RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEURE
ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ MUSTAPHA BEN BOULAID DE BATNA 2
FACULTÉ DES MATHÉMATIQUES ET DE L'INFORMATIQUE
DÉPARTEMENT DE MATHÉMATIQUES
Filière : Statistique et Analyse des Données

Modèles linéaires généralisés
Chapitre 1,2,3
(Cours-TD-TP)

Par

FATEH MERAHI

Première année Master Mathématiques (SAD)

Table des matières

Introduction

Coefficients : 3.

Crédits : 6.

Objectifs de l'enseignement : Les modèles linéaires généralisés permettent d'étudier la liaison entre une variable dépendante ou réponse Y et un ensemble de variables explicatives ou prédicteurs X_1,...., X_p

Contenu de la matière :

Chapitre 1 : Introduction

- o Modèles linéaires pour les données continues
- o Modèles linéaires pour les données discrètes

Chapitre 2 : Régression logistique

- o Rappels, vocabulaire
- o Distribution de BernoulliLien logit
- o Distribution binomiale et modèle logistique

Chapitre 3 : Inférence pour le modèle logistique

- o Maximum de vraisemblance
- o Prédiction et intervalles de confiance
- o Qualité d'ajustement
- o Exemple 1 : comparaison de 2 groupes
- o Exemple 2 : comparaison de plus de 2 groupes
- o Exemple 3 : modèle à une variable
- o Exemple 4 : modèle à deux préditeurs

Chapitre 4 : Diagnostiques de régression pour les données binaires

Chapitre 5 : Variantes des modèles logistiques

- o Autres fonctions de lien
- o Loi multinomiale
- o Modèle logistique conditionnel
- o Modèle logistique hiérarchique
- o Modèles pour une réponse ordinale

Chapitre 6 : Régression de Poisson

- o Distribution de Poisson
- o Modèle log-linéaire
- o Données hétéroscédastiques
- o Inférence
- o Sur-dispersion

Chapitre 7 : Validation, sélection de modèles

- o Performance en prédiction
- o Estimation par validation croisée
- o Méthodes boostrap, "out of bag"
- o Sélection de variables

Mode d'évaluation : Contrôle continu (40%) , examen (60%)

Références bibliographiques

— Hardin, J. and Hilbe, J. (2012). Generalized Linear Models and Extensions, 3rdEdition. College Station, Texas : Stata Press. Un livre avec des exemples et des applications incluant des analyses avec Stata.

— Hosmer, D.W. and Lemeshow, S. (2013). Applied Logistic Regression, 3rd Edition.New York : John Wiley and Sons. Une discussion détaillée sur le modèle logistique avec des applications.

— McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models, 2nd Edition.London : Chapman and Hall. La "bible" des modèles linéaires généralisés. Trèsintéressant, mais plutôt destinée à des étudiants avancés.

— Notes de cours de G. Rodriguez et exemples de codes. R :http ://data.princeton.edu/wws509/

— Notes de cours de L. Rouvière. http ://perso.univrennes2.fr/system/files/users/rouviere_l/poly_logistique_web.pdf.

— Notes de cours de F. Bertrand. http ://www.irma.ustrasbg.fr/~fbertran/enseignement/

Ecole_Doctorale_SVS_Automne_ 2008/ED_RegLog.pdf

— Arthur Charpentier. (2013). Partie 4-modèles linéaires généralisés. Université du Québec à Montréal. http ://freakonometrics.hypotheses.org/.
— Nelder et Wedderburn. (1972). Modèles linéaires généralisés. Présentation.
— Mc Cullagh et Nelder. (1989). Modèles linéaires généralisés. Présentation.

Chapitre 1

Introduction

1.1 Modèles linéaires pour les données continues

• Les modèles linéaires sont utilisés pour étudier comment une variable continue dépend d'un ou plusieurs prédicteurs ou variables explicatives. Les prédicteurs peuvent être quantitatifs ou qualitatifs, ils permettent d'étudier la liaison entre une variable dépendante ou réponse Y et ensemble de variables explicatives ou prédicteurs $X_1, X_2, ..., X_p$.

Exemple 1. *données des efforts des plannings familiaux en Amérique du Sud (Mauldin and Berelson, 1978). Le niveau social et les efforts des plannings familiaux sont mesurés par une combinaison d'indices. Plus l'indice est élevé plus le niveau social (resp. l'effort) est élevé.*

	$Niv.social$	$Effort$	$Déclin\ du\ tx\ nat.$
Bolivia	46	0	1
Brazil	74	0	10
Chile	89	16	29
Colombia	77	16	25
CostaRica	84	21	29
Cuba	89	15	40
Dominican Rep	68	14	21
Ecuador	70	6	0
El Salvador	60	13	13
.	.	.	.

Dans cet exemple on cherche à comprendre comment le niveau social et les efforts de planification influent sur le taux de natalité.

Notons Y le taux de natalité et X_1, X_2 respectivement le niveau social et l'effort de planification.

On dispose de $n = 20$ observations $\{(x_{i1}, x_{i2}, y_i)\}, i = 1, ..., 20$.

On observe que le déclin du taux de natalité augmente avec le niveau social et avec l'effort de planification.

• Si on suppose que y_i est une variable gaussienne de moyenne μ_i et et de variance σ^2, on pourra écrire le modèle

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} \\ &= \beta_0 + \mathbf{x}_i^T \beta, \end{aligned}$$

pour $i = 1, ..., n$.

• La densité de probabilité de Y_i est donnée par

$$\varphi(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y-\mu_i)}{\sigma^2}\right).$$

• Une manière plus standard d'écrire ce modèle linéaire est

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon = \beta_0 + \mathbf{X}^T \beta + \epsilon,$$

avec ϵ une variable aléatoire de loi de Gauss centrée et de variance σ^2.

• $\mathbf{X}$ est **la matrice de design** et $\mathbf{X}^T\beta$ est **le prédicteur linéaire.**

• Les variables aléatoires $y_1, y_2, ..., y_n$ étant supposées indépendantes.

• Certaines des variables X_j peuvent se déduire de variables initiales utilisées dans le modèle, par exemple :

◦ $X_3 = X_1 X_2$ de façon àétudier l'interaction entre X_1 et X_2.

◦ $X_4 = X_1^2$ de façon à prendre en compte un effet non linéaire de la variable X_1.

• La loi de probabilité de **la composante aléatoire** Y appartient à la famille exponentielle dont la fonction de densité s'écrit

$$\varphi(y|\theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right),$$

où $a(\,\cdot\,), b(\,\cdot\,)$ et $c(\,\cdot\,)$ sont des fonctions, et où θ est appelé **paramètre naturel (ou canonique).** Le paramètre θ est *le paramètre d'intérêt* tandi que ϕ est considéré comme un **paramètres de nuisance** (et supposé connu, dans un premier temps).

• Chacune des lois de la famille exponentielle possède une fonction de lien spécifique, dite **fonction de lien canonique**, permettant de relier l'espérance de Y au paramètre naturel (ou canonique) θ. Le lien canonique est tel $g(\mu) = \theta$, par exemple :

Exemple :

• La loi **Gaussienne** de moyenne μ et de variance $\sigma^2, \mathcal{N}(\mu, \sigma^2)$ appartient à cette famille, avec $\theta = \mu$, $\phi = \sigma^2, a(\phi) = \phi, b(\theta) = \frac{\theta^2}{2}$ et

$$c(y, \phi) = -\frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log\left(2\pi\sigma^2\right)\right), \ y \in \mathbb{R}.$$

• La loi de **Bernoulli** de moyenne π , $\mathcal{B}(\pi)$ correspond au cas $\theta = \log\left(\frac{\pi}{1-\pi}\right)$, $\phi = 1, a(\phi) = 1, b(\theta) = \log(1 + \exp(\theta))$ et $c(y, \phi) = 0$.

• La loi de **Binomiale** de moyenne $n\pi$, $\mathcal{B}(n, \pi)$ correspond au cas $\theta = \log\left(\frac{\pi}{1-\pi}\right)$, $\phi = 1, a(\phi) = 1, b(\theta) = n\log(1 + \exp(\theta))$ et $c(y, \phi) = \log\begin{pmatrix} n \\ y \end{pmatrix} = \log\left(\frac{n!}{y!(n-y)!}\right)$.

• La loi Poisson de moyenne λ , $\mathcal{P}(\lambda)$ appartient à cette famille, avec $\theta = \log\lambda$, $\phi = 1, a(\phi) = 1, b(\theta) = \exp\theta = \lambda$ et $c(y, \phi) = -\log y!$.

1.1.1 Inférence

• L'estimation des paramètres du modèle β_0, β, σ se fait par minimisation d'un critère des moindres carrées (LS)

$$(\beta_0^*, \beta^*) = \arg\min_{(\beta_0, \beta)} \sum_{i=1}^{n} \left(y_i - \beta_0 - \mathbf{x}_i^T\beta\right)^2,$$

et

$$\sigma^* = \sqrt{Var\left\{Y - \beta_0 - \mathbf{X}^T\beta\right\}},$$

ou par maximum de vraisemblance (ML).

• Dans le cas du modèle linéaire, ces deux estimateurs sont identiques et

$$\widehat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}\mathbf{y}.$$

• On dispose ensuite de tests (test de Fisher, test de Wald, etc) pour valider et interpréter le modèle.

• Le test t permet de tester l'hypothèse $H_0 : \beta_j = 0$ pour chaque variable j.

• Le test de Fisher permet de tester plusieurs paramètres simultanément. C'est notamment important quand on travaille avec des variables catégorielles recodées en variables binaires.

1.1.2 Cas des prédicteurs qualitatifs : analyse de la variance

• Dans certains cas, les prédicteurs sont tous qualitatifs.

• Les notations sont un peu différentes de celles du modèle de régression. On note K le nombre de niveaux dans le facteur et n_k le nombre d'observations dans le niveau k et y_{ki} est la réponse de l'individu i au niveau k.

• On écrit alors le modèle de l'analyse de la variance à un facteur :

$$y_{ki} \rightsquigarrow \mathcal{N}\left(\mu_{ki}, \sigma^2\right),$$

avec

$$\mu_{ki} = \mu + \alpha_k$$

où μ est un effet moyen (commun à tous les niveaux du facteur) et α_k représente l'effet spécifique du niveau k.

1.2 La fonction de lien

• **La fonction de lien** $g(.)$ d'un modèle linéaire généraliséest le lien entre **la composante aléatoire** Y et **la composante déterministe** $(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)$.

• Le lien spécifie comment l'espérance mathématique de Y notée μ est liée au prédicteur linéaire construit à partir des variables explicatives.

• On peut modéliser l'espérance μ directement (régression linéaire usuelle) ou modéliser une fonction monotone et dérivable $g(\mu)$ de l'espérance μ :

$$g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p.$$

• A chaque loi (ou type de variable Y), on associe une fonction de lien canonique

Modèle	*Logistique*	$Log - linéaire$	*Linéaire*
Loi de $Y\|X = x$	*Bernoulli*	*Poisson*	*Gauss*
$E(Y\|X = x)$	$\log it\left(E(Y\|X = x)\right) = \mathbf{X}^T\beta$	$\log\left(E(Y\|X = x)\right) = \mathbf{X}^T\beta$	$E(Y\|X = x)$

1.3 Modèles linéaires pour les données discrètes

• Quand la variable à prédire est discrète, elle ne suit plus une *loi de Gauss* mais une **loi de Bernoulli**, une **loi binomiale** (ou multinomiale), une **loi de Poisson**, etc.

• De façon similaire au modèle linéaire, on va écrire que le paramètre de localisation de la loi varie avec les prédicteurs.

• Exemple *des variables dichotomiques* : **loi de Bernoulli** $y_i \in \{0,1\}$ avec $y_i \rightsquigarrow \mathcal{B}(\pi_i)$ avec

$$g(\pi_i) = \beta_0 + \mathbf{x}_i^T \beta$$

où

$$\pi_i = g^{-1}\left(\beta_0 + \mathbf{x}_i^T \beta\right).$$

La fonction g est **une fonction de lien** définie de $[0,1]$ sur $\mathbb{R}$. En effet, $\beta_0 + \mathbf{x}_i^T \beta \in \mathbb{R}$ alors que $0 < \pi_i < 1$ est une probabilité.

• Exemple *des variables* de *comptage* : loi de Poisson $y_i \in \{0,1,2,...\}$ avec $y_i \rightsquigarrow \mathcal{P}(\lambda_i)$ avec

$$g(\lambda_i) = \beta_0 + \mathbf{x}_i^T \beta.$$

En effet, $\beta_0 + \mathbf{x}_i^T \beta \in \mathbb{R}$ alors que $\lambda_i \in \mathbb{R}^+$.

• Pour choisir un modèle linéaire généralisé (**GLM**) il faut :

◦ choisir la loi de $Y|X = x$ dans la famille exponentielle des **GLM**.

◦ choisir une fonction de lien inversible g.

• Pour utiliser un modèle **GLM** il faudra donc estimer les paramètres $\mathbf{X}^T\beta$. Une fois cette estimation réalisée, on disposera d'une estimation de $\eta(x)$ ainsi que de

$$E\{Y|X = x\} = g^{-1}\left(\mathbf{X}^T\beta\right).$$

1.3.1 Alternative : introduction d'une variable latente

• Une alternative à l'approche précédente consiste à introduire une variable latente qui suit un modèle linéaire.

• Exemple de la loi de Bernoulli. On définit alors

$$y_i^* = \beta_0 + \mathbf{x}_i^T \beta + u_i$$

avec u_i une variable à densité symétrique et de fonction de répartition F. On n'observe pas directement y_i^* mais on observe y_i et

$$\begin{aligned} y_i &= 1 \text{ si } y_i^* > 0 \\ y_i &= 0 \text{ si } y_i^* \leq 0 \end{aligned}$$

• Les deux formulations sont équivalentes. En effet, ce second modèle permet d'écrire la probabilité

$$\pi_i = P(y_i = 1) = P(y_i^* > 0) = P\left(u_i > -\beta_0 - \mathbf{x}_i^T\beta\right) = 1 - F\left(-\beta_0 - \mathbf{x}_i^T\beta\right).$$

F définit de façon unique la fonction de lien.

1.3.2 Données censurées (ou tronquées)

• Remarque : dans le cas de données censurées on utilise naturellement la formulation avec variable latente.

• Exemple de données censurées : Un exemple typique de données censurées à droite sont des données de suivi de cohorte (ex : cancer) avec lesquelles on cherche à estimer une durée de survie. On n'observe pas la date de décès des patients encore vivant à la fin de l'étude.

1.3.3 Inférence

• Dans les modèles linéaires généralisés, on n'a pas accès aux moindres carrés (y_i^* n'est pas observée) et l'estimation des paramètres du modèle se fait par maximum de vraisemblance.

• Comme dans le cas du modèle linéaire, on dispose d'outils statistiques pour valider et interpréter un modèle ou pour comparer des modèles entre eux.

Travaux Pratiques

Exemple 1 :

Pour illustrer, on utilise le jeu de données « housingprices » (issu du package DAAG), composé de quinze observations et trois variables :

• sale.price = prix de vente de la maison (en milliers de dollars australiens).

• area = surface au sol de la maisonarea = surface au sol de la maison.

• bedrooms = nombre de chambres dans la maison.

```
### Installe + charge le package DAAG
install.packages("DAAG")
library(DAAG)
### Charge les données
data(houseprices)
### Aperçu des premières lignes
head(houseprices)
```

Objectif :

- On peut vouloir expliquer le prix de vente des maisons, en fonction de leur surface en présumant que plus la surface est élevée, et plus le prix de vente sera élevé. Il s'agit là d'une regression dite "simple" car elle ne comporte qu'une seule variables explicative.
- De plus, on peut aussi supposer que le nombre de chambres influence le prix de la maison à la hausse ; il s'agira là d'une regression "multiple" avec deux variables explicatives.
- L'objectif est d'évaluer si chacune des deux variables influence le prix, et, si tel est le cas, de tenter de quantifier cet effet.

Regression linéaire simple :

Dans ce cas, on considère une seule variable explicative. Ici, on souhaite donc estimer les

coefficients du modèle :

$$sale.price = b0 + b1 * area + e$$

La commande à utiliser dans R est : **lm()**

y = b0 + b1*x1

Variable à expliquer : y = prix de vente de la maison

Une variable explicative : x1 = Surface au sol de la maison

pricereg <- lm(sale.price ~area, data=houseprices)

Call :

lm(formula = sale.price ~area, data = houseprices)

Residuals :

Min	$1Q$	$Median$	$3Q$	Max
−92.499	−19.302	2.406	28.019	80.607

Coefficients :

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	70.7504	60.3477	1.172	0.2621
area	0.1878	0.0664	2.828	0.0142 *

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error : 48.18 on 13 degrees of freedom

Multiple R-squared : 0.3809, Adjusted R-squared : 0.3333

F-statistic : 7.997 on 1 and 13 DF, p-value : 0.01425

On a donc l'équation de la droite de régression :

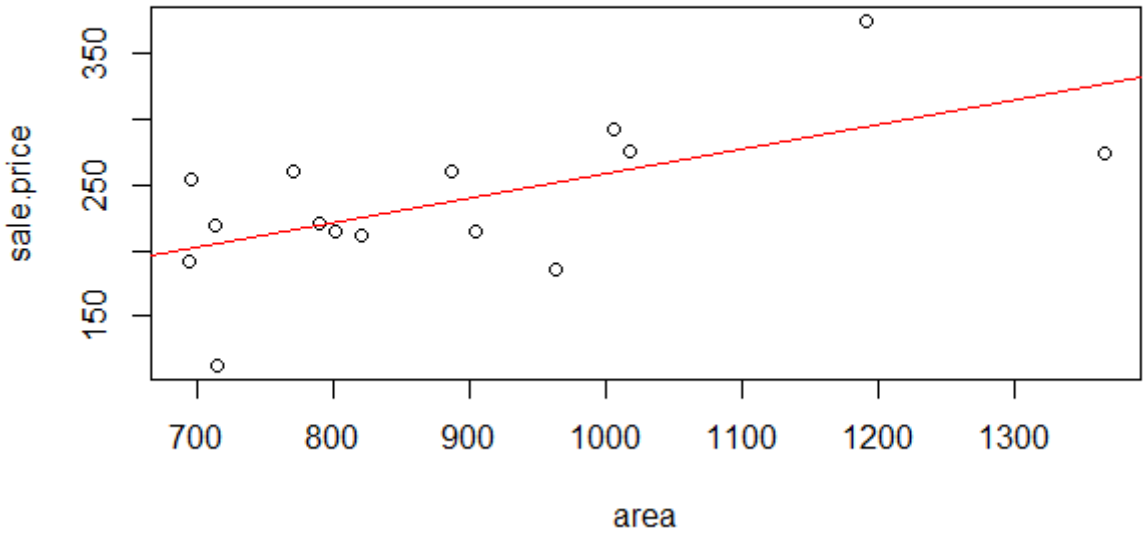
$$sale.price = 70.75 + 0.188 * area + e,$$

que l'on peut tracer avec le nuage de points :

plot : "vraies" valeurs et droite de regression

plot(sale.price ~area, data=houseprices)

abline(pricereg, col = "red")



On peut tester l'effet de la variable "$x_1 = area$" et la commande à utiliser dans R est : **anova()**

> anova(pricereg)

Analysis of Variance Table :

Response : sale.price

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
area	1	18566	18566.0	7.9974	0.01425 *
Residuals	13	30179	2321.5		

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- $F\text{-}value = \frac{Mean\ Sq(area)}{Mean\ Sq(Residuals)}$.

Regression linéaire multiple :

Ici, on considère deux (ou plus) variables explicatives. On souhaite donc estimer les coefficients du modèle :

$$sale.price = b0 + b1 * area + b2 \times bedroom + e$$

Variables explicatives : x1 = Surface au sol de la maison, x2 = nombre de chambres

pricereg2 <-lm(sale.price ~area + bedrooms , data=houseprices)

Call :

lm(formula = sale.price ~area + bedrooms, data = houseprices)

Residuals :

Min	$1Q$	$Median$	$3Q$	Max
-80.897	-4.247	1.539	13.249	42.027

Coefficients :

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-141.76132	67.87204	-2.089	0.05872
area	0.14255	0.04697	3.035	0.01038 *
bedrooms	58.32375	14.75962	3.952	0.00192 **

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error : 33.06 on 12 degrees of freedom

Multiple R-squared : 0.731, Adjusted R-squared : 0.6861

F-statistic : 16.3 on 2 and 12 DF, p-value : 0.0003792

Interprétation :

L'équation de la droite de regression est :

$$sale.price = -141.76 + 0.14 * area + 58.32 * bedrooms + e$$

.

