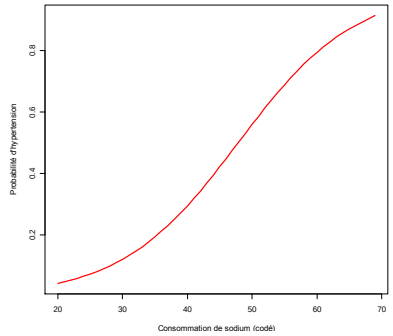
MAT7381 Chapitre 8
Régression logistique et autre modèles

8.1 Régression logistique

La régression linéaire ordinaire est conçue pour prédire, à partir de certaines variables exogènes, la valeur d'une variable *y quantitative*. Comme, par exemple, lorsque *y* est une mesure de la pression artérielle et *x* est la consommation quotidienne de sodium. Mais pour des raisons aussi bien médicales que statistiques on pourrait préférer noter simplement que la pression est normale ou excessive. Si *m* est le nombre de personnes dont la consommation de sodium est *x* ; et *y* est le nombre de ceux, parmi eux, qui souffrent d'hypertension, alors *y* est une variable binomiale de paramètres *m* et π. Une régression linéaire stipule que E(*y*) = *m*π est une fonction linéaire de *x* [$\pi = \pi(x) = \beta_o+\beta_1 x$] et que Var(*y*) est fixe. La deuxième hypothèse est manifestement fausse, puisque Var(*y*) = *m*π(1-π). La première (un taux d'accroissement fixe) est rarement vérifiée en pratique. Généralement, on observera plutôt un accroissement faible aux extrémités et fort aux valeurs intermédiaires de *x* — une fonction en forme de S, comme celle-ci :



La fonction *logistique*, $f(x) = \dfrac{e^x}{1+e^x}$, est une telle fonction et pourrait servir à modéliser π(*x*), avec forcément un décalage et une dilatation (ou contraction) des valeurs de *x* : $\pi(x) = \dfrac{e^{\beta_o+\beta_1 x}}{1+e^{\beta_o+\beta_1 x}}$. Les données prennent la forme de *k* variables binomiales $y_1, \ldots, y_k$ de paramètres $m_1, \ldots, m_k$ et $\pi(x_1), \ldots, \pi(x_k)$, respectivement.

Multiples variables exogènes

La généralisation au cas de multiples variables exogènes est immédiate : $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ sont des vecteurs de dimension *q*, $\boldsymbol{\beta}$ est un vecteur de paramètre de même dimension, et $\pi(\boldsymbol{x}) = \dfrac{e^{\boldsymbol{\beta}'\boldsymbol{x}}}{1+e^{\boldsymbol{\beta}'\boldsymbol{x}}}$, un vecteur dont la i^e composante est $\dfrac{e^{\boldsymbol{\beta}'\boldsymbol{x}_i}}{1+e^{\boldsymbol{\beta}'\boldsymbol{x}_i}}$, $i = 1, ;\ldots, k$.

Le paramètre $\boldsymbol{\beta}$ sera estimé par la méthode du maximum de vraisemblance.

8.1.1 Le modèle à une variable exogène

Les données suivantes proviennent d'une étude sur la relation entre l'âge et l'incidence de maladies cardiovasculaires. Les variables sont

x_i : l'âge au dernier anniversaire
m_i : nombre de personne d'âge x_i
y_i : nombre de personnes d'âge x_i souffrant d'une maladie cardiovasculaire

Les données sont présentées au tableau 8.1.1.

Tableau 8.1.1

Âge	m_i	y_i	Âge	m_i	y_i	Âge	m_i	y_i	Âge	m_i	y_i
20	1	0	35	2	0	46	2	1	57	6	4
23	1	0	36	3	1	47	3	1	58	3	2
24	1	0	37	3	1	48	3	2	59	2	2
25	2	1	38	2	0	49	3	1	60	2	1
26	2	0	39	2	1	50	2	1	61	1	1
28	2	0	40	2	1	51	1	0	62	2	2
29	1	0	41	2	0	52	2	1	63	1	1
30	6	1	42	4	1	53	2	2	64	2	1
32	2	0	43	3	1	54	1	1	65	1	1
33	2	0	44	4	2	55	3	2	69	1	1
34	5	1	45	2	1	56	3	3			

Le modèle

Les y_i sont des variables indépendantes de loi binomiale :

$$y_i \sim \mathcal{B}(m_i\,;\pi_i),\ \ i = 1, \ldots, k,$$

où π_i est la probabilité qu'une personne d'âge x_i soit atteinte d'une maladie cardiovasculaire. Le modèle dit *saturé* n'impose aucune contrainte sur les π_i, outre celle qui définit une probabilité, soit $1 < \pi_i < 1$.

Le *modèle logistique* stipule que la probabilité π d'une maladie cardio-vasculaire est une fonction logistique de l'âge x :

$$\pi_i = \pi_i(\beta_o\,;\beta_1) = \frac{e^{\beta_o+\beta_1 x_i}}{1+e^{\beta_o+\beta_1 x_i}}\ ,$$

où $\boldsymbol{\beta} = [\beta_o\ ;\beta_1]'$. Les *logits* η_i, définis par $\eta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$ sont donc fonctions linéaires des x_i :

$$\eta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_o + \beta_1 x_i\,,\ i = 1, \ldots k$$

Le vecteur $\boldsymbol{y} = [y_1\ ; y_2\ ; \ldots\ ; y_k]'$ est d'espérance $\boldsymbol{\mu} = \mathbf{M}\boldsymbol{\pi}$ et de matrice de covariance

$$\mathbf{V} = \mathbf{V}(\boldsymbol{\pi}) = \begin{pmatrix} m_1\pi_1(1-\pi_1) & 0 & \cdots & 0 \\ 0 & m_2\pi_2(1-\pi_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & \cdots & \cdots & m_k\pi_k(1-\pi_k) \end{pmatrix}$$

où $\mathbf{M}$ est la matrice diagonale dont les éléments sont les m_i.

La fonction de vraisemblance est

$$L(\boldsymbol{y}\,;\boldsymbol{\pi}(\beta_o\ ;\beta_1)) = \prod_{i=1}^{k}\binom{m_i}{y_i}[\pi_i(\beta_o;\beta_1)]^{y_i}[1-\pi_i(\beta_o;\beta_1)]^{m_i-y_i}\ .$$

L'estimateur du maximum de vraisemblance $\hat{\boldsymbol{\beta}}$ est la valeur de $\boldsymbol{\beta}$ qui maximise L (ou son logarithme). Le

logarithme de L est

$$\ell(\boldsymbol{y}\,;\boldsymbol{\pi}(\beta_o\,;\beta_1)) \;=\; \sum_{i=1}^{k}\ln\binom{m_i}{y_i} \;+\; \sum_{i=1}^{k}\left[y_i\left(\beta_o+\beta_1 x_i\right)-m_i\ln\left(1+e^{\beta_o+\beta_1 x_i}\right)\right].$$

où $\boldsymbol{\pi}(\beta_o+\beta_1 x_i) = [\pi_1(\beta_o+\beta_1 x_i), \ldots, \pi_k(\beta_o+\beta_1 x_i)]$.

