab}^2}{v_{bb}} \right) \right\}^{1/2} \right] / (1 - g),$$

où $g = z_{1-\alpha}^2 \sigma^2 v_{bb} / \hat{b}^2$, et $\hat{r} = \hat{a} / \hat{b}$ est l'estimateur de r . Dans un problème de régression, a et b sont souvent des coefficients, les valeurs de v_{aa} , v_{ab} et v_{bb} sont fournies par les programmes ("matrice de covariance sans échelle") et σ^2 est estimé par $\hat{\sigma}^2$ (avec $n - 2$ degrés de liberté). Il faut alors remplacer $z_{1-\alpha}$ par $t_{1-\alpha, n-2}$. Dans l'exemple (Section 6) la valeur $t_{97.5\%, 39} = 2.0226$ a été utilisée.

Chapitre 18

Régression multiple: introduction

La régression multiple est l'une des méthodes les plus importantes en statistique. Son but est d'étudier et modéliser la relation entre une variable réponse Y et plusieurs variables explicatives X_1, X_2, \dots, X_p .

18.1 Modèle de régression multiple: exemples

Ajustement d'un polynôme. La Table 1 contient des mesures de concentration (pmol/ml) du peptide C en relation avec l'âge pour $n = 43$ enfants diabétiques.

Table 1. Concentrations de peptide C et âge de 43 enfants

Age	Conc.	Age	Conc.
5.2	4.8	11.3	5.1
8.8	4.1	1.0	3.9
10.5	5.2	14.5	5.7
10.6	5.5	11.9	5.1
10.4	5.0	8.1	5.2
1.8	3.4	13.8	3.7
12.7	3.4	15.5	4.9
15.6	4.9	9.8	4.8
5.8	5.6	11.0	4.4
1.9	3.7	12.4	5.2
2.2	3.9	11.1	5.1
4.8	4.5	5.1	4.6
7.9	4.8	4.8	3.9
5.2	4.9	4.2	5.1
0.9	3.0	6.9	5.1
11.8	4.6	13.2	6.0
7.9	4.8	9.9	4.9
11.5	5.5	12.5	4.1
10.6	4.5	13.2	4.6
8.5	5.3	8.9	4.9
11.1	4.7	10.8	5.1
12.8	6.6		

La Figure 1 représente les logarithmes des concentrations en fonction d'Age. Comme la relation n'a pas une allure linéaire, on peut penser de la décrire à l'aide d'un polynôme de deuxième degré (fonction quadratique). Plus précisément, nous considérons la variable réponse $Y = \ln(\text{Concentration})$, la variable explicative $X_1 = \text{Age}$, ainsi que son carré $X_2 = \text{Age}^2$ et nous ajustons le modèle

$$Y \approx \theta_0 + \theta_1 X_1 + \theta_2 X_2 \quad (1)$$

aux données. (Le signe " \approx " indique que la relation n'est pas parfaite: une "erreur" sera introduite par la suite). Il faut donc déterminer les coefficients θ_0 , θ_1 et θ_2 à l'aide des données.

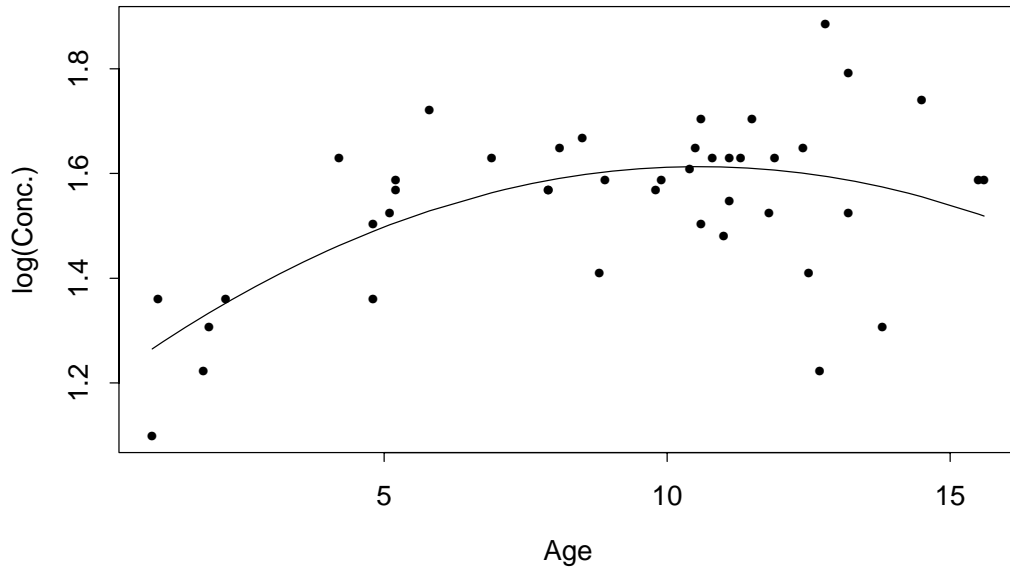


Figure 1. Log(concentration) du peptide C et âge de 43 enfants diabétiques

Nous indiquons par y_i ($i = 1, \dots, n$) les valeurs de la variable réponse, par x_{i1} les valeurs de la variable explicative $X_1 = \text{Age}$ et par x_{i2} les valeurs de la deuxième variable explicative $X_2 = \text{Age}^2$. Dans l'exemple,

$$\begin{aligned} y_1 &= \ln(4.8), & y_2 &= \ln(4.1), & \dots, & y_{43} &= \ln(5.1); \\ x_{11} &= 5.2, & x_{21} &= 8.8, & \dots, & x_{43,1} &= 10.8; \\ x_{12} &= (5.2)^2, & x_{22} &= (8.8)^2, & \dots, & x_{43,2} &= (10.8)^2. \end{aligned}$$

Alors, une méthode fréquemment utilisée pour déterminer θ_0 , θ_1 , et θ_2 consiste à les choisir de façon que la somme

$$\sum_{i=1}^n (y_i - \theta_0 - \theta_1 x_{i1} - \theta_2 x_{i2})^2$$

soit minimale. C'est la *méthode des moindres carrés*. Les valeurs $\hat{\theta}_0 = 1.197$, $\hat{\theta}_1 = 0.079$ et $\hat{\theta}_2 = -0.004$ ont été obtenues de cette façon. Avec ces valeurs on obtient la courbe représentée dans la Figure 1.

Variables explicatives quantitatives et qualitatives. Un certain type d'appareil médical administrant de façon automatique et continue une hormone anti-inflammatoire a été testé sur 27 sujets. La Table 2 donne les quantités d'hormone ("Quantité" en mmg) qui restent dans 27 appareils – un par sujet – après un certain nombre d'heures ("Hrs") d'utilisation.

Table 2. Quantités d'hormone dans 27 appareils

Lot	Hrs	Quantité	Lot	Hrs	Quantité	Lot	Hrs	Quantité
A	99	25.8	B	376	16.3	C	119	28.8
A	152	20.5	B	385	11.6	C	188	22.0
A	293	14.3	B	402	11.8	C	115	29.7
A	155	23.2	B	29	32.5	C	88	28.9
A	196	20.6	B	76	32.0	C	58	32.8
A	53	31.1	B	296	18.0	C	49	32.5
A	184	20.9	B	151	24.1	C	150	25.4
A	171	20.9	B	177	26.5	C	107	31.7
A	52	30.4	B	209	25.8	C	125	28.5

Les appareils ont été échantillonnés dans trois groupes ("Lot") provenant de trois fabricants: A, B, C. Il faut étudier la relation entre la variable réponse "Quantité" et les variables explicatives "Hrs" et "Lot". Les données sont représentées dans la Figure 2.

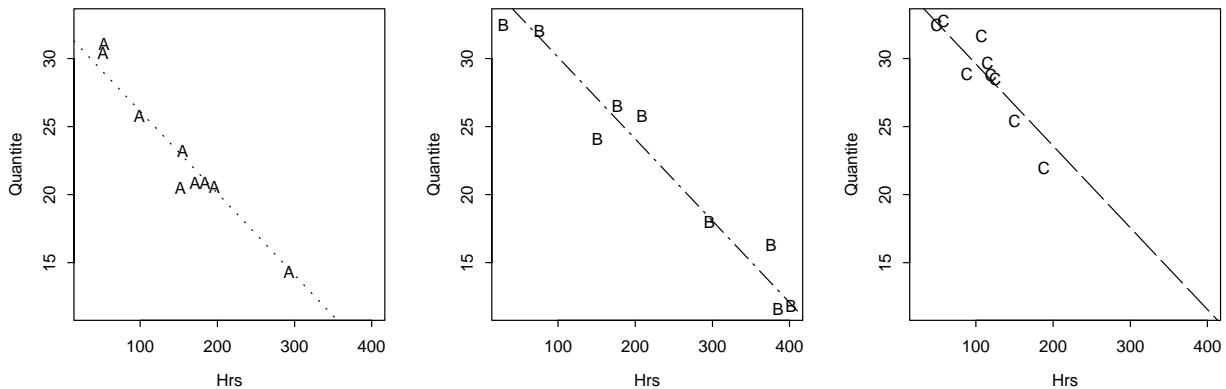


Figure 2. "Quantité" versus "Hrs" pour les groupes A, B et C et modèles ajustés.

La variable Lot est qualitative: on dit que c'est un *facteur* en trois classes. Les facteurs doivent être codés comme variables numériques et il y a plusieurs façons de faire. La plus simple utilise trois *variables indicatrices* X_1 , X_2 et X_3 définies par leurs valeurs observées x_{i1} , x_{i2} et x_{i3} :

$$\begin{aligned}
 x_{i1} &= 1 \text{ si l'appareil } i \text{ appartient au Lot A,} \\
 &= 0 \text{ si l'appareil } i \text{ n'appartient pas au Lot A,} \\
 x_{i2} &= 1 \text{ si l'appareil } i \text{ appartient au Lot B,} \\
 &= 0 \text{ si l'appareil } i \text{ n'appartient pas au Lot B,} \\
 x_{i3} &= 1 \text{ si l'appareil } i \text{ appartient au Lot C,} \\
 &= 0 \text{ si l'appareil } i \text{ n'appartient pas au Lot C.}
 \end{aligned}$$

En définissant $Y = \text{Quantité}$ et $X_4 = \text{Hrs}$, on peut alors décrire les données par le modèle

$$Y \approx \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3 + \theta_4 X_4 \quad (2)$$

et déterminer les coefficients θ_1 , θ_2 , θ_3 et θ_4 à l'aide des données. Ces coefficients sont les intercepts des trois droites parallèles $Y = \theta_1 + \theta_4 X_4$, $Y = \theta_2 + \theta_4 X_4$ et $Y = \theta_3 + \theta_4 X_4$ qui décrivent les relations entre Quantité et Hrs pour les trois groupes.

La méthode des moindres carrés pour déterminer θ_1 , θ_2 , θ_3 , et θ_4 consiste à les choisir de façon que la somme

$$\sum_{i=1}^n (y_i - \theta_1 x_{i1} - \theta_2 x_{i2} - \theta_3 x_{i3} - \theta_4 x_{i4})^2$$

soit minimale. Dans l'exemple on trouve $\hat{\theta}_1 = 32.13$, $\hat{\theta}_2 = 36.11$, $\hat{\theta}_3 = 35.60$, $\hat{\theta}_4 = -0.06$, ce qui signifie que pour les données du groupe A, $Y \approx 32.13 - 0.06X_4$, pour les données du groupe B, $Y \approx 36.11 - 0.06X_4$, et pour celles du groupe C, $Y \approx 35.60 - 0.06X_4$. Ces trois droites sont indiquées dans les diagrammes de la Figure 2.

Une autre façon de "paramétriser" le modèle utilise seulement deux variables indicatrices, par exemple, X_1 et X_2 :

$$Y \approx \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_4 X_4. \quad (3)$$

Dans ce cas, θ_0 est l'intercept de la droite du groupe C, tandis que θ_1 et θ_2 sont les écarts entre les intercepts de A et B par rapport à C, qui constitue le *niveau de référence*. La méthode des moindres carrés donne $\hat{\theta}_0 = 35.60$, $\hat{\theta}_1 = -3.47$, $\hat{\theta}_2 = 0.51$ et $\hat{\theta}_4 = -0.06$. L'intercept de A est donc $35.60 - 3.47 = 32.13$ et celui de B est $35.60 + 0.51 = 36.11$. En général, pour coder un facteur à deux niveaux il suffit d'utiliser une seule variable indicatrice (pour la présence ou l'absence de l'une des deux caractéristiques). Pour coder un facteur à k niveaux il suffit d'utiliser $k - 1$ variables indicatrices.

Remarque. Les coefficients $\theta_0, \dots, \theta_4$ du modèle

$$Y \approx \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3 + \theta_4 X_4$$

ne peuvent pas être déterminés de façon unique. En effet, une infinité de valeurs de $\theta_0, \theta_1, \theta_2$ et θ_3 peuvent fournir $\theta_0 + \theta_1 = 32.13$, $\theta_0 + \theta_2 = 36.11$ et $\theta_0 + \theta_3 = 35.60$.

Modèle avec interactions. Pour décrire les quantités d'hormone distribuées par trois appareils nous avons utilisé trois droites parallèles. On peut se demander si un modèle plus souple, avec trois droites non nécessairement parallèles, ne serait pas plus avantageux. La modélisation des trois droites peut se faire en utilisant les variables indicatrices X_1 , X_2 et X_3 ainsi que trois variables supplémentaires X_5 , X_6 , X_7 définies comme suit:

$$\begin{aligned} x_{i5} &= \text{Hrs de l'appareil } i, \text{ si } i \text{ appartient au Lot A,} \\ &= 0 \text{ si l'appareil } i \text{ n'appartient pas au Lot A,} \\ x_{i6} &= \text{Hrs de l'appareil } i, \text{ si } i \text{ appartient au Lot B,} \\ &= 0 \text{ si l'appareil } i \text{ n'appartient pas au Lot B,} \\ x_{i7} &= \text{Hrs de l'appareil } i, \text{ si } i \text{ appartient au Lot C,} \\ &= 0 \text{ si l'appareil } i \text{ n'appartient pas au Lot C.} \end{aligned}$$

Si Y indique la réponse, on peut décrire les données par le modèle

$$Y \approx \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3 + \theta_5 X_5 + \theta_6 X_6 + \theta_7 X_7. \quad (4)$$

Les coefficients θ_1 , θ_2 et θ_3 sont les intercepts, tandis que θ_5 , θ_6 et θ_7 sont les pentes des trois droites. Si les pentes sont différentes, on dit qu'il y a *interaction* entre le facteur Lot et la variable Hrs: dans ce cas, la variable Hrs explique Y de façon différente selon le Lot. Notons que $X_5 = X_1 X_4$, $X_6 = X_2 X_4$ et qu'une autre façon de paramétriser (5) est

$$Y \approx \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_4 X_4 + \theta_5 X_1 X_4 + \theta_6 X_2 X_4. \quad (5)$$

Dans cette paramétrisation θ_4 est la pente de la droite du Lot C (*pente de référence*). On dit que θ_5 mesure l'interaction entre Hrs et Lot A et que θ_6 est l'interaction entre Hrs et Lot B. L'utilisation de produits entre deux variables est la façon habituelle d'introduire des interactions dans un modèle.