Les coefficients associés aux variables "area" et "bedrooms" sont significatifs (respectivement à 95% et à 99%, cf : sortie précédente), on peut donc les interpréter. Toutes choses égales par ailleurs, une chambre supplémentaire augmente par exemple le prix de la maison de 58 milles dollars environ, et 100 unités de surface (inconnue) supplémentaires vont l'augmenter de 14 mille dollars, toutes choses égales par ailleurs.

Fonctions :

Coefficient : coef() permet de n'extraire que les coefficients estimés.

Extraction des coefficients

>coef(pricereg2)

(Intercept)	area	bedrooms
-141.7613221	0.1425469	58.3237508

Confint : confint() permet d'afficher l'intervalle de confiance à 95% pour les coefficients estimés.

>confint(pricereg2)

	2.5 %	97.5 %
(Intercept)	-289.64179643	6.1191523
area	0.04019939	0.2448945
bedrooms	26.16530118	90.4822003

Fitted : fitted() permet d'extraire les valeurs prédites.

Extraction des valeurs prédites

> fitted(pricereg2)

9	10	11	12	13	14	15	16	17
190.4612	220.5387	205.8563	286.2528	193.5973	228.8064	208.5647	193.3122	236.6465
18	19	20	21	22	23			
217.9728	204.1458	249.0701	259.7611	293.2596	377.9546			

Residuals : resid() permet d'extraire les résidus (Valeur prédite - Valeur réelle).

Extraction des résidus

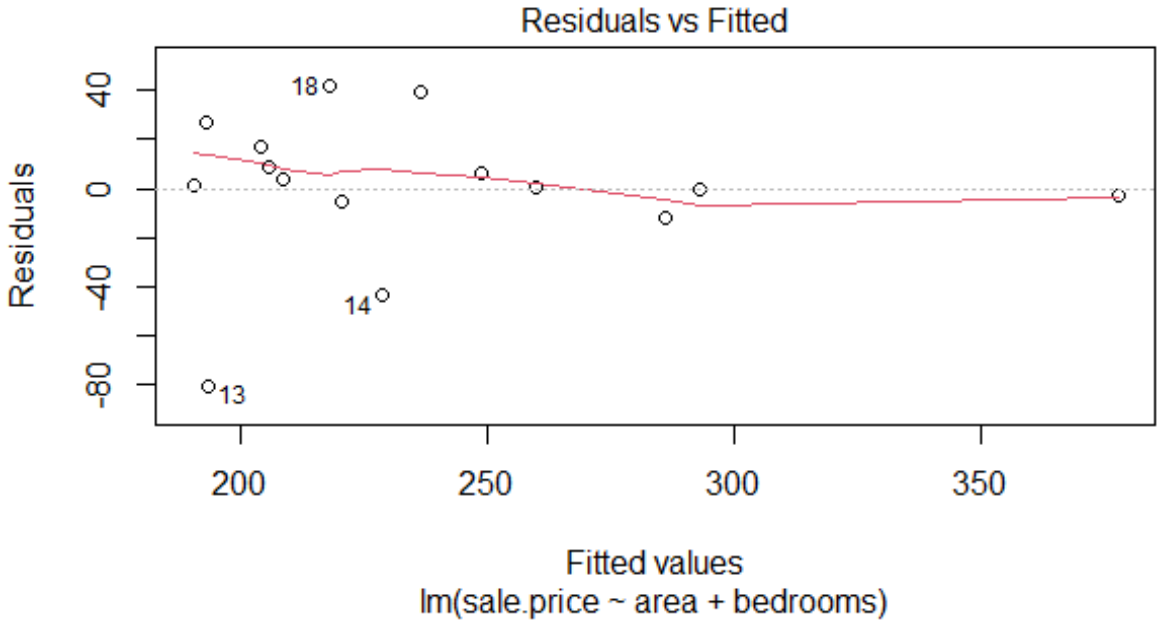
resid(pricereg2)

> resid(pricereg2)

9	10	11	12	13	14	15	16
1.5387504	−5.5386516	9.1436820	−12.2527859	−80.8972821	−43.8063735	3.4352904	26.6878
17	18	19	20	21	22	23	
39.3535454	42.0271931	17.3542452	5.9299058	0.2388861	−0.2596422	−2.9545748	

plot Residuals vs Fitted

> plot(pricereg2)



Predict : La fonction predict() permet de prédire la valeur de y (i.e du prix) pour de nouvelles données (des variables explicatives).

Prédiction du prix de la maison en fonction de sa superficie et de son nombre de chambres

predict(pricereg2, newdata=data.frame(area=800,bedrooms=2), se.fit=TRUE, interval = "prediction", level = 0.99)

Test de significativité globale du modèle :

• $H0$: absence de significativité globale des variables, i.e au moins une variable n'est pas significativement différente de zéro.

• Le test est basé sur la statistique de Fisher (F-test) :

◦ Comme la $p-value$ associée est **inférieure** à $\alpha < 1\%$, on peut dire que l'on rejète fortement H_0.

> anova(pricereg2)

Analysis of Variance Table :

Response : sale.price

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
area	1	18566	18566	16.988	0.001416 **
bedrooms	1	17065	17065.0	15.615	0.001922 **
Residuals	12	13114	1092.9		

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$$F\text{-}statistic\,(area) = \frac{Mean\ Sq\,(area)}{Mean\ Sq\,(Residuals)} = \frac{18566}{1092.9} = 16.98783$$

$F\text{-}statistic\,(area) = 16.98783$ on 1 and 12 DF.

$$F\text{-}statistic\,(bedrooms) = \frac{Mean\ Sq\,(bedrooms)}{Mean\ Sq\,(Residuals)} = \frac{17065.0}{1092.9} = 15.61442$$

$F\text{-}statistic\,(bedrooms) = 15.61442$ on 1 and 12 DF.

$$F\text{-}statistic = \frac{(18566 + 17065)/2}{13114/12} = 16.30113$$

$F\text{-}statistic = 16.30113$ on 2 and 12 DF.

Significativité des coefficients :

• Un coefficient est significatif, autrement dit, qu'il est significativement différent de zéro.

• Pour cela on utilise un test de Student :

◦ On calcule la statistique t pour chaque variable :

$$t = \frac{\widehat{\beta}_j}{\widehat{SE}(\widehat{\beta}_j)},$$

$\widehat{SE}(\widehat{\beta}_j) = \sqrt{Var\left(\widehat{\beta}_j\right)}$: l'estimateur de l'erreur standard (l'écart-type) de $\widehat{\beta}_j$.

◦ t assumée suivre une **loi de Student** que l'on compare ensuite avec **la valeur théorique** issue d'une table de Student (déterminée par le niveau du test, et la nombre d'observations).

◦ On utilise souvent un niveau de $\alpha = 5\%$ (soit un intervalle de confiance de 95%).

◦ Si la *t-value* calculée est **supérieure** (en valeur absolue) à la valeur théorique déterminée T, alors on **rejète** H_0 :Le coefficient est bien significativement différent de zéro.

◦ Soit $p\text{-}value = P\,(T > |t|)$, si

$$p - value < \alpha,$$

alors **le coefficient β_j est significatif.**

Magnitude du coefficient :

• Dans un modèle linéaire basique, on peut interpréter le coefficient β_j associé à la variable X_j comme l'effet de X_j sur Y (variable à expliquer).

Qualité du modèle :

• On peut regarder la qualité de la régression (au regard des données), mesurée par **le coefficient de détermination** (R-$Squared$ ou R^2),

$$R^2 = 1 - \frac{SSR}{SST},$$

où

◦SSR : représente l**a somme des carrés des résidus.**

◦ SST : représente **la somme des écarts à la moyenne des valeurs observées**.

◦ Plus la valeur de R^2 est proche de 1, et plus l'adéquation entre le modèle et les données observées va être forte.

• Le R^2 **ajusté (Adjusted R-Squared) :**

$$Adjusted\ R - Squared = 1 - \frac{SSR/(n-p)}{SST/(n-1)}$$

Travaux Dirigés

Exercice 1. Dans le modèle de régression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon = \beta_0 + \mathbf{X}^T \beta + \epsilon,$$

(a) Déterminer les éléments de la matrice de design $\mathbf{X}$.

(b) Calculer le prédicteur linéaire $\mathbf{X}^T \beta$.

Exercice 2.

(a) Utiliser la méthode des mondres carrées pour montrer que l'estimateure du vecteur $\widehat{\beta}$ est donné par

$$\widehat{\beta} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}\mathbf{y}$$

(b) Calculer σ^2 la variance de variable aléatoire Gaussienne ϵ.

Exercice 3.

(a) Soit $X \rightsquigarrow \mathcal{B}(p)$ "la loi de Bernoulli". Déterminer la loi de X.

(b) Calculer la moyenne et la variance de X.

Exercice 4.

(a) Soit $X \rightsquigarrow \mathcal{B}(n, p)$ "la loi binomiale (ou multinomiale),". Déterminer la loi de X.

(b) Calculer la moyenne et la variance de X.

Exercice 5.

(a) Soit $X \rightsquigarrow \mathcal{P}(\lambda)$ "la loi de Poisson". Déterminer la loi de X.

(b) Calculer la moyenne et la variance de X.

Exercice 6.

- Pour des raisons de santé publique, on s'intéresse à la concentration d'ozone O_3 dans l'air (en microgrammes par millilitre). En particulier, on cherche à savoir s'il est possible d'expliquer le taux maximal d'ozone de la journée par la température T_{12} à midi. Les données sont :

T_{12}	23.8	16.3	27.2	7.1	25.1	27.5	19.4	19.8	32.2	20.7
O_3	115.4	76.8,	113.8	81.6,	115.4	125	83.6	75.2	136.8	102.8

◦ On cherche à expliquer la variable $Y = O_3$ à partir de la variable $X_1 = T_{12}$.

• Ecrire le modèle de régression linéaire correspondant et déterminer la matrice de design.

• Représentez, à l'aide d'une graphe adapté, la relation entre la variable X_1 et Y.

• Donner une estimation des différents paramètres du modèle.

• Calculer le coefficient de détermination $r^2 = \rho_{2(X_1,Y)}$, avec

$$\rho_{(X_1,Y)} = \frac{cov\,(Y, X_1)}{\sqrt{Var\,\{Y\}}\sqrt{Var\,\{X_1\}}}$$

• Importer les données sous R et étudier l'effet de la variable X_1 sur la variable Y.

Exercice 7. Les données utilisées sont le taux de décès par attaque cardiaque chez les hommes de 55 à 59 ans dans différents pays.

$Y = (124, 49, 181, 4, 22, 152, 75, 54, 43, 41, 17, 22, 16, 10, 63, 170, 125, 15, 221, 171, 97, 254)$

$X_1 = (33, 31, 38, 17, 20, 39, 30, 29, 35, 31, 23, 21, 8, 23, 37, 40, 38, 25, 39, 33, 38, 39)$

$X_2 = (8, 6, 8, 2, 4, 6, 7, 7, 6, 5, 4, 3, 3, 3, 6, 8, 6, 4, 7, 7, 6, 8)$

$X_3 = (81, 55, 80, 24, 78, 52, 52, 45, 50, 69, 66, 45, 24, 43, 38, 72, 41, 38, 52, 52, 66, 89)$

Les variables sont les suivantes :

• Y : $100 * log$(nombre de décès par crise cardiaque pour 100000 hommes de 55 à 59 ans).

• X_1 : nombre de téléphones pour 1000 habitants.

• X_2 : calories grasses en pourcentage du total de calories.

• X_3 : calories protéines animales en pourcentage du total de calories.

a) On cherche tout d'abord à expliquer la variable Y à partir de la variable X_1, X_2 et X_3.

• Ecrire le modèle de régression linéaire correspondant et déterminer la matrice de design.

• Représentez, à l'aide d'une graphe adapté, la relation entre les variables X_1 et Y, X_2 et Y, X_3 et Y.

• Donner une estimation des différents paramètres du modèle.

• Importer les données sous R et étudier l'effet des variables X_1, X_2 et X_3 sur la variable Y.

Solutions

Exercice 1.

(a)

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

(b) Produit matrice vecteur :

$$\mathbf{X}^T\beta = \begin{pmatrix} x_{11} & x_{12} & ... & x_{1p} \\ x_{21} & x_{22} & ... & x_{2p} \\ : & : & ... & : \\ x_{n1} & x_{n2} & ... & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ : \\ \beta_p \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{p} x_{1j}\beta_j \\ \sum_{j=1}^{p} x_{2j}\beta_j \\ : \\ \sum_{j=1}^{p} x_{nj}\beta_j \end{pmatrix}$$

Exercice 2.

(a) Par la méthode des moindres carrées (MC), on minimise sur (β_0, β) la quantité $G(\beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \mathbf{x}_i^T\beta\right)^2$, on obtient

$$\begin{aligned} (\beta_0^*, \beta^*) &= \arg\min_{(\beta_0,\beta)} \sum_{i=1}^{n} \left(y_i - \beta_0 - \mathbf{x}_i^T\beta\right)^2 \\ &= \arg\min_{(\beta_0,\beta)} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_{i1}\beta_1 - x_{i2}\beta_2 - ... - x_{ij}\beta_j - ... - x_{ip}\beta_p\right)^2, \end{aligned}$$

alors

$$\left\{\begin{array}{l} \frac{\partial G(\beta)}{\partial \beta_0} = \sum\limits_{i=1}^{n} 2\,(-1)\,(y_i - \beta_0 - x_{i1}\beta_1 - x_{i2}\beta_2 - ... - x_{ij}\beta_j - ... - x_{ip}\beta_p) = 0 \\ \frac{\partial G(\beta)}{\partial \beta_1} = \sum\limits_{i=1}^{n} 2\,(-x_{i1})\,(y_i - \beta_0 - x_{i1}\beta_1 - x_{i2}\beta_2 - ... - x_{ij}\beta_j - ... - x_{ip}\beta_p) = 0 \\ \vdots \\ \frac{\partial G(\beta)}{\partial \beta_j} = \sum\limits_{i=1}^{n} 2\,(-x_{ij})\,(y_i - \beta_0 - x_{i1}\beta_1 - x_{i2}\beta_2 - ... - x_{ij}\beta_j - ... - x_{ip}\beta_p) = 0 \\ \vdots \\ \frac{\partial G(\beta)}{\partial \beta_p} = \sum\limits_{i=1}^{n} 2\,(-x_{ip})\,(y_i - \beta_0 - x_{i1}\beta_1 - x_{i2}\beta_2 - ... - x_{ij}\beta_j - ... - x_{ip}\beta_p) = 0 \end{array}\right. ,$$

d'où

$$\left\{\begin{array}{l} \sum\limits_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - x_{i2}\beta_2 - ... - x_{ij}\beta_j - ... - x_{ip}\beta_p) = 0 \\ \sum\limits_{i=1}^{n} x_{i1}\,(y_i - \beta_0 - x_{i1}\beta_1 - x_{i2}\beta_2 - ... - x_{ij}\beta_j - ... - x_{ip}\beta_p) = 0 \\ \vdots \\ \sum\limits_{i=1}^{n} x_{ij}\,(y_i - \beta_0 - x_{i1}\beta_1 - x_{i2}\beta_2 - ... - x_{ij}\beta_j - ... - x_{ip}\beta_p) = 0 \\ \vdots \\ \sum\limits_{i=1}^{n} x_{ip}\,(y_i - \beta_0 - x_{i1}\beta_1 - x_{i2}\beta_2 - ... - x_{ij}\beta_j - ... - x_{ip}\beta_p) = 0 \end{array}\right.$$

$$\left\{\begin{array}{r} n\beta_0 + \left(\sum\limits_{i=1}^{n} x_{i1}\right)\beta_1 + \left(\sum\limits_{i=1}^{n} x_{i1}\ x_{i2}\right)\beta_2 + ... + \left(\sum\limits_{i=1}^{n} x_{i1}\ x_{ij}\right)\beta_j + ... + \left(\sum\limits_{i=1}^{n} x_{i1}\ x_{ip}\right)\beta_p = \sum\limits_{i=1}^{n} y_i \\ n\beta_0 + \left(\sum\limits_{i=1}^{n} x_{i1}\ x_{i1}\right)\beta_1 + \left(\sum\limits_{i=1}^{n} x_{i2}\ x_{i1}\right)\beta_2 + ... + \left(\sum\limits_{i=1}^{n} x_{ij}\ x_{i1}\right)\beta_j + ... + \left(\sum\limits_{i=1}^{n} x_{pi}\ x_{i1}\right)\beta_p = \sum\limits_{i=1}^{n} x_{i1}\ y_i \\ \vdots \\ n\beta_0 + \left(\sum\limits_{i=1}^{n} x_{1i}\ x_{ij}\right)\beta_1 + \left(\sum\limits_{i=1}^{n} x_{i2}\ x_{ij}\right)\beta_2 + ... + \left(\sum\limits_{i=1}^{n} x_{ij}\ x_{ij}\right)\beta_j + ... + \left(\sum\limits_{i=1}^{n} x_{ip}\ x_{ij}\right)\beta_p = \sum\limits_{i=1}^{n} x_{ij}\ y_i \\ \vdots \\ n\beta_0 + \left(\sum\limits_{i=1}^{n} x_{i1}\ x_{ip}\right)\beta_1 + \left(\sum\limits_{i=1}^{n} x_{i2}\ x_{ip}\right)\beta_2 + ... + \left(\sum\limits_{i=1}^{n} x_{ij}\ x_{ip}\right)\beta_j + ... + \left(\sum\limits_{i=1}^{n} x_{ip}\ x_{ip}\right)\beta_p = \sum\limits_{i=1}^{n} x_{ip}\ y_i \end{array}\right. ,$$

on peut écrir

$$\begin{pmatrix} n & \left(\sum\limits_{i=1}^{n} x_{i1}\right) & \left(\sum\limits_{i=1}^{n} x_{i1}\ x_{i2}\right) & .. & \left(\sum\limits_{i=1}^{n} x_{i1}\ x_{ij}\right) & .. & \left(\sum\limits_{i=1}^{n} x_{i1}\ x_{ip}\right) \\ n & \left(\sum\limits_{i=1}^{n} x_{i1}\ x_{i1}\right) & \left(\sum\limits_{i=1}^{n} x_{i2}\ x_{i1}\right) & .. & \left(\sum\limits_{i=1}^{n} x_{ij}\ x_{i1}\right) & .. & \left(\sum\limits_{i=1}^{n} x_{ip}\ x_{i1}\right) \\ : & : & : & : & : & : & : \\ n & \left(\sum\limits_{i=1}^{n} x_{i1}\ x_{ij}\right) & \left(\sum\limits_{i=1}^{n} x_{i2}\ x_{ji}\right) & .. & \left(\sum\limits_{i=1}^{n} x_{ij}\ x_{ij}\right) & .. & \left(\sum\limits_{i=1}^{n} x_{ip}\ x_{ij}\right) \\ : & : & : & : & : & : & : \\ n & \left(\sum\limits_{i=1}^{n} x_{i1}\ x_{ip}\right) & \left(\sum\limits_{i=1}^{n} x_{i2}\ x_{ip}\right) & .. & \left(\sum\limits_{i=1}^{n} x_{ij}\ x_{ip}\right) & .. & \left(\sum\limits_{i=1}^{n} x_{ip}\ x_{ip}\right) \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ : \\ \beta_j \\ : \\ \beta_p \end{pmatrix} = \begin{pmatrix} \sum\limits_{i=1}^{n} y_i \\ \sum\limits_{i=1}^{n} x_{i1}\ y_i \\ \sum\limits_{i=1}^{n} x_{i2}\ y_i \\ : \\ \sum\limits_{i=1}^{n} x_{ij}\ y_i \\ : \\ \sum\limits_{i=1}^{n} x_{ip}\ y_i \end{pmatrix},$$

et

$$\begin{pmatrix} 1 & 1 & .. & 1 \\ x_{11} & x_{21} & .. & x_{n1} \\ x_{12} & x_{22} & .. & x_{n2} \\ : & : & : & : \\ x_{1p} & x_{2p} & .. & x_{np} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{12} & ... & x_{1p} \\ 1 & x_{21} & x_{22} & ... & x_{2p} \\ : & : & : & ... & : \\ 1 & x_{n1} & x_{n2} & ... & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ : \\ \beta_j \\ : \\ \beta_p \end{pmatrix} = \begin{pmatrix} 1 & 1 & .. & 1 \\ x_{11} & x_{21} & .. & x_{n1} \\ x_{12} & x_{22} & .. & x_{n2} \\ : & : & : & : \\ x_{1p} & x_{2p} & .. & x_{np} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ : \\ y_i \\ : \\ y_n \end{pmatrix},$$

d'où

$$\left(\mathbf{X}^T\mathbf{X}\right)\beta = \mathbf{X}^T y,$$

et la solution $\widehat{\beta}$ est donnée par

$$\widehat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T y.$$

(b) On estime la variance σ^2,

$$\begin{aligned} \widehat{\sigma}^2 &= Var\{\varepsilon\} \\ &= Var\left\{Y - \widehat{\beta}_0 - \mathbf{X}^T\widehat{\beta}\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{\beta}_0 - \mathbf{x}_i^T\widehat{\beta}\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \beta_0 - x_{1i}\beta_1 - x_{2i}\beta_2 - ... - x_{ji}\beta_j - ... - x_{pi}\beta_p\right)^2. \end{aligned}$$

.