La valeur de $\boldsymbol{\beta}$ qui maximise la vraisemblance doit satisfaire l'équation $\dfrac{\partial \ell(\boldsymbol{y}\,;\boldsymbol{\pi}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0}$.

Cette équation n'a pas de solution explicite. La méthode de Newton-Raphson est une des méthodes itératives qui, sous certaines conditions, converge vers $\hat{\boldsymbol{\beta}}$. À partir d'une valeur initiale $\boldsymbol{\beta}_o$, on calcule la suite

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t - \mathbf{H}^{-1}(\boldsymbol{\beta}_t)\boldsymbol{h}(\boldsymbol{\beta}_t),$$

où $\boldsymbol{h}(\boldsymbol{\beta})$ est le vecteur des dérivées partielles de $\ell(\boldsymbol{y}\,;\boldsymbol{\pi}(\boldsymbol{\beta}))$ par rapport à $\boldsymbol{\beta}$, et qui ici est égale à

$$\boldsymbol{h}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}}\ell(\boldsymbol{y}\,;\boldsymbol{\pi}(\boldsymbol{\beta})) = \mathbf{X}'(\boldsymbol{y}-\mathbf{M}\boldsymbol{\pi}).$$

$\mathbf{H}$ est la matrice des dérivées secondes $\mathbf{H}(\boldsymbol{\beta}) = \dfrac{\partial}{\partial\beta_i\partial\beta_j}\ell(\boldsymbol{y}\,;\boldsymbol{\pi}(\boldsymbol{\beta}))$. Elle est égale à

$$\mathbf{H}(\boldsymbol{\beta}) = -\mathbf{X}'\mathbf{V}\mathbf{X}$$

où $\mathbf{X} = \begin{bmatrix}1 & x_1\\ \vdots & \vdots\\ 1 & x_k\end{bmatrix}$. La suite est donc

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t + (\mathbf{X}'\mathbf{V}_t\mathbf{X})^{-1}\mathbf{X}'(\boldsymbol{y}-\mathbf{M}\boldsymbol{\pi}_t)$$

où $\mathbf{V}_t = \mathbf{V}(\boldsymbol{\pi}_t) = \mathbf{V}(\boldsymbol{\pi}(\boldsymbol{\beta}_t))$.

Propriétés asymptotiques de $\hat{\boldsymbol{\beta}}$

$\hat{\boldsymbol{\beta}}$ est approximativement sans biais et sa matrice de covariance est approximativement $[\mathbf{X'V}(\boldsymbol{\pi})\mathbf{X}]^{-1}$. Elle est estimée par $[\mathbf{X}'\mathbf{V}(\hat{\boldsymbol{\pi}})\mathbf{X}]^{-1}$.

Commande R

La commande suivante permet d'obtenir les estimateurs et leur variance.

```
> modèle<-glm(cbind(y,mi-y)~age,family=binomial)
Coefficients:
(Intercept)          age
    -5.3095       0.1109
Degrees of Freedom: 42 Total (i.e. Null);  41 Residual
Null Deviance:      53.06
Residual Deviance: 23.75        AIC: 62.72
```

On estime donc que la probabilité d'avoir des problèmes cardiovasculaires à l'âge x est $\pi = \dfrac{e^{-5{,}3095+0{,}1109x}}{1+e^{-5{,}3095+0{,}1109x}}$. Elle serait donc de $\dfrac{e^{-5{,}3095+0{,}1109(75)}}{1+e^{-5{,}3095+0{,}1109(75)}} \approx 95\,\%$ pour une personne âgée de 75 ans et de $\dfrac{e^{-5{,}3095+0{,}1109(25)}}{1+e^{-5{,}3095+0{,}1109(25)}} \approx 7\,\%$ pour une personne âgée de 25 ans. Ces probabilités semblent à première vue élevées, mais elles pourraient s'expliquer s'il s'avérait que l'échantillon a été tiré d'une population particu-

lière, comme par exemple, des personnes ayant des antécédents, personnels ou parentaux, de maladie cardiaques.
La commande `summary()` permet de tester des hypothèses sur les coefficients. Les hypothèses testées ci-dessous sont $\beta_o = 0$ (généralement sans intérêt particulier) et $\beta_1 = 0$.

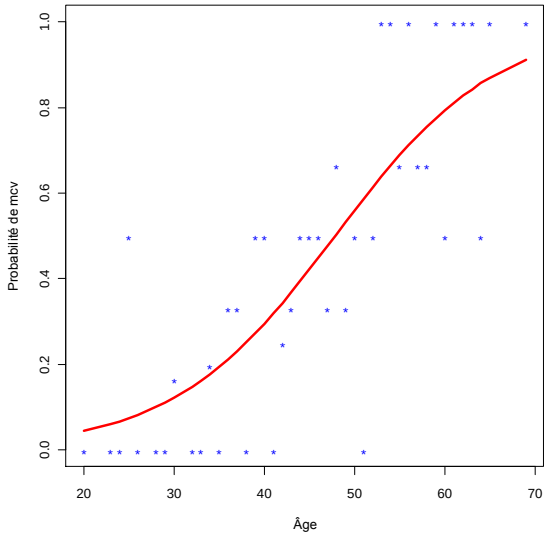
```
> summary(modèle)
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.30945    1.13363  -4.684 2.82e-06 ***
age          0.11092    0.02406   4.610 4.02e-06 ***

    Null deviance: 53.064  on 42  degrees of freedom
Residual deviance: 23.754  on 41  degrees of freedom
AIC: 62.724
```

Ces tests sont basés sur les statistiques $\hat{\beta}_o/\hat{\sigma}_{\hat{\beta}_o}$ et $\hat{\beta}_1/\hat{\sigma}_{\hat{\beta}_1}$ qui, sous les hypothèses respectives, suivent à peu près une loi normale centrée-réduite. L'hypothèse $\beta_1 = 0$, rejetée ici, signifierait que toutes les probabilités sont égales, c'est-à-dire, elle ne dépend pas de l'âge. Les écarts-types estimés $\hat{\sigma}_{\hat{\beta}_o}$ et $\hat{\sigma}_{\hat{\beta}_1}$ proviennent de la matrice de covariance estimée $[\mathbf{X}'\mathbf{V}(\hat{\boldsymbol{\pi}})\mathbf{X}]^{-1}$. On l'obtient par la commande suivante :

```
> summary(modèle)$cov.unscaled
             (Intercept)           age
(Intercept)  1.28512208 -0.0266759891
age         -0.02667599  0.0005788544
```