Le degré d'ajustement des modèles (4) et (5) est certainement supérieur à celui des modèles (2) et (3). Toutefois, il n'est pas certain que la complexité accrue de (4) et (5) justifie ce gain. La question du choix entre ces modèles sera abordée au Chapitre 20.

18.2 Définitions et propriétés

Plusieurs concepts et propriétés de la régression simple s'étendent à la régression multiple. Considérons par exemple le modèle

$$Y \approx \theta_0 + \theta_1 X_1 + \dots + \theta_p X_p.$$

On dit que $\theta_0, \theta_1, \dots, \theta_p$ sont les *coefficients* et que θ_0 est la *constante additive* du modèle. Les coefficients sont habituellement estimés selon la méthode des moindres carrés par un programme d'ordinateur. Nous indiquons les *estimations* par $\hat{\theta}_0, \hat{\theta}_1$, etc. Alors

$$\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_{i1} + \dots + \hat{\theta}_p x_{ip}, \quad i = 1, \dots, n$$

sont les *réponses calculées* et

$$e_i = y_i - \hat{y}_i \quad i = 1, \dots, n$$

les *résidus*. Evidemment

$$y_i = \hat{y}_i + e_i$$

d'où découle la décomposition

$$s^2(Y) = s^2(\hat{Y}) + s^2(E),$$

où \hat{Y} est le vecteur des réponses calculées et E celui des résidus. Cette décomposition de $s^2(Y)$ est connue comme *analyse de la variance* (Chapitre 3). Le premier terme est la *variance expliquée* par le modèle et le deuxième la *variance résiduelle*. En outre, la somme des résidus est nulle: $\sum e_i = 0$. (Mais ceci n'est pas certain si la constante additive est absente.)

Le *coefficient de détermination* est défini par

$$R^2 = s^2(\hat{Y})/s^2(Y).$$

Il jouit des propriétés habituelles

- $0 \leq R^2 \leq 1$
- Si R^2 est proche de 1 (par exemple $R^2 = 0.8$) le modèle explique très bien la variation de Y . Si R^2 est proche de 0, les variables X_1, X_2 , etc. ne contiennent pas d'information utile pour expliquer la variation de Y .

L'*écart type de l'erreur* (ou *erreur standard des résidus*) noté s_E ou $\hat{\sigma}$ est défini par

$$s_E = \sqrt{\frac{1}{n-p-1} \sum_{i=1}^n e_i^2}.$$

18.3 Notation matricielle

La notation matricielle est très utile dans la régression multiple car les données ont la forme d'une matrice dont les colonnes correspondent aux variables et les lignes aux observations. Nous introduisons cette notation en reprenant d'abord le cas de la régression simple (Chapitre 3). On définit le *vecteur des réponses observées* y , et le *vecteur des erreurs* u par

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}.$$

(On utilise d'habitude les minuscules y et u dans ce contexte.) La *matrice du modèle* X (ou *matrice de design*) et le *vecteur des paramètres* θ sont définis par

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \theta = \begin{pmatrix} a \\ b \end{pmatrix}.$$

La première colonne contient des "1" et sera associée à a ; la deuxième contient les valeurs de la variable explicative. On obtient ainsi

$$X\theta = \begin{pmatrix} a + bx_1 \\ a + bx_2 \\ \vdots \\ a + bx_n \end{pmatrix}$$

et les n équations caractérisant la structure modèle sont exprimées d'un seul coup par:

$$y = X\theta + u. \tag{6}$$

Cette équation représente aussi les modèles de régression multiple si on définit de façon appropriée la matrice X et le vecteur θ . Par exemple, pour exprimer le modèle (1) définissons

$$X = \begin{pmatrix} 1 & 5.2 & 5.2^2 \\ 1 & 8.8 & 8.8^2 \\ \vdots & \vdots & \vdots \\ 1 & 10.8 & 10.8^2 \end{pmatrix}, \quad \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{pmatrix}.$$

La première colonne sera associée à θ_0 , la deuxième contient les valeurs de X_1 et la troisième les valeurs de X_2 . On obtient ainsi l'équation (6). Dans le cas du modèle (3) définissons

$$X = \begin{pmatrix} 1 & 1 & 0 & 99 \\ 1 & 1 & 0 & 152 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 376 \\ 1 & 0 & 1 & 385 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 119 \\ 1 & 0 & 0 & 188 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 125 \end{pmatrix}, \quad \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_4 \end{pmatrix}.$$

La première colonne est associée à θ_0 , la deuxième contient les valeurs de X_1 , la troisième les valeurs de X_2 et la quatrième les valeurs de X_4 . On obtient encore l'équation (6).

En général, on considérera un vecteur y de n réponses observées, un vecteur u de n erreurs (non observées), une matrice de modèle X à n lignes et p colonnes

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

et un vecteur θ de p paramètres

$$\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_p \end{pmatrix}$$

tels que

$$y = X\theta + u. \quad (7)$$

Les colonnes de X contiennent les valeurs des variables explicatives. Si une constante additive est présente, la première colonne est formée de "1". On notera par $x_1^T, x_2^T, \dots, x_n^T$ les lignes de la matrice X et par X_1, X_2, \dots, X_p ses colonnes. (Ici, $(\cdot)^T$ indique l'opération matricielle de transposition.) L'équation matricielle (7) résume les n équations

$$y_i = \theta_1 x_{i1} + \dots + \theta_p x_{ip} + u_i, \quad i = 1, \dots, n$$

que l'on peut aussi écrire de la façon suivante:

$$y_i = x_i^T \theta + u_i, \quad i = 1, \dots, n.$$

Pour spécifier le modèle, on écrira aussi

$$Y \approx \theta_1 X_1 + \dots + \theta_p X_p.$$

Chapitre 19

Ajustement du modèle de régression multiple

Ce chapitre considère quelques aspects du calcul des coefficients d'une régression par la méthode des moindres carrés ainsi qu'une interprétation géométrique de cette méthode. Il n'est pas requis pour les chapitres suivants.

Dans ce chapitre, la longueur d'un vecteur y sera notée par $|y|$: donc, $|y| = (y^T y)^{1/2}$.

19.1 La méthode des moindres carrés

Nous considérons un modèle de régression multiple défini par sa matrice de modèle X , (n lignes et p colonnes) son vecteur de paramètres θ (p composantes), son vecteur de réponses observées y (n composantes) et l'équation structurelle

$$y = X\theta + u,$$

où u est le vecteur des erreurs. On dit que cette équation représente un *modèle linéaire* dans les coefficients $\theta_1, \dots, \theta_p$.

Selon la méthode des moindres carrés, une estimation du vecteur θ est obtenue en minimisant la fonction

$$Q(\theta) = |y - X\theta|^2.$$

Dans ce but, il faut résoudre le système de p équations $\partial Q / \partial \theta_j = 0$, $j = 1, \dots, p$, où $\partial Q / \partial \theta_j$ indique la dérivée partielle de Q par rapport à θ_j . On obtient l'équation matricielle

$$X^T X \theta = X^T y$$

connue comme le système des *équations normales*.

Le calcul de la solution est normalement réalisé par un programme d'ordinateur. La solution est un vecteur noté $\hat{\theta}$. On définit:

- le vecteur des réponses calculées $\hat{y} = X\hat{\theta}$;
- le vecteur des résidus $e = y - \hat{y}$.

Remarques sur le calcul de $\hat{\theta}$

1. Si le rang de la matrice X est égal à p (c'est à dire, si les colonnes de X sont linéairement indépendantes) alors le rang de $X^T X$ est égal à p et le problème de minimisation a une solution unique

$$\hat{\theta} = (X^T X)^{-1} X^T y,$$

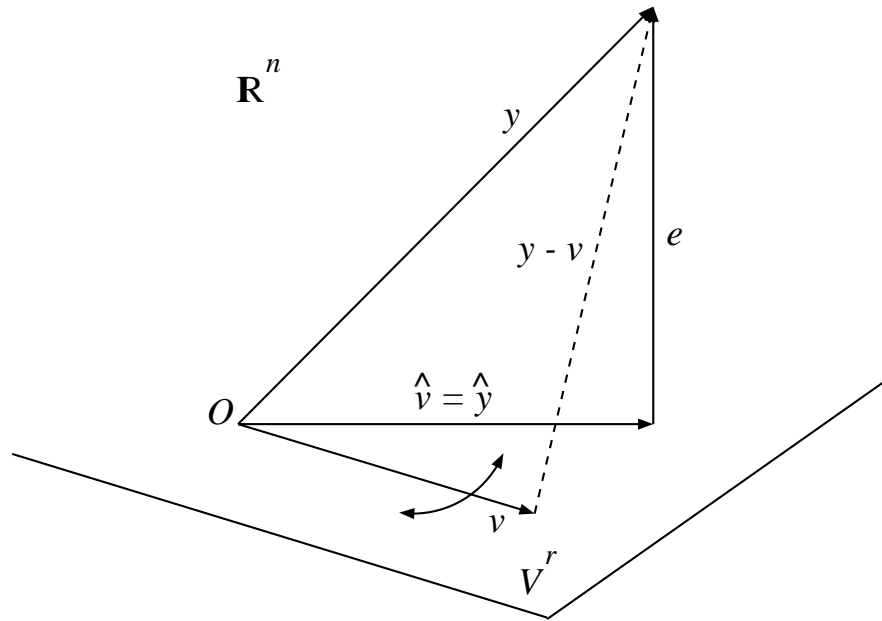
où $(\cdot)^{-1}$ indique l'opération d'inversion d'une matrice.

2. Si le rang de X est inférieur à p il y a une infinité de vecteurs θ qui minimisent $Q(\theta)$. Pour en choisir une, on réduit d'habitude le nombre de coefficients en posant des conditions supplémentaires. Par exemple, on supprime certaines variables explicatives (c'est à dire, on pose leur coefficient égal à zéro). Il y a aussi des procédés qui permettent de choisir la solution de *longueur* $\sqrt{\theta^T \theta}$ minimale.

19.2 Propriétés algébriques et géométriques

Nous supposons que le rang de X est $r \leq p$.

- Les composantes de $\hat{\theta}$ sont des fonctions linéaires en y_1, \dots, y_n .
- Les colonnes de X génèrent un sous-espace de dimension r de \mathbb{R}^n . Notons ce sous-espace par V^r . Pour tout θ , le vecteur $v = \theta_1 X_1 + \dots + \theta_p X_p = X\theta$ obtenu comme combinaison linéaire des colonnes X_1, \dots, X_p de X appartient donc à V^r .
- La méthode des moindres carrés minimise la longueur du vecteur $y - X\theta = y - v$. Le vecteur $\hat{v} = X\hat{\theta}$ est donc la projection de y en V^r . Il coïncide avec le vecteur \hat{y} des réponses calculées.



- Le vecteur des résidus e est orthogonal à V^r . Par conséquent, $X^T e = 0$. Cette dernière équation vectorielle coïncide avec les équations normales.
- Evidemment

$$y = \hat{y} + e,$$

et, par le théorème de Pythagore,

$$|y|^2 = |\hat{y}|^2 + |e|^2.$$

Si \bar{y} indique le vecteur avec n composantes égales à la moyenne arithmétique des y_i , nous avons aussi $y - \bar{y} = \hat{y} - \bar{y} + e$ et, comme \bar{y} est orthogonale à e ($e^T \bar{y} = 0$):

$$|y - \bar{y}|^2 = |\hat{y} - \bar{y}|^2 + |e|^2.$$

Ceci signifie que la variance de y est la somme de deux parties: la première est la "partie expliquée par le modèle" et la deuxième est la variance résiduelle (*analyse de la variance*). Donc,

$$R^2 = \frac{|\hat{y} - \bar{y}|^2}{|y - \bar{y}|^2}.$$

Chapitre 20

Inférence classique pour la régression multiple

Ce chapitre étend les résultats du Chapitre 17 à la régression multiple. Nous considérons la relation

$$Y \approx \theta_1 X_1 + \dots + \theta_p X_p$$

entre une réponse Y et p variables explicatives X_1, \dots, X_p ; X_1 pourrait être identique à 1, auquel cas, θ_1 serait une constante additive.

20.1 Modèle classiques pour l'inférence

Comme dans le cas de la régression simple, l'inférence classique pour la régression multiple se fonde sur un ensemble de conditions concernant la distribution de la variable réponse Y en relation avec les variables explicatives X_1, \dots, X_p . Il n'est pas nécessaire de supposer que les observations des variables explicatives sont obtenues de façon aléatoire. Les conditions suivantes forment le *modèle de Gauss* pour la régression multiple.

1. $Y_i = \theta_1 x_{i1} + \dots + \theta_p x_{ip} + U_i$, $i = 1, \dots, n$, où $\theta_1, \dots, \theta_p$ sont des paramètres.
2. Les erreurs U_i sont i.i.d. et indépendents de X_1, \dots, X_p .
3. $U_i \sim \mathcal{N}(0, \sigma^2)$ où σ^2 est un paramètre.

La condition 1 correspond à n équations pour les réponses observées:

$$y_i = \theta_1 x_{i1} + \dots + \theta_p x_{ip} + u_i, \quad i = 1, \dots, n.$$

Les erreurs u_i ne sont pas observables. La condition 1 caractérise la structure du modèle; les condition 2 et 3 la partie aléatoire.