Exercice 3.

(a) Une variable aléatoire X suivant la loi de Bernoulli est appelée variable de Bernoulli et on note $X \rightsquigarrow \mathcal{B}(p)$. La variable aléatoire X suit la loi de Bernoulli de probabilité $p \in \{0,1\}$ si

$$P(X = k) = \begin{cases} p & \text{si} \quad k = 1 \\ 1 - p & \text{si} \quad k = 0 \end{cases}$$

ou, de manière équivalente,

$$P(X = k) = p^k (1-p)^{1-k}, \quad k \in \{0,1\}$$

(b) L'espérance mathématique de X est donnée par

$$E\{X\} = \sum_{k=0}^{1} kP(X = k) = p.$$

La variance de X est donnée par

$$\begin{aligned} Var\{X\} &= E\left\{(X - E\{X\})^2\right\} \\ \sum_{k=0}^{1} (k - E\{X\})^2 P(X = k) &= p(1-p). \end{aligned}$$

Exercice 4.

(a) Si Y_1, Y_2, ... , Y_n sont des variables aléatoires de Bernoulli de paramètre p, indépendantes et identiquement distribuées, alors leur somme $X = \sum_{k=1}^{n} Y_k$ est aussi une variable aléatoire, qui suit la loi binomiale, alors

$$X \rightsquigarrow \mathcal{B}(n, p)$$

si

$$P(X = k) = C_n^k \, p^k (1-p)^{1-k}, \quad k \in \{0, 1, 2, ..., k\}$$

$C_n^k = \frac{n!}{k!(n-k)!}$.

(b) L'espérance mathématique de X est donnée par

$$\begin{aligned} E\{X\} &= E\left\{\sum_{k=1}^{n} Y_k\right\} \\ &= \sum_{k=1}^{n} E\{Y_k\} = \sum_{k=1}^{n} p = np. \end{aligned}$$

La variance de X est donnée par

$$\begin{aligned} Var\{X\} &= Var\left\{\sum_{k=1}^{n} Y_k\right\} \\ &= \sum_{k=1}^{n} Var\{Y_k\} = \sum_{k=1}^{n} p(1-p) = np(1-p) \end{aligned}$$

Exercice 5. .

(a) Une variable aléatoire X suivant la loi de Poisson est appelée variable de Poisson et on note $X \rightsquigarrow \mathcal{P}(\lambda)$. La variable aléatoire X suit la loi de Poisson $\lambda > 0$, si

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, ...$$

(b) L'espérance mathématique de X est donnée par

$$\begin{aligned} E\{X\} &= \sum_{k\in\mathbb{N}} kP(X=k) \\ &= \sum_{k=0}^{\infty} ke^{-\lambda}\frac{\lambda^k}{k!} \\ &= e^{-\lambda}\sum_{k=1}^{\infty} k\frac{\lambda^k}{k!} \\ &= e^{-\lambda}\sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} \\ &= \lambda e^{-\lambda}\sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda}\sum_{m=0}^{\infty}\frac{\lambda^m}{m!} = \lambda e^{-\lambda}e^{\lambda} = \lambda \end{aligned}$$

La variance mathématique de X est donnée par

$$\begin{aligned} Var\{X\} &= E\left\{(X - E\{X\})^2\right\} \\ &= E\left\{X^2\right\} - (E\{X\})^2 = \lambda. \end{aligned}$$

Exercice 6.

T_{12}	23.8	16.3	27.2	7.1	25.1	27.5	19.4	19.8	32.2	20.7
O_3	115.4	76.8,	113.8	81.6,	115.4	125	83.6	75.2	136.8	102.8

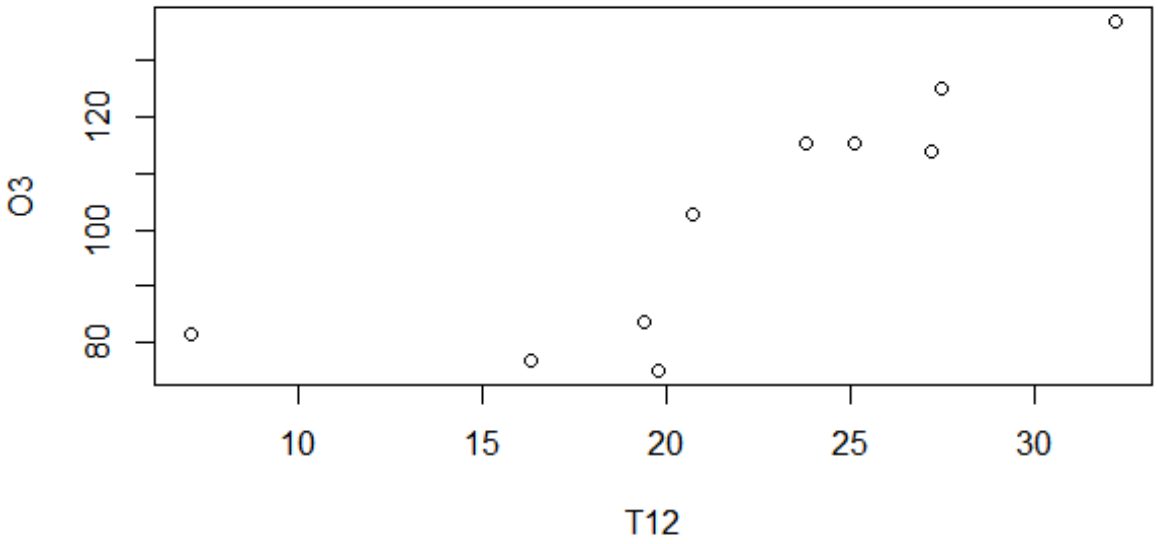
∘ On cherche à expliquer la variable $Y = O_3$ à partir de la variable $X_1 = T_{12}$:

• Le modèle de régression est donné comme suit :

$$Y = \beta_0 + \beta_1 X_1 + \epsilon,$$

avec $\epsilon \rightsquigarrow \mathcal{N}(0,1)$ et la matrice de design $\mathbf{X}$ est donnée par

$$\mathbf{X} = \begin{pmatrix} 1 & 23.8 \\ 1 & 16.3 \\ 1 & 27.2 \\ 1 & 7.1 \\ 1 & 25.1 \\ 1 & 27.5 \\ 1 & 19.4 \\ 1 & 19.8 \\ 1 & 32.2 \\ 1 & 20.7 \end{pmatrix}$$



- La *représentation graphique :*

- *L'estimation des paramètres :*

$$\left(\widehat{\beta}_0, \widehat{\beta}_1\right) = \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{10} \left(y_i - \beta_0 - \beta_1 x_{i1}\right)^2,$$

alors

$$\begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T Y$$
$$= \begin{pmatrix} 45.004407 \\ 2.630561 \end{pmatrix},$$

alors

$$\widehat{Y} = 45.004407 + 2.630561 * X_1 + \epsilon.$$

◦ *Deuxième méthode :*

On a

$$Y = \beta_0 + \beta_1 X_1 + \epsilon,$$

alors

$$\begin{aligned} E\{Y\} &= \beta_0 + \beta_1 E\{X_1\} + E\{\epsilon\} \\ \overline{Y} &= \beta_0 + \beta_1 \overline{X}_1 \quad \text{car } E\{\epsilon\} = 0. \end{aligned}$$

D'autre part

$$\begin{aligned} Cov(Y, X_1) &= Cov(\beta_0, X_1) + \beta_1 Cov(X_1, X_1) + Cov(\epsilon, X_1) \\ &= 0 + \beta_1 Cov(X_1, X_1) + 0 \\ &= \beta_1 Var\{X_1\} \end{aligned}$$

donc

$$\begin{cases} \widehat{\beta}_1 = \frac{Cov(Y,X_1)}{Var\{X_1\}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})(x_{i1}-\overline{x}_1)}{\frac{1}{n}\sum_{i=1}^{n}(x_{i1}-\overline{x}_1)} \\ \widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}_1. \end{cases},$$

avec $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ et $\overline{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_{i1}$.

• Calculer le coefficient de détermination $r^2 = \rho^2_{(X_1,Y)}$, avec

$$\begin{aligned} \rho_{(X_1,Y)} &= \frac{cov(Y, X_1)}{\sqrt{Var\{Y\}}\sqrt{Var\{X_1\}}} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})(x_{i1}-\overline{x}_1)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{i1}-\overline{x}_1)}} = \frac{\sum_{i=1}^{n}(y_i-\overline{y})(x_{i1}-\overline{x}_1)}{\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}\sqrt{\sum_{i=1}^{n}(x_{i1}-\overline{x}_1)}} \\ &= \frac{\sum_{i=1}^{n} x_{1i}(y_i-\overline{y}) - \sum_{i=1}^{n}\overline{x}_1(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}\sqrt{\sum_{i=1}^{n}(x_{i1}-\overline{x}_1)}} = \frac{\sum_{i=1}^{n}(y_i-\overline{y})x_{1i} - \overline{x}_1\sum_{i=1}^{n} y_i + \sum_{i=1}^{n}\overline{x}_1\overline{y}}{\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}\sqrt{\sum_{i=1}^{n}(x_{i1}-\overline{x}_1)}} \\ &= \frac{\sum_{i=1}^{n}(y_i-\overline{y})x_{1i} - \overline{x}_1(n\overline{y}) + n\overline{x}_1\overline{y}}{\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}\sqrt{\sum_{i=1}^{n}(x_{i1}-\overline{x}_1)}} = \frac{\sum_{i=1}^{n}(y_i-\overline{y})x_{1i}}{\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}\sqrt{\sum_{i=1}^{n}(x_{i1}-\overline{x}_1)}} = 0.8390984. \end{aligned}$$

• *L'étude de l'effet de la variable X_1 sur la variable Y sous R*:

```
T12 <- c(23.8,16.3,27.2,7.1,25.1,27.5,19.4,19.8,32.2,20.7)
O3 <- c(115.4,76.8,113.8,81.6,115.4,125,83.6,75.2,136.8,102.8)
A <- lm(O3 ~T12)
summary(A)
```

```
model.matrix(A) # matrice de design
coef(A)
confint(A)
fitted(A) # or : z <- predict(A,type="response")
resid(A)
anova(A)
# Représentation graphique :
plot(T12,O3)
# coefficient de détermination :
r <- cov(T12,O3)/sqrt(var(O3)*var(T12))
print(r)
beta1 <- cov(T12,O3)/var(T12)
print(beta1)
beta0 <- mean(O3)-beta2*mean(T12)
print(beta0)
> summary(A)
Call :
lm(formula = O3 ~T12)
Residuals :
```

Min	$1Q$	$Median$	$3Q$	Max
−21.890	−9.001	3.856	7.514	17.919

Coefficients :

	$Estimate$	$Std.Error$	$t-value$	$Pr(>\|t\|)$
$(Intercept)$	45.0044	13.8050	3.260	0.0115∗
$T12$	2.6306	0.6029	4.363	0.0024 ∗ ∗

—

```
Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error : 12.67 on 8 degrees of freedom
Multiple R-squared : 0.7041,      Adjusted R-squared : 0.6671
F-statistic : 19.03 on 1 and 8 DF, p-value : 0.002403
> model.matrix(A) # matrice de design
```

$(Intercept)$	$T12$
1	23.8
1	16.3
1	27.2
1	7.1
1	25.1
1	27.5
1	19.4
1	19.8
1	32.2
1	20.7

attr(,"assign")

[1] 0 1

> coef(A)

$(Intercept)$	$T12$
45.004407	2.630561

> confint(A)

	2.5%	97.5%
$(Intercept)$	13.169971	76.83884
$T12$	1.240182	4.02094

> fitted(A)

1	2	3	4	5	6	7	8	9	10
107.611	87.882	116.555	63.681	111.031	117.344	96.037	97.089	129.708	99.457

> resid(A)

1	2	3	4	5	6	7	8	9	10
7.788	−11.082	−2.755	17.918	4.368	7.655	−12.437	−21.889	7.091	3.342

> anova(A)

Analysis of Variance Table

Response : O3

	Df	$Sum\ Sq$	$Mean\ Sq$	$F\ value$	$Pr(>F)$
$T12$	1	3057.8	3057.81	19.035	0.002403 ∗ ∗
Residuals	8	1285.1	160.64		