Voici le nuage de points avec la courbe logistique :



La déviance

La *déviance* est une mesure basée sur la fonction de vraisemblance

$$L = \prod_{i=1}^{43} \binom{m_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i} .$$

Elle compare deux modèles imbriqués, c'est-à-dire, des modèles dont les espaces paramétriques sont sous-ensemble l'un de l'autre. Si $\mathcal{M}$ est un modèle donné et $\mathcal{M}_o$ est un sous-modèle de $\mathcal{M}$ dans le sens que $\mathcal{M}_o$ impose certaines restrictions aux paramètres de $\mathcal{M}$, alors la déviance est

$$D = -2(\log L_o/L),$$

où L et L_o sont les fonctions de vraisemblance maximisées sous $\mathcal{M}$ et $\mathcal{M}_o$, respectivement.

Si $\mathcal{M}_o$ est un modèle défini par une hypothèse H_o énoncée dans le cadre d'un modèle $\mathcal{M}$, alors la déviance fournit un test de l'ajustement du modèle. Si H_o est vraie, L et L_o seront proches et la déviance résiduelle sera faible. Sous H_o,

$$D \sim \chi^2_\nu$$

où ν est le nombre de contraintes imposées au modèle par H_o.

La déviance résiduelle

La *déviance résiduelle* D_r compare un modèle $\mathcal{M}$ au *modèle saturé* $\mathcal{M}_s$. Elle est définie par

$$D_r = \textit{Déviance résiduelle} = -2\log L_m - (-2\log L_s)$$

où L_m et L_s sont les fonctions de vraisemblance maximisée sous $\mathcal{M}$ et $\mathcal{M}_s$, respectivement. Puisque le modèle saturé n'impose aucune restriction sur les π_i, le vecteur $\boldsymbol{\pi} = (\pi_1; \pi_2;\ldots;\pi_{43})$ est estimé simplement par $\hat{\boldsymbol{p}} = [\hat{p}_1; \hat{p}_2; \ldots \hat{p}_{43}] = (y_1/m_1; y_2/m_2; \ldots; y_{43}/m_{43})$. La vraisemblance maximale dans le modèle saturé est

$$L_s = \prod_{i=1}^{43} \binom{m_i}{y_i} \hat{p}_i^{y_i} (1-\hat{p}_i)^{m_i - y_i}$$

Si $\mathcal{M}$ est le modèle logistique, les π_i sont soumis à la restriction $\pi_i = \dfrac{e^{\beta_o+\beta_1 x_i}}{1+e^{\beta_o+\beta_1 x_i}}$ et la vraisemblance maximale sous le modèle est $L_m = \prod_{i=1}^{43} \binom{m_i}{y_i} \hat{\pi}_i^{y_i} (1-\hat{\pi}_i)^{m_i - y_i}$, où $\hat{\pi}_i = \dfrac{e^{\hat{\beta}_o+\hat{\beta}_1 x_i}}{1+e^{\hat{\beta}_o+\hat{\beta}_1 x_i}}$ et $\hat{\beta}_o$ et $\hat{\beta}_1$ sont les estimateurs du maximum de vraisemblance de β_o et β_1, respectivement. Dans l'exemple,

$$-2\log L_m = 58{,}72404$$

et

$$-2\log L_s = 34{,}96972.$$

La déviance résiduelle est

$$D_r = \textit{Déviance résiduelle} = -2\log L_m - (-2\log L_s) = -2\log\frac{L_m}{L_s} = 58{,}72404 - 34{,}96972 = 23{,}754.$$

La déviance résiduelle teste l'adéquation du modèle. Si le modèle est adéquat, L_m et L_s seront proches et la déviance résiduelle sera faible. Sous le modèle, D_r suit asymptotiquement une loi χ^2 à ν_r degrés de liberté, où ν_r est le nombre de contraintes imposées aux π_i par le modèle $\mathcal{M}$. Dans l'exemple, le nombre de paramètres est 43 dans le modèle saturé et 2 (β_o et β_1) dans le modèle logistique. Donc $\nu_r = 41$.

La déviance nulle

La *déviance nulle* teste, dans un modèle saturé, une hypothèse H_o qui stipule l'égalité des π_i. Dans l'exemple,

$$H_o : \pi_1 = \pi_2 = \ldots, \pi_{43}$$

ce qui est équivalent à supposer que les π_i sont tous égaux (= π, disons). Dans ce modèle, l'estimateur du maximum de vraisemblance est $\hat{\pi} = \Sigma y_i / \Sigma m_i$ et la vraisemblance maximale est

$$L_o = \prod_{i=1}^{43} \binom{m_i}{y_i} \hat{\pi}^{\Sigma y_i} (1-\hat{\pi})^{\Sigma m_i - \Sigma y_i}$$

$$-2\log L_o = 88{,}03393$$

La *déviance nulle* est définie par

$$D_o = \textit{Déviance nulle} = -2\log\frac{L_o}{L_s} = 88{,}03393 - 34{,}96972 = 53{,}064$$

avec pour nombre de degrés de liberté la différence entre les degrés de libertés, 42-41 = 1.

Test d'une hypothèse dans le cadre d'un modèle général (pas nécessairement saturé)

Les deux déviances calculées ci-dessus permettent de tester deux modèles, le modèle logistique et le modèle nul, chacun dans le cadre du modèle saturé. Normalement, lorsqu'on accepte le modèle logistique, on le tient pour vrai et on teste H_o dans le cadre du modèle logistique. Ce test est effectué à l'aide de la différence entre les deux déviances, soit

$$G^2 = \left(-2\log\frac{L_o}{L_s}\right) - \left(-2\log\frac{L_m}{L_s}\right) = 53{,}064 - 23{,}754 = 29{,}31.$$

Remarquez que

$$G^2 = -2\log\frac{L_o}{L_m} = -2\log\lambda,$$

où λ est le rapport des maximums de vraisemblance pour tester H_o dans le modèle logistique. Dans l'exemple, on a $-2\log\lambda = 29{,}31$, et puisque cette statistique suit à peu près une loi χ^2 à 1 degré de liberté sous H_o, on rejette H_o.