20.2 Distributions des estimateurs

Les résultats suivants s'obtiennent sous le modèle de Gauss.

- $\hat{\theta}$ suit une distribution de Gauss multivariée avec vecteur de moyennes θ et matrice de covariance $\Sigma^2(\theta)$:

$$\hat{\theta} \sim \mathcal{N}(\theta, \Sigma^2(\hat{\theta})), \quad \text{avec} \quad \Sigma^2(\hat{\theta}) = \sigma^2 (X^T X)^{-1}.$$

En outre, si $x = (x_1, \dots, x_p)^T$ est un vecteur (colonne) contenant des valeurs données des variables explicatives, nous considérons la réponse calculée $\hat{y}_x = \hat{\theta}^T x$ ainsi que $y_x = \theta^T x$. Alors,

$$\hat{y}_x \sim \mathcal{N}(y_x, \sigma^2(\hat{y}_x)), \quad \text{avec} \quad \sigma^2(\hat{y}_x) = x^T \Sigma^2(\hat{\theta}) x.$$

Ces résultats pourraient permettre de réaliser des inférences si σ^2 était connu. En pratique, il faut presque toujours estimer σ^2 et, dans ce but, on utilise l'estimateur

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^n e_i^2.$$

Des estimations $\hat{\Sigma}^2(\hat{\theta})$ et $\hat{\sigma}^2(\hat{y}_x)$ sont alors obtenues en remplaçant σ^2 par $\hat{\sigma}^2$ dans les expressions de $\Sigma^2(\hat{\theta})$ et $\sigma^2(\hat{y}_x)$. On démontre que:

- La variable aléatoire $(n-p)\hat{\sigma}^2/\sigma^2$ suit une distribution χ^2 à $n-p$ degrés de liberté.

– Les estimateurs standardisés

$$(\hat{\theta}_j - \theta_j)/\hat{\sigma}(\hat{\theta}_j), \quad j = 1, \dots, p \quad \text{et} \quad (\hat{y}_x - y_x)/\hat{\sigma}(\hat{y}_x)$$

suivent une distribution t à $n - p$ degrés de liberté.

Note. La matrice $(X^T X)^{-1}$ est parfois appelée *matrice de covariance sans échelle* des coefficients estimés.

20.3 Intervalles de confiance et test usuels

Les résultats précédents permettent d'obtenir les intervalles de confiance pour les coefficients θ_j ($j = 1, \dots, p$) et pour $y_x = x^T \theta$. Soit α une probabilité préfixée (par exemple $\alpha = 2.5\%$). Alors, des intervalles de confiance bilatéraux avec coefficient de couverture $1 - 2\alpha$ sont:

$$\begin{aligned} & [\hat{\theta}_j - \hat{\sigma}(\hat{\theta}_j) t_{1-\alpha, n-p}, \quad \hat{\theta}_j + \hat{\sigma}(\hat{\theta}_j) t_{1-\alpha, n-p}], \quad j = 1, \dots, p, \\ & [\hat{y}_x - \hat{\sigma}(\hat{y}_x) t_{1-\alpha, n-p}, \quad \hat{y}_x + \hat{\sigma}(\hat{y}_x) t_{1-\alpha, n-p}], \end{aligned}$$

où $t_{1-\alpha, n-p}$ est le percentile $1 - \alpha$ de la distribution t à $n - p$ degrés de liberté. En outre, un intervalle de confiance avec coefficient de couverture $1 - 2\alpha$ pour σ^2 est donné par

$$[(n - 2)\sigma^2/\chi_{1-\alpha, n-p}^2, \quad (n - 2)\sigma^2/\chi_{\alpha, n-p}^2],$$

où $\chi_{\alpha, n-p}^2$ est le percentile α de la distribution χ^2 à $n - 2$ degrés de liberté.

Pour un certain k , l'hypothèse

$$H_0 : \theta_k = c_0,$$

où c_0 est une valeur donnée, peut être rejetée au niveau α , en faveur de $H_1 : \theta_k \neq c_0$, si la statistique

$$T = (\hat{\theta}_k - c_0)/\hat{\sigma}(\hat{\theta}_k)$$

n'appartient pas à l'intervalle $[-t_{\alpha/2, n-p}, t_{\alpha/2, n-p}]$.

De façon équivalente, on peut rejeter H_0 en faveur de H_1 au niveau α si l'intervalle de confiance avec coefficient de couverture $1 - \alpha$ pour θ_k ne contient pas la valeur préfixée c_0 .

Remarques

1. Les logiciels de statistique courants fournissent les valeurs de $\hat{\sigma}(\hat{\theta}_j)$ ($j = 1, \dots, p$) ainsi que celles des statistiques $\hat{\theta}_j/\hat{\sigma}(\hat{\theta}_j)$ et les P-values correspondantes. Par exemple, R et S-plus donnent

$$P(|t_{n-p}| > |\hat{\theta}_j/\hat{\sigma}(\hat{\theta}_j)|), \quad j = 1, \dots, p$$

où t_{n-p} indique une variable aléatoire suivant une distribution t à $n - p$ degrés de liberté et $\hat{\theta}_j/\hat{\sigma}(\hat{\theta}_j)$ désigne la valeur observée de la statistique correspondante.

2. Si $[A, B]$ et $[C, D]$ sont des intervalles de confiance avec coefficient de couverture $1 - \alpha$ pour θ_1 et θ_2 (par exemple) on ne peut pas affirmer que le rectangle $[A, B] \times [C, D]$ couvre le point (θ_1, θ_2) avec probabilité $1 - \alpha$! Nous effleurons ici un problème d'inférence statistique simultanée que nous n'approfondirons pas.

20.4 Analyse des résidus

Si le modèle de Gauss est approprié, les résidus ont approximativement une distribution de Gauss. Il faut donc examiner cette condition à l'aide d'un qq-plot. En outre, la variance des résidus ne doit pas dépendre des variables explicatives. Il est donc opportun de représenter graphiquement les résidus en fonction des valeurs observées de X_1, \dots, X_p . Aucune relation (relation non linéaire, variance non homogène) ne doit apparaître. Enfin, on peut représenter les résidus en fonction des réponses calculées. Si une relation apparaît le modèle de Gauss et les inférences obtenues avec son appui doivent être mis en doute.

20.5 Exemples

Ajustement d'un polynôme. Nous considérons les données de la Table 1, Chapitre 18 et ajustons le modèle

$$\log(\text{Concentration}) \approx \theta_0 + \theta_1 \text{Age} + \theta_2 \text{Age}^2.$$

On a les résultats suivants:

Coefficients:

	Value	Std.Error	t value	Pr(> t)
theta0	1.1973	0.0767	15.6040	0.0000
theta1	0.0787	0.0204	3.8673	0.0004
theta2	-0.0037	0.0012	-3.0406	0.0042

Residual standard error: 0.1299 on 40 degrees of freedom

Multiple R-Squared: 0.3686

Correlation of Coefficients:

	theta0	theta1
theta1	-0.8880	
theta2	0.7678	-0.9696

Donc, dans les notations des sections précédentes,

$$\begin{aligned} \hat{\theta}_0 &= +1.1973, & \hat{\sigma}(\hat{\theta}_0) &= 0.0767, \\ \hat{\theta}_1 &= +0.0787, & \hat{\sigma}(\hat{\theta}_1) &= 0.0204, \\ \hat{\theta}_2 &= -0.0037, & \hat{\sigma}(\hat{\theta}_2) &= 0.0012. \end{aligned}$$

L'erreur standard des résidus est $\hat{\sigma} = 0.1299$ et $R^2 = 0.3686$. (La corrélation entre $\hat{\theta}_1$ et $\hat{\theta}_2$ est -0.9696 , celle entre $\hat{\theta}_1$ et $\hat{\theta}_0$ est -0.8880 et celle entre $\hat{\theta}_0$ et $\hat{\theta}_2$ est 0.7678 .)

Si le modèle de Gauss peut être retenu, on obtient les inférences suivantes:

$$\begin{aligned} \frac{\hat{\theta}_0}{\hat{\sigma}(\theta_0)} &= 15.6040 & \text{et} & & P(|t_{40}| > 15.6040) &= 0.0000, \\ \frac{\hat{\theta}_1}{\hat{\sigma}(\theta_1)} &= +3.8673 & \text{et} & & P(|t_{40}| > 3.8673) &= 0.0004, \\ \frac{\hat{\theta}_2}{\hat{\sigma}(\theta_2)} &= -3.0406 & \text{et} & & P(|t_{40}| > 3.0406) &= 0.0042. \end{aligned}$$

En outre, les intervalles de confiance avec coefficient de couverture 95% pour θ_0 , θ_1 et θ_2 sont (avec $t_{40,0.975} = 2.0211$):

$$\begin{aligned} [1.1973 - 2.0211 \cdot 0.0767, 1.1973 + 2.0211 \cdot 0.0767] &= [1.0422, 1.3524], \\ [0.0787 - 2.0211 \cdot 0.0204, 0.0787 + 2.0211 \cdot 0.0204] &= [0.0376, 0.1198], \\ [-.0037 - 2.0211 \cdot 0.0012, -.0037 + 2.0211 \cdot 0.0012] &= [-.0062, -.0012]. \end{aligned}$$

Selon cette analyse, $\hat{\theta}_0$, $\hat{\theta}_1$ et $\hat{\theta}_2$ sont significativement différentes de 0 (au niveau 1%). La courbe dessinée dans la Figure 1, Chapitre 18, soulève toutefois quelques doutes à propos du modèle polynomial de deuxième degré, comme description de la relation entre $\log(\text{Conc.})$ et Age. En effet, on ne voit pas clairement pour quelle raison biologique la relation devrait être décroissante pour $\text{Age} > 10$. Enfin, l'analyse graphique des résidus fournie dans la Figure 1 suggère que la variance des erreurs croît en fonction de l'âge. Il s'agit d'une violation du modèle de Gauss qui soulève quelques doutes supplémentaires sur la validité de l'inférence.

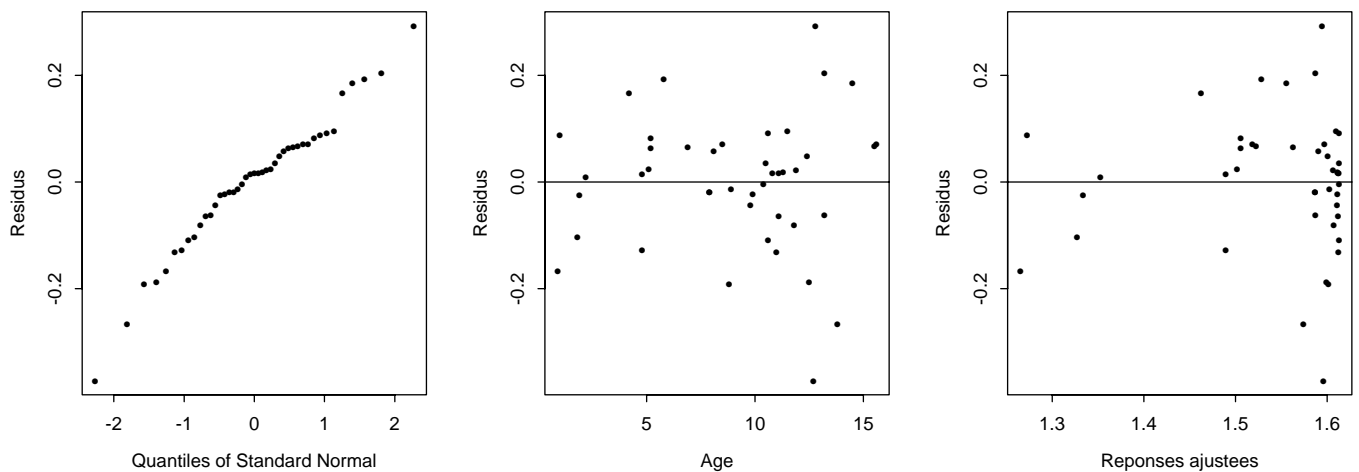


Figure 1. Analyse des résidus de la régression polynomiale. (1) Quantile-quantile plot des résidus; (2) résidus versus âge; (3) résidus versus réponses calculées.

Variables explicatives quantitatives et qualitatives. En ajustant le modèle (3), Chapitre 18, aux données de la Table 2, Chapitre 18, on obtient:

Coefficients:

	Value	Std.Error	t value	Pr(> t)
theta0	35.5973	0.6596	53.9698	0.0000
theta1	-3.4657	0.7691	-4.5061	0.0002
theta2	0.5078	0.8681	0.5849	0.5643
theta4	-0.0601	0.0035	-17.3095	0.0000

Residual standard error: 1.605 on 23 degrees of freedom

Multiple R-Squared: 0.945

Correlation of Coefficients:

	theta0	theta1	theta2
theta1	-0.4600		
theta2	-0.2136	0.5164	
theta4	-0.5847	-0.1787	-0.4900

L'analyse graphique des résidus fournie dans la Figure 2 ne contredit pas les hypothèses classiques pour l'inférence. L'écart 0.5078 entre l'intercept du groupe B et l'intercept de référence C (35.5973) n'est donc pas significativement différent de zéro. En d'autres termes, l'hypothèse $H_0 : \theta_2 = 0$ ne peut pas être rejetée, car $P(|t_{23}| > 0.5849) = 0.5643$. Par contre, l'intercept de A est significativement plus petit que celui de C, car $P(|t_{40}| > 4.5061) = 0.0002$. La différence est visible dans la Figure 2 du Chapitre 18.