—

Signif. codes : 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```
# coefficient de détermination :
> r <- cov(T12,O3)/sqrt(var(O3)*var(T12))
> print(r)
[1] 0.8390984
> beta1 <- cov(T12,O3)/var(T12)
> print(beta1)
[1] 2.630561
> beta0 <- mean(O3)-beta2*mean(T12)
> print(beta0)
[1] 45.00441
```

Chapitre 2

Régression logistique : Estimation et inférence statistique.

La régression logistique ou **modèle logit** est un modèle de régression binomiale. Comme pour tous les modèles de régression binomiale, il s'agit de modéliser au mieux un modèle mathématique simple à des observations réelles nombreuses. En d'autres termes d'associer à un vecteur de variables aléatoires $(x_1, ..., x_p)$ une variable aléatoire binomiale notée y. La régression logistique constitue un cas particulier de modèle linéaire généralisé. Elle est largement utilisée en apprentissage automatique.

La régression logistique est largement répandue dans de nombreux domaines, comme par exemple :

• En médecine, elle permet par exemple de trouver les facteurs qui caractérisent un groupe de sujets malades par rapport à des sujets sains.

• Dans le domaine bancaire, pour détecter les groupes à risque lors de la souscription d'un crédit.

2.1 Définition du modèle et Notations

Soit Y la variable à prédire (variable expliquée) et $X = (X_1, ..., X_p)$ les variables prédictives (variables explicatives).

Dans le cadre de la régression logistique binaire, la variable Y prend deux modalités possible $\{1, 0\}$. Les variables X_j sont exclusivement continues ou binaires.

• Soit Ω un ensemble de n échantillons, comportant n_1, n_0 observations correspondant à la modalité $1, 0$ respectivement de Y.

• $Y \rightsquigarrow \mathcal{B}(\pi)$, $P(Y=1) = \pi$, $P(Y=0) = 1-\pi$.

• $P(X|Y=1), P(X|Y=0)$ est la distribution conditionnelle des X sachant la valeur prise par Y.

• $P(Y=1|X), P(Y=0|X)$ est la La probabilité a posteriori d'obtenir la modalité $1, 0$ respectivement de Y.

2.2 Le modèle LOGIT

• Soient $y_1, y_2, ..., y_n$, n observations indépendantes de Y telle que $Y_i \rightsquigarrow \mathcal{B}(n_i, \pi_i)$, $n_i \geq 1$.

la fonction de lien $g(\mu) = logit(\pi) = \ln\left(\frac{\pi}{1-\pi}\right)$, alors

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi},$$

i.e ;

$$logit(\pi_i) = X_i^T \beta$$

• Il s'agit bien d'une « régression » car on veut montrer une relation de dépendance entre une variable à expliquer et une série de variables explicatives.

• Il s'agit d'une régression « logistique » car la loi de probabilité est modélisée à partir d'une loi logistique.

En effet, après transformation de l'équation ci-dessus, nous obtenons

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip})},$$

• Pour les prédicteurs continus, on peut regarder la dérivée de cette expression

$$\frac{\partial \pi_i}{\partial x_{ji}} = \beta_j \pi_i (1 - \pi_i),$$

On remarque que l'effet de la variable j dépend de la valeur du prédicteur et de celle de la probabilité.

•

2.3 Odds et odds ratio

• Les coefficients du modèle logistique sont souvent interprétés en terme d'odds ratio.

• L'odds (**chance**) pour un individu x d'obtenir la réponse $Y = 1$ est défini par :

$$odds(x) = \frac{P(Y = 1|X = x)}{1 - P(Y = 1|X = x)}.$$

• L'odds ratio OR (**rapport des odds**) (**rapport des chances**) entre deux individus x et $\widetilde{x}$ est

$$OR(x, \widetilde{x}) = \frac{odds(x)}{odds(\widetilde{x})}.$$

Propriétés :

•
$$OR(x, \widetilde{x}) > 1 \Longleftrightarrow P(Y = 1|X = x) > P(Y = 1|X = \widetilde{x}).$$

•
$$OR(x, \widetilde{x}) = 1 \Longleftrightarrow P(Y = 1|X = x) = P(Y = 1|X = \widetilde{x}).$$

•
$$OR(x, \widetilde{x}) < 1 \Longleftrightarrow P(Y = 1|X = x) < P(Y = 1|X = \widetilde{x}).$$

• si $P(Y = 1|X = x)$ et $P(Y = 1|X = \widetilde{x})$ sont très petits par rapport à 1, on peut faire l'approximation

$$OR(x, \widetilde{x}) = \frac{P(Y = 1|X = x)}{P(Y = 1|X = \widetilde{x})}.$$

• $OR(x, \widetilde{x})$ mesure de l'impact d'une variable : il est facile de vérifier que

$$OR(x, \widetilde{x}) = \exp(\beta_1(x_1 - \widetilde{x}_1) + \beta_2(x_2 - \widetilde{x}_2) + ... + \beta_p(x_p - \widetilde{x}_2)).$$

Si on considère deux observations qui diffèrent seulement par la $j^{ième}$ variable, alors

$$OR(x, \widetilde{x}) = \exp(\beta_j(x_j - \widetilde{x}_j)),$$

mesure l'influence de cette variable.

2.4 Estimation et inférence pour le modèle logistique

2.4.1 Maximum de vraisemblance

• Soit Y_i une variable aléatoire de loi $P(Y_i = y) = f(y, \beta)$.

• Soit $y_1, y_2, ..., y_n$, n réalisations indépendantes de $Y_1, Y_2, ..., Y_n$.

• La **vraisemblance du modèle** $f(y, \beta)$ sachant l'échantillon $y_1, y_2, ..., y_n$ est la probabilité d'observer cet échantillon pour un vecteur β donné, donc on peut écrir

$$\mathcal{L}(\beta) = P_\beta(Y_1 = y_1, ..., Y_n = y_n).$$

Par l'indépendance des variables $Y_1, Y_2, ..., Y_n$, on a

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} P_\beta(Y_i = y_i) = \prod_{i=1}^{n} f(y_i, \beta_i).$$

Comme le log est une fonction croissante qui ne change pas la position des optima locaux d'une fonction, alors en prenant le logarithme, on obtient

$$\log \mathcal{L}(\beta) = \sum_{i=1}^{n} \log P_\beta(Y_i = y_i) = \sum_{i=1}^{n} \log f(y_i, \beta_i)$$

• Le vecteur $\widehat{\beta}$ qui réalise **le maximum de vraisemblance** (MV) annule le gradient (dérivée première) de la fonction de **vraisemblance** et à vérifier que son hessien (dérivée seconde) est défini négatif :

◦ Pour obtenir l'estimateur du maximum de vraisemblance on résout donc **le système d'équations du score** suivant :

$$\begin{cases} \frac{\partial \log \mathcal{L}(\beta)}{\partial \beta_1} = 0 \\ \vdots \\ \frac{\partial \log \mathcal{L}(\beta)}{\partial \beta_p} = 0 \end{cases}$$

• Comme dans *le modèle logistique* **LOGIT,** les équations sont non linéaires, donc ce système n'admet pas de solution analytique et on approche sa solution en utilisant un algorithme itératif.

2.4.2 Propriétés de l'estimateur du maximum de vraisemblance (EMV)

• Pour les modèles linéaires généralisés, on peut montrer que l'estimateur du maximum de vraisemblance existe et qu'il est unique.

• Il a les propriétés suivantes :

◦ **Il est consistent :** il tend vers la vraie valeur du paramètre quand le nombre d'observations n tend vers l'infini.

$$\lim_{n\to\infty}\widehat{\beta}\overset{p.s}{=}\beta.$$

∘ **Il est asymptotiquement efficace** : c'est l'estimateur qui a la plus petit variance possible, sous réserve que le nombre d'observations n soit assez grand.

$$\lim_{n\to\infty}Var\left\{\widehat{\beta}\right\}=0.$$

∘ **Il est asymptotiquement distribué suivant une loi de Gauss :** quand le nombre d'observations n tend vers l'infini, l'estimateur du maximum de vraisemblance tend, en loi, vers une variable gaussienne.

$$\lim_{n\to\infty}\widehat{\beta}\overset{\mathcal{L}}{=}\mathcal{N}\left(E\left\{\widehat{\beta}\right\},Var\left\{\widehat{\beta}\right\}\right),$$

$Var\left\{\widehat{\beta}\right\}$ et la matrice de variance de covariance de $\widehat{\beta}$.

2.4.3 Le modèle logistique

- Soient $(y_1,x_1),(y_2,x_2),...,(y_n,x_n)$, n observations indépendantes du vecteur (Y,X).
- **La log vraisemblance** est donnée par

$$\begin{aligned}\log\mathcal{L}(\beta) &= \sum_{i=1}^{n}\log P_\beta(Y_i=y_i)\\ &= \sum_{i=1}^{n}\log\left[C_{n_i}^{y_i}\ \pi_i^{y_i}(1-\pi_i)^{(n_i-y_i)}\right]\\ &= \sum_{i=1}^{n}\left[\log C_{n_i}^{y_i}+y_i\log\pi_i+(n_i-y_i)\log(1-\pi_i).\right]\end{aligned}$$

$Y_i\rightsquigarrow\mathcal{B}(n_i,\pi_i)$, $n_i\geq 1$, le terme $\log C_{n_i}^{y_i}$ ne dépend des paramètres $\widehat{\beta}$, **La log vraisemblance** $\log\mathcal{L}(\beta)$ a la forme suivante

$$\log\mathcal{L}(\beta)=\sum_{i=1}^{n}\left[y_i\log\pi_i+(n_i-y_i)\log(1-\pi_i)\right].$$

le paramètre π_i dépend des covariables x_i et d'un vecteur de paramètres $\widehat{\beta}$, i.e ;

$$\pi_i=\frac{\exp(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+...+\beta_px_{ip})}{1+\exp(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+...+\beta_px_{ip})}$$

• Pour maximiser l**a log-vraisemblance** $\log \mathcal{L}(\beta)$, on peut calculer les dérivées premières et secondes.

• Et utiliser l'algorithme de **Newton-Raphson** pour approcher **l'estimateur du maximum de vraisemblance** et en pratique, on utilise l'algorithme du "**iteratively re-weighted least square**" (**IRLS**) :

IRLS pour le modèle **LOGIT :**

· Choisir un β^0 initial.

Répéter jusqu'à convergence :

Calculer $\widehat{\eta} = X^T \beta^{k-1}$ avec k le numéro d'itération.

Calculer $\widehat{\mu} = logit^{-1}(\widehat{\eta})$, probabilité prédite de $Y_i = 1$.

Calculer $z_i = \widehat{\eta}_i + (y_i - \widehat{\mu}_i) \frac{d\eta_i}{d\mu_i} = \widehat{\eta}_i + \frac{n_i}{\widehat{\mu}_i(n_i - \widehat{\mu}_i)} (y_i - \widehat{\mu}_i)$.

Régresser z sur les covariables

$$\beta^k \longleftarrow \left(X^T W X\right)^{-1} X^T W z,$$

avec une matrice de poids diagonale de terme

$$w_{ii} = \frac{\widehat{\mu}_i (n_i - \widehat{\mu}_i)}{n_i}.$$

• **L'estimateur obtenu est consistent** : il tend en moyenne vers la vraie valeur du paramètre quand n tend vers l'infini

$$\lim_{n \to \infty} E\left\{\widehat{\beta}\right\} = \beta,$$

et sa variance est donnée par

$$\lim_{n \to \infty} Var\left\{\widehat{\beta}\right\} = \left(X^T W X\right)^{-1}.$$

2.5 Prédiction

• L'estimation de la moyenne de y est $\widehat{\mu}$ avec $g(\widehat{\mu}) = X^T \widehat{\beta}$.

• **Intervalle de confiance :** On peut construire un intervalle de confiance de cet estimateur pour indiquer sa précision. Pour obtenir cet intervalle de confiance, on a besoin de la loi de $\widehat{\mu}$.

◦ Un intervalle de confiance $[\mu_-, \mu_+]$ au risque α ($\alpha = 1\%$ ou $\alpha = 5\%$) est obtenu pour $\widehat{\mu}$ via

$$logit(\mu_-) = X^T \widehat{\beta} - Z_{1-\frac{\alpha}{2}} \sqrt{Var\left\{X^T \widehat{\beta}\right\}},$$

et

$$logit\left(\mu_{+}\right)=X^{T}\widehat{\beta}+Z_{1-\frac{\alpha}{2}}\sqrt{Var\left\{X^{T}\widehat{\beta}\right\}},$$

avec la variance du prédicteur linéaire est $X^{T}\widehat{\beta}$ est

$$Var\left\{X^{T}\widehat{\beta}\right\}=X^{T}\left(X^{T}WX\right)^{-1}X,$$

où $Z_{1-\frac{\alpha}{2}}=\Phi^{-1}\left(1-\frac{\alpha}{2}\right)$ le quantile de la loi de Gauss centrée réduite : $\Phi\left(Z_{1-\frac{\alpha}{2}}\right)=1-\frac{\alpha}{2}$, $\Phi\left(.\right)$ la fonction de répartition de la loi de Gauss centrée réduite.

2.6 Testes de significatifs

2.6.1 Évaluation statistique de la régression

Pour vérifier la significativité globale du modèle, nous pouvons introduire un test analogue à l'évaluation de la régression linéaire multiple. Ces tests reposent sur la distribution asymptotique des estimateurs du maximum de vraisemblance. L'hypothèse nulle s'écrit

$$H_{0}:\beta_{1}=\beta_{2}=...=\beta_{p}=0.$$

que l'on oppose à l'hypothèse alternative

$$H_{1}:un\ des\ coefficients\ au\ moins\ est\ non\ nul.$$

La statistique du rapport de vraisemblance suit une loi du *chi-deux* à p degrés de libertés, sa statistique de test s'écrit

$$\mathcal{X}_{p}^{2}=\sum_{i=1}^{n}\frac{n_{i}\left(y_{i}-\widehat{\mu}_{i}\right)^{2}}{\widehat{\mu}_{i}\left(n_{i}-\widehat{\mu}_{i}\right)}$$

• Si la probabilité critique (*p-value*) est inférieure au niveau de signification α que l'on s'est fixé, on peut considérer que le modèle est globalement significatif.

2.6.2 Évaluation individuelle des coefficients

• Pour savoir quelles sont les variables qui jouent réellement un rôle dans cette relation, alors on cherche à tester le rôle significatif d'une variable. Nous réalisons le test suivant

$$H_{0}:\beta_{j}=0,\quad \text{contre}\quad H_{1}:\beta_{j}\neq 0.$$

La statistique de **WALD** répond à ce test, elle s'écrit

$$W = \frac{\widehat{\beta}_j - \beta_j}{\sqrt{Var\left(\widehat{\beta}_j\right)}},$$

elle suit une loi de chi-deux $dl = 1$ ou encore elle suit une de Gauss centrée et réduite $\mathcal{N}(0,1)$ quand $n \to \infty$.

• **Les intervalles de confiance :** La statistique du test de Wald est aussi utilisée pour calculer des intervalles de confiance de niveau de confiance $1-\alpha$ pour les paramètres β_j :

$$\left[\widehat{\beta}_j - Z_{1-\frac{\alpha}{2}}\sqrt{Var\left(\widehat{\beta}_j\right)}, \widehat{\beta}_j - Z_{1-\frac{\alpha}{2}}\sqrt{Var\left(\widehat{\beta}_j\right)}\right],$$

avec $Z_{1-\frac{\alpha}{2}} = \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$ le quantile de la loi de Gauss centrée réduite : $\Phi\left(Z_{1-\frac{\alpha}{2}}\right) = 1-\frac{\alpha}{2}$, $\Phi(.)$ la fonction de répartition de la loi de Gauss centrée réduite.

2.7 Déviance

• La déviance mesure un écart entre les valeurs observées y_i et $n_i - y_i$ et les valeurs estimées $\widehat{\mu}_i$ et $n_i - \widehat{\mu}_i$:

$$D = 2\sum_{i=1}^{n}\left\{y_i \log\left(\frac{y_i}{\widehat{\mu}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \widehat{\mu}_i}\right)\right\},$$

où y_i est la valeur observée et $\widehat{\mu}_i$ la valeur prédite pour l'observation i.

• Si l'ajustement est parfait le rapport des valeurs observées sur les valeurs prédites est égal à 1 et son log est nul. L'ajustement est donc d'autant meilleur que la déviance est faible.

2.8 Résidus basés sur la déviance

• Les résidus sont alors définis par

$$d_i = \sqrt{2\left[y_i \log\left(\frac{y_i}{\widehat{\mu}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \widehat{\mu}_i}\right)\right]}.$$

• Les observations telles que $d_i > 2$ peuvent indiquer un défaut d'ajustement.

2.9 Résidus de Pearson

• L'approche la plus simple pour obtenir des résidus est de calculer la différence entre les valeurs observées et les valeurs prédites et de diviser par l'écart-type des valeurs observées :

$$p_i = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\mu}_i \left(n_i - \widehat{\mu}_i\right)/n_i}},$$

où la $\widehat{\mu}_i$ sont les valeurs prédites et le dénominateur est donné par

$$var\left\{y_i\right\} = n_i \pi_i \left(1 - \pi_i\right) \simeq n_i \frac{\widehat{\mu}_i}{n_i}\left(1 - \frac{\widehat{\mu}_i}{n_i}\right).$$

• Dans les deux cas, un individu qui a un résidu p_i doit nécessiter une attention particulière.

2.10 Critères basés sur la vraisemblance

2.10.1 Critère d'Akaike (AIC)

• Le critère d'Akaike est défini par

$$AIC = -2logL + 2k,$$

avec k le nombre de paramètres à estimer.

• Si l'on considère un ensemble de modèles candidats, le modèle choisi est celui qui aura la plus faible valeur d'**AIC**.

2.10.2 Critère "Bayes Information criterion (BIC)

• Le critère **BIC** est défini par

$$BIC = -2logL + k\log(n),$$

avec n le nombre d'observations.

Travaux Pratiques

Example 1 :

• Dans cet exemple on utilise les packages suivants :

· **aod**

· **ggplot2**

Assurez-vous que vous pouvez les charger avant d'essayer d'exécuter les exemples.

Si aucun package n'est installé, exécutez :

install.packages("aod"),

install.packages("ggplot2"),

• Un chercheur s'intéresse à la manière dont les variables, telles que le **GRE** ("Graduate Record Exam scores" : scores aux examens d'études supérieures), la moyenne cumulative **GPA** ("grade point average" :moyenne pondérée cumulative) et le prestige de l'établissement de premier cycle, affectent l'admission aux études supérieures. La variable de réponse, admettre/ne pas admettre, est une variable binaire.

> library(aod)

> library(ggplot2)

> mydata <- read.csv("https ://stats.idre.ucla.edu/stat/data/binary.csv")

> ## view the first few rows of the data

> head(mydata)

admit	gre	gpa	rank
0	380	3	3
1	660	3	3
1	800	4	1
1	640	3	4
0	520	2	4
1	760	3	2

Description des données :

admit : une variable de réponse binaire (résultat, dépendante).

Il existe trois variables prédictives : : **gre**, **gpa** et **rank**

Nous traiterons les variables **gre** et **gpa** comme continues.

La variable **rang** prend les valeurs de 1 à 4.

Les établissements de **rang** 1 ont le prestige le plus élevé, tandis que ceux de **rang** 4 ont le plus bas.

• On utilise les commandes :

· **summary :** pour obtenir des descriptions de base pour l'ensemble des données.

· **sapply :** pour appliquer la fonction **sd** à chaque variable de l'ensemble de données.

> summary(mydata)

	admit	gre	gpa	rank
Min.	0.000	220	2.26	1.00
1st Qu.	0.000	520	3.13	2.00
Median.	0.000	580	3.40	2.00
Mean.	0.318	588	3.39	2.48
3rd Qu.	1.000	660	3.67	3.00
Max.	1.000	800	4.00	4.00

> sapply(mydata, sd)

admit gre gpa rank

0.466 115.517 0.381 0.944

admit	gre	gpa	rank
0.466	115.516	0.380	0.944

two-way contingency table of categorical outcome and predictors we want

to make sure there are not 0 cells

xtabs(~admit + rank, data = mydata)

	rank			
admit	1	2	3	4
0	28	97	93	55
1	33	54	28	12

Utilisation du modèle logit :

Le code ci-dessous estime un modèle de régression logistique à l'aide de la fonction **glm** (modèle linéaire généralisé).

Premièrement, nous convertissons **rank** en facteur pour indiquer que **rank** doit être traité comme une variable catégorielle.

> mydata$rank <- factor(mydata$rank)

> mylogit <- glm(admit ~gre + gpa + rank, data = mydata, family = "binomial")

> summary(mylogit)

Call :

glm(formula = admit ~gre + gpa + rank, family = "binomial",

data = mydata)

Deviance Residuals :

Min	1Q	Median	3Q	Max
-1.6268	-0.8662	-0.6388	1.1490	2.0790

Coefficients :

	Estimate	Std. Error	z value	$Pr(>\|z\|)$
(Intercept)	-3.989979	1.139951	-3.500	0.000465 ***
gre	0.002264	0.001094	2.070	0.038465 *
gpa	0.804038	0.331819	2.423	0.015388 *
rank2	-0.675443	0.316490	-2.134	0.032829 *
rank3	-1.340204	0.345306	-3.881	0.000104 ***
rank4	-1.551464	0.417832	-3.713	0.000205 ***

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance : 499.98 on 399 degrees of freedom

Residual deviance : 458.52 on 394 degrees of freedom

AIC : 470.52

Number of Fisher Scoring iterations : 4

• On utilise la fonction : **confint** pour obtenir des intervalles de confiance pour les estimations de coefficients.

> ## CIs using profiled log-likelihood

> confint(mylogit)

> ## Waiting for profiling to be done...

	2.5 %	97.5 %
(Intercept)	-6.271620	-1.79255
gre	0.000138	0.00444
gpa	0.160296	1.46414
rank2	-1.300889	-0.05675
rank3	-2.027671	-0.67037
rank4	-2.400027	-0.75354

• *Test de Wald :*

wald.test : c'est une fonction de la bibliothèque **aod** pour tester un effet global de **rank**.

Sigma : fournit la matrice de covariance de variance des termes d'erreur.

Terms : indique à R quels termes du modèle doivent être testés, dans ce cas, les termes 4, 5 et 6 sont les trois termes pour les niveaux de rank.

> wald.test(b = coef(mylogit), Sigma = vcov(mylogit), Terms = 4 :6)

Wald test :

———-

Chi-squared test :

X2 = 20.9, df = 3, P(> X2) = 0.00011

La statistique de test du chi-deux de 20,9, avec 3 degrés de liberté (df), est associée à une p-value =0,00011 indiquant que l'effet global du rank est statistiquement significatif.

Tester la différence entre le coefficient de rank=2 et le coefficient de rank=3 :

l <- cbind(0, 0, 0, 1, -1, 0)

> wald.test(b = coef(mylogit), Sigma = vcov(mylogit), L = l)

Wald test :

Chi-squared test :

X2 = 5.5, df = 1, P(> X2) = 0.019

La statistique de test du chi-deuxde 5,5 avec 1 degré de liberté est associée à p-value=0,019,

indiquant que la différence entre le coefficient de rank=2 et le coefficient de rank=3 est statistiquement

significatif.

• odds-ratios (OR) :

> ## odds-ratios (OR) only :

> exp(coef(mylogit))

(Intercept)	gre	gpa	rank2	rank3	rank4
0.0185001	1.0022670	2.2345448	0.5089310	0.2617923	0.2119375

> ## odds ratios and 95% CI

> exp(cbind(OR = coef(mylogit), confint(mylogit)))

	OR	2.5 %	97.5 %
(Intercept)	0.0185001	0.001889165	0.1665354
gre	1.0022670	1.000137602	1.0044457
gpa	2.2345448	1.173858216	4.3238349
rank2	0.5089310	0.272289674	0.9448343
rank3	0.2617923	0.131641717	0.5115181
rank4	0.2119375	0.090715546	0.4706961

Maintenant, nous pouvons dire que pour une augmentation d'une unité de gpa, les chances d'être admis à l'université (par rapport à ne pas être admis) augmentent d'un facteur de 2,23.

• *Probabilités prédites :*

Nous commencerons par calculer la probabilité d'admission prédite à chaque valeur de **rank**, en tenant **gre** et **gpa** à leur moyenne. Nous créons et visualisons d'abord le bloc de données.

> newdata1 <- with(mydata, data.frame(gre = mean(gre), gpa = mean(gpa), rank = factor(1 :4)))

> newdata1

gre	gpa	rank
588	3.39	1
588	3.39	2
588	3.39	3
588	3.39	4

· **newdata1$rankP :** indique à R que nous voulons créer une nouvelle variable dans l'ensemble de données (data frame) **newdata1** appelée **rankP**.

· le reste de la commande indique à R que les valeurs de **rankP** doivent être des prédictions faites à l'aide de la fonction **predict()**.

> newdata1$rankP <- predict(mylogit, newdata = newdata1, type = "response")

> newdata1

gre	gpa	rank	rankP
588	3.39	1	0.517
588	3.39	2	0.352
588	3.39	3	0.219
588	3.39	4	0.185

Dans le résultat ci-dessus, nous voyons que la probabilité prédite d'être accepté dans un programme d'études supérieures est de **0,52** pour les étudiants des établissements de premier cycle les plus

prestigieux (**rank=1**), et **0,18** pour les étudiants des établissements les moins bien classes (**rank=4**).

· Nous pouvons faire quelque chose de très similaire pour créer une table de probabilités prédites faisant varier la valeur de **gre** et de **rank**.

· Nous allons les tracer, nous allons donc créer **100** valeurs de **gre** entre **200** et **800**, à chaque valeur de **rank** (c'est-à-dire 1, 2, 3 et 4).

> newdata2 <- with(mydata, data.frame(gre = rep(seq(from = 200, to = 800, length.out = 100), 4), gpa = mean(gpa), rank = factor(rep(1 :4, each = 100))))

· Le code pour générer les probabilités prédites (la première ligne ci-dessous) est le même que précédemment, sauf que nous allons également demander des erreurs standard afin de pouvoir tracer un

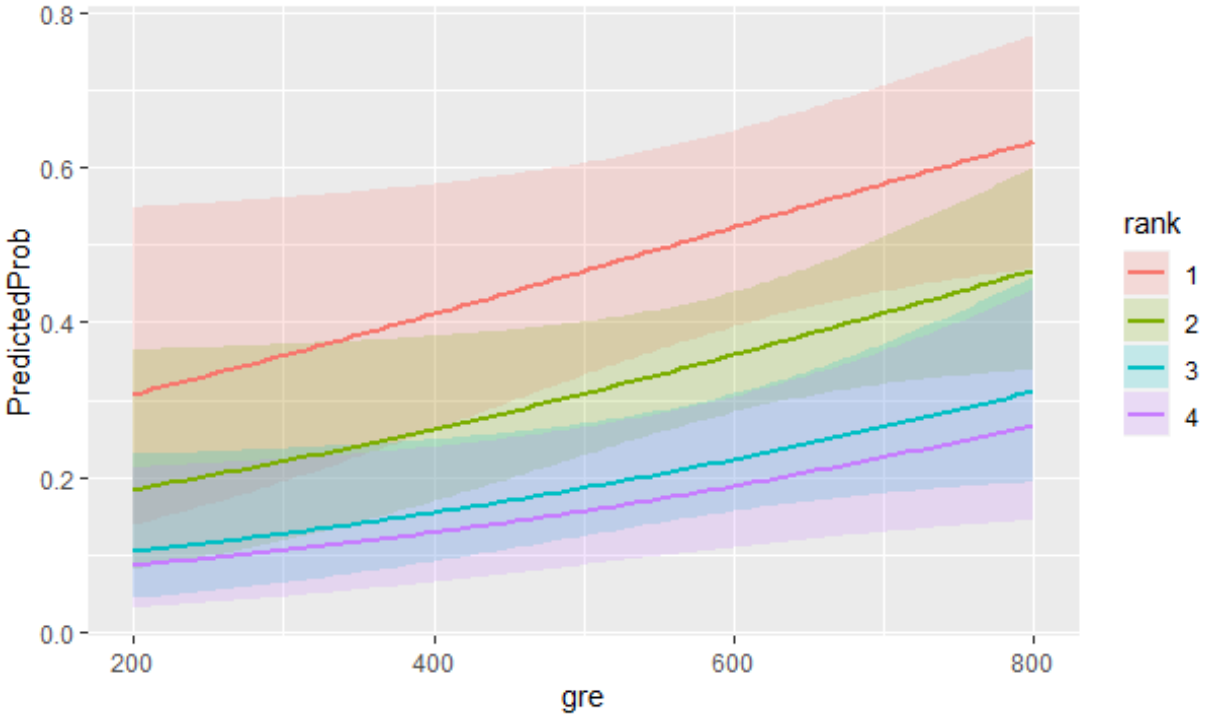
intervalle de confiance. Nous obtenons les estimations sur l'échelle des liens et transformons à la fois les valeurs prédites et les limites de confiance en probabilités :