Résumé des hypothèses et des déviances

Considérons k valeurs *distinctes* de x

<table>
<tr><td colspan="2">Modèle</td><td colspan="3"></td></tr>
<tr><td>Nom</td><td>Description</td><td colspan="3">Déviances et tests</td></tr>
<tr><td>H_o : Nul</td><td>$\pi_1=\ldots=\pi_k$</td><td rowspan="2">$G^2 = -2\log(L_o/L_m)$
$= 29{,}31$, $\nu = 1$
Teste H_o dans le modèle logistique</td><td rowspan="2">Déviance résiduelle :
$-2\log(L_m/L_s) = 23{,}754$
$\nu = k-2 = 41$
Teste le modèle dans le modèle saturé</td><td rowspan="2">Déviance nulle
$-2\log(L_o/L_s) = 53{,}064$
$\nu = k-1 = 42$
Teste le modèle nul dans le modèle saturé</td></tr>
<tr><td>H_m : Logistique</td><td>$\pi_i = \frac{e^{\beta_o+\beta_1 x_i}}{1+e^{\beta_o+\beta_1 x_i}}$</td></tr>
<tr><td>H_s : Saturé</td><td>π_i sans contrainte</td><td></td><td></td><td></td></tr>
</table>

Autre présentation R

La commande **R** peut être exprimée différemment si chaque observation correspond à un seul individu. Les variables, z_i, disons, au nombre de 100 (et non 43), sont toutes bernoulliennes. Les valeurs x_i de la variable indépendante x sont également au nombre de 100. Alors la commande suivante donne les mêmes résultats que tantôt :

```
> summary(glm(z~x,family=binomial(logit)))
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.30945    1.13365  -4.683 2.82e-06 ***
x            0.11092    0.02406   4.610 4.02e-06 ***
    Null deviance: 136.66  on 99  degrees of freedom
Residual deviance: 107.35  on 98  degrees of freedom
AIC: 111.35
```

On remarque cependant que les valeurs `Null deviance` et `Residual deviance` ne sont pas les mêmes dans les deux tableaux. En voici un résumé :

Modèle		Déviances et tests		
Nom	Description			
H_o : Nul	$\pi_1=\ldots=\pi_{100}$	$G^2 = -2\log(L_o/L_m)$ = 29,31, $\nu = 1$ Teste H_o dans le modèle logistique	Déviance résiduelle : $-2\log(L_m/L_s)$ = 107,35 $\nu = k-2$ Teste le modèle dans le modèle saturé	Déviance nulle $-2\log(L_o/L_s)$ = 136,66 $\nu = n-1= 99$ Teste le modèle nul dans le modèle saturé
H_m : Logistique	$\pi_i = \frac{e^{\beta_o+\beta_1 x_i}}{1+e^{\beta_o+\beta_1 x_i}}$			
H_s : Saturé	π_i sans contrainte			

La différence tient au fait que le modèle saturé dans la deuxième présentation comprend beaucoup plus de paramètres, dans le sens qu'on ne suppose même pas que les observations correspondant à une même valeur de x ont la même probabilité de succès : chaque individu a sa probabilité propre. Il y a donc 100 probabilités (et non 43) dans le modèle saturé. Elles sont estimées par 0 ou 1 selon que $y_i = 0$ ou 1 et la probabilité des z_i sous ces valeurs est 1. La vraisemblance maximale sous ce modèle est donc $L_s = 1$. La fonction de vraisemblance sous le modèle logistique ne change pas; et la fonction de vraisemblance sous l'hypothèse nulle non plus.

On a donc $-2\log L_s = 0$;

$$L_m = \prod_{i=1}^{100} \hat{\pi}_i^{z_i} (1-\hat{\pi}_i)^{1-z_i} ,$$

où

$$\hat{\pi}_i = \frac{e^{\hat{\beta}_o+\hat{\beta}_1 x_i}}{1+e^{\hat{\beta}_o+\hat{\beta}_1 x_i}}$$

et $\hat{\beta}_o$ et $\hat{\beta}_1$, étant les estimateurs du maximum de vraisemblance de β_o et β_1 dans le modèle logistique, ne changent pas.

Alors

$-2\log L_m = 107{,}35 \Rightarrow$ Déviance résiduelle = 107,35, puisque $-2\log L_s = 0$.

Finalement, sous l'hypothèse nulle, $\hat{\pi} = \Sigma z_i/100$, $L_o = \prod_{i=1}^{100} \hat{\pi}^{z_i} (1-\hat{\pi})^{1-z_i}$ et

$-2\log L_o = 136{,}66 \Rightarrow$ Déviance nulle = 136,66.

Le test de l'hypothèse nulle ne change pas : $-2\log L_0 - (-2\log L_m) = -2\log(L_o/L_m) = 136{,}66 - 107{,}35 = 29{,}31$, comme avant.

Le D de Somer

La mesure D de Somer joue le même rôle qu'un coefficient de corrélation, et partage certaines de ses propriétés : ses valeurs s'étendent de –1 à 1; D prend la valeur 1 lorsque la relation est parfaite et positive; D = -1 lorsque la relation est parfaite et négative; et $D = 0$ lorsqu'il n'y a pas de relation entre y et la variable exogène. La *mesure D* est basée sur les notions de concordance et discordance. On considère toutes les paires $(y_i ; y_j)$ pour lesquelles $y_i = 0$ et $y_j = 1$. Une telle paire $(y_i ; y_j)$ est dite *concordante* si $\hat{\pi}_i < \hat{\pi}_j$; elle est dite *discordante* si $\hat{\pi}_i > \hat{\pi}_j$. D est définie par

$$D = \frac{\text{Nombre de concordances} - \text{Nombre de discordances}}{\text{Nombre de paires}}$$

À titre d'exemple, considérons quelques observations des données de Lemeshow :

		y_j =1				
		$\hat{\pi}_j$				
	$\hat{\pi}_i$	0,073344	0,121125	0,176807	0,211436	0,230521
	0,043479	C	C	C	C	C
y_i = 0	0,059621	C	C	C	C	C
	0,066153	C	C	C	C	C
	0,073344	E	C	C	C	C
	0,081248	D	C	C	C	C
	0,099422	D	C	C	C	C
	0,109804	D	C	C	C	C
	0,121125	D	E	C	C	C
	0,146793	D	D	C	C	C
	0,161237	D	D	C	C	C
	0,176807	D	D	E	C	C

Il y a 42 concordances, 10 discordances, et 3 ex aequo. Alors D = (42-10)/55 = 0,58

Calcul du D de Somer par `R`. Si `y` est le vecteur d'observations, `p` le vecteur des probabilités estimées, et ***e*** un vecteur de n « 1 », alors la commande suivante retournera le D de Somer

```
> 0.5*sum(sign((outer(y,e)-outer(e,y))*((outer(p,e)-outer(e,p))))))/(sum(y)*(100-sum(y))
```