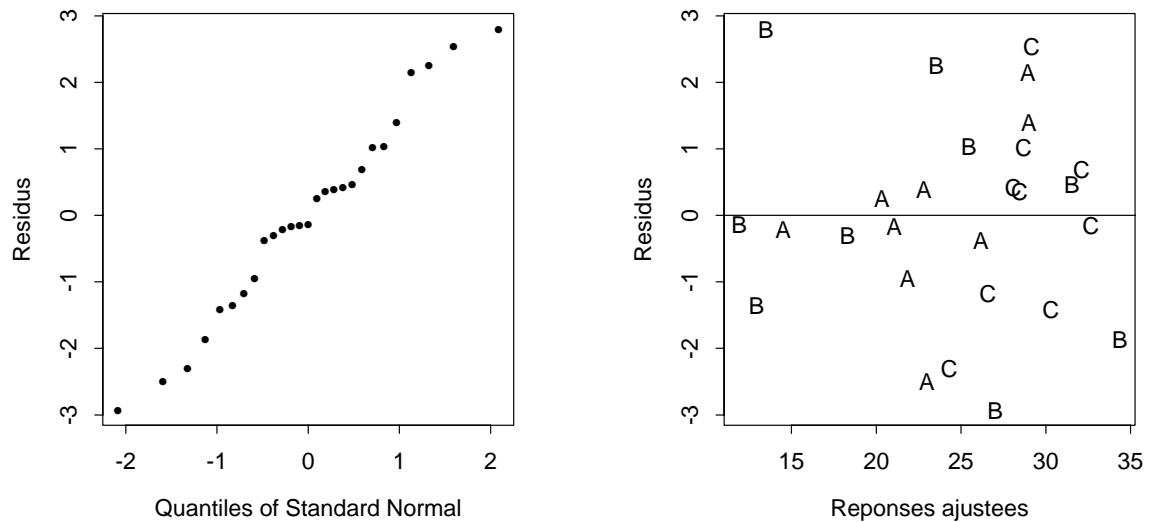


Figure 2. Analyse des résidus du modèle pour les quantité d'hormone. (1) Quantile-quantile plot des résidus; (2) résidus versus réponses calculées.

20.6 Le test F d'une hypothèse linéaire

Nous avons considéré des hypothèses du type $H_0 : \theta_k = 0$. Ce type d'hypothèse concerne l'un ou l'autre des paramètres pris individuellement. Dans la suite de ce chapitre, nous considérons des hypothèses plus complexes qui concernent plusieurs paramètres à la fois.

Exemple: test de parallélisme

Pour décrire les quantités d'hormone distribuées par trois appareils nous avons utilisé trois droites parallèles. Un modèle avec trois droites non parallèles aurait un degré d'ajustement supérieur (R^2 plus élevé); toutefois, il n'est pas certain que la complexité accrue du modèle justifie ce gain. Considérons le modèle (4) du Chapitre 18:

$$\Omega : \quad Y \approx \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3 + \theta_5 X_5 + \theta_6 X_6 + \theta_7 X_7.$$

Ce modèle Ω sera appelé le *modèle complet*. Il sera comparé au *modèle réduit*

$$\omega : \quad Y \approx \eta_1 Z_1 + \eta_2 Z_2 + \eta_3 Z_3 + \eta_4 Z_4,$$

où $Z_1 = X_1$, $Z_2 = X_2$, $Z_3 = X_3$ et $Z_4 = X_5 + X_6 + X_7$. Le modèle ω est obtenu de Ω en utilisant les deux équations

$$H_0 : \quad \theta_5 = \theta_6 = \theta_7,$$

qui représentent l'hypothèse de parallélisme. Dans le modèle ω , η_4 représente la pente commune des trois droites exprimées comme fonctions de $Z_4 = \text{Hrs}$. L'hypothèse H_0 est un système de deux équations linéaires dans les coefficients:

$$\begin{aligned} \theta_5 - \theta_6 &= 0, \\ \theta_5 - \theta_7 &= 0. \end{aligned}$$

On dit que H_0 est une *hypothèse linéaire*.

Le coefficient R^2 de Ω vaut 0.9971, tandis que celui de ω vaut 0.9966. La différence est minime: en d'autres termes le gain en ajustement ne semble pas justifier le modèle plus complexe.

Cas général

En général, soit

$$\Omega : \quad Y \approx \theta_1 X_1 + \dots + \theta_p X_p$$

un modèle de régression multiple. Nous appellerons Ω le *modèle complet*. Nous supposons que les conditions de Gauss s'appliquent à Ω . Une *hypothèse linéaire* est un système de r équations indépendantes dans les coefficients, c'est à dire,

$$H_0 : \quad A\theta = 0$$

où A est une matrice $p \times p$ (de constantes) de rang r et $\theta = (\theta_1, \dots, \theta_p)^T$. En utilisant ces équations il est possible d'exprimer r coefficients à l'aide des autres et d'obtenir ainsi un modèle réduit

$$\omega : \quad Y \approx \eta_1 Z_1 + \dots + \eta_q Z_q,$$

où $q = p - r$ et Z_1, \dots, Z_q sont des combinaisons linéaires de X_1, \dots, X_p .

Statistique de test

Les ajustement de Ω et ω aux données fournissent les vecteurs de résidus r_Ω et r_ω . Indiquons par $|r_\Omega|^2$ et $|r_\omega|^2$ les sommes des carrés de leurs composantes, et soit

$$f = \frac{n-p}{p-q} \cdot (|r_\omega|^2 - |r_\Omega|^2) / |r_\Omega|^2.$$

Sous H_0 , la variable aléatoire f suit une distribution F à $p-q$ degrés de liberté (dans le numérateur) et $n-p$ degrés de liberté (dans le dénominateur). On peut donc rejeter H_0 au niveau α si la valeur observée de f est supérieure au percentile $1-\alpha$ de la distribution F à $p-q$ et $n-p$ degrés de liberté.

Remarque. Une expression équivalente de f est

$$f = \frac{(R_\Omega^2 - R_\omega^2)/(p-q)}{(1 - R_\Omega^2)/(n-p)},$$

où R_Ω^2 et R_ω^2 indiquent les coefficients de détermination des modèles Ω et ω .

Exemple: continuation

Pour Ω on obtient

Coeff.	Value	Std.Error	t value	Pr(> t)
theta1	33.3601	1.2116	27.5343	0.0000
theta2	35.2061	1.0645	33.0726	0.0000
theta3	37.1937	1.5063	24.6918	0.0000
theta5	0.0062	0.0147	0.4241	0.6758
theta6	0.0182	0.0133	1.3659	0.1864
theta7	-0.0745	0.0127	-5.8490	0.0000

Residual standard error: 1.556 on 21 degrees of freedom

Multiple R-Squared: 0.9971

Pour ω on obtient

Coeff.	Value	Std.Error	t value	Pr(> t)
eta1	32.1316	0.7483	42.9408	0.0000
eta2	36.1051	0.9716	37.1588	0.0000
eta3	35.5973	0.6596	53.9698	0.0000
eta4	-0.0601	0.0035	-17.3095	0.0000

Residual standard error: 1.605 on 23 degrees of freedom

Multiple R-Squared: 0.9966

Les sommes des carrés des résidus sont obtenues à partir des erreurs standards des résidus:

$$|r_\Omega|^2 = 50.8691 \approx 21 \cdot 1.556^2 \quad \text{et} \quad |r_\omega|^2 = 59.2709 \approx 23 \cdot 1.605^2.$$

Ainsi,

$$f = \frac{27-6}{6-4} \cdot (59.2709 - 50.8691) / 50.8691 = 1.7342.$$

Le percentile 95% de la distribution F à 2 et 21 degrés de liberté se situe à 3.4668. Il n'est donc pas possible de rejeter l'hypothèse de parallélisme au niveau 5%.

20.7 Recherche et validation d'un modèle

La recherche et la validation d'un modèle sont parmi les domaines les plus difficiles de la statistique. Construire un modèle est, en partie, un art. Dans ce qui suit, seules les idées principales sont présentées. On peut distinguer deux catégories de techniques:

- (1) Techniques exploratoires, habituellement basées sur l'analyse graphique des données et des résidus
- (2) Techniques d'inférence basées sur les tests.

Toute analyse de données devrait commencer par une analyse exploratoire pour obtenir une bonne compréhension des données et repérer des valeurs et des tendances particulières.

Sélection de variables. Plusieurs études font intervenir un grand nombre de prédicteurs X_j , mais on ne pourrait pas tous les inclure dans le modèle qui deviendrait trop complexe et l'ajustement trop imprécis. En outre, certains prédicteurs sont fortement corrélés. Les principes suivants devraient être observés:

- (i) Inclure les variables qui sont pertinentes dans le domaine d'application;
- (ii) Réduire au maximum le nombre de variables;
- (iii) Utiliser l'analyse exploratoire comme guide.

Si le nombre k de variables est élevé et les connaissances préalables sont faibles, une méthode de sélection pas-à-pas ("stepwise") peut être envisagée. La méthode "forward selection" commence avec β_0 et inclut les variables au fur et à mesure selon un ordre dicté par leur signification statistique (par exemple, le p -value de la statistique T). La méthode "backward selection" débute avec le modèle le plus complet et élimine une à une les variables de moindre importance (la possibilité de récupérer certaines variables éliminées est prise en considération). En principe, la "backward selection" est préférable, mais elle n'est faisable que si le nombre total de variables est modéré. Le principe d'élimination est le suivant: calculer la signification statistique (p -value) de chaque variable X_j ($j = 1, \dots, k$) tout en gardant les autres; éliminer la variable X_j la moins significative. Certaines variables particulièrement importantes peuvent être retenues obligatoirement dans le modèle. Si un groupe de variables indicatrices représentent la codification d'une variable catégorielle, elles doivent être retenues ou éliminées en bloc.

Linéarité. Jusqu'ici nous avons considéré des modèles avec des variables explicatives non modifiées; mais parfois il convient de les transformer. Supposons avoir une bonne raison pour penser qu'une certaine variable X_j agit de manière quadratique sur la réponse. Il est alors possible de créer une nouvelle variable X_j^2 et de tester son utilité (H_0 : le coefficient de X_j^2 est nul). D'autres outils sont disponibles pour détecter et étudier des éventuelles non-linéarités, par exemple: le "plot des résidus partiels" (Collett (1991), p.135), les transformations de Box-Cox (Carroll and Ruppert, 1988) de la variable réponse, la "modélisation additive généralisée" (Hasties et Tibshirani (1990)).

Interactions. Le nombre d'interactions deux à deux entre k variables est $k(k-1)/2$. D'autre part, les interactions sont relativement rares, mais elles méritent de l'attention. Si k n'est pas trop élevé, pour écarter les interactions clairement inutiles, on peut les modéliser toutes et appliquer une "backward selection" tout en gardant les termes linéaires dans le modèle.

Outliers et points influents. Pour détecter les outliers et les points influents (cas qui déterminent en grande partie les résultats de l'analyse) on peut utiliser des "procédés statistiques robustes" (Hampel et al., 1986; Rousseeuw et Leroy, 1987).

Chapitre 21

Inférence par bootstrap pour la régression

Le bootstrap (Chapitre 16) permet de réaliser l'inférence sans faire appel à un modèle mathématique de la distribution des données. Ce chapitre décrit le bootstrap pour la régression multiple. Les notations du Chapitre 18 seront utilisées; en particulier, le symbole x_i^T indiquera la i -ème ligne de la matrice X du modèle et y_i la i -ème réponse observée. On utilisera aussi l'abréviation $z_i = (x_i^T, y_i)$.

21.1 Rappel des idées de base

Supposons que z_1, \dots, z_n soient les observations dans un problème de régression et qu'elles proviennent d'une population de distribution multivariée F : z_i i.i.d. $\sim F$. Nous souhaitons calculer la distribution d'une statistique $s(z_1, \dots, z_n)$. L'approche classique utilise la description de F fournie par un modèle mathématique (par exemple, la distribution de Gauss) dont les paramètres sont ajustés aux données. Cette approche repose donc sur l'hypothèse que le modèle est adéquat. L'approche bootstrap remplace F par la fonction de distribution empirique F_n , qui associe une probabilité $1/n$ à chaque observation. On dérive ensuite la distribution de s en fonction de F_n , appelée la *distribution bootstrap* de s . Le calcul est effectué par simulation: plusieurs échantillons sont générés à partir de la distribution F_n ; la statistique s est évaluée à l'aide de chaque échantillon simulé; la distribution empirique des valeurs simulées de s (distribution bootstrap) est alors considérée comme une approximation de la distribution de s . On l'utilise, par exemple, pour calculer les intervalles de confiance percentiles (Chapitre 16). Pour la régression, il y a deux schémas de simulation possible: le bootstrap des paires et le bootstrap des résidus.

21.2 Bootstrap des paires et bootstrap des résidus

Le *bootstrap des paires* est particulièrement approprié lorsque les lignes de X caractérisent des individus qui ont été échantillonnés et pour lesquels on a observé la réponse conjointement aux variables explicatives. On obtient k échantillons simulés (par exemple, $k = 1000$) par tirage au sort, avec remplacement, de n paires $(x_1^{*T}, y_1^*), \dots, (x_n^{*T}, y_n^*)$ de l'ensemble des n lignes de X et des réponses y correspondantes. Par exemple, pour les données de la Table 2, Chapitre 18, chaque échantillon simulé est obtenu en tirant au sort 27 triades (Lot, Hrs, Quantité). Un de ces échantillons figure dans la table suivante:

Lot	Hrs	Quantité	Lot	Hrs	Quantité	Lot	Hrs	Quantité
B	29	32.5	A	155	23.2	C	88	28.9
C	125	28.5	A	52	30.4	B	209	25.8
B	177	26.5	C	107	31.7	A	171	20.9
B	29	32.5	A	184	20.9	B	177	26.5
A	99	25.8	C	58	32.8	C	119	28.8
C	107	31.7	A	53	31.1	B	296	18.0
B	385	11.6	B	402	11.8	C	115	29.7
C	119	28.8	B	76	32.0	C	58	32.8
A	171	20.9	C	107	31.7	A	99	25.8

En ajustant le modèle à chacun des k échantillons simulés, on obtient k vecteurs de coefficients simulés et leur distribution bootstrap conjointe.