> newdata3 < - cbind(newdata2, predict(mylogit, newdata = newdata2, type = "link",se = TRUE))

> newdata3 <- within(newdata3, { PredictedProb <- plogis(fit)

LL <- plogis(fit - (1.96 * se.fit))

UL <- plogis(fit + (1.96 * se.fit))})



> ## view first few rows of final dataset

> head(newdata3)

gre	gpa	rank	fit	se.fit	residual.scale	UL	LL	PredictedProb
200.00	3.38	1	-0.81148	0.51477	1	0.54920	0.13938	0.30757
206.06	3.38	1	-0.79776	0.50909	1	0.54985	0.14238	0.31050
212.12	3.38	1	-0.78403	0.50344	1	0.55050	0.14544	0.31344
218.18	3.38	1	-0.77031	0.49782	1	0.55117	0.14854	0.31641
224.24	3.38	1	-0.75659	0.49222	1	0.55185	0.15169	0.31938
230.30	3.38	1	-0.74286	0.48664	1	0.55254	0.15489	0.32237

• *La représentation graphique :*

· Il peut également être utile d'utiliser des graphiques de probabilités prédites pour comprendre et/ou présenter le modèle.

· Nous utiliserons le package **ggplot2** pour la représentation graphique.

· Ci-dessous, nous faisons un graphique avec les probabilités prédites et des intervalles de confiance à 95%.

> ggplot(newdata3, aes(x = gre, y = PredictedProb)) + geom_ribbon(aes(ymin = LL, ymax = UL, fill = rank), alpha = 0.2) + geom_line(aes(colour = rank), size = 1)

• *La statistique de test de la différence de déviance pour les deux modèles :*

· Trouver la différence de déviance pour les deux modèles (c'est-à-dire la statistique de test), nous pouvons utiliser la commande :

> with(mylogit, null.deviance - deviance)

[1] 41.5

· Les degrés de liberté pour la différence entre les deux modèles sont égaux au nombre de variables prédictives dans le mode et peuvent être obtenus en utilisant :

> with(mylogit, df.null - df.residual)

[1] 5

• Enfin, **p-value** peut être obtenue en utilisant :

> with(mylogit, pchisq(null.deviance - deviance, df.null - df.residual, lower.tail = FALSE))

[1] 7.58e-08

· Le **chi-deux** de **41,46** avec **5** degrés de liberté (**dl**) et une **p-value** associée **inférieure** à **0,001** nous indique que notre modèle dans son ensemble s'adapte nettement mieux qu'un modèle vide. C'est ce qu'on appelle parfois un **test de rapport de vraisemblance** (le résidu de déviance est de **-2*log de vraisemblance**).

· Pour voir le **log de vraisemblance** du modèle, nous tapons :

>logLik(mylogit)

'log Lik.' -229.2587 (df=6)

Travaux Dirigés

• Méthode du maximum de vraisemblance MV (maximum likelihood ML) :

◦ Cette méthode permet de calculer, à partir d'un échantillon observé, la (les) meilleure(s) valeur(s) d'un paramètre d'une loi de probabilité.

◦ *Le principe de la méthode MV :* Si un phénomène X a été l'objet de n observations indépendantes $x_1, x_2, ..., x_n$ les unes des autres, sa loi de probabilité $P(X = x)$(dans le cas discret : loi binomiale, loi de Poisson) ou sa densité (en cas de loi continue, comme la loi normale) est une fonction $f(x; \theta_1, ..., \theta_n)$ où $\theta_1, ..., \theta_n$ sont les paramètres de la loi.

· *Dans le cas discret :* on définit la fonction de **MV** est la probabilité de l'échatillon observé en fonction des paramètres $\theta = (\theta_1, ..., \theta_n)$:

$$\begin{aligned} L(x_1, x_2, ..., x_n; \theta) &= P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n; \theta) \\ &= P_X\ (x_1, x_2, ..., x_n; \theta), \end{aligned}$$

avec $X = (X_1, X_2, ..., X_n)$.

· *Dans le cas continu :* on définit la fonction de **MV** est définie par

$$L(x_1, x_2, ..., x_n; \theta) = P_X\ (x_1, x_2, ..., x_n; \theta).$$

· On maximise la fonction $L(x_1, x_2, ..., x_n; \theta)$ sur l'ensemble des paramètres θ pour trouver $\widehat{\theta}$ l'estimateur de MV,

$$\widehat{\theta} = \arg\max_{\theta} L(x_1, x_2, ..., x_n; \theta).$$

Exercice 1. Soit la variable aléatoire $X \rightsquigarrow \mathcal{B}(n, p)$ et $(X_1, X_2, ..., X_n)$ un échatillon $i.i.d$ de même loi que X.

a) Calculer la fonction de **MV** de cette échatillon.

b) Estimer le paramètre p de cette loi.

Exercice 2. Soit la variable aléatoire $X \rightsquigarrow \mathcal{N}(m, \sigma)$ et $(X_1, X_2, ..., X_n)$ un échatillon $i.i.d$ de même loi que X.

a) Calculer la fonction de **MV** de cette échatillon.

b) Estimer les paramètres de cette loi.

Exercice 3.

a) Montrer les propriétés suivantes :

•
$$OR(x, \widetilde{x}) > 1 \Longleftrightarrow P(Y = 1|X = x) > P(Y = 1|X = \widetilde{x}).$$

•
$$OR(x, \widetilde{x}) = 1 \Longleftrightarrow P(Y = 1|X = x) = P(Y = 1|X = \widetilde{x}).$$

•
$$OR(x, \widetilde{x}) < 1 \Longleftrightarrow P(Y = 1|X = x) < P(Y = 1|X = \widetilde{x}).$$

• si $P(Y = 1|X = x)$ et $P(Y = 1|X = \widetilde{x})$ sont très petits par rapport à 1, on peut faire l'approximation

$$OR(x, \widetilde{x}) = \frac{P(Y = 1|X = x)}{P(Y = 1|X = \widetilde{x})}.$$

b) Quel est le meilleurs modèles parmi les trois modèles suivant :

· Modèle 1 : AIC = 51.09.

· Modèle 2 : AIC = 42.87.

· Modèle 3 : AIC = 43.10.

$$odds(x) = \frac{P(Y = 1|X = x)}{1 - P(Y = 1|X = x)}.$$

• L'odds ratio OR (**rapport des odds**) (**rapport des chances**) entre deux individus x et $\widetilde{x}$ est

$$OR(x, \widetilde{x}) = \frac{odds(x)}{odds(\widetilde{x})}.$$

Exercice 4. Données sur maladie coronarienne :

AGE : age.

CHD : diagnostic de maladie coronarienne.

$CHD \setminus Age$	$x = 1$ $(Age \geq 55)$	$x = 0$ $(Age < 55)$	$Total$
$y = 1$ (Yes)	21	22	43
$y = 0$ (No)	6	51	57
$Total$	27	73	100

a) Quelle est la loi de la variable y ="CHD" ?

b) Déterminer les paramètres y.

c) Quelle est l'espérance de y ?

d) Déterminer l'expression de la fonction reliant l'espérace de la variable y ="CHD" et la variable explicative x ="Age".

e) Calculer les "*odds*" de $x = 1$ et de $x = 0$. Déduire "Odds Ratio" $OR\,(x = 1, x = 0)$.

Solutions

Exercice 1.

a) La fonction de **MV** de cette échatillon est donnée par

$$\begin{aligned} L(x_1, x_2, ..., x_n; p) &= P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) \\ &= P(X_1 = x_1)(X_2 = x_2)...(X_n = x_n) \text{ car } X_i (i = 1.., n) \text{ sont indépendantes.} \\ &= \prod_{i=1}^{n} C_n^{x_i} p^{x_i} (1-p)^{1-x_i} \\ &= \left(\prod_{i=1}^{n} C_n^{x_i} \right) p^{\sum_{i=1}^{n} x_i} (1-p)^{n - \sum_{i=1}^{n} x_i}. \end{aligned}$$

b) On a

$$\widehat{p} = \arg\max_{\pi} L(x_1, x_2, ..., x_n; p),$$

on maximise la fonction $L(x_1, x_2, ..., x_n; p)$ sur l'ensembles des paramètres, alors

$$\begin{aligned} \frac{\partial L}{\partial p} &= \left(\prod_{i=1}^{n} C_n^{x_i} \right) \left[\left(\sum_{i=1}^{n} x_i \right) p^{\left(\sum_{i=1}^{n} x_i \right) - 1} (1-p)^{n - \sum_{i=1}^{n} x_i} - \left(n - \sum_{i=1}^{n} x_i \right) p^{\sum_{i=1}^{n} x_i} (1-p)^{\left(n - \sum_{i=1}^{n} x_i \right) - 1} \right] \\ &= \left(\prod_{i=1}^{n} C_n^{x_i} \right) (1-p)^{n - \sum_{i=1}^{n} x_i} p^{\sum_{i=1}^{n} x_i} \left[\left(\sum_{i=1}^{n} x_i \right) p^{-1} - \left(n - \sum_{i=1}^{n} x_i \right) (1-p)^{-1} \right], \end{aligned}$$

$\frac{\partial L}{\partial \pi} = 0$, implique

$$\frac{\sum_{i=1}^{n} x_i}{p} = \frac{n - \sum_{i=1}^{n} x_i}{1-p},$$

alors

$$p \left(n - \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} x_i \right) = \sum_{i=1}^{n} x_i,$$

$$p = \frac{1}{n} \sum_{i=1}^{n} x_i$$

donc l'estimateur de MV du paramètre p est donné par

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}.$$

Exercice 2.

a) La fonction de **MV** de cette échatillon est donnée par

$$\begin{aligned} L(x_1, x_2, ..., x_n; m, \sigma) &= P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) \\ &= f(x_1) f(x_2) ... f(x_n) \text{ car les } X_1, X_2, ..., X_n \text{ sont indépendantes.} \\ &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i - m)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \left(\prod_{i=1}^{n} \exp\left(-\frac{(x_i - m)^2}{2\sigma^2}\right)\right) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2\right). \end{aligned}$$

b) Prenons le logarithme népérien $\mathcal{L}(m, \sigma) = \ln L(x_1, x_2, ..., x_n; m, \sigma)$ du produit, on obtient

$$\mathcal{L}(m, \sigma) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2 - n\ln\sigma\sqrt{2\pi}.$$

Les dérivées partielles par rapport à m et à σ sont respectivement

$$\frac{\partial\mathcal{L}(m, \sigma)}{\partial m} = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m),$$

et

$$\frac{\partial\mathcal{L}(m, \sigma)}{\partial m} = \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - m)^2 - \frac{n}{\sigma},$$

Ces dérivées s'annulent lorsque

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

et

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2,$$

donc l'estimateur de MV du paramètre p est donné par

$$\widehat{m} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

et

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2.$$

Exercice 3.

a)

• Si $P(Y=1|X=x) > P(Y=1|X=\widetilde{x})$, alors

$$\begin{aligned} P(Y=1|X=x) \quad > \quad & P(Y=1|X=\widetilde{x}) \Longleftrightarrow 1-P(Y=1|X=x) < 1-P(Y=1|X=\widetilde{x}) \\ \Longleftrightarrow \quad & \frac{1}{1-P(Y=1|X=x)} > \frac{1}{1-P(Y=1|X=\widetilde{x})} \\ \Longleftrightarrow \quad & \frac{P(Y=1|X=x)}{1-P(Y=1|X=x)} > \frac{P(Y=1|X=\widetilde{x})}{1-P(Y=1|X=\widetilde{x})} \\ \Longleftrightarrow \quad & odds(x) > odds(\widetilde{x}) \\ \Longleftrightarrow \quad & OR(x,\widetilde{x}) = \frac{odds(x)}{odds(\widetilde{x})} > 1. \end{aligned}$$

• Si $P(Y=1|X=x) = P(Y=1|X=\widetilde{x})$, alors

$$\begin{aligned} P(Y=1|X=x) \quad = \quad & P(Y=1|X=\widetilde{x}) \Longleftrightarrow 1-P(Y=1|X=x) = 1-P(Y=1|X=\widetilde{x}) \\ \Longleftrightarrow \quad & \frac{P(Y=1|X=x)}{1-P(Y=1|X=x)} = \frac{P(Y=1|X=\widetilde{x})}{1-P(Y=1|X=\widetilde{x})} \\ \Longleftrightarrow \quad & odds(x) = odds(\widetilde{x}) \\ \Longleftrightarrow \quad & OR(x,\widetilde{x}) = \frac{odds(x)}{odds(\widetilde{x})} = 1 \end{aligned}$$

• Si $P(Y=1|X=x) < P(Y=1|X=\widetilde{x})$, alors

$$\begin{aligned} P(Y=1|X=x) \quad < \quad & P(Y=1|X=\widetilde{x}) \Longleftrightarrow 1-P(Y=1|X=x) < 1-P(Y=1|X=\widetilde{x}) \\ \Longleftrightarrow \quad & \frac{1}{1-P(Y=1|X=x)} < \frac{1}{1-P(Y=1|X=\widetilde{x})} \\ \Longleftrightarrow \quad & \frac{P(Y=1|X=x)}{1-P(Y=1|X=x)} < \frac{P(Y=1|X=\widetilde{x})}{1-P(Y=1|X=\widetilde{x})} \\ \Longleftrightarrow \quad & odds(x) < odds(\widetilde{x}) \\ \Longleftrightarrow \quad & OR(x,\widetilde{x}) = \frac{odds(x)}{odds(\widetilde{x})} < 1. \end{aligned}$$

• si $P(Y=1|X=x)$ et $P(Y=1|X=\widetilde{x})$ sont très petits par rapport à 1, alors

$$1-P(Y=1|X=x) \simeq 1 \quad \text{et} \quad 1-P(Y=1|X=\widetilde{x}) \simeq 1,$$

donc

$$odds(x) = \frac{P(Y=1|X=x)}{1-P(Y=1|X=x)} \simeq P(Y=1|X=x),$$

et

$$odds(\widetilde{x}) = \frac{P(Y=1|X=\widetilde{x})}{1-P(Y=1|X=\widetilde{x})} \simeq P(Y=1|X=\widetilde{x}),$$

ce qui implique

$$OR\left(x,\widetilde{x}\right)=\frac{odds(x)}{odds(\widetilde{x})}\simeq\frac{P\left(Y=1|X=x\right)}{P\left(Y=1|X=\widetilde{x}\right)},$$

donc dans ce cas, on peut faire l'approximation

$$OR\left(x,\widetilde{x}\right)=\frac{P\left(Y=1|X=x\right)}{P\left(Y=1|X=\widetilde{x}\right)}.$$

b) le modèle choisi est celui qui aura la plus faible valeur d'AIC : le modèle choisi est "Modèle 2 : AIC = 42.87".

· Modèle 1 : AIC = 51.09.

· Modèle 2 : AIC = 42.87$\longrightarrow Meilleurs.$

· Modèle 3 : AIC = 43.10.

Exercice 4.

On note $Y=CHD$, et $X_1=Age$.

$CHD \setminus Age$	$x=1$ $(Age\geq 55)$	$x=0$ $(Age<55)$	$Total$
$y=1\ (Yes)$	21	22	43
$y=0\ (No)$	6	51	57
$Total$	27	73	100

,

on peut écrie

$$y_i=\beta_0+\beta_1x_{i1}+\epsilon_i,$$

avec $\epsilon_i\rightsquigarrow\mathcal{N}\left(0,\sigma^2\right)$.

a) La loi de la variable Y ="CHD" est : $Bernouli$, i.e ; $Y\rightsquigarrow\mathcal{B}\left(\pi\right)$

b) $P\left(Y=1|x_1=1\right)=\frac{21}{43}=0.488$, $P\left(Y=1|x_1=0\right)=\frac{22}{43}=0.511$, $P\left(Y=0|x_1=1\right)=\frac{6}{57}=0.10$, $P\left(Y=0|x_1=0\right)=\frac{51}{57}=0.89$.

c) $E\left\{Y\right\}=\pi$.

d) Comme $Y\rightsquigarrow\mathcal{B}\left(\pi\right)$, le modèle de régression est logistique, donc la fonction de lien $g\left(\pi\right)=\log it\left(\pi\right)=\log\left(\frac{\pi}{1-\pi}\right)$, alors

$$\log\left(\frac{\pi_i}{1-\pi_i}\right)=\beta_0+\beta_1x_{i1},$$

ce qui implique

$$\pi_i=\frac{\exp\left(\beta_0+\beta_1x_{i1}\right)}{1+\exp\left(\beta_0+\beta_1x_{i1}\right)}.$$

e) On a

$$\begin{aligned} odds\,(x_1 = 1) &= \frac{P\,(Y = 1|x_1 = 1)}{1 - P\,(Y = 1|x_1 = 1)} \\ &= \frac{\frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}}{1 - \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}} \\ &= \exp\left(\beta_0 + \beta_1\right), \end{aligned}$$

et d'autre part

$$odds\,(x_1 = 1) = \frac{P\,(Y = 1|x_1 = 1)}{1 - P\,(Y = 1|x_1 = 1)} = \frac{\frac{21}{27}}{\frac{6}{27}} = \frac{21}{6} = 3.5.$$

On aussi

$$\begin{aligned} odds\,(x_1 = 0) &= \frac{P\,(Y = 1|x_1 = 0)}{1 - P\,(Y = 1|x_1 = 0)} \\ &= \frac{\frac{\exp(\beta_0)}{1+\exp(\beta_0)}}{1 - \frac{\exp(\beta_0)}{1+\exp(\beta_0)}} \\ &= \exp\left(\beta_0\right), \end{aligned}$$

et

$$odds\,(x_1 = 0) = \frac{P\,(Y = 1|x_1 = 0)}{1 - P\,(Y = 1|x_1 = 0)} = \frac{\frac{22}{73}}{\frac{51}{73}} = \frac{22}{51} = 0.431,$$

on peut déduire

$$OR\,(x_1 = 1, x_1 = 0) = \frac{\exp\left(\beta_0 + \beta_1\right)}{\exp\left(\beta_0\right)} = \exp\left(\beta_1\right).$$

on aussi

$$OR\,(x_1 = 1, x_1 = 0) = \frac{odds\,(x_1 = 1)}{odds\,(x_1 = 0)} = \frac{3.5}{0.431} = 8.12,$$

ce qui implique

$$\exp\left(\beta_1\right) = 8.12,$$

alors

$$\beta_1 = \log\left(8.12\right) = 2.094,$$

et

$$\exp\left(\beta_0\right) = odds\,(x_1 = 0) = 0.431,$$

donc

$$\beta_0 = \log\left(0.431\right) = -0.841.$$

Chapitre 3

Régression de Poisson

CH3]

3.1 Distribution de Poisson

• On dit qu'une variable Y a une distribution de Poisson de paramètre $\lambda > 0$ si elle prend des valeurs entières $y = 0, 1, 2, ...$avec la probabilité

$$P(Y = y) = e^{-\lambda}\frac{\lambda^y}{y!},$$

et on note

$$Y \rightsquigarrow \mathcal{P}(\lambda)$$

• La moyenne et la variance d'une loi de Poisson sont égales à λ

$$E\{Y\} = Var\{Y\} = \lambda.$$

Propriétés :

• Si $Y \rightsquigarrow \mathcal{B}(n, \pi)$ alors $Y \rightsquigarrow \mathcal{P}(\lambda)$ quand $n \longrightarrow \infty$ avec $\lambda = n\pi$.

• Si $Y_j \rightsquigarrow \mathcal{P}(\lambda_j)$ pour $j = 1, ..., m$ et $Y_1, ..., Y_m$ sont indépendantes, alors $Y_1 + Y_2 + ... + Y_m \rightsquigarrow \mathcal{P}(\lambda_1 + \lambda_2 + ... + \lambda_m)$.

3.2 Modèle log-linéaire

• Soit $Y_1, Y_2, ..., Y_n$ un échantillon de variables aléatoires indépendantes de Poisson telles que

$$Y_i \rightsquigarrow \mathcal{P}(\lambda_i) \quad \text{pour } i = 1, ..., n.$$

- On suppose de plus que la moyenne λ_i dépend de covariables $\mathbf{x}_i$.
- On pose le modèle

$$\lambda_i = \mathbf{x}_i^T \beta,$$

la moyenne d'une loi de Poisson est positive mais le terme de droite peut prendre des valeurs négatives, on pose alors

$$\log(\lambda_i) = \mathbf{x}_i^T \beta.$$

- En prenant l'exponentiel, on obtient un modèle multiplicatif pour la moyenne

$$\lambda_i = \exp\left(\mathbf{x}_i^T \beta\right).$$

Quand $\mathbf{x}_j$ augmente de 1 point, la moyenne est multipliée par $\exp(\beta_j)$.

- Un des effets avantageux du *log* est qu'il ramène les données d'évènements rares à une échelle plus linéaire.

3.3 Inférence

- **La vraisemblance** de n observations de Poisson indépendantes s'écrit

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} P(Y_i = y_i) \\ &= \prod_{i=1}^{n} e^{-\lambda_i} \frac{\lambda_i^{y_i}}{y_i!} \\ &= e^{-\sum_{i=1}^{n} \lambda_i} \prod_{i=1}^{n} \frac{\lambda_i^{y_i}}{y_i!}. \end{aligned}$$

- **La log-vraisemblance** de n observations de Poisson indépendantes est donnée par

$$\begin{aligned} \log L(\beta) &= -\sum_{i=1}^{n} \lambda_i + \sum_{i=1}^{n} \log\left(\frac{\lambda_i^{y_i}}{y_i!}\right) \\ &= -\sum_{i=1}^{n} \lambda_i + \sum_{i=1}^{n} \left[y_i \log(\lambda_i) - \log(y_i!)\right] \\ &= \sum_{i=1}^{n} \left(y_i \log(\lambda_i) - \lambda_i\right) - \log(y_i!) \\ &= \sum_{i=1}^{n} \left(y_i \ \left(\mathbf{x}_i^T \beta\right) - \exp\left(\mathbf{x}_i^T \beta\right)\right) - \log(y_i!), \end{aligned}$$

avec

$$\log(\lambda_i) = \mathbf{x}_i^T \beta = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$$

• En dérivant **la log-vraisemblance** en fonction de β on obtient

$$\begin{cases} \frac{\partial \log L(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} \left(y_i - \exp\left(\mathbf{x}_i^T \beta\right)\right), \\ \frac{\partial \log L(\beta)}{\partial \beta_1} = \sum_{i=1}^{n} x_{i1} \left(y_i - \exp\left(\mathbf{x}_i^T \beta\right)\right), \\ \vdots \\ \frac{\partial \log L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij} \left(y_i - \exp\left(\mathbf{x}_i^T \beta\right)\right), \\ \vdots \\ \frac{\partial \log L(\beta)}{\partial \beta_p} = \sum_{i=1}^{n} x_{ip} \left(y_i - \exp\left(\mathbf{x}_i^T \beta\right)\right), \end{cases}$$

en annulant ces dérivées, et comme $\exp\left(\mathbf{x}_i^T \beta\right) = \lambda_i$ on obtient les équations d'estimation

$$\begin{cases} \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \lambda_i \\ \sum_{i=1}^{n} x_{i1} y_i = \sum_{i=1}^{n} x_{i1} \lambda_i \\ \vdots \\ \sum_{i=1}^{n} x_{ij} y_i = \sum_{i=1}^{n} x_{ij} \lambda_i \\ \vdots \\ \sum_{i=1}^{n} x_{ip} y_i = \sum_{i=1}^{n} x_{ip} \lambda_i \end{cases}$$

On peut écrir

$$\begin{pmatrix} 1 & 1 & ... & 1 & ... & 1 \\ x_{11} & x_{21} & ... & x_{i1} & ... & x_{n1} \\ x_{12} & x_{22} & ... & x_{i2} & ... & x_{n2} \\ : & : & : & : & : & : \\ x_{1j} & x_{2j} & ... & x_{ij} & ... & x_{nj} \\ : & : & : & : & : & : \\ x_{1p} & x_{2p} & ... & x_{ip} & ... & x_{np} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ : \\ y_i \\ : \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & 1 & ... & 1 & ... & 1 \\ x_{11} & x_{21} & ... & x_{i1} & ... & x_{n1} \\ x_{12} & x_{22} & ... & x_{i2} & ... & x_{n2} \\ : & : & : & : & : & : \\ x_{1j} & x_{2j} & ... & x_{ij} & ... & x_{nj} \\ : & : & : & : & : & : \\ x_{1p} & x_{2p} & ... & x_{ip} & ... & x_{np} \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ : \\ \lambda_i \\ : \\ \lambda_n \end{pmatrix},$$

et la forme matricielle de ces équations est donnée par

$$\mathbf{X}^T y = \mathbf{X}^T \, \widehat{\lambda}$$

où $\mathbf{X}$ est la matrice de design (incluant une colonne de 1), y est la réponse observée et $\widehat{\lambda}$ est la réponse prédite avec l'estimateur du maximum de vraisemblance (**MV**) .

3.4 Estimation en pratique

En général, l'estimation se fait en utilisant l'algorithme numérique **Iterated Re-weighted Least Square** (**IRLS**) comme pour la régression logistique avec ici

$$z_i = \eta_i + \frac{y_i - \widehat{\lambda}_i}{\widehat{\lambda}_i},$$

et les poids

$$w_{ii} = \widehat{\lambda}_i.$$

3.5 Déviance

- La déviance permet de mesurer l'écart entre le modèle et l'observation.

$$D = 2\sum_{i=1}^{n}\left\{y_i \log\left(\frac{y_i}{\widehat{\lambda}_i}\right) + \left(\widehat{\lambda}_i - y_i\right)\right\}.$$

- L'ajustement est donc d'autant meilleur que D est faible.

3.6 tests de Pearson

- Pour tester la qualité d'ajustement,

H_0 : Le modèle permet de reproduire les observations.

H_1 : Le modèle ne permet pas de reproduire les observations.

- On peut utiliser la statistique de **Pearson**

$$X_p^2 = \sum_{i=1}^{n}\frac{\left(\widehat{\lambda}_i - y_i\right)^2}{\widehat{\lambda}_i},$$

qui suit approximativement une loi du chi-deux à $n - p$ degrés de libertés si n est grand.

3.7 Sur-dispersion

• Une des caractéristiques de la distribution de Poisson est l'égalité de la moyenne et de la variance :

$$E\{Y\} = Var\{Y\} = \lambda.$$

• Cependant, les données réelles présentent parfois de la **sur-dispersion** i.e ; une variance plus importante que la moyenne.

• Il existe plusieurs modèles/solutions permettant de prendre ne compte la **sur-dispersion :**

◦ modèle quasi-Poisson,

◦ modèle de Poisson avec estimation robuste de la variance,

◦ modèle à réponse binomiale négative.

• Supposons que la variance est proportionnelle à la moyenne

$$Var\{Y\} = \phi E\{Y\} = \phi\lambda.$$

◦ Si $\phi > 1$ on a une **sur-dispersion**.

◦ Si $\phi < 1$ un **sous-dispersion** (mais ce second cas est rare en pratique).

• Dans le cas où la variance est proportionnelle à la moyenne $Var\{Y\} = \phi E\{Y\} = \phi\lambda$:

◦ Pour adapter l'algorithme **IRLS**, on remplace les poids w_{ii} par

$$w_{ii} = \frac{\widehat{\lambda}}{\phi},$$

Ce sont les poids du modèle de Poisson divisés par ϕ. La constante ϕ disparait quand on calcule

$$\left(\mathbf{X}^T W \mathbf{X}\right)^{-1} \mathbf{X}^T W z.$$

• La variance de $\widehat{\beta}$ est alors

$$Var\left\{\widehat{\beta}\right\} = \phi\left(\mathbf{X}^T W \mathbf{X}\right)^{-1},$$

avec $W = diag(\lambda_1, ..., \lambda_n)$.

• L'estimation de la constante ϕ s'appuie sur la statistique du **chi-deux** de **Pearson**

$$\mathcal{X}_p^2 = \sum_{i=1}^{n} \frac{(y_i - \lambda_i)^2}{Var\{y_i\}} = \sum_{i=1}^{n} \frac{(y_i - \lambda_i)^2}{\phi\lambda_i},$$

et

$$E\left\{\mathcal{X}_p^2\right\} = n - p.$$

Par la méthode des moments, on a donc

$$\widehat{\phi} = \frac{\mathcal{X}_p^2}{n-p}.$$

Travaux Pratiques

Exemple :

• Le nombre de bourses obtenues par les élèves d'une école secondaire : Les prédicteurs du nombre de bourses obtenues suivant le type de programme dans lequel l'étudiant était inscrit (par exemple, professionnel, général ou académique) et le résultat de son examen final en mathématiques.

• *Description des données* :

Dans cet exemple :

num_awards : la variable de réponse et indique le nombre de boursses obtenues par les élèves d'un lycée au cours d'une année.

math : la variable prédictive continue et représente les notes des élèves à leur examen final de mathématiques.

prog : la variable prédictive catégorique avec trois niveaux indiquant le type de programme dans lequel les étudiants étaient inscrits. Il est codé comme **1 = "General"**, **2 = "Academic"** et **3 = "Vocational"**.

• Commençons par charger les données et examiner quelques paramètres de la statistique descriptive.