Matrice de covariance

La matrice de covariance de $\hat{\boldsymbol{\beta}}$ est à peu près égale à

$$V(\hat{\boldsymbol{\beta}}) = [\mathbf{X'VX}]^{-1}, \text{où } \mathbf{V} = \begin{pmatrix} m_1\pi_1(1-\pi_1) & 0 & \cdots & 0 \\ 0 & m_2\pi_2(1-\pi_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & \cdots & \cdots & m_k\pi_k(1-\pi_k) \end{pmatrix}$$

On l'estime en remplaçant $\boldsymbol{\pi}$ par $\hat{\boldsymbol{\pi}}$ dans **V.**

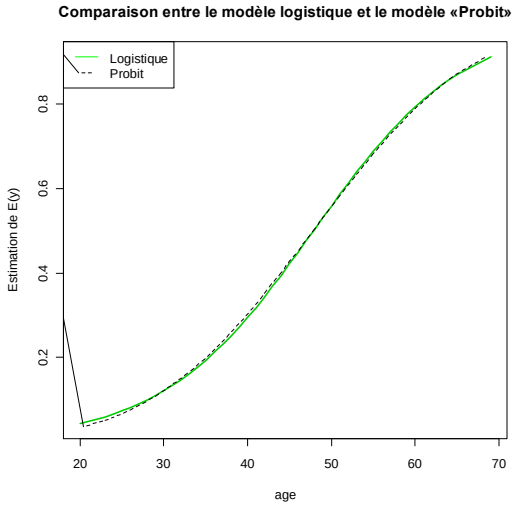
Un autre modèle plausible : les probits.

Si l'hypothèse fondamentale du modèle logistique, soit que $\pi_i = \dfrac{e^{\beta_o+\beta_1 x_i}}{1+e^{\beta_o+\beta_1 x_i}}$, semble à première vue gratuite et quelque peu abusive, c'est qu'elle est une expression un peu trop précise d'une hypothèse qu'il n'est pas par ailleurs difficile à accepter : l'hypothèse que la probabilité d'une maladie cardiovasculaire croît graduellement avec l'âge, à un taux faible d'abord puis rapide et finalement, à l'approche de 1, faible encore. Une courbe en forme d'un S allongé. Plusieurs fonctions présentent de telles caractéristiques, et on admet volontiers que la fonction logistique n'est pas plus raisonnable qu'une autre. Mais le fait est qu'il importe peu qu'on choisisse la fonction logistique ou une autre. À condition de bien choisir les valeurs de β_o et de β_1 on trouvera que plusieurs alternatives à la fonction logistique donneront des courbes très semblables à celle-ci. La fonction de répartition d'une loi continue pourrait servir, en particulier celle d'une normale centrée-réduite $\pi(x) = \Phi(\beta_o+\beta_1 x)$. La fonction de lien ici est $\Phi^{-1}(\pi) = \beta_o+\beta_1 x$. Voici le traitement **R.**

```
> mcvprobit<-glm(cbind(y,mi-y)~age,binomial(probit))
> summary(mcvprobit)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.14573    0.62460  -5.036 4.74e-07 ***
age          0.06580    0.01335   4.930 8.21e-07 ***
    Null deviance: 53.064  on 42  degrees of freedom
```

```
Residual deviance: 23.900  on 41  degrees of freedom
AIC: 62.87
```

Ces résultats sont étonnamment proches de ceux du modèle logistique quant à la qualité de l'ajustement. Le graphique suivant montre les valeurs estimées de *y* pour les deux modèles. On constate que les deux séries de prédictions sont presque confondues.

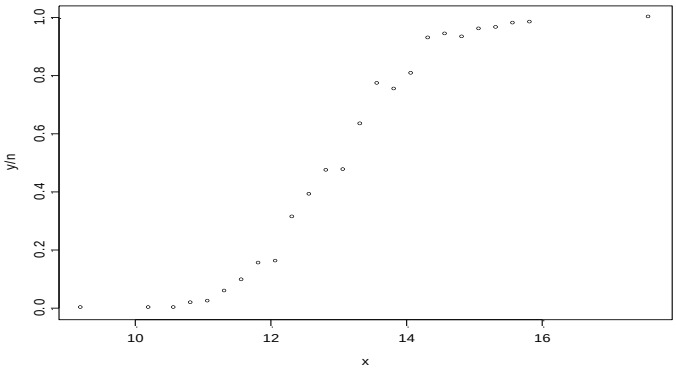


Exemple — Âge du début de la puberté

Pourcentage de filles de Varsovie ayant eu leur puberté, selon l'âge

#	Âge(*x*)	Effectif (*n*)	Nombre ayant atteint la puberté (*y*)
1	9,21	376	0
2	10,21	200	0
3	10,58	93	0
4	10,83	120	2
5	11,08	90	2
6	11,33	88	5
7	11,58	105	10
8	11,83	111	17
9	12,08	100	16
10	12,33	93	29
11	12,58	100	39
12	12,83	108	51
13	13,08	99	47
14	13,33	106	67
15	13,58	105	81
16	13,83	117	88
17	14,08	98	79
18	14,33	97	90
19	14,58	120	113
20	14,83	102	95
21	15,08	122	117
22	15,33	111	107
23	15,58	94	92
24	15,83	114	112
25	17,58	1049	1049

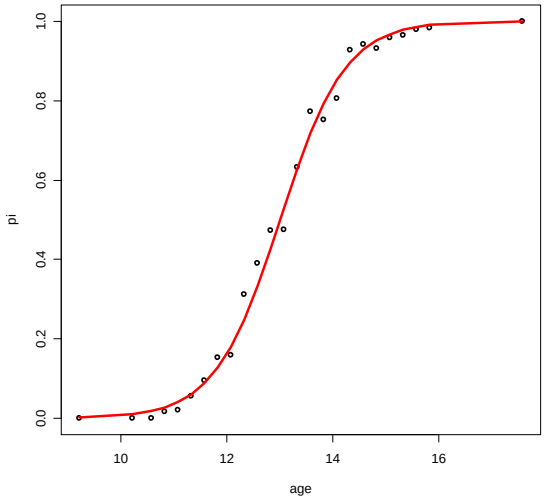
Pourcentage de filles ayant atteint la puberté en fonction de l'âge



Commande R

```
> a<-glm(se~age,binomial)
> summary(a)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -21.22639    0.77068  -27.54   <2e-16 ***
age           1.63197    0.05895   27.68   <26 ***
    Null deviance: 3693.884  on 24  degrees of freedom
Residual deviance:   26.703  on 23  degrees of freedom
AIC: 114.76
```

Ajustement logistique



Interprétation de la fonction logistique comme fonction de répartition

Soit X l'âge à laquelle un fille choisie au hasard atteint sa puberté — une variable aléatoire dont la fonction de répartition $P(X \leq x)$ représente la probabilité qu'une fille observée à l'âge x ait déjà atteint sa puberté. Selon le modèle logistique, $\pi(x) = \dfrac{e^{\beta_o+\beta_1 x}}{1+e^{\beta_o+\beta_1 x}}$. Donc la fonction logistique est la fonction de répartition de X.