Si on admet les hypothèses 1 et 2 de la Section 17.1, mais qu'on se méfie de l'hypothèse 3, on peut se limiter à "simuler les erreurs". Celles-ci sont représentées par les résidus

$$r_i = y_i - \hat{\theta}_1 x_{i1} + \dots + \hat{\theta}_p x_{ip}, \quad i = 1, \dots, n.$$

Le *bootstrap des résidus* utilise la distribution empirique des résidus comme estimation de la distribution des erreurs e_i : on obtient donc un échantillon simulé d'erreurs en tirant au sort, avec remplacement, n résidus r_1^*, \dots, r_n^* de l'ensemble $\{r_1, \dots, r_n\}$. On construit ensuite n réponses simulées

$$y_i^* = \hat{\theta}_1 x_{i1} + \dots + \hat{\theta}_p x_{ip} + r_i^*, \quad i = 1, \dots, n.$$

L'ajustement du modèle à $(x_1, y_1^*), \dots, (x_n, y_n^*)$ fournit alors un vecteur de coefficients simulés. Le procédé est répété k fois (par exemple, $k = 1000$) pour obtenir la distribution bootstrap du vecteur des coefficients estimés.

21.3 Exemples

Exemple 1. Considérons les données de la Table 2, Chapitre 18, et le modèle ω du Chapitre 20, Section 6. La Figure 1 montre les histogrammes des 1000 valeurs simulées par bootstrap des paires de $\hat{\eta}_1, \hat{\eta}_2, \hat{\eta}_3$ et $\hat{\eta}_4$.

Les estimations bootstrap des erreurs standard de $\hat{\eta}_1, \hat{\eta}_2, \hat{\eta}_3$ et $\hat{\eta}_4$ sont: 0.798, 1.252, 0.645 et 0.004. Ces valeurs sont assez proches de celles fournies par la méthode classique et reportées au Chapitre 20, Section 6. L'estimation bootstrap la moins semblable à l'estimation classique est celle de l'erreur standard de $\hat{\eta}_2$; or, la distribution bootstrap de $\hat{\eta}_2$ est clairement asymétrique.

Les intervalles percentiles de couverture 95% pour η_1, η_2, η_3 et η_4 sont respectivement:

$$[30.54, 33.74], \quad [34.34, 39.46], \quad [34.45, 37.02], \quad [-0.07, -0.05].$$

Ils sont indiqués par des segments verticaux continus dans la Figure 1. Les intervalles classiques correspondants, obtenus par la méthode décrite au Chapitre 20, Section 3, sont:

$$[30.58, 33.68], \quad [34.09, 38.11], \quad [34.23, 36.96], \quad [-0.07, -0.05].$$

Ils sont indiqués par des segments verticaux en traitillé dans la Figure 1. Les intervalles classiques et les intervalles percentile sont assez semblables dans cet exemple. Ce fait n'est pas surprenant car, comme nous l'avons remarqué grâce à l'analyse graphique du Chapitre 20, Section 5, les hypothèses classiques pour l'inférence sont plausibles.

Exemple 2. Le bootstrap des résidus du modèle ω produit les histogrammes de la Figure 2. Les estimations des erreurs standard de $\hat{\eta}_1, \hat{\eta}_2, \hat{\eta}_3$ et $\hat{\eta}_4$ sont respectivement 0.695, 0.926, 0.609, 0.003 et les intervalles percentiles de couverture 95% sont:

$$[30.78, 33.44], \quad [34.26, 38.01], \quad [34.44, 36.84], \quad [-0.07, -0.05].$$

Les extrémités de ces intervalles sont indiquées par des segments verticaux dans la Figure 2.

Remarque. Le bootstrap fournit une approximation de la distribution conjointe de $\hat{\eta}_1, \hat{\eta}_2, \hat{\eta}_3$ et $\hat{\eta}_4$. Cette distribution nous permet d'estimer, par exemple, la corrélation entre les coefficients estimés. Les histogrammes dans la Figure 1 et Figure 2 représentent les distributions marginales.

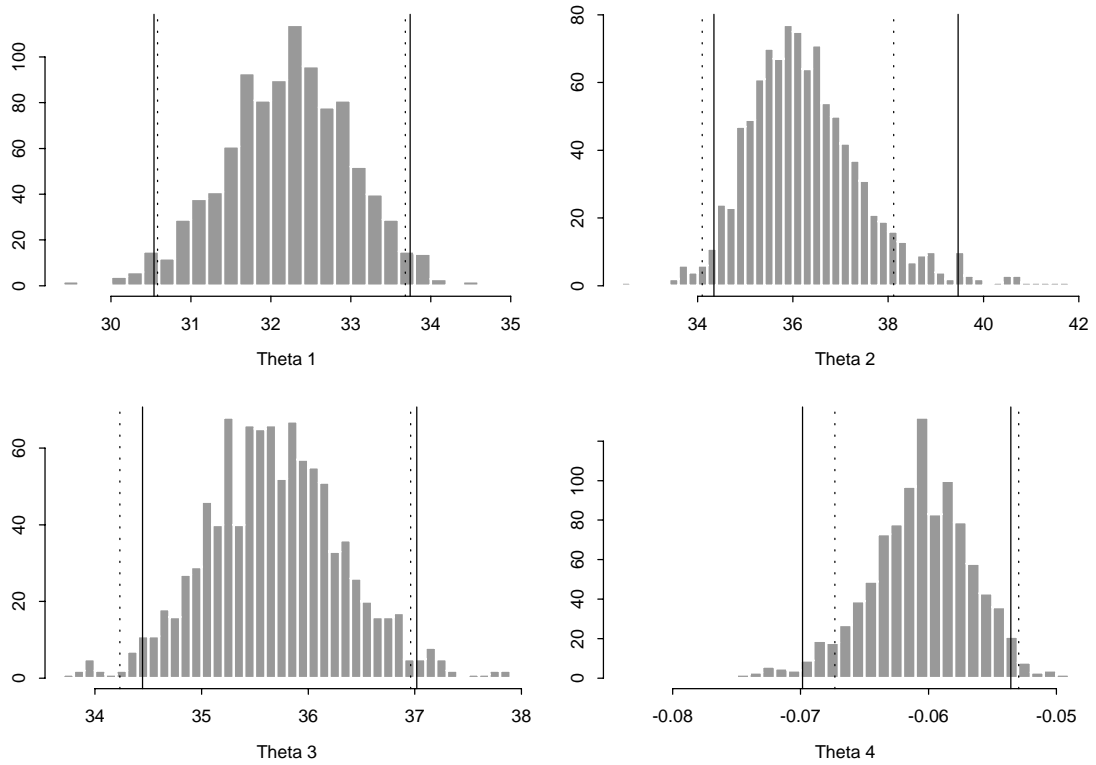


Figure 1. Histogrammes de 1000 valeurs simulées par bootstrap des paires des coefficients de ω . Les traits verticaux continus indiquent les intervalles percentiles et les traits en traitillé les intervalles de confiance classiques. La couverture est de 95% dans tous les cas.

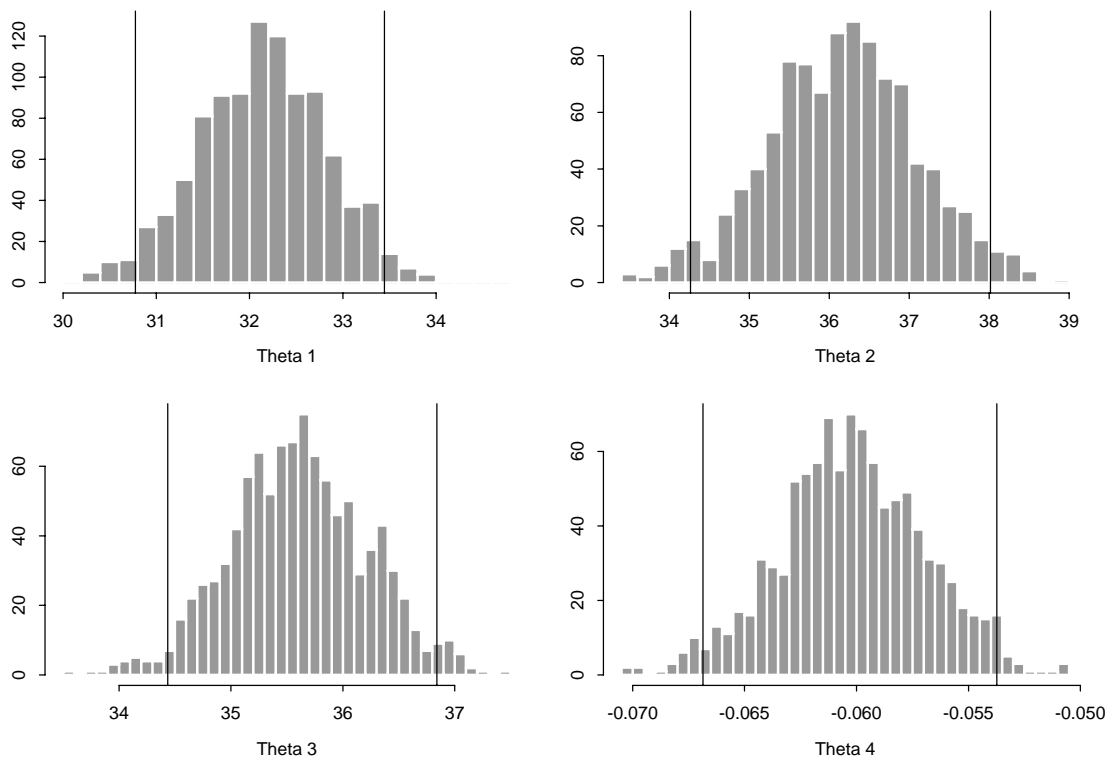


Figure 2. Histogrammes de 1000 valeurs simulées par bootstrap des résidus des coefficients de ω . Les segments verticaux indiquent les intervalles percentiles de couverture 95%.

21.4 Test bootstrap d'une hypothèse linéaire

En général, pour effectuer un test statistique, il faut choisir une statistique de test s et déterminer sa distribution sous l'hypothèse nulle. La méthode bootstrap estime la distribution des données sous l'hypothèse nulle à l'aide d'une distribution empirique cohérente avec l'hypothèse nulle, et dérive la distribution de s par simulation. Dans le cas de la régression, nous considérons un modèle complet

$$\Omega : Y \approx \theta_1 X_1 + \dots + \theta_p X_p$$

avec p paramètres, une hypothèse linéaire $H : A\theta = 0$, où A est une matrice $p \times p$ de rang r , et un modèle réduit

$$\omega : Y \approx \eta_1 Z_1 + \dots + \eta_q Z_q$$

($q = p - r$) obtenu de Ω en utilisant H (voir Chapitre 20, Section 6). L'ajustement de Ω aux données fournit le vecteur de résidus $r_\Omega = (r_{\Omega,1}, \dots, r_{\Omega,n})$ et l'ajustement de ω fournit les estimations $\hat{\theta}_{\omega,1}, \dots, \hat{\theta}_{\omega,n}$ et le vecteur des résidus r_ω . Comme Ω est retenu en tant que modèle adéquat, les résidus r_Ω ne contiennent pas de biais et peuvent servir à estimer la distribution des erreurs e_i . On dérive donc la distribution de la statistique de test par bootstrap des résidus r_Ω . Plus précisément, soit $r_{\Omega,1}^*, \dots, r_{\Omega,n}^*$ un échantillon simulé d'erreurs tirées, avec remise, de $\{r_{\Omega,1}, \dots, r_{\Omega,n}\}$. Un échantillon de réponses simulées sous H est

$$y_i^* = \hat{\theta}_{\omega,1} x_{i,1} + \dots + \hat{\theta}_{\omega,q} x_{i,q} + r_{\Omega,i}^*, \quad i = 1, \dots, n.$$

En ajustant Ω et ω à $(x_1, y_1^*), \dots, (x_n, y_n^*)$, on obtient deux vecteurs de résidus r_Ω^* et r_ω^* , qui permettent de calculer une valeur simulée

$$f^* = [(n - p)/(p - q)](|r_\omega^*|^2 - |r_\Omega^*|^2)/|r_\Omega^*|^2$$

de la statistique de test f . Avec k valeurs simulées f^* on estime le p -value du test par (Nombre de $f^* > f_0$)/ k , où f_0 est la valeur observée de f .

Exemple 3. La Figure 3 montre l'histogramme de 1000 valeurs simulées de f pour tester l'hypothèse de parallélisme des droites représentées par le modèle Ω , Chapitre 20, Section 6. La densité de la distribution F à 2 et 21 degrés de liberté, indiquée dans la figure, est très proche de l'histogramme. La valeur observée de f est $f_0 = 1.7342$ (segment vertical); le p -value bootstrap est 0.215, tandis que celui classique est 0.2009.

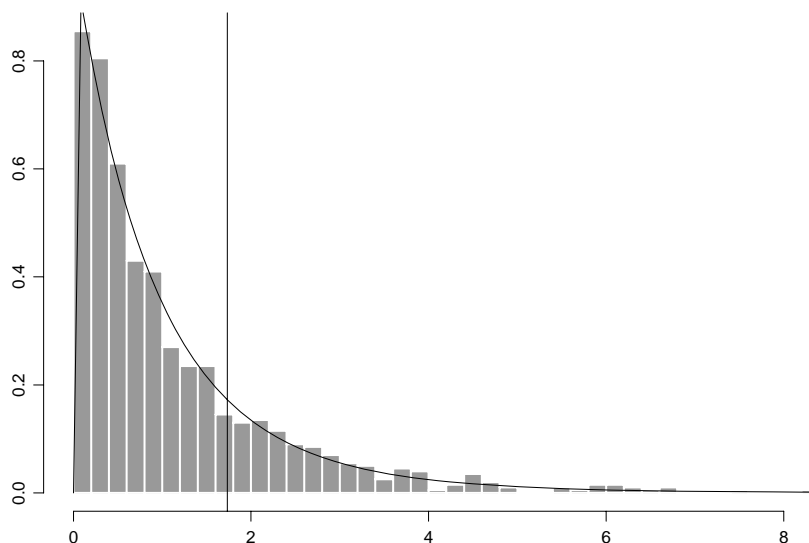


Figure 3. Histogramme de 1000 valeurs simulées de la statistique f et densité de la distribution F à 2 et 21 degrés de liberté. Le trait vertical indique $f_0 = 1.7342$.

Complément

Bootstrap d'un lissage non paramétrique.