```
p <- read.csv("https ://stats.idre.ucla.edu/stat/data/poisson_sim.csv")
p <- within(p, {
prog <- factor(prog, levels=1 :3, labels=c("General", "Academic",
"Vocational"))
id <- factor(id)
})
summary(p)
```

	id		num_awards	prog	math	phat
1	:	1	Min. :0.00	General : 0	Min. :33.00	Min. :0.06131
2	:	1	1st Qu. :0.00	Academic : 0	1st Qu. :45.00	1st Qu. :0.18874
3	:	1	Median :0.00	Vocational : 0	Median :52.00	Median :0.37991
4	:	1	Mean :0.63	NA's :200	Mean :52.65	Mean :0.63000
5	:	1	3rd Qu. :1.00		3rd Qu. :59.00	3rd Qu. :0.84825
6	:	1	Max. :6.00		Max. :75.00	Max. :2.99866

(Other) :194

• La moyenne inconditionnelle et la variance de la variable de réponse **num_awards** ne sont pas extrêmement différentes.

· Le modèle suppose que ces valeurs, conditionnées par les variables prédictives, seront égales (ou du moins approximativement).

• Nous pouvons utiliser la fonction **tapply** pour afficher la moyenne"M" et l'écart-type"SD" par type de programme.

· Le tableau ci-dessous montre le nombre moyen de bourses par type de programme **prog** et semble suggérer que le type de programme **prog** est un bon candidat pour prédire le nombre de bourses (variable de réponse **num_awards**) car la valeur moyenne de **num_awards** semble varier selon le programme **prog**. De plus, les moyennes et les variances au sein de chaque niveau de prog – les moyennes et les variances conditionnelles – sont similaires.

· Un histogramme conditionnel séparé par type de programme **prog** est tracé pour montrer la distribution.

```
with(p, tapply(num_awards, prog, function(x) {
sprintf("M (SD) = %1.2f (%1.2f)", mean(x), sd(x))
}))
```

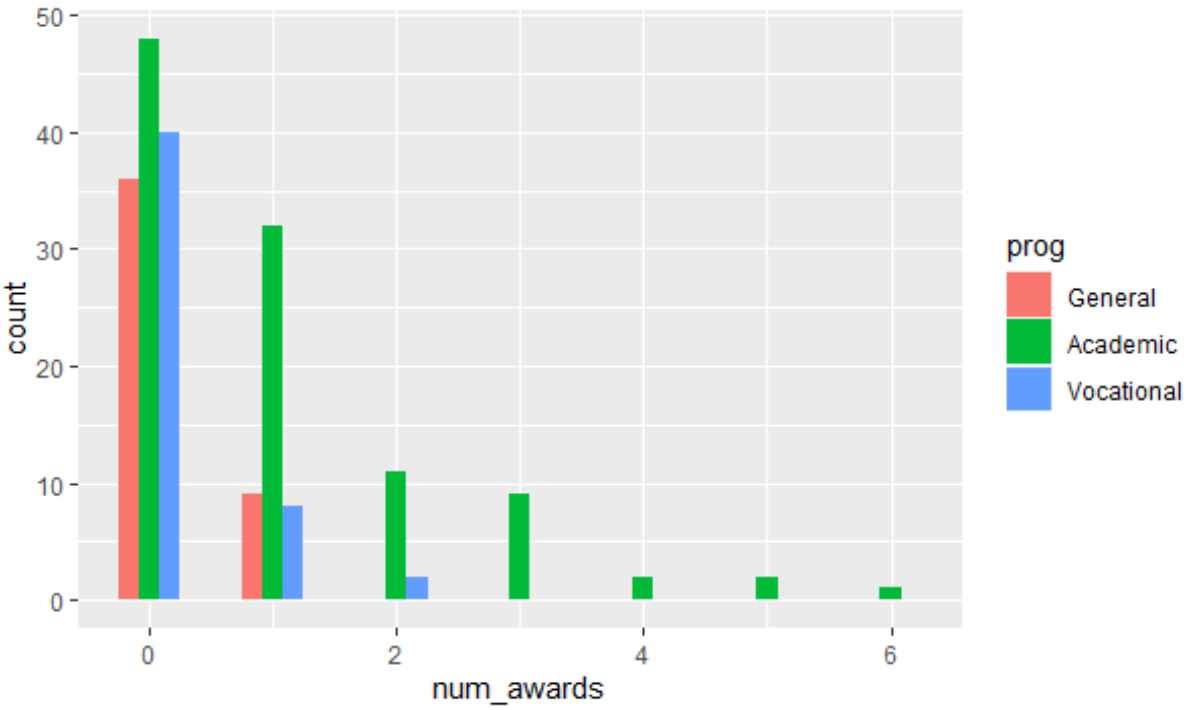
General	*Academic*	*Vocational*
"$M(SD) = 0.20(0.40)$"	"$M(SD) = 1.00(1.28)$"	"$M(SD) = 0.24(0.52)$"