Matrice de covariance du vecteur $\hat{\boldsymbol{\beta}}$

```
> V<-summary(a)$cov.unscaled
> V
             (Intercept)          age
(Intercept)  0.59395485 -0.045281754
age         -0.04528175  0.003475466
```

Parmi les paramètres qu'on pourrait vouloir estimer, considérons la médiane x_o , le point tel que $F(x_o) = 0{,}5$. C'est une fonction des paramètres :

$$x_o = -\beta_o/\beta_1.$$

On estime x_o par $\hat{x}_o = -\hat{\beta}_o/\hat{\beta}_1$. Les calculs donnent $\hat{x}_o = 13{,}0065940797$.

En utilisant la méthode delta, on a à peu près

$$V(-\hat{\beta}_o/\hat{\beta}_1) \approx \frac{1}{\beta_1^2}\mathrm{Var}(\hat{\beta}_o) + \left(\frac{\beta_o}{\beta_1^2}\right)^2 \mathrm{Var}(\hat{\beta}_1) - 2\left(\frac{\beta_o}{\beta_1^3}\right)\mathrm{Cov}(\hat{\beta}_o, \hat{\beta}_1)$$

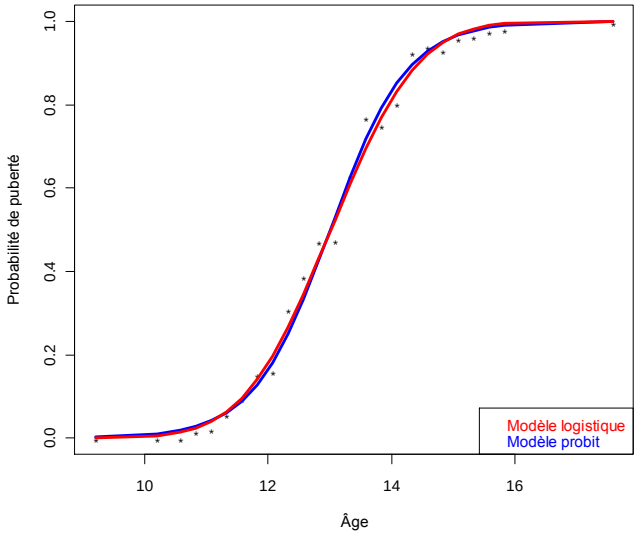
En remplaçant dans la formule ci-dessus β_1 et β_o par leurs estimateurs, nous obtenons $V(-\hat{\beta}_o/\hat{\beta}_1) = 0{,}0014946$.

Probits

Si on avait suppose que la loi de X (l'âge à laquelle une fille choisie au hasard atteint sa puberté), alors $\pi(x) = \Phi(\beta_o + \beta_1 x)$. Voici la commande qui permet de traiter ce modèle :

```
> b<-glm(se~age,family=binomial(probit))
> summary(b)
             Estimate  Std. Error z value Pr(>|z|)
(Intercept) -11.81894    0.38702  -30.54   <2e-16 ***
age           0.90782    0.02955   30.72   <2e-16 ***
    Null deviance: 3693.884  on 24  degrees of freedom
Residual deviance:   22.887  on 23  degrees of freedom
```

Ici aussi on trouve que l'ajustement que le modèle logistique et le modèle *probit* sont également adéquats :



8.1.2 Plusieurs variables exogènes

Nous illustrons à présent certains modèles logistiques à plus d'une variable exogène. Les idées principales sont essentiellement les mêmes. Les observations $y_1, \ldots, y_k$ sont indépendantes, $y_i \sim \mathcal{B}(m_i\ ; \pi_i)$, où

$$\pi_i = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}}$$

où $\mathbf{x}_i = [x_{i1}\ ; x_{i2}\ ; \ldots\ ; x_{iq}]'$ est le vecteur des valeurs des variables exogènes correspondant au groupe i. La première composante x_{i1} étant généralement égale à 1 ; et $\boldsymbol{\beta}$ est vecteur de paramètres, $\boldsymbol{\beta} = \begin{bmatrix} \beta_o \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$ ($p = q$-1).

La fonction de vraisemblance est

$$L(\mathbf{y}\ ; \boldsymbol{\pi}(\boldsymbol{\beta})) = \prod_{i=1}^{k} \binom{m_i}{y_i} [\pi_i(\boldsymbol{\beta})]^{y_i} [1-\pi_i(\boldsymbol{\beta})]^{m_i - y_i}\ ,$$

et son logarithme est

$$\ell(\mathbf{y}\ ; \boldsymbol{\pi}(\boldsymbol{\beta})) \ = \ \sum_{i=1}^{k} \ln \binom{m_i}{y_i} \ + \ \sum_{i=1}^{k} \left[y_i \left(\mathbf{x}_i'\boldsymbol{\beta} \right) - m_i \ln\left(1+e^{\mathbf{x}_i'\boldsymbol{\beta}}\right) \right].$$

Exemple : variable exogène qualitative
Supposons qu'on traitait la variable âge dans le premier exemple comme une variable qualitative, en classant ses valeurs en 10 groupes. On observerait des données comme celles-ci :

	Groupe										
Groupe	1	2	3	4	5	6	7	8	9	10	Total
Nombre de Succès	1	1	2	3	4	5	5	10	8	4	43
Nombre de sujets	10	10	10	11	11	10	10	13	10	5	100

La matrice de succès/échecs est

```
> se
      succes echecs
 [1,]      1      9
 [2,]      1      9
 [3,]      2      8
 [4,]      3      8
 [5,]      4      7
 [6,]      5      5
 [7,]      5      5
 [8,]     10      3
 [9,]      8      2
[10,]      4      1
```

On définit le groupe comme une variable x ordonnée prenant les valeurs 1 à 10.

```
> x<-as.ordered(1:10)
```

Le modèle est saturé, mais afin de le réduire, nous invoquons la paramétrisation « poly », ce qui permet de choisir une séquence d'hypothèses réductrices. Voici les commandes appropriées et leurs résultats :