La Figure 4 montre à nouveau les données de la Table 1, Chapitre 18. La ligne foncée, qui représente une fonction $\ell(\text{Age})$, est obtenue par un procédé de *lissage non paramétrique* (“loess”). La description de ce procédé est en dehors du cadre de ce cours; voir, par exemple, Chambers et Hastie, Eds., “Statistical Models in S”, Wadworth & Brooks/Cole, 1992; Chapitre 8. La ligne s’adapte aux données de façon locale, sans faire appel à un modèle paramétrique unique pour l’ensemble des valeurs d’Age. Il faut la comparer au polynôme de deuxième degré de la Figure 1, Chapitre 18. Le polynôme décroît pour $\text{Age} > 10$, tandis que le lissage s’aplatit à partir de $\text{Age} = 5$ ou 6.

Les lignes en traitillé représentent 20 lissages $\ell^*(\text{Age})$ calculés, par le même procédé ℓ , sur autant d’échantillons simulés. Chaque échantillon a été obtenu en tirant au sort, avec remise, 43 paires $(\text{Age}, \ln(\text{Conc.}))$ de la Table 1. Le nuage donne une image palpable de la variabilité du lissage et confirme visuellement son aplatissement.

Pour tester l’existence d’une éventuelle flexion pour $\text{Age} > 10$, nous avons simulé 500 échantillons et, pour chaque échantillon, nous avons calculé un lissage simulé ℓ^* . Pour chaque lissage, nous avons cherché à calculer la pente

$$(\ell^*(15.6) - \ell^*(10))/(15.6 - 10),$$

mais seuls 329 échantillons s’étendaient jusqu’à la valeur 15.6. Ainsi, nous n’avons obtenu que 329 pentes simulées. Les percentiles 5% et 95% de la distribution des 329 pentes ainsi obtenus sont respectivement -0.0183 et 0.0153 . On ne peut donc pas rejeter l’hypothèse que la flexion est nulle.

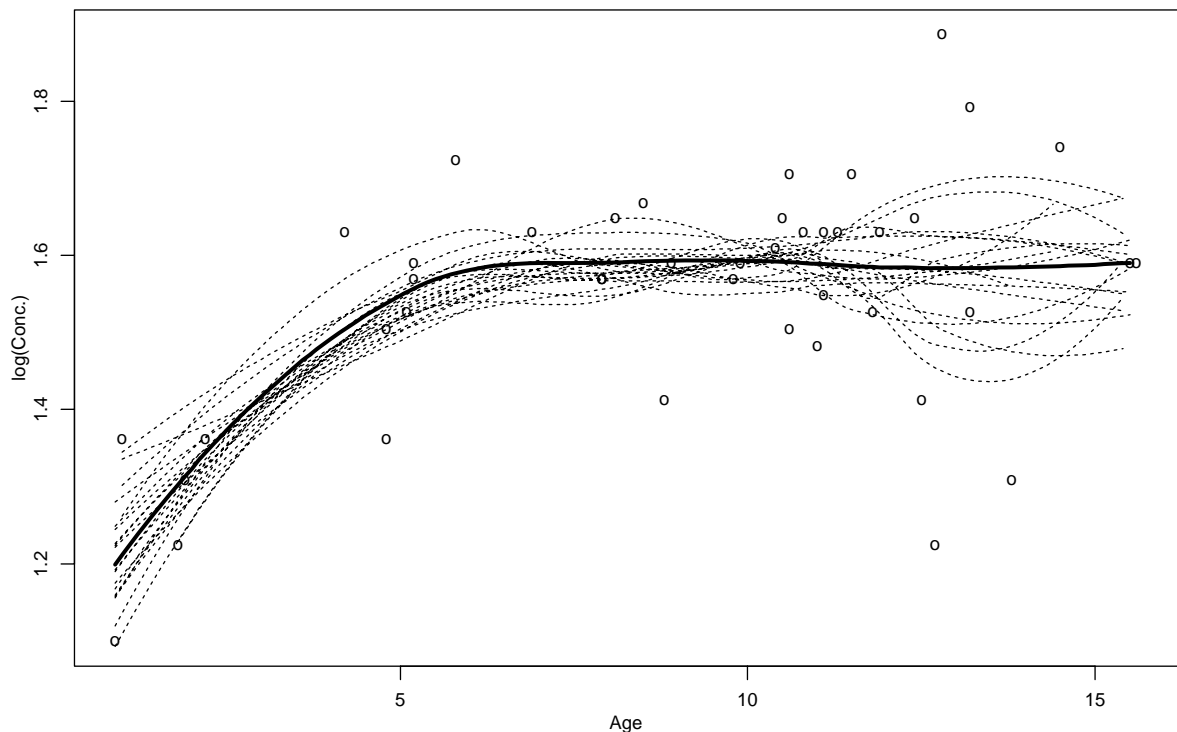


Figure 4. Lissage non-paramétrique (ligne foncée) des données de la Table 1, Chapitre 18, et 20 lissages non-paramétriques simulés (lignes traitillées).

Chapitre 22

Introduction à la régression logistique

La régression ordinaire permet d'analyser une variable réponse quantitative en fonction d'une ou plusieurs variables explicatives. Souvent, c'est un *résultat binaire* (ou *dichotomique*) d'une expérience ou d'une observation que l'on souhaite mettre en relation avec des variables explicatives; par exemple:

- des patients peuvent survivre ou décéder; les différentes thérapies et les facteurs de risque peuvent être considérés comme des variables qui contribuent à expliquer la survie ou le décès;
- des personnes peuvent être atteintes par une maladie. On souhaite étudier la relation entre les chances d'être atteint et certains facteurs explicatifs ou facteurs de risque (par exemple, âge, fumée, sexe);
- des personnes peuvent avoir ou ne pas avoir un emploi selon leur âge, sexe, type de formation;
- un appareil peut fonctionner ou ne pas fonctionner; cet état peut être mis en relation avec son âge, les conditions de l'environnement, etc.

La *régression logistique* permet d'étudier la relation entre une variable réponse binaire et plusieurs variables explicatives. Ce chapitre donne une brève introduction à la régression logistique. On trouvera un traitement plus approfondi dans le livre de Hosmer et Lemeshow (1989), duquel cette introduction est tirée.

22.1 Introduction

En général, le résultat d'une observation binaire est appelé "succès" ou "échec". Il est représenté mathématiquement par une variable aléatoire Y telle que $Y = 1$ s'il y a succès et $Y = 0$ s'il y a échec. Cette variable a une distribution de Bernoulli et on note par $p = P(Y = 1)$ la probabilité de succès; donc $P(Y = 0) = 1 - p$. L'espérance mathématique et la variance de Y sont, respectivement, $E(Y) = p$ et $\sigma^2(Y) = p(1 - p)$. Le résultat Y peut dépendre des valeurs assumées par k variables explicatives X_1, \dots, X_k au moment de l'observation et nous souhaitons étudier cette relation. L'exemple suivant montre que les techniques de régression ordinaire ne sont pas adaptées à ce type d'analyse.

Exemple 1. La Table 1 concerne un échantillon de 100 personnes, pour lesquels la présence (CHD = 1) ou l'absence (CHD = 0) d'une maladie cardiovasculaire a été observée. On souhaite étudier la relation entre CHD et la variable explicative âge (AGE). La Figure 1 montre un diagramme de dispersion de CHD versus AGE. Evidemment, ce diagramme ne donne pas une information très utile même si on remarque une proportion plus élevée de "cas" (CHD = 1) pour les personnes âgées que pour les jeunes. Il n'est pas opportun d'adapter une droite à ce diagramme. Il est, toutefois, raisonnable de décrire la relation entre la probabilité de CHD = 1 pour une valeur donnée a de AGE, c'est à dire, la probabilité conditionnelle $P(\text{CHD} = 1 | \text{AGE} = a)$, par une fonction mathématique simple (modèle) de la variable AGE. La Figure 2, qui représente les fréquences relatives de CHD = 1 selon les catégories d'âge définies par la variable AGRP de la Table 1, nous suggère l'allure de cette fonction.

Figure 1. Diagramme de CHD et AGE.

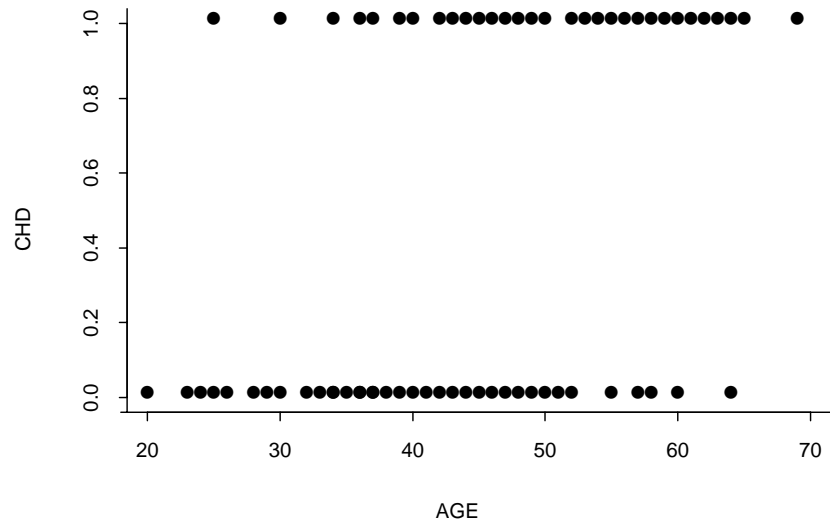
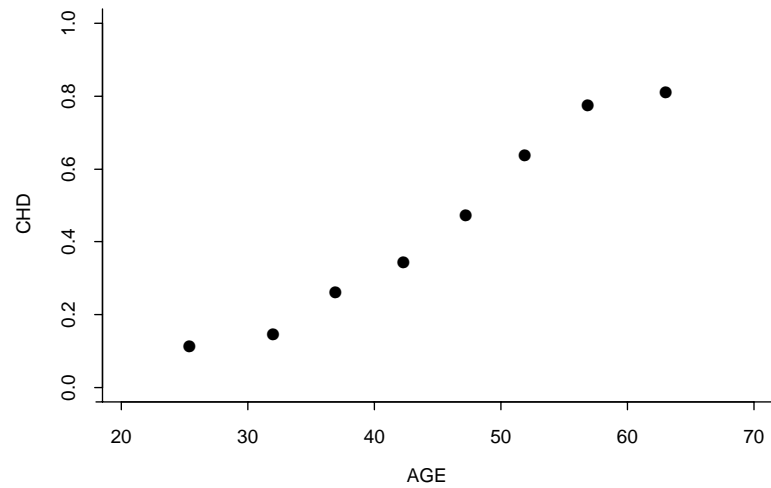
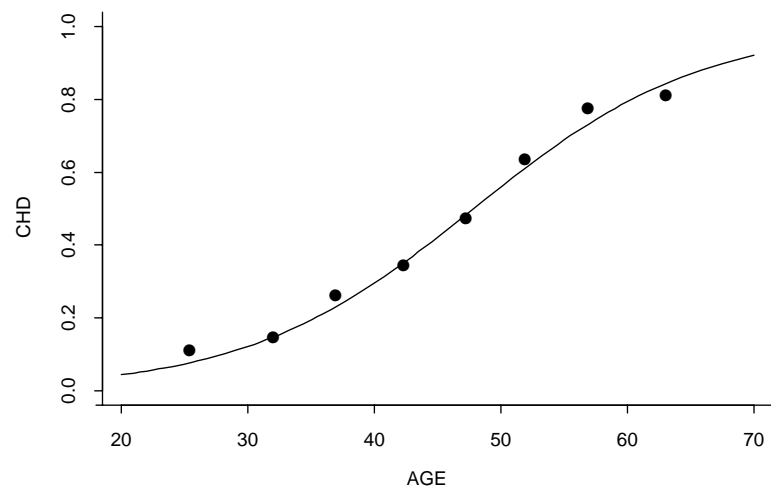


Figure 2. Diagramme des proportions de personnes avec CHD = 1 selon AGE en groupes.

Figure 3. $\hat{p}(\text{AGE}) = \exp(-5.31 + 0.111 \cdot \text{AGE}) / (1 + \exp(-5.31 + 0.111 \cdot \text{AGE}))$ 

Considérons d'abord le cas d'une seule variable explicative quantitative X . Nous nous proposons d'utiliser une fonction mathématique $p(x)$ simple comme modèle pour $P(Y = 1|X = x)$. S'agissant d'une probabilité, la fonction $p(x)$ doit être bornée par les valeurs 0 et 1. Elle ne peut donc pas être linéaire. L'Exemple 1 suggère que $p(x)$ a une forme sigmoïdale qui peut être approchée par une fonction de distribution cumulative, par exemple, la fonction de distribution normale $F = \Phi$. Plus précisément, on peut utiliser le modèle

$$p(x) = \Phi(\beta_0 + \beta_1 x).$$

Ici, β_0 et β_1 sont les *paramètres* du modèle ou *coefficients*. Si Φ^{-1} est la fonction inverse de Φ (*transformation probit*), on obtient

$$\Phi^{-1}(p(x)) = \beta_0 + \beta_1 x,$$

c'est-à-dire, une relation linéaire. Ce modèle, connu comme le *modèle probit*, a joui d'une certaine popularité dans l'essai biologique (Finney, 1978).

Toutefois, la forme la plus utilisée est celle de la fonction de distribution logistique F_L , c'est-à-dire:

$$F_L(\beta_0 + \beta_1 x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

On pose donc le modèle

$$p(x) = F_L(\beta_0 + \beta_1 x)$$

appelé *modèle logit* ou *logistique*. La transformation inverse

$$F_L^{-1}(y) = \ln(y/(1 - y)), \quad 0 < y < 1,$$

est appelée la *transformation logit* et l'expression $\ln(p/(1 - p))$ est appelé le *logit de p*, noté $\text{logit}(p)$. Donc,

$$F_L^{-1}(p(x)) = \text{logit}(p(x)) = \ln\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x$$

est une fonction linéaire. La fonction $K(x) = \text{logit}(p(x))$ est aussi appelée une *link function* dans la théorie des modèles linéaires généralisés (McCullagh et Nelder, 1989). On observe qu'elle peut varier entre $-\infty$ et $+\infty$.