```
ggplot(p, aes(num_awards, fill = prog)) +
geom_histogram(binwidth=.5, position="dodge")
```

• **Régression de Poisson :**

· La régression de Poisson est souvent utilisée pour modéliser les données de comptage.

· L'analyse du modèle de Poisson à l'aide de la fonction **glm**.



· Adapter le modèle et le stocker dans l'objet **m1** et obtenons en même temps les résultats du modèle.

summary(m1 <- glm(num_awards ~prog + math, family="poisson", data=p))

Call :

glm(formula = num_awards ~prog + math, family = "poisson", data = p)

Deviance Residuals :

Min	$1Q$	$Median$	$3Q$	Max
-2.2043	-0.8436	-0.5106	0.2558	2.6796

Coefficients :

	$Estimate$	$Std.Error$	$z-value$	$Pr(>\|z\|)$
$(Intercept)$	-5.2471	0.6585	-7.97	$1.6e-15***$
$progAcademic$	1.0839	0.3583	3.03	$0.0025**$
$progVocational$	0.3698	0.4411	0.84	0.4018
$math$	0.0702	0.0106	6.62	$3.6e-11***$

—

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance : 287.67 on 199 degrees of freedom

Residual deviance : 189.45 on 196 degrees of freedom

AIC : 373.5

Number of Fisher Scoring iterations : 6.

• Nous pouvons également tester l'effet global de prog en comparant la déviance du modèle complet avec la déviance du modèle hors **prog**.

· Le test du chi2 à deux degrés de liberté indique que **prog** est un prédicteur statistiquement significatif de **num_awards**.

update m1 model dropping prog

m2 <- update(m1, . ~. - prog)

test model differences with chi square test

anova(m2, m1, test="Chisq")

Analysis of Deviance Table

Model 1 : num_awards ~math

Model 2 : num_awards ~prog + math

	$Resid.Df$	$Resid.Dev$	Df	$Deviance$	$Pr(>Chi)$
1	198	204			
2	196	189	2	14.6	0.0006852 ∗ ∗

—

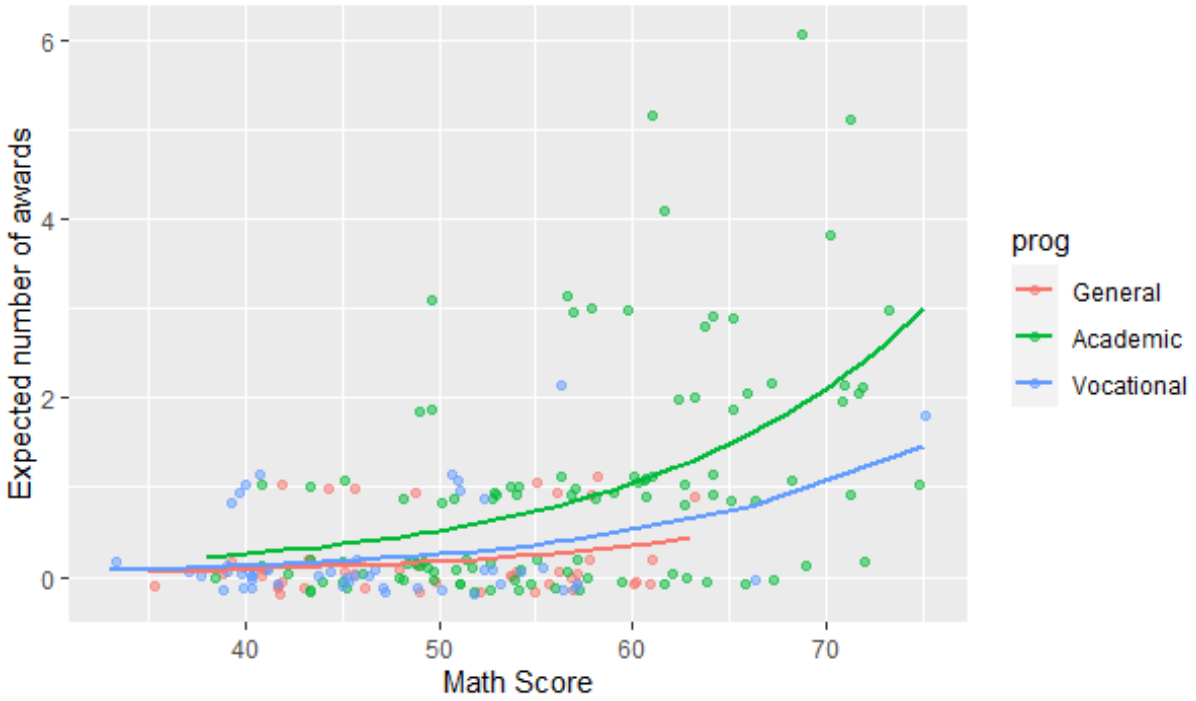
Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

• Nous pouvons également représenter graphiquement le nombre d'événements prévu avec les commandes ci-dessous.

· Le graphique indique que le plus grand nombre de bourses est prévu pour ceux du programme **académique** (**prog = 2**), surtout si l'étudiant a un score élevé en mathématiques.

· Le plus petit nombre de bourses prévues concerne les étudiants du programme **général** (**prog = 1**).

· Le graphique superpose les lignes des valeurs attendues sur les points réels.


Expected number of awards
6
4
2
0
40
50
60
70
Math Score
prog
General
Academic
Vocational

Tables Statistiques usuelles

Table 1

Loi Binomiale

$$P(X = k) = C_n^k p^k (1-p)^{n-k}$$

(k le nombre d'occurrences parmi n)

n = 10

		p									
		0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
k	0	0,5987	0,3487	0,1969	0,1074	0,0563	0,0282	0,0135	0,0060	0,0025	0,0010
	1	0,9139	0,7361	0,5443	0,3758	0,2440	0,1493	0,0860	0,0464	0,0233	0,0107
	2	0,9885	0,9298	0,8202	0,6778	0,5256	0,3828	0,2616	0,1673	0,0996	0,0547
	3	0,9990	0,9872	0,9500	0,8791	0,7759	0,6496	0,5138	0,3823	0,2660	0,1719
	4	0,9999	0,9984	0,9901	0,9672	0,9219	0,8497	0,7515	0,6331	0,5044	0,3770
	5	1,0000	0,9999	0,9986	0,9936	0,9803	0,9527	0,9051	0,8338	0,7384	0,6230
	6	1,0000	1,0000	0,9999	0,9991	0,9965	0,9894	0,9740	0,9452	0,8980	0,8281
	7	1,0000	1,0000	1,0000	0,9999	0,9996	0,9984	0,9952	0,9877	0,9726	0,9453
	8	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9995	0,9983	0,9955	0,9893
	9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9997	0,9990
	10	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

n=20

		p									
		0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
k	0	0,3585	0,1216	0,0388	0,0115	0,0032	0,0008	0,0002	0,0000	0,0000	0,0000
	1	0,7358	0,3917	0,1756	0,0692	0,0243	0,0076	0,0021	0,0005	0,0001	0,0000
	2	0,9245	0,6769	0,4049	0,2061	0,0913	0,0355	0,0121	0,0036	0,0009	0,0002
	3	0,9841	0,8670	0,6477	0,4114	0,2252	0,1071	0,0444	0,0160	0,0049	0,0013
	4	0,9974	0,9568	0,8298	0,6296	0,4148	0,2375	0,1182	0,0510	0,0189	0,0059
	5	0,9997	0,9887	0,9327	0,8042	0,6172	0,4164	0,2454	0,1256	0,0553	0,0207
	6	1,0000	0,9976	0,9781	0,9133	0,7858	0,6080	0,4166	0,2500	0,1299	0,0577
	7	1,0000	0,9996	0,9941	0,9679	0,8982	0,7723	0,6010	0,4159	0,2520	0,1316
	8	1,0000	0,9999	0,9987	0,9900	0,9591	0,8867	0,7624	0,5956	0,4143	0,2517
	9	1,0000	1,0000	0,9998	0,9974	0,9861	0,9520	0,8782	0,7553	0,5914	0,4119
	10	1,0000	1,0000	1,0000	0,9994	0,9961	0,9829	0,9468	0,8725	0,7507	0,5881
	11	1,0000	1,0000	1,0000	0,9999	0,9991	0,9949	0,9804	0,9435	0,8692	0,7483
	12	1,0000	1,0000	1,0000	1,0000	0,9998	0,9987	0,9940	0,9790	0,9420	0,8684
	13	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9985	0,9935	0,9786	0,9423
	14	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9984	0,9936	0,9793
	15	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9985	0,9941
	16	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9987
	17	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998
	18	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Loi Binomiale (suite)

$$P(X = k) = C_n^k p^k (1-p)^{n-k}$$

(k le nombre d'occurrences parmi n)

n = 25

		p									
		0,05	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**
k	**0**	0,2774	0,0718	0,0172	0,0038	0,0008	0,0001	0,0000	0,0000	0,0000	0,0000
	1	0,6424	0,2712	0,0931	0,0274	0,0070	0,0016	0,0003	0,0001	0,0000	0,0000
	2	0,8729	0,5371	0,2537	0,0982	0,0321	0,0090	0,0021	0,0004	0,0001	0,0000
	3	0,9659	0,7636	0,4711	0,2340	0,0962	0,0332	0,0097	0,0024	0,0005	0,0001
	4	0,9928	0,9020	0,6821	0,4207	0,2137	0,0905	0,0320	0,0095	0,0023	0,0005
	5	0,9988	0,9666	0,8385	0,6167	0,3783	0,1935	0,0826	0,0294	0,0086	0,0020
	6	0,9998	0,9905	0,9305	0,7800	0,5611	0,3407	0,1734	0,0736	0,0258	0,0073
	7	1,0000	0,9977	0,9745	0,8909	0,7265	0,5118	0,3061	0,1536	0,0639	0,0216
	8	1,0000	0,9995	0,9920	0,9532	0,8506	0,6769	0,4668	0,2735	0,1340	0,0539
	9	1,0000	0,9999	0,9979	0,9827	0,9287	0,8106	0,6303	0,4246	0,2424	0,1148
	10	1,0000	1,0000	0,9995	0,9944	0,9703	0,9022	0,7712	0,5858	0,3843	0,2122
	11	1,0000	1,0000	0,9999	0,9985	0,9893	0,9558	0,8746	0,7323	0,5426	0,3450
	12	1,0000	1,0000	1,0000	0,9996	0,9966	0,9825	0,9396	0,8462	0,6937	0,5000
	13	1,0000	1,0000	1,0000	0,9999	0,9991	0,9940	0,9745	0,9222	0,8173	0,6550
	14	1,0000	1,0000	1,0000	1,0000	0,9998	0,9982	0,9907	0,9656	0,9040	0,7878
	15	1,0000	1,0000	1,0000	1,0000	1,0000	0,9995	0,9971	0,9868	0,9560	0,8852
	16	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9992	0,9957	0,9826	0,9461
	17	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998	0,9988	0,9942	0,9784
	18	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9984	0,9927
	19	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9996	0,9980
	20	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9995
	21	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999
	22	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Loi Binomiale (suite)

$$P(X = k) = C_n^k \, p^k \, (1-p)^{n-k}$$

(k le nombre d'occurrences parmi n)

n = 50

	p	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
k	0	0,0769	0,0052	0,0003	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,2794	0,0338	0,0029	0,0002	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	2	0,5405	0,1117	0,0142	0,0013	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000
	3	0,7604	0,2503	0,0460	0,0057	0,0005	0,0000	0,0000	0,0000	0,0000	0,0000
	4	0,8964	0,4312	0,1121	0,0185	0,0021	0,0002	0,0000	0,0000	0,0000	0,0000
	5	0,9622	0,6161	0,2194	0,0480	0,0070	0,0007	0,0001	0,0000	0,0000	0,0000
	6	0,9882	0,7702	0,3613	0,1034	0,0194	0,0025	0,0002	0,0000	0,0000	0,0000
	7	0,9968	0,8779	0,5188	0,1904	0,0453	0,0073	0,0008	0,0001	0,0000	0,0000
	8	0,9992	0,9421	0,6681	0,3073	0,0916	0,0183	0,0025	0,0002	0,0000	0,0000
	9	0,9998	0,9755	0,7911	0,4437	0,1637	0,0402	0,0067	0,0008	0,0001	0,0000
	10	1,0000	0,9906	0,8801	0,5836	0,2622	0,0789	0,0160	0,0022	0,0002	0,0000
	11	1,0000	0,9968	0,9372	0,7107	0,3816	0,1390	0,0342	0,0057	0,0006	0,0000
	12	1,0000	0,9990	0,9699	0,8139	0,5110	0,2229	0,0661	0,0133	0,0018	0,0002
	13	1,0000	0,9997	0,9868	0,8894	0,6370	0,3279	0,1163	0,0280	0,0045	0,0005
	14	1,0000	0,9999	0,9947	0,9393	0,7481	0,4468	0,1878	0,0540	0,0104	0,0013
	15	1,0000	1,0000	0,9981	0,9692	0,8369	0,5692	0,2801	0,0955	0,0220	0,0033
	16	1,0000	1,0000	0,9993	0,9856	0,9017	0,6839	0,3889	0,1561	0,0427	0,0077
	17	1,0000	1,0000	0,9998	0,9937	0,9449	0,7822	0,5060	0,2369	0,0765	0,0164
	18	1,0000	1,0000	0,9999	0,9975	0,9713	0,8594	0,6216	0,3356	0,1273	0,0325
	19	1,0000	1,0000	1,0000	0,9991	0,9861	0,9152	0,7264	0,4465	0,1974	0,0595
	20	1,0000	1,0000	1,0000	0,9997	0,9937	0,9522	0,8139	0,5610	0,2862	0,1013
	21	1,0000	1,0000	1,0000	0,9999	0,9974	0,9749	0,8813	0,6701	0,3900	0,1611
	22	1,0000	1,0000	1,0000	1,0000	0,9990	0,9877	0,9290	0,7660	0,5019	0,2399
	23	1,0000	1,0000	1,0000	1,0000	0,9996	0,9944	0,9604	0,8438	0,6134	0,3359
	24	1,0000	1,0000	1,0000	1,0000	0,9999	0,9976	0,9793	0,9022	0,7160	0,4439
	25	1,0000	1,0000	1,0000	1,0000	1,0000	0,9991	0,9900	0,9427	0,8034	0,5561
	26	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9955	0,9686	0,8721	0,6641
	27	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9981	0,9840	0,9220	0,7601
	28	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9993	0,9924	0,9556	0,8389
	29	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9966	0,9765	0,8987
	30	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9986	0,9884	0,9405
	31	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9995	0,9947	0,9675
	32	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998	0,9978	0,9836
	33	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9991	0,9923
	34	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9967
	35	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9987
	36	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9995
	37	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998
	38	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Table 2

Loi de Poisson

$$P(X = k) = e^{-\mu} \frac{\mu^k}{k!}$$

(μ le nombre d'occurrences moyen)

	μ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
x	0	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488	0,4966	0,4493	0,4066	0,3679
	1	0,9953	0,9825	0,9631	0,9384	0,9098	0,8781	0,8442	0,8088	0,7725	0,7358
	2	0,9998	0,9989	0,9964	0,9921	0,9856	0,9769	0,9659	0,9526	0,9371	0,9197
	3	1,0000	0,9999	0,9997	0,9992	0,9982	0,9966	0,9942	0,9909	0,9865	0,9810
	4	1,0000	1,0000	1,0000	0,9999	0,9998	0,9996	0,9992	0,9986	0,9977	0,9963
	5	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9998	0,9997	0,9994
	6	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999
	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

	μ	1,5	2	3	4	5	6	7	8	9	10
x	0	0,2231	0,1353	0,0498	0,0183	0,0067	0,0025	0,0009	0,0003	0,0001	0,0000
	1	0,5578	0,4060	0,1991	0,0916	0,0404	0,0174	0,0073	0,0030	0,0012	0,0005
	2	0,8088	0,6767	0,4232	0,2381	0,1247	0,0620	0,0296	0,0138	0,0062	0,0028
	3	0,9344	0,8571	0,6472	0,4335	0,2650	0,1512	0,0818	0,0424	0,0212	0,0103
	4	0,9814	0,9473	0,8153	0,6288	0,4405	0,2851	0,1730	0,0996	0,0550	0,0293
	5	0,9955	0,9834	0,9161	0,7851	0,6160	0,4457	0,3007	0,1912	0,1157	0,0671
	6	0,9991	0,9955	0,9665	0,8893	0,7622	0,6063	0,4497	0,3134	0,2068	0,1301
	7	0,9998	0,9989	0,9881	0,9489	0,8666	0,7440	0,5987	0,4530	0,3239	0,2202
	8	1,0000	0,9998	0,9962	0,9786	0,9319	0,8472	0,7291	0,5925	0,4557	0,3328
	9	1,0000	1,0000	0,9989	0,9919	0,9682	0,9161	0,8305	0,7166	0,5874	0,4579
	10	1,0000	1,0000	0,9997	0,9972	0,9863	0,9574	0,9015	0,8159	0,7060	0,5830
	11	1,0000	1,0000	0,9999	0,9991	0,9945	0,9799	0,9467	0,8881	0,8030	0,6968
	12	1,0000	1,0000	1,0000	0,9997	0,9980	0,9912	0,9730	0,9362	0,8758	0,7916
	13	1,0000	1,0000	1,0000	0,9999	0,9993	0,9964	0,9872	0,9658	0,9261	0,8645
	14	1,0000	1,0000	1,0000	1,0000	0,9998	0,9986	0,9943	0,9827	0,9585	0,9165
	15	1,0000	1,0000	1,0000	1,0000	0,9999	0,9995	0,9976	0,9918	0,9780	0,9513
	16	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998	0,9990	0,9963	0,9889	0,9730
	17	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9996	0,9984	0,9947	0,9857
	18	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9993	0,9976	0,9928
	19	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9989	0,9965
	20	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9996	0,9984
	21	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998	0,9993
	22	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9997
	23	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999
	24	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Table 3

Loi Normale Centrée Réduite

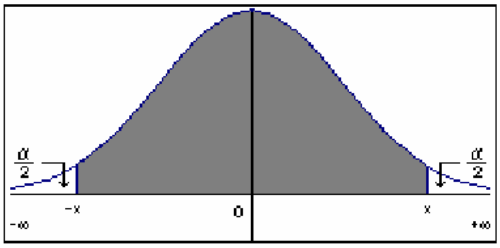
Fonction de répartition F(z)=P(Z<z)

Exemple : P(Z<1.96)= 0.97500 se trouve en ligne 1.9 et colonne 0.06

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,50000	0,50399	0,50798	0,51197	0,51595	0,51994	0,52392	0,52790	0,53188	0,53586
0,1	0,53983	0,54380	0,54776	0,55172	0,55567	0,55962	0,56356	0,56750	0,57142	0,57535
0,2	0,57926	0,58317	0,58706	0,59095	0,59484	0,59871	0,60257	0,60642	0,61026	0,61409
0,3	0,61791	0,62172	0,62552	0,62930	0,63307	0,63683	0,64058	0,64431	0,64803	0,65173
0,4	0,65542	0,65910	0,66276	0,66640	0,67003	0,67365	0,67724	0,68082	0,68439	0,68793
0,5	0,69146	0,69498	0,69847	0,70194	0,70540	0,70884	0,71226	0,71566	0,71904	0,72241
0,6	0,72575	0,72907	0,73237	0,73565	0,73891	0,74215	0,74537	0,74857	0,75175	0,75490
0,7	0,75804	0,76115	0,76424	0,76731	0,77035	0,77337	0,77637	0,77935	0,78231	0,78524
0,8	0,78815	0,79103	0,79389	0,79673	0,79955	0,80234	0,80511	0,80785	0,81057	0,81327
0,9	0,81594	0,81859	0,82121	0,82382	0,82639	0,82894	0,83147	0,83398	0,83646	0,83891
1,0	0,84135	0,84375	0,84614	0,84850	0,85083	0,85314	0,85543	0,85769	0,85993	0,86214
1,1	0,86433	0,86650	0,86864	0,87076	0,87286	0,87493	0,87698	0,87900	0,88100	0,88298
1,2	0,88493	0,88686	0,88877	0,89065	0,89251	0,89435	0,89617	0,89796	0,89973	0,90148
1,3	0,90320	0,90490	0,90658	0,90824	0,90988	0,91149	0,91309	0,91466	0,91621	0,91774
1,4	0,91924	0,92073	0,92220	0,92364	0,92507	0,92647	0,92786	0,92922	0,93056	0,93189
1,5	0,93319	0,93448	0,93575	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408
1,6	0,94520	0,94630	0,94738	0,94845	0,94950	0,95053	0,95154	0,95254	0,95352	0,95449
1,7	0,95544	0,95637	0,95728	0,95819	0,95907	0,95994	0,96080	0,96164	0,96246	0,96327
1,8	0,96407	0,96485	0,96562	0,96638	0,96712	0,96784	0,96856	0,96926	0,96995	0,97062
1,9	0,97128	0,97193	0,97257	0,97320	0,97381	0,97441	0,97500	0,97558	0,97615	0,97670
2,0	0,97725	0,97778	0,97831	0,97882	0,97933	0,97982	0,98030	0,98077	0,98124	0,98169
2,1	0,98214	0,98257	0,98300	0,98341	0,98382	0,98422	0,98461	0,98500	0,98537	0,98574
2,2	0,98610	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,98840	0,98870	0,98899
2,3	0,98928	0,98956	0,98983	0,99010	0,99036	0,99061	0,99086	0,99111	0,99134	0,99158
2,4	0,99180	0,99202	0,99224	0,99245	0,99266	0,99286	0,99305	0,99324	0,99343	0,99361