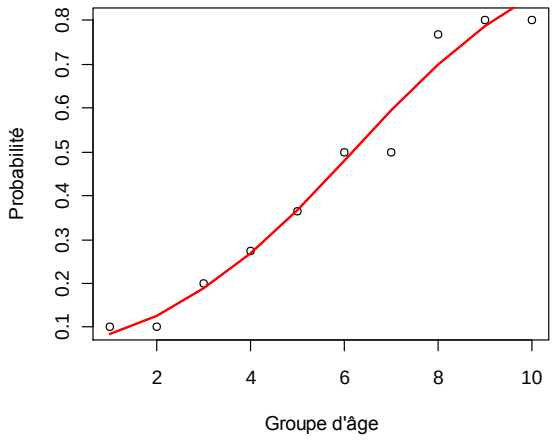
```
> modèle3<-glm(se~x,binomial)
> summary(modèle3)
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.33446    0.26090  -1.282    0.200
x.L          4.06199    0.97125   4.182 2.89e-05 ***
x.Q         -0.09785    0.92693  -0.106    0.916
x.C         -0.29626    0.88636  -0.334    0.738
```

```
x^4          -0.12477    0.87221  -0.143    0.886
x^5          -0.42720    0.84636  -0.505    0.614
x^6           0.12675    0.79658   0.159    0.874
x^7           0.08230    0.73721   0.112    0.911
x^8           0.38983    0.68426   0.570    0.569
x^9           0.23902    0.64390   0.371    0.710
    Null deviance: 2.9557e+01  on 9  degrees of freedom
Residual deviance: 2.4425e-15  on 0  degrees of freedom
AIC: 44.211
```

La sortie montre que seul le coefficient du terme linéaire est significativement différent de 0 et suggère par conséquent que l'on peut se permettre de traiter *x* comme une variable quantitative. On examine donc le modèle qui la traite ainsi :

```
> modèle4<-glm(se~x,binomial)
> summary(modèle4)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.8519     0.6243  -4.568 4.92e-06 ***
x             0.4624     0.1004   4.605 4.12e-06 ***
    Null deviance: 29.55679  on 9  degrees of freedom
Residual deviance:  0.92855  on 8  degrees of freedom
AIC: 29.139
```

On remarque que, bien que le deuxième modèle impose de sévères contraintes sur les paramètres du premier, sa valeur AIC lui est bien inférieure. Ceci reflète la pénalité que le critère AIC fait payer pour chaque paramètre du modèle.



Exemple : Cancer du sein

[Morrison, A. S., Black M.M, Lowe, C.R., MacMahon, B. and Yuasa, S [1973] Some international differences in histology and survival in breast cancer. *Int. J. Can*cer 11, 261-267.]

succes:	décédé
age	1 ≤ 50; 2 = 50-69; 3 ≥ 70
inflammation:	1= grosse inflammation; 0 = faible inflammation
apparence:	1 = maligne; 0 = bénigne
centre :	1= Tokyo; 2= Boston; 3= Glamorgan

succes	echecs	age	inflammation	apparence	centre	succes	echecs	age	inflammation	apparence	centre
9	26	1	0	1	1	3	10	2	1	1	2
7	68	1	0	0	1	2	3	2	1	0	2
4	25	1	1	1	1	9	15	3	0	1	2
3	9	1	1	0	1	18	26	3	0	0	2
9	20	2	0	1	1	3	1	3	1	1	2
9	46	2	0	0	1	0	1	3	1	0	2
11	18	2	1	1	1	16	16	1	0	1	3
2	5	2	1	0	1	7	20	1	0	0	3
2	1	3	0	1	1	3	8	1	1	1	3
3	6	3	0	0	1	0	1	1	1	0	3
1	5	3	1	1	1	14	27	2	0	1	3
0	1	3	1	0	1	12	39	2	0	0	3
6	11	1	0	1	2	3	10	2	1	1	3
7	24	1	0	0	2	0	4	2	1	0	3
6	4	1	1	1	2	3	12	3	0	1	3
0	1	1	1	0	2	7	11	3	0	0	3
8	18	2	0	1	2	3	4	3	1	1	3
20	58	2	0	0	2	0	1	3	1	0	3

```
> options("contrasts")
        unordered           ordered
"contr.treatment"      "contr.poly"
```

La variable « centre », ayant été déclarée comme facteur, cette paramétrisation comparera Boston et Glamorgan à Tokyo.

```
> modèle1<-glm(se~apparence+age+centre,family=binomial)
```

La matrice de design prend cette forme

```
e apparence age2 age3 centre2 centre3
1     1      0     0      0      0
1     0      0     0      0      0
.........................
1     1      0     1      0      1
1     0      0     1      0      1
```

```
> summary(modèle1)
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.65287    0.19496  -8.478  < 2e-16 ***
apparence    0.52071    0.16679   3.122  0.00180 **
age2         0.05431    0.18982   0.286  0.77477
age3         0.43149    0.24042   1.795  0.07269 .
centre2      0.55319    0.20899   2.647  0.00812 **
centre3      0.43758    0.21090   2.075  0.03801 *
    Null deviance: 58.230  on 35  degrees of freedom
Residual deviance: 33.679  on 30  degrees of freedom
```

Il ne semble pas y avoir de différence entre les groupes d'âge 1 et 2, mais il y en une entre le groupe 3 et le groupe 1. Le tableau ne permet par de comparer les groupes 2 et 3 entre eux.

De même, Boston et Glamorgan sont significativement différents de Tokyo, mais le tableau ne fournit pas un test comparant ces deux centres entre eux.

Voici quelques données que l'on peut extraire de l'objet « modèle1 »

Le vecteur des $\hat{\boldsymbol{\beta}}$

```
> modèle1$coefficients
(Intercept)   apparence        age2        age3     centre2     centre3
-1.65286865  0.52071260  0.05431372  0.43149188  0.27659624  0.43757814
```

Le vecteur des $\hat{\pi}_i$

```
> pichp<-modèle1$fitted.values
0.2437634 0.1607216 0.2437634 0.1607216 0.2539146 0.1681837 0.2539146 0.1681837
0.3316650 0.2276943 0.3316650 0.2276943 0.3591711 0.2498006 0.3591711 0.2498006
0.3717656 0.2601166 0.3717656 0.2601166 0.4631987 0.3389035 0.4631987 0.3389035
0.3330155 0.2287663 0.3330155 0.2287663 0.3451868 0.2384898 0.3451868 0.2384898
0.4346052 0.3135018 0.4346052 0.3135018
```