Le modèle peut être étendu à l'analyse d'une variable réponse binaire Y en fonction de plusieurs variables explicatives X_1, \dots, X_k , qui peuvent être quantitatives, en catégories ordonnées, ou qualitatives (exprimées de façon numérique). Dans ce cas, on cherche une fonction $p(x_1, \dots, x_k)$ à plusieurs variables comme modèle pour la probabilité conditionnelle $P(Y = 1|X_1 = x_1, \dots, X_k = x_k)$. Le modèle logit utilise la fonction

$$p(x_1, \dots, x_k) = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)},$$

c'est-à-dire la relation linéaire

$$K(x_1, \dots, x_k) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k,$$

avec link function

$$K(x_1, \dots, x_k) = \ln(p(x_1, \dots, x_k)/(1 - p(x_1, \dots, x_k))).$$

En pratique, les coefficients $\beta_0, \beta_1, \dots, \beta_k$ doivent être déterminés à l'aide des données. On utilise la méthode du maximum de vraisemblance (Chapitre 8). En général, cette méthode fournit des estimateurs avec de bonnes propriétés statistiques: les estimateurs ont approximativement une distribution normale et leurs variances sont relativement petites. Toutefois, ces propriétés ne sont valables que si la taille n de l'échantillon est grande et que le nombre de paramètres est petit (McCullagh et Nelder, 1989).

Les estimations sont souvent associées à des tests d'hypothèses du type

$$H_0 : \beta_h = \beta_{h+1} = \dots = \beta_k = 0$$

avec $1 \leq h \leq k$. L'hypothèse H_0 affirme que X_h, X_{h+1}, \dots, X_k ne sont pas utiles pour expliquer la probabilité conditionnelle de succès $P(Y = 1 | X_1 = x_1, \dots, X_k = x_k)$. À l'aide de ces tests, le problème de la construction d'un modèle adéquat – c'est-à-dire, avec un bon degré d'ajustement et un faible nombre de paramètres – peut être abordé. Enfin, on peut calculer des intervalles de confiance pour les coefficients $\beta_0, \beta_1, \dots, \beta_p$.

Remarque. En général, les modèles logit et probit fournissent des valeurs très proches. Toutefois, l'interprétation des paramètres du modèle logit est avantageuse, car elle s'appuie sur des importants concepts utilisés en épidémiologie (Section 4, ci-dessous).

22.2 Estimation et tests: cas d'une seule variable explicative

Nous allons esquisser la méthode du maximum de vraisemblance pour le cas d'une seule variable explicative X , c'est-à-dire la *régression logistique simple*. La vraisemblance d'un échantillon (x_i, y_i) , $i = 1, \dots, n$ (où les x_i sont les valeurs observées de X et les y_i celles de Y – donc $y_i = 0$ ou 1) est

$$p(x_i)^{y_i} (1 - p(x_i))^{1-y_i},$$

où

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

dépend de β_0 et β_1 . Comme on admet que les observations sont indépendantes, la vraisemblance de l'échantillon selon le modèle est

$$L(\beta_0, \beta_1) = \prod_{i=1}^n p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}.$$

Le critère du maximum de vraisemblance détermine les valeurs de β_0 et β_1 qui rendent maximale cette vraisemblance. Dans ce but, il convient de considérer l'opposé de son logarithme, c'est-à-dire, la fonction *log-likelihood*

$$\begin{aligned} \ell(\beta_0, \beta_1) &= -\ln L(\beta_0, \beta_1) \\ &= -\sum_{i=1}^n [y_i \ln p(x_i) + (1 - y_i) \ln(1 - p(x_i))]. \end{aligned}$$

On minimise alors cette fonction en annulant ses dérivées partielles selon β_0 et β_1 . On obtient ainsi les conditions

$$\sum_{i=1}^n (y_i - p(x_i)) = 0 \quad \text{et} \quad \sum_{i=1}^n x_i (y_i - p(x_i)) = 0.$$

Les solutions $\hat{\beta}_0$ et $\hat{\beta}_1$ de ces équations sont les *estimateurs du maximum de vraisemblance* de β_0 et β_1 . En général, elles sont calculées à l'aide de programmes de calcul numérique.

A l'aide des estimations $\hat{\beta}_0$ et $\hat{\beta}_1$, on peut estimer les probabilités de succès pour différentes valeurs x de la variable explicative:

$$\hat{p}(x) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x)}.$$

Les valeurs de la fonction $\hat{p}(x)$ sont parfois appelées les *probabilités ajustées*.

Exemple 2. Avec les données de la Table 1, on obtient les coefficients estimés indiqués dans la Table 2, c'est-à-dire, $\hat{\beta}_0 = -5.310$ et $\hat{\beta}_1 = 0.111$ et donc

$$\hat{p}(x) = \frac{\exp(-5.31 + 0.111 \times \text{AGE})}{1 + \exp(-5.31 + 0.111 \times \text{AGE})}.$$

La Figure 3 donne le graphique de cette fonction qui s'adapte assez bien aux fréquences relatives de CHD selon AGE (en groupes). La valeur du log likelihood $\ell(\hat{\beta}_0, \hat{\beta}_1)$ est -53.677 .

Table 2. Résultats de l'ajustement d'un modèle logistique à une seule variable explicative $X = \text{AGE}$ aux données de la Table 1.

Variable	Estimation Coefficient	Erreur Standard	Coeff./ $\hat{\sigma}$
AGE	0.111	0.024	4.61
Constante	-5.310	1.134	-4.68

Log-likelihood=-53.677

Les programmes usuels fournissent aussi les écarts types $\hat{\sigma}(\hat{\beta}_0)$ et $\hat{\sigma}(\hat{\beta}_1)$ de $\hat{\beta}_0$ et $\hat{\beta}_1$. Grâce au fait que la distribution des estimateurs est approximativement normale on peut construire des intervalles de confiance avec coefficient de couverture $1 - 2\alpha$:

$$[\hat{\beta}_j - \hat{\sigma}(\hat{\beta}_j)z_{1-\alpha}, \quad \hat{\beta}_j + \hat{\sigma}(\hat{\beta}_j)z_{1-\alpha}], \quad j = 0, 1,$$

où $z_{1-\alpha}$ est le quantile $1 - \alpha$ de la distribution normale standard (par exemple, $\alpha = 0.025$ et $z_{0.975} = 1.96$).

Enfin, on peut aussi tester l'hypothèse

$$H_0 : \beta_j = 0$$

($j = 1$ ou $j = 2$) contre l'une des deux alternatives

$$H_1 : \beta_j > 0 \quad (\text{unilatérale}) \quad \text{ou} \quad H_1 : \beta_j \neq 0 \quad (\text{bilatérale}).$$

Le procédé le plus simple utilise la statistique

$$T = \hat{\beta}_j / \hat{\sigma}(\hat{\beta}_j).$$

Sous l'hypothèse, la statistique T a approximativement une distribution normale standard. Au niveau α , on rejette donc H_0 en faveur d'une alternative unilatérale H_1 (par exemple) si $T > z_{1-\alpha}$. De façon équivalente, on rejette H_0 si la valeur observée t_0 de T est telle que $P(T > t_0) < \alpha$. Ce test est connu comme le *test de Wald*. Un autre test sera présenté dans la section suivante.

Exemple 3. Les écarts types et les valeurs de la statistique T pour les coefficients β_0 et β_1 de l'Exemple 1 sont donnés dans la Table 2. Pour l'hypothèse $H_0 : \beta_1 = 0$ (β_1 est le coefficient de la variable AGE) on obtient $t_0 = 0.111/0.024 = 4.610$. A l'aide d'une table de la distribution normale on trouve que $P(T > 4.610) < 0.0001$ et on conclut que la variable AGE est importante pour expliquer la probabilité de CHD=1.

22.3 Estimation et tests: cas de plusieurs variables explicatives

Un des buts principaux de la régression logistique est celui d'examiner les effets conjoints de plusieurs variables explicatives et de leurs interactions.

Exemple 4. Comme un petit poids à la naissance (LBW = Low Birth Weight) a une influence négative sur le développement de l'enfant, les facteurs de risque de LBW sont de grand intérêt en médecine préventive. Dans une étude de 189 cas, 8 facteurs de risque potentiels (âge maternel, fumée, hypertension, etc.) ont été enregistrés. Les données figurent dans Hosmer et Lemeshow (1989). $n_1 = 59$ bébés avaient un poids au-dessous de la normale et $n_0 = 130$ un poids normal. Quatre variables ont été choisies comme prédicteurs: l'âge de la mère (AGE), son poids aux dernières règles (PDS), le nombre de visites médicales qu'elle a eues durant le premier trimestre (VST) et sa race, en 3 catégories, codées à l'aide de deux variables indicatrices RACE1 et RACE2.

Souvent, comme dans l'Exemple 4, des informations concernant un grand nombre de variables explicatives X_1, \dots, X_k sont disponibles. Comme dans le cas de la régression multiple ordinaire, elles forment une matrice du modèle X dont les lignes sont les vecteurs $(1, x_{i1}, \dots, x_{ik})$ et x_{ik} indique la i -ème observation (observation du cas i) de la variable k . Le modèle

$$K(x_1, \dots, x_k) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k,$$

est alors ajusté par la méthode du maximum de vraisemblance. Dans ce but, on résout un système de $(k + 1)$ équations pour les coefficients β_0 et β_1, \dots, β_k , que l'on obtient en annulant les dérivées partielles de la fonction log likelihood $\ell(\beta_0, \beta_1, \dots, \beta_p)$:

$$\frac{\partial \ell(\beta_0, \beta_1, \dots, \beta_k)}{\partial \beta_0} = \sum_{i=1}^n (y_i - p(x_{i1}, \dots, x_{ip})) = 0,$$

$$\frac{\partial \ell(\beta_0, \beta_1, \dots, \beta_k)}{\partial \beta_j} = \sum_{i=1}^n x_{ij} (y_i - p(x_{i1}, \dots, x_{ip})) = 0, \quad j = 1, \dots, k.$$

L'interprétation des données fournie par la régression multiple est supérieure à celle fournie par la régression simple. La régression multiple tient compte des éventuelles associations entre les variables explicatives. Les coefficients de chaque variable sont épurés des contributions fournies par les autres variables et représentent, donc, des effets propres.

Exemple 4 (continuation). La Table 3 donne les coefficients estimés d'une régression logistique de LBW en fonction de AGE, PDS, RACE (RACE1 et RACE2) et VST. La dernière colonne donne les valeurs de la statistique $\hat{\beta}_j/\hat{\sigma}(\hat{\beta}_j)$ pour le test de Wald de chaque coefficient. On voit immédiatement que les effets de PDS et RACE1 sont significatifs ($P < 0.05$). Au contraire, les effets de AGE et de VST sont nettement non-significatifs et ces variables peuvent être écartées du modèle. Toutefois, RACE2 ne peut pas être éliminée puisqu'elle est utilisée en combinaison avec RACE1.

Table 3. Estimation des coefficients d'une régression logistique multiple sur des données concernant des bébés de faible poids à la naissance.

Variable	Estimation Coefficient	Erreur Standard $\hat{\sigma}$	Coeff./ $\hat{\sigma}$
AGE	-0.024	0.034	-0.71
PDS	-0.014	0.00652	-2.14
RACE1	1.004	0.497	2.02
RACE2	0.433	0.362	1.20
VST	-0.049	0.167	-0.30
Constante	1.295	1.069	1.21

Log-Likelihood=-111.286

Pour tester une *hypothèse linéaire* qui concerne plusieurs coefficients on utilise le *test du rapport de vraisemblance*. Supposons que le modèle courant (ou complet) soit

$$K(x_1, \dots, x_k) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

et que l'hypothèse à tester soit

$$H_0 : \beta_h = \beta_{h+1} = \dots = \beta_k = 0$$

avec $1 \leq h \leq k$ (c'est le type d'hypothèse linéaire le plus fréquent). Le modèle réduit est donc

$$K(x_1, \dots, x_k) = \beta_0 + \beta_1 x_1 + \dots + \beta_{h-1} x_{h-1}.$$

On définit d'abord la *déviante* du modèle courant par rapport au modèle saturé (voir note ci-dessous):

$$D(\text{modèle courant}) = -2 \ln \left(\frac{\text{vraisemblance du modèle courant}}{\text{vraisemblance du modèle saturé}} \right).$$

La déviante est une mesure de comparaison entre les probabilités $\hat{p}(x_i^*)$ ajustées à l'aide du modèle courant et celles ajustées à l'aide du modèle saturé, c'est-à-dire, les fréquences observées.

La statistique du test du rapport de vraisemblance est

$$\begin{aligned} G &= -2 \ln \left(\frac{\text{vraisemblance du modèle réduit}}{\text{vraisemblance du modèle complet}} \right) \\ &= D(\text{modèle réduit}) - D(\text{modèle complet}) \\ &= -2 [\ln(\text{vraisemblance du modèle réduit}) - \ln(\text{vraisemblance du modèle complet})]. \end{aligned}$$

Dans son esprit, ce calcul est similaire à la différence des sommes des carrés des résidus dans la régression ordinaire. Sous l'hypothèse H_0 , la statistique G a approximativement une distribution χ^2 avec $k - h + 1$ degré de liberté. On rejette donc H_0 , au niveau α , si la valeur observée g_0 de G dépasse le quantile $1 - \alpha$ de la distribution χ^2 à $k - h + 1$ degré de liberté.

Exemple 4 (continuation). On peut tester si l'ensemble des 5 variables de la Table 3 explique la probabilité d'une réponse positive de façon significative. L'hypothèse est:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0.$$

La vraisemblance du modèle complet (à 6 coefficients) doit être comparée à celle du modèle réduit $K(x_1, \dots, x_5) = \beta_0$. On trouve

$$\begin{aligned} \ln(\text{vraisemblance du modèle complet}) &= -111.29, \\ \ln(\text{vraisemblance du modèle réduit}) &= -117.34. \end{aligned}$$

Donc

$$g_0 = -2((-117.34) - (-111.29)) = 12.1$$

et $P(G > 12.1) = 0.033$ (G a 5 = 6 - 1 degrés de liberté); le modèle complet est donc significatif. Par analogie, on pourrait tester s'il est opportun d'inclure les variables VST et AGE en supposant que PDS, RACE1 et RACE2 soient incluses de toute façon. La vraisemblance d'un modèle à 6 coefficients (5 variables et un intercept) devrait être comparée à celle d'un modèle à 3 variables; G aurait 6 - 3 = 3 degrés de liberté.