2,5	0,99379	0,99396	0,99413	0,99430	0,99446	0,99461	0,99477	0,99492	0,99506	0,99520
2,6	0,99534	0,99547	0,99560	0,99573	0,99585	0,99598	0,99609	0,99621	0,99632	0,99643
2,7	0,99653	0,99664	0,99674	0,99683	0,99693	0,99702	0,99711	0,99720	0,99728	0,99736
2,8	0,99744	0,99752	0,99760	0,99767	0,99774	0,99781	0,99788	0,99795	0,99801	0,99807
2,9	0,99813	0,99819	0,99825	0,99831	0,99836	0,99841	0,99846	0,99851	0,99856	0,99861
3,0	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99897	0,99900
3,1	0,99903	0,99906	0,99910	0,99913	0,99916	0,99918	0,99921	0,99924	0,99926	0,99929
3,2	0,99931	0,99934	0,99936	0,99938	0,99940	0,99942	0,99944	0,99946	0,99948	0,99950
3,3	0,99952	0,99953	0,99955	0,99957	0,99958	0,99960	0,99961	0,99962	0,99964	0,99965
3,4	0,99966	0,99968	0,99969	0,99970	0,99971	0,99972	0,99973	0,99974	0,99975	0,99976

Table 4

Loi de Student



α	1	0,8	0,6	0,4	0,2	0,1	0,05	0,02	0,01	0,002	0,001
1 − α	0	0,2	0,4	0,6	0,8	0,9	0,95	0,98	0,99	0,998	0,999
ν = ddl											
1	0,0000	0,3249	0,7265	1,3764	3,0777	6,3137	12,706	31,821	63,656	318,29	636,58
2	0,0000	0,2887	0,6172	1,0607	1,8856	2,9200	4,3027	6,9645	9,9250	22,328	31,600
3	0,0000	0,2767	0,5844	0,9785	1,6377	2,3534	3,1824	4,5407	5,8408	10,214	12,924
4	0,0000	0,2707	0,5686	0,9410	1,5332	2,1318	2,7765	3,7469	4,6041	7,1729	8,6101
5	0,0000	0,2672	0,5594	0,9195	1,4759	2,0150	2,5706	3,3649	4,0321	5,8935	6,8685
6	0,0000	0,2648	0,5534	0,9057	1,4398	1,9432	2,4469	3,1427	3,7074	5,2075	5,9587
7	0,0000	0,2632	0,5491	0,8960	1,4149	1,8946	2,3646	2,9979	3,4995	4,7853	5,4081
8	0,0000	0,2619	0,5459	0,8889	1,3968	1,8595	2,3060	2,8965	3,3554	4,5008	5,0414
9	0,0000	0,2610	0,5435	0,8834	1,3830	1,8331	2,2622	2,8214	3,2498	4,2969	4,7809
10	0,0000	0,2602	0,5415	0,8791	1,3722	1,8125	2,2281	2,7638	3,1693	4,1437	4,5868
11	0,0000	0,2596	0,5399	0,8755	1,3634	1,7959	2,2010	2,7181	3,1058	4,0248	4,4369
12	0,0000	0,2590	0,5386	0,8726	1,3562	1,7823	2,1788	2,6810	3,0545	3,9296	4,3178
13	0,0000	0,2586	0,5375	0,8702	1,3502	1,7709	2,1604	2,6503	3,0123	3,8520	4,2209
14	0,0000	0,2582	0,5366	0,8681	1,3450	1,7613	2,1448	2,6245	2,9768	3,7874	4,1403
15	0,0000	0,2579	0,5357	0,8662	1,3406	1,7531	2,1315	2,6025	2,9467	3,7329	4,0728
16	0,0000	0,2576	0,5350	0,8647	1,3368	1,7459	2,1199	2,5835	2,9208	3,6861	4,0149
17	0,0000	0,2573	0,5344	0,8633	1,3334	1,7396	2,1098	2,5669	2,8982	3,6458	3,9651
18	0,0000	0,2571	0,5338	0,8620	1,3304	1,7341	2,1009	2,5524	2,8784	3,6105	3,9217
19	0,0000	0,2569	0,5333	0,8610	1,3277	1,7291	2,0930	2,5395	2,8609	3,5793	3,8833
20	0,0000	0,2567	0,5329	0,8600	1,3253	1,7247	2,0860	2,5280	2,8453	3,5518	3,8496
21	0,0000	0,2566	0,5325	0,8591	1,3232	1,7207	2,0796	2,5176	2,8314	3,5271	3,8193
22	0,0000	0,2564	0,5321	0,8583	1,3212	1,7171	2,0739	2,5083	2,8188	3,5050	3,7922
23	0,0000	0,2563	0,5317	0,8575	1,3195	1,7139	2,0687	2,4999	2,8073	3,4850	3,7676
24	0,0000	0,2562	0,5314	0,8569	1,3178	1,7109	2,0639	2,4922	2,7970	3,4668	3,7454
25	0,0000	0,2561	0,5312	0,8562	1,3163	1,7081	2,0595	2,4851	2,7874	3,4502	3,7251
26	0,0000	0,2560	0,5309	0,8557	1,3150	1,7056	2,0555	2,4786	2,7787	3,4350	3,7067
27	0,0000	0,2559	0,5306	0,8551	1,3137	1,7033	2,0518	2,4727	2,7707	3,4210	3,6895
28	0,0000	0,2558	0,5304	0,8546	1,3125	1,7011	2,0484	2,4671	2,7633	3,4082	3,6739
29	0,0000	0,2557	0,5302	0,8542	1,3114	1,6991	2,0452	2,4620	2,7564	3,3963	3,6595
30	0,0000	0,2556	0,5300	0,8538	1,3104	1,6973	2,0423	2,4573	2,7500	3,3852	3,6460
40	0,0000	0,2550	0,5286	0,8507	1,3031	1,6839	2,0211	2,4233	2,7045	3,3069	3,5510
50	0,0000	0,2547	0,5278	0,8489	1,2987	1,6759	2,0086	2,4033	2,6778	3,2614	3,4960
60	0,0000	0,2545	0,5272	0,8477	1,2958	1,6706	2,0003	2,3901	2,6603	3,2317	3,4602
70	0,0000	0,2543	0,5268	0,8468	1,2938	1,6669	1,9944	2,3808	2,6479	3,2108	3,4350
80	0,0000	0,2542	0,5265	0,8461	1,2922	1,6641	1,9901	2,3739	2,6387	3,1952	3,4164
90	0,0000	0,2541	0,5263	0,8456	1,2910	1,6620	1,9867	2,3685	2,6316	3,1832	3,4019
100	0,0000	0,2540	0,5261	0,8452	1,2901	1,6602	1,9840	2,3642	2,6259	3,1738	3,3905
200	0,0000	0,2537	0,5252	0,8434	1,2858	1,6525	1,9719	2,3451	2,6006	3,1315	3,3398
∞	0,0000	0,2533	0,5244	0,8416	1,2816	1,6449	1,9600	2,3263	2,5758	3,0903	3,2906

Table 5

Loi du χ^2

$$P(\chi_\nu^2 \geq \chi_{\nu,\alpha}^2) = \alpha$$

$1-\alpha$	0,001	0,005	0,01	0,025	0,05	0,1	0,5	0,9	0,95	0,975	0,99	0,995	0,999
α	0,999	0,995	0,99	0,975	0,95	0,9	0,5	0,1	0,05	0,025	0,01	0,005	0,001
ν = ddl													
1	0,00	0,00	0,00	0,00	0,00	0,02	0,45	2,71	3,84	5,02	6,63	7,88	10,83
2	0,00	0,01	0,02	0,05	0,10	0,21	1,39	4,61	5,99	7,38	9,21	10,60	13,82
3	0,02	0,07	0,11	0,22	0,35	0,58	2,37	6,25	7,81	9,35	11,34	12,84	16,27
4	0,09	0,21	0,30	0,48	0,71	1,06	3,36	7,78	9,49	11,14	13,28	14,86	18,47
5	0,21	0,41	0,55	0,83	1,15	1,61	4,35	9,24	11,07	12,83	15,09	16,75	20,51
6	0,38	0,68	0,87	1,24	1,64	2,20	5,35	10,64	12,59	14,45	16,81	18,55	22,46
7	0,60	0,99	1,24	1,69	2,17	2,83	6,35	12,02	14,07	16,01	18,48	20,28	24,32
8	0,86	1,34	1,65	2,18	2,73	3,49	7,34	13,36	15,51	17,53	20,09	21,95	26,12
9	1,15	1,73	2,09	2,70	3,33	4,17	8,34	14,68	16,92	19,02	21,67	23,59	27,88
10	1,48	2,16	2,56	3,25	3,94	4,87	9,34	15,99	18,31	20,48	23,21	25,19	29,59
11	1,83	2,60	3,05	3,82	4,57	5,58	10,34	17,28	19,68	21,92	24,73	26,76	31,26
12	2,21	3,07	3,57	4,40	5,23	6,30	11,34	18,55	21,03	23,34	26,22	28,30	32,91
13	2,62	3,57	4,11	5,01	5,89	7,04	12,34	19,81	22,36	24,74	27,69	29,82	34,53
14	3,04	4,07	4,66	5,63	6,57	7,79	13,34	21,06	23,68	26,12	29,14	31,32	36,12
15	3,48	4,60	5,23	6,26	7,26	8,55	14,34	22,31	25,00	27,49	30,58	32,80	37,70
16	3,94	5,14	5,81	6,91	7,96	9,31	15,34	23,54	26,30	28,85	32,00	34,27	39,25
17	4,42	5,70	6,41	7,56	8,67	10,09	16,34	24,77	27,59	30,19	33,41	35,72	40,79
18	4,90	6,26	7,01	8,23	9,39	10,86	17,34	25,99	28,87	31,53	34,81	37,16	42,31
19	5,41	6,84	7,63	8,91	10,12	11,65	18,34	27,20	30,14	32,85	36,19	38,58	43,82
20	5,92	7,43	8,26	9,59	10,85	12,44	19,34	28,41	31,41	34,17	37,57	40,00	45,31
21	6,45	8,03	8,90	10,28	11,59	13,24	20,34	29,62	32,67	35,48	38,93	41,40	46,80
22	6,98	8,64	9,54	10,98	12,34	14,04	21,34	30,81	33,92	36,78	40,29	42,80	48,27
23	7,53	9,26	10,20	11,69	13,09	14,85	22,34	32,01	35,17	38,08	41,64	44,18	49,73
24	8,08	9,89	10,86	12,40	13,85	15,66	23,34	33,20	36,42	39,36	42,98	45,56	51,18
25	8,65	10,52	11,52	13,12	14,61	16,47	24,34	34,38	37,65	40,65	44,31	46,93	52,62
26	9,22	11,16	12,20	13,84	15,38	17,29	25,34	35,56	38,89	41,92	45,64	48,29	54,05
27	9,80	11,81	12,88	14,57	16,15	18,11	26,34	36,74	40,11	43,19	46,96	49,65	55,48
28	10,39	12,46	13,56	15,31	16,93	18,94	27,34	37,92	41,34	44,46	48,28	50,99	56,89
29	10,99	13,12	14,26	16,05	17,71	19,77	28,34	39,09	42,56	45,72	49,59	52,34	58,30
30	11,59	13,79	14,95	16,79	18,49	20,60	29,34	40,26	43,77	46,98	50,89	53,67	59,70

Pour $\nu > 30$, La loi du χ^2 peut –être approximée par la loi normale $N(\nu, \sqrt{\nu})$

Table 6

Loi de Fisher F

$$P(F_{\nu_1,\nu_2} < \mathsf{f}_{\nu_1,\nu_2,\alpha}) = \alpha$$

$\alpha = 0{,}975$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	15	20	30	50	100	200	500	•
1	648	800	864	900	922	937	948	957	963	969	985	993	1001	1008	1013	1016	1017	1018
2	38,5	39,0	39,2	39,3	39,3	39,4	39,4	39,4	39,4	39,4	39,4	39,4	39,5	39,5	39,5	39,5	39,5	39,5
3	17,4	16,0	15,4	15,1	14,9	14,7	14,6	14,5	14,5	14,4	14,3	14,2	14,1	14,0	14,0	13,9	13,9	13,9
4	12,2	10,6	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,66	8,56	8,46	8,38	8,32	8,29	8,27	8,26
5	10,0	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,43	6,33	6,23	6,14	6,08	6,05	6,03	6,02
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,27	5,17	5,07	4,98	4,92	4,88	4,86	4,85
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,57	4,47	4,36	4,28	4,21	4,18	4,16	4,14
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,10	4,00	3,89	3,81	3,74	3,70	3,68	3,67
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,77	3,67	3,56	3,47	3,40	3,37	3,35	3,33
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,52	3,42	3,31	3,22	3,15	3,12	3,09	3,08
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53	3,33	3,23	3,12	3,03	2,96	2,92	2,90	2,88
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,18	3,07	2,96	2,87	2,80	2,76	2,74	2,72
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25	3,05	2,95	2,84	2,74	2,67	2,63	2,61	2,60
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15	2,95	2,84	2,73	2,64	2,56	2,53	2,50	2,49
15	6,20	4,76	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	2,86	2,76	2,64	2,55	2,47	2,44	2,41	2,40
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99	2,79	2,68	2,57	2,47	2,40	2,36	2,33	2,32
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92	2,72	2,62	2,50	2,41	2,33	2,29	2,26	2,25
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87	2,67	2,56	2,44	2,35	2,27	2,23	2,20	2,19
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82	2,62	2,51	2,39	2,30	2,22	2,18	2,15	2,13
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,57	2,46	2,35	2,25	2,17	2,13	2,10	2,09
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70	2,50	2,39	2,27	2,17	2,09	2,05	2,02	2,00
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64	2,44	2,33	2,21	2,11	2,02	1,98	1,95	1,94
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59	2,39	2,28	2,16	2,05	1,97	1,92	1,90	1,88
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55	2,34	2,23	2,11	2,01	1,92	1,88	1,85	1,83
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,31	2,20	2,07	1,97	1,88	1,84	1,81	1,79
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39	2,18	2,07	1,94	1,83	1,74	1,69	1,66	1,64
50	5,34	3,98	3,39	3,06	2,83	2,67	2,55	2,46	2,38	2,32	2,11	1,99	1,87	1,75	1,66	1,60	1,57	1,55
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,06	1,94	1,82	1,70	1,60	1,54	1,51	1,48
80	5,22	3,86	3,28	2,95	2,73	2,57	2,45	2,36	2,28	2,21	2,00	1,88	1,75	1,63	1,53	1,47	1,43	1,40
100	5,18	3,83	3,25	2,92	2,70	2,54	2,42	2,32	2,24	2,18	1,97	1,85	1,71	1,59	1,48	1,42	1,38	1,35
200	5,10	3,76	3,18	2,85	2,63	2,47	2,35	2,26	2,18	2,11	1,90	1,78	1,64	1,51	1,39	1,32	1,27	1,23
500	5,05	3,72	3,14	2,81	2,59	2,43	2,31	2,22	2,14	2,07	1,86	1,74	1,60	1,46	1,34	1,25	1,19	1,14
•	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11	2,05	1,83	1,71	1,57	1,43	1,30	1,21	1,13	1,00

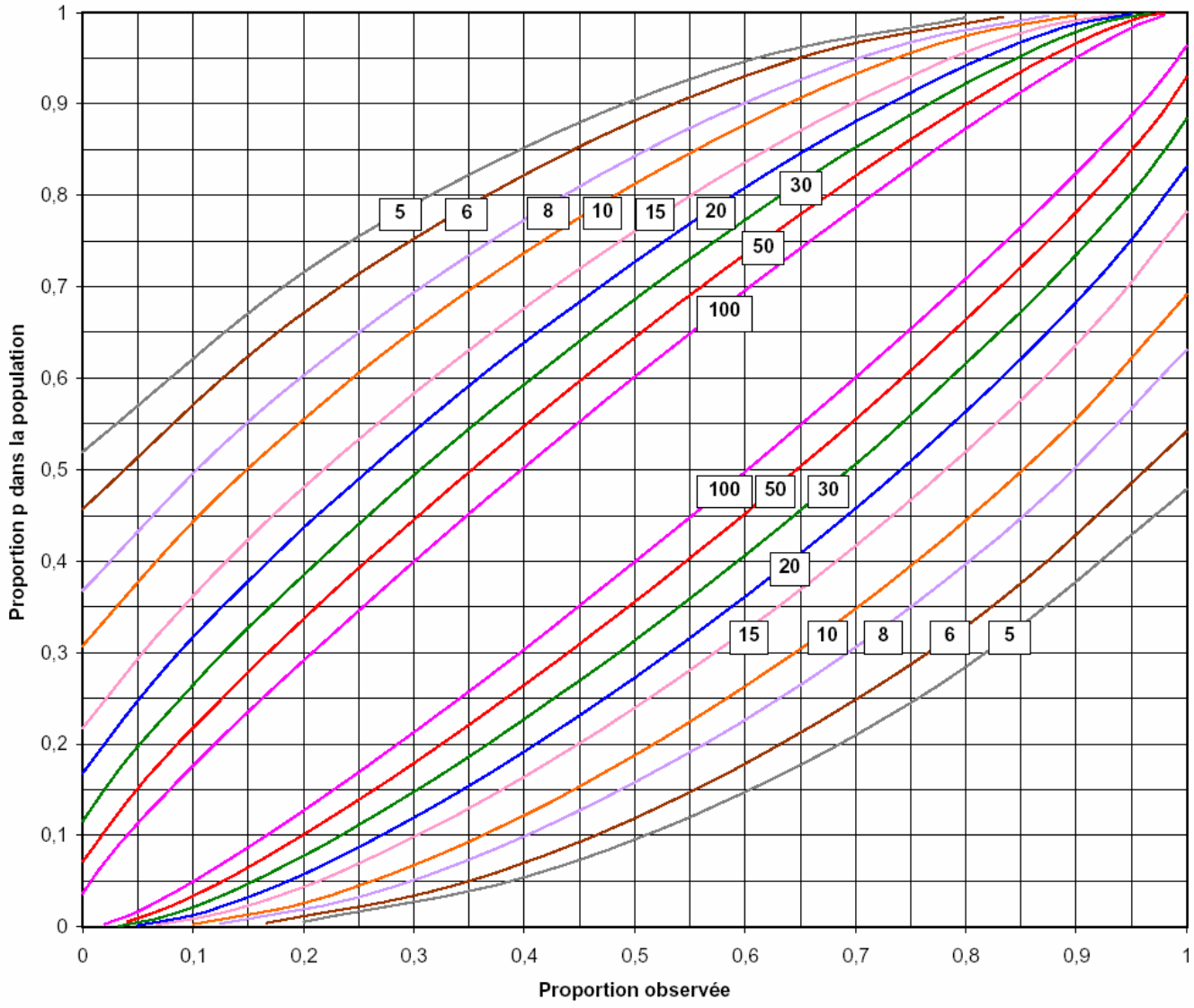
Loi de Fisher F (suite)

$$P(F_{\nu_1,\nu_2} < \mathsf{f}_{\nu_1,\nu_2,\alpha}) = \alpha$$

$\alpha = 0{,}95$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	15	20	30	50	100	200	500	•
1	161	200	216	225	230	234	237	239	241	242	246	248	250	252	253	254	254	254
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5	19,5
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,70	8,66	8,62	8,58	8,55	8,54	8,53	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,86	5,80	5,75	5,70	5,66	5,65	5,64	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,62	4,56	4,50	4,44	4,41	4,39	4,37	4,37
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	3,94	3,87	3,81	3,75	3,71	3,69	3,68	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,51	3,44	3,38	3,32	3,27	3,25	3,24	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,22	3,15	3,08	3,02	2,97	2,95	2,94	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,01	2,94	2,86	2,80	2,76	2,73	2,72	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,85	2,77	2,70	2,64	2,59	2,56	2,55	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,72	2,65	2,57	2,51	2,46	2,43	2,42	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,62	2,54	2,47	2,40	2,35	2,32	2,31	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,53	2,46	2,38	2,31	2,26	2,23	2,22	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,46	2,39	2,31	2,24	2,19	2,16	2,14	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,40	2,33	2,25	2,18	2,12	2,10	2,08	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,35	2,28	2,19	2,12	2,07	2,04	2,02	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,31	2,23	2,15	2,08	2,02	1,99	1,97	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,27	2,19	2,11	2,04	1,98	1,95	1,93	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,23	2,16	2,07	2,00	1,94	1,91	1,89	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,20	2,12	2,04	1,97	1,91	1,88	1,86	1,84
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,15	2,07	1,98	1,91	1,85	1,82	1,80	1,78
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,11	2,03	1,94	1,86	1,80	1,77	1,75	1,73
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,07	1,99	1,90	1,82	1,76	1,73	1,71	1,69
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,04	1,96	1,87	1,79	1,73	1,69	1,67	1,65
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,01	1,93	1,84	1,76	1,70	1,66	1,64	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	1,92	1,84	1,74	1,66	1,59	1,55	1,53	1,51
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03	1,87	1,78	1,69	1,60	1,52	1,48	1,46	1,44
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,84	1,75	1,65	1,56	1,48	1,44	1,41	1,39
80	3,96	3,11	2,72	2,49	2,33	2,21	2,13	2,06	2,00	1,95	1,79	1,70	1,60	1,51	1,43	1,38	1,35	1,32
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93	1,77	1,68	1,57	1,48	1,39	1,34	1,31	1,28
200	3,89	3,04	2,65	2,42	2,26	2,14	2,06	1,98	1,93	1,88	1,72	1,62	1,52	1,41	1,32	1,26	1,22	1,19
500	3,86	3,01	2,62	2,39	2,23	2,12	2,03	1,96	1,90	1,85	1,69	1,59	1,48	1,38	1,28	1,21	1,16	1,11
•	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,67	1,57	1,46	1,35	1,24	1,17	1,11	1,00

Intervalle de confiance bilatéral à 95 % d'une proportion



Bibliographie

[1] Hardin, J. and Hilbe, J. (2012). Generalized Linear Models and Extensions, 3rdEdition. College Station, Texas : Stata Press. Un livre avec des exemples et des applications incluant des analyses avec Stata.

[2] Hosmer, D.W. and Lemeshow, S. (2013). Applied Logistic Regression, 3rd Edition.New York : John Wiley and Sons. Une discussion détaillée sur le modèle logistique avec des applications.

[3] McCullagh,P. and Nelder, J.A. (1989). Generalized Linear Models, 2nd Edition.London : Chapman and Hall. La "bible" des modèles linéaires généralisés. Trèsintéressant, mais plutôt destinée à des étudiants avancés.

[4] Notes de cours de G. Rodriguez et exemples de codes R :http ://data.princeton.edu/wws509/.

[5] Notes de cours de L. Rouvière.
http ://perso.univrennes2.fr/system/files/users/rouviere_l/poly_logistique_web.pdf.

[6] Notes de cours de F. Bertrand. http ://www.irma.ustrasbg.fr/~fbertran/enseignement/Ecole_Doctorale_SVS_Automne_2008/ED_RegLog.pdf.

[7] Valérie Monbet. Modèles linéaires généralisés. IRMAR, Université de Rennes 1.

[8] Arthur Charpentier. (2013). Partie 4-modèles linéaires généralisés. Université du Québec à Montréal. http ://freakonometrics.hypotheses.org/.

[9] Nelder et Wedderburn. (1972). Modèles linéaires généralisés. Présentation.

[10] Mc Cullagh et Nelder. (1989). Modèles linéaires généralisés. Présentation