Exemple 5. Pour le cas d'une seule variable explicative, il n'y a que trois "modèles courants" possibles: le modèle $K(x) = \beta_0 + \beta_1 x$, le modèle sans intercept $K(x) = \beta_1 x$ et le modèle constant $K(x) = \beta_0$. Si $H_0 : \beta_1 = 0$ on a $k = h = 1$, $k - h + 1 = 1$ et on obtient

$$D(\text{modèle courant}) = -2 \sum_{i=1}^{n^*} [y_i^* \ln(\hat{p}(x_i^*)/y_i^*) + (1 - y_i^*) \ln((1 - \hat{p}(x_i^*))/(1 - y_i^*))].$$

Le signe * indique que des "cas similaires" (avec la même valeur de la variable explicative) ont été regroupés (comme dans la Figure 2). En d'autres termes, y_i^* est la fréquence relative de succès pour $X = x_i$; c'est aussi l'estimation de $p(x_i^*)$ sous le modèle saturé. Avec les données de la Table 1 et $H_0 : \beta_1 = 0$ on trouve $g_0 = 29.31$. Comme G a approximativement une distribution χ^2 à 1 degré de liberté, $P(G > 29.31)$ est inférieur à 0.001.

Note. Un *modèle saturé* est un modèle qui a autant de paramètres que de points qu'il doit ajuster; par exemple, une droite de régression lorsque les données représentées dans le diagramme de dispersion sont regroupées dans deux seuls points.

22.4 Interprétation des coefficients

Dans le cas de la régression ordinaire simple, une variation unitaire dans la valeur x de la variable X produit un changement de β_1 unités dans l'espérance conditionnelle $E(Y|X = x)$ de Y . Pour la régression logistique à une seule variable explicative la relation entre $p(x)$ et x est donnée par le logit:

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x.$$

Donc, un incrément unitaire en x produit une variation de “ β_1 logits”. Nous allons préciser ce que cette expression signifie pour différents types de variables explicatives X .

Variable explicative binaire. Nous considérons la régression logistique simple, mais la généralisation au cas multiple est possible. Une variable explicative binaire est utilisée pour indiquer la présence ($X = 1$) ou l'absence ($X = 0$) d'une certaine condition X . Pour mesurer l'association entre X et Y , où $Y = 1$ indique la présence d'une maladie, on utilise en épidémiologie le *odds ratio* ou *rapport des cotes* (Fleiss (1981)). La *cote* (*odds*) de $Y = 1$ pour les individus avec $X = 0$ est définie comme

$$\Omega(0) = \frac{P(Y = 1|X = 0)}{1 - P(Y = 1|X = 0)} = \frac{p(0)}{1 - p(0)}.$$

Par analogie, on définit la cote de $Y = 1$ en présence de $X = 1$:

$$\Omega(1) = \frac{P(Y = 1|X = 1)}{1 - P(Y = 1|X = 1)} = \frac{p(1)}{1 - p(1)}.$$

La cote est donc le rapport entre la probabilité d'être malade et la probabilité d'être sain et son logarithme est le logit. Enfin, l'*odds ratio* pour comparer la présence et l'absence de X est le rapport

$$o(1, 0) = \Omega(1)/\Omega(0).$$

Si l'association entre X et Y est faible, $P(Y = y|X = 0) \approx P(Y = y|X = 1)$ et $o(1, 0)$ est proche de 1. Inversement, un odds ratio supérieur ou inférieur à 1 indique une association entre X et Y . Avec $p(x) = \exp(\beta_0 + \beta_1 x)/(1 + \exp(\beta_0 + \beta_1 x))$ on obtient

$$o(1, 0) = \exp(\beta_1)$$

et donc

$$\beta_1 = \ln(o(1, 0)) = \text{logit}(p(1)) - \text{logit}(p(0)).$$

Le coefficient β_1 indique donc de combien le logit de devenir malade est augmenté par l'exposition à la condition X .

Remarques

1. On peut estimer $o(1, 0)$ par $\hat{o}(1, 0) = \exp(\hat{\beta}_1)$ et obtenir un intervalle de confiance pour $o(1, 0)$ en prenant l'exponentielle ($\exp(\cdot)$) des limites d'un intervalle de confiance pour β_1 .
2. Si les valeurs de $P(Y = 1|X = 0)$ et de $P(Y = 1|X = 1)$ sont très petites, les odds $\Omega(1)$ et $\Omega(0)$ sont proches de leur numérateur et la valeur numérique de l'odds ratio est proche de celle du *risque relatif* $r(1, 0) = P(Y = 1|X = 1)/P(Y = 1|X = 0)$. L'approximation suivante du risque relatif est toutefois meilleure: $r \approx o + o[1 - o]p(0)$.

Variable explicative qualitative à plusieurs niveaux. Pour l'interprétation d'une variable explicative qualitative (facteur) à plusieurs niveaux, nous nous servons d'un exemple.

Exemple 6. La Table 4 fournit les fréquences de $Y = 1$ (CHD présent) et de $Y = 0$ (CHD absent) selon les 4 catégories de la variable Race à 4 niveaux: Blanche, Noire, Hispanique, Autre.

Table 4. Classification de données hypothétiques selon CHD et Race, pour 100 sujets.

CHD	Blanche	Noire	Hispanique	Autre	Total
Présent	5	20	15	10	50
Absent	20	10	10	10	50
Total	25	30	25	20	100
Odds ratio ($\hat{\delta}$)	1.0	8.0	6.0	4.0	
$\ln(\hat{\delta})$	0.0	2.08	1.79	1.39	
Int. conf. à 95%		(2.3,27.6)	(1.7,21.3)	(1.1,14.9)	

Sans utiliser de modèles, les odds ratios pour comparer chaque niveau de Race à Race Blanche peuvent être estimés à l'aide des tableaux 2×2 correspondants.

Pour utiliser le modèle de régression, il faut coder numériquement la variable Race à 4 niveaux. Le codage usuel utilise 3 variables indicatrices D_1 , D_2 et D_3 , par exemple celles définies dans la Table 5, où Blanche est le niveau de référence. (Comme pour la régression multiple, pour coder un facteur à k niveaux, il faut utiliser $k - 1$ variables indicatrices.)

Table 5. Codage du facteur Race avec niveau de référence Blanche.

Race	Variables		
	D_1	D_2	D_3
Blanche	0	0	0
Noire	1	0	0
Hispanique	0	1	0
Autre	0	0	1

Les coefficients estimés $\hat{\beta}_1$, $\hat{\beta}_2$ et $\hat{\beta}_3$ de D_1 , D_2 et D_3 sont respectivement les logarithmes des odds ratios qui figurent dans la Table 4. Par exemple:

$$\begin{aligned} \ln(\hat{\delta}(\text{Noire,Blanche})) &= \text{logit}(\hat{p}(\text{Noire})) - \text{logit}(\hat{p}(\text{Blanche})) \\ &= [\hat{\beta}_0 + \hat{\beta}_1(1) + \hat{\beta}_2(0) + \hat{\beta}_3(0)] - [\hat{\beta}_0 + \hat{\beta}_1(0) + \hat{\beta}_2(0) + \hat{\beta}_3(0)] = \hat{\beta}_1 \end{aligned}$$

Donc $\hat{\beta}_1 = 2.079$, $\hat{\beta}_2 = 1.792$, $\hat{\beta}_3 = 1.386$. En outre,

$$p(\text{Blanche}) = \exp(\beta_0)/(1 + \exp(\beta_0)) = 1/5$$

d'où $\beta_0 = \ln(1/4) = -1.386$.

Variable explicative continue. Soit X une variable explicative continue et soit $p(x) = P(Y = 1|X = x)$. Considérons l'odds ratio correspondant à deux valeurs x_1 et x_0 de X :

$$o(x_1, x_0) = \frac{p(x_1)/(1 - p(x_1))}{p(x_0)/(1 - p(x_0))}.$$

Si

$$K(x) = \beta_0 + \beta_1 x,$$

alors β_1 est le log de l'odds ratio correspondant à un incrément unitaire:

$$\beta_1 = \ln(o(x + 1, x)).$$

Si on s'intéresse à un incrément de c unités, on obtient évidemment,

$$K(x + c) - K(x) = c\beta_1, \quad \text{c'est-à-dire,} \quad o(x + c, x) = \exp(c\beta_1).$$

Remarque. On peut facilement obtenir un intervalle de confiance avec coefficient de couverture $1 - 2\alpha$ pour $o(x + c, x)$. L'intervalle est:

$$[\exp(c\hat{\beta}_1 - z_{1-\alpha}c\hat{\sigma}(\hat{\beta}_1)), \exp(c\hat{\beta}_1 + z_{1-\alpha}c\hat{\sigma}(\hat{\beta}_1))].$$

Exemple 7. Avec les données de la Table 1 on avait obtenu $\hat{K}(\text{AGE}) = -5.310 + 0.111 \times \text{AGE}$. L'odds ratio pour un incrément de AGE de 10 ans est alors $o(\text{AGE}+10, \text{AGE}) = 3.03$ et un intervalle de confiance de couverture 95% est

$$[\exp(10 \times 0.111 - 1.96 \times 10 \times 0.024), \exp(10 \times 0.111 + 1.96 \times 10 \times 0.024)] = [1.90, 4.86].$$

Variable explicative en catégories ordonnées. Une variable en catégories ordonnées (ou variable ordinale) est une variable dont les modalités ne sont pas numériques mais peuvent être ordonnées. Un exemple est une variable avec modalités Bon, Satisfaisant, Suffisant, Insuffisant. Si le nombre de modalités est supérieur à 3, il convient généralement de traiter une variable ordinale comme si elle était quantitative (et coder les modalités avec leur rang); dans le cas contraire, il faut la traiter comme un facteur.

Interactions. Dans la régression logistique multiple, l'effet d'une variable explicative X_j sur la réponse moyenne est ajusté en tenant compte des autres variables X_k , avec $k \neq j$, comme dans la régression multiple ordinaire. Supposons, par exemple, que le modèle

$$K(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

soit utilisé pour expliquer $Y = \text{CHD}$ à l'aide de $X_1 = \text{AGE}$ et de $X_2 = \text{SEXE}$. (Evidemment, il faudrait connaître le sexe de chaque sujet, mais la Table 1 ne donne pas cette information.) Si AGE et SEXE étaient associés, l'effet d'AGE constaté dans l'analyse univariée (Exemples 1, 2, 3) pourrait être dû au sexe. En effet, CHD est plus fréquent chez les hommes que chez les femmes, mais les chances de CHD augmentent aussi avec l'âge, et les femmes atteignent en moyenne un âge plus élevé. La régression multiple permet d'évaluer l'effet propre du sexe en ayant pris en compte celui propre à l'âge.

Ce qu'on vient d'affirmer est valable s'il n'y a pas d'interaction entre X_1 et X_2 . Dans notre exemple, une *interaction* impliquerait que l'effet du sexe varie en fonction de l'âge (il serait donc spécifique à l'âge). La Figure 4 illustre ce point: si les logits de CHD en fonction de AGE pour SEXE=hommes et SEXE=femmes sont parallèles (lignes l_1 et l_2), l'effet du sexe ne dépend pas de l'âge: il n'y a pas d'interaction. Si les logits ne sont pas parallèles (lignes l_2 et l_3), l'effet du sexe varie selon l'âge et il y a interaction. (Dans ce cas, l'odds ratio pour comparer les sexes est aussi dépendant de l'âge.)

Pour inclure cette interaction dans le modèle, on utilise une variable explicative supplémentaire définie comme le produit $X_1 \cdot X_2$, donc:

$$K(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2.$$

La présence de l'interaction peut être vérifiée par un test de l'hypothèse $H_0 : \beta_{12} = 0$.

En définitive, la meilleure façon d'interpréter une régression logistique multiple est de calculer et de comparer les valeurs de $\hat{p}(x_1, \dots, x_p)$ pour différents jeux de valeurs (x_1, \dots, x_p) . Par exemple, on pourrait comparer les probabilités de CHD pour les fumeurs-hommes-obèses et pour les non-fumeurs-femmes-obèses.

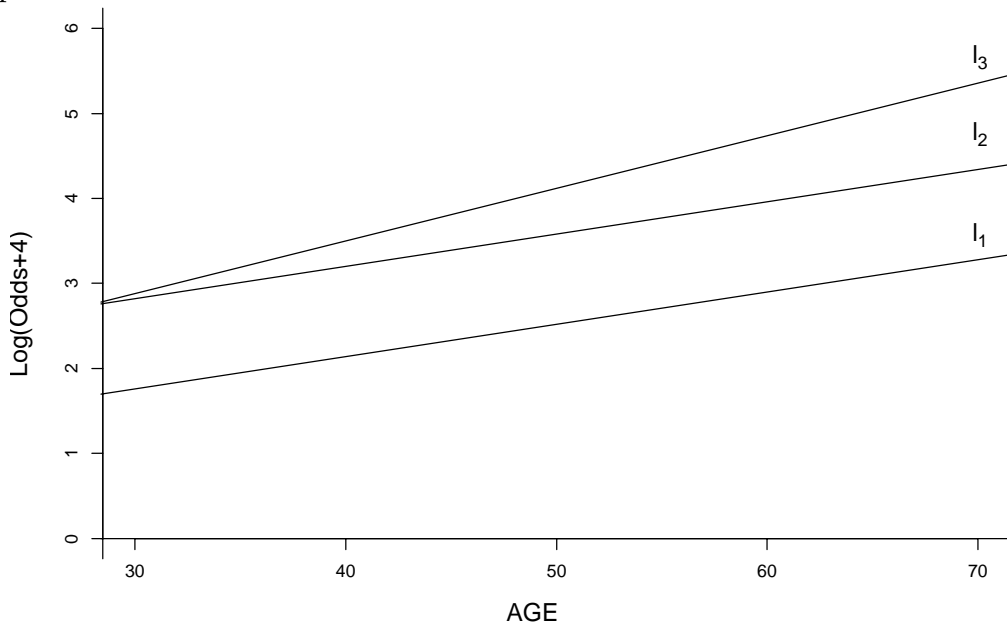


Figure 4. Logit en fonction de AGE pour 3 modèles différents.