Le vecteur $\hat{\boldsymbol{\eta}} = \mathbf{X}\hat{\boldsymbol{\beta}}$

```
> eta<-modèle1$linear.predictors
-1.1321561 -1.6528687 -1.1321561 -1.6528687 -1.0778423 -1.5985549 -1.0778423
-1.5985549 -0.7006642 -1.2213768 -0.7006642 -1.2213768 -0.5789636 -1.0996762
-0.5789636 -1.0996762 -0.5246499 -1.0453625 -0.5246499 -1.0453625 -0.1474717
-0.6681843 -0.1474717 -0.6681843 -0.6945779 -1.2152905 -0.6945779 -1.2152905
-0.6402642 -1.1609768 -0.6402642 -1.1609768 -0.2630860 -0.7837986 -0.2630860
-0.7837986
```

La matrice de covariance estimée de $\hat{\boldsymbol{\beta}}$, $\mathbf{S} = \mathbf{X'VX}$, où $\mathbf{V}$ est la matrice diagonale dont les éléments sont $m_i\hat{\pi}_i(1-\hat{\pi}_i)$:

```
> round(summary(modèle1)$cov.scaled,5)
            (Intercept) apparence     age2     age3 centre2  centre3
(Intercept)     0.03801  -0.01542 -0.01768 -0.01456 -0.00948 -0.01745
apparence      -0.01542   0.02782  0.00154  0.00118  0.00147 -0.00192
age2           -0.01768   0.00154  0.03603  0.02225 -0.00328 -0.00512
age3           -0.01456   0.00118  0.02225  0.05780 -0.00716 -0.00871
centre2        -0.00948   0.00147 -0.00328 -0.00716  0.01092  0.01166
centre3        -0.01745  -0.00192 -0.00512 -0.00871  0.01166  0.04448
```

Ces données peuvent servir à tester des hypothèses non testées dans la sortie **R**. Une hypothèse linéaire $\mathbf{L}'\boldsymbol{\beta} = \mathbf{0}$ concernant les β peut être testée par la statistique $Q = [\mathbf{L}'\hat{\boldsymbol{\beta}}]'[\mathbf{L}'\mathbf{S}\mathbf{L}]^{-1}\mathbf{L}'\hat{\boldsymbol{\beta}}$, qui suit, sous l'hypothèse, une loi χ^2 à r degrés de liberté, où r est le nombre de colonnes de $\mathbf{L}$. Par exemple,

- L'hypothèse que les groupes d'âge 2 et 3 sont comparable est $\mathbf{L}'\boldsymbol{\beta} = 0$, où $\mathbf{L}'$ = [0,0,1,-1,0,0]. On obtient Q = 2,88, à 1 degré de liberté, ce qui correspond à une valeur p de 0,09.
- L'hypothèse que les centres Boston et Glamorgan sont comparables l'un à l'autre est $\mathbf{L}'\boldsymbol{\beta} = 0$, où $\mathbf{L}'$ = [0,0,0,0,1,-1]. On obtient Q = 0,417, à 1 degré de liberté, ce qui correspond à une valeur p de 0,519.
- L'hypothèse plus globale que l'âge n'est pas significatif est $\mathbf{L}'\boldsymbol{\beta}$, avec $\mathbf{L}' = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$. On obtient Q = 9,76 à 2 degrés de liberté, une valeur p de 0,0076, ce qui, bien sûr mène au rejet de l'hypothèse.

On considère un modèle dans lequel on ne fait pas de distinction entre Boston et Glamorgan.

```
> modèle2<-glm(se~apparence+age2+age3+I(centre2+centre3),binomial)
> summary(modèle2)
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.61443    0.18895  -8.544  < 2e-16 ***
apparence             0.53806    0.16567   3.248  0.00116 **
age2                  0.06341    0.18943   0.335  0.73781
age3                  0.43964    0.24038   1.829  0.06741 .
I(centre2 + centre3)  0.27310    0.10339   2.641  0.00825 **
```

Le deuxième groupe d'âge ne semblant pas trop s'éloigner du premier, on considère ces deux groupes comme un seul : on distingue donc seulement ceux de mons de 70 ans de ceux de 70 ans ou plus.

```
> modèle3<-glm(se~apparence+age3+I(centre2+centre3),binomial)
> summary(modèle3)
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.5828     0.1629  -9.715  < 2e-16 ***
```

```
apparence              0.5358     0.1655   3.238  0.00121 **
age3                   0.4007     0.2101   1.907  0.05649 .
I(centre2 + centre3)   0.2789     0.1019   2.736  0.00623 **

    Null deviance: 58.230  on 35  degrees of freedom
Residual deviance: 34.593  on 32  degrees of freedom
```

Voici les estimations des probabilités de cancer dans chaque catégorie d'apparence (maligne=1, bénigne = 0), d'âge (<70 = 0; ≥ 70 = 1), et de centre (Tokyo =0; Boston ou Glamorgan = 1).

```
> tapply(pichap,data.frame(L[,-1]),mean)
, , BosGlam = 0
            Âge
App..maligne         0         1
           0 0.1704039 0.2346810
           1 0.2598144 0.3438393
, , BosGlam = 1
            Âge
App..maligne         0         1
           0 0.2135111 0.2883957
           1 0.3168999 0.4091806
```

Matrice de covariance de $\hat{\boldsymbol{\beta}}$

```
> round(summary(modèle3)$cov.scaled,4)
                     (Intercept) apparence    age3 I(centre2 + centre3)
(Intercept)               0.0265   -0.0156 -0.0038               -0.0107
apparence                -0.0156    0.0274  0.0002                0.0017
age3                     -0.0038    0.0002  0.0441               -0.0051
I(centre2 + centre3)     -0.0107    0.0017 -0.0051                0.0104
```

Déviances

Trois modèles:

$\mathcal{M}_s$: le modèle saturé : $E(y_i/m_i) = \pi_i$

$\mathcal{M}_o$: le modèle nul : $E(y_i/m_i) = \pi$

$\mathcal{M}$: le modèle considéré : $E(y_i/m_i) = \dfrac{e^{\boldsymbol{\beta}'\mathbf{x}_i}}{1+e^{\boldsymbol{\beta}'\mathbf{x}_i}}$

La déviance compare un modèle donné au modèle saturé

`Null deviance: 58.230 on 35 degrees of freedom`	Teste le modèle nul $\mathcal{M}_o$ (1 paramètre) dans le cadre du modèle saturé $\mathcal{M}_s$ (36 cases, donc paramètres). Valeur p = 0,0081. L'hypothèse rejetée est l'hypothèse que les probabilités de cancer ne dépendent d'aucun des facteurs considérées.
`Residual deviance: 34.593 on 32 degrees of freedom`	On teste le modèle $\mathcal{M}$ (4 paramètres) dans le cadre du modèle saturé $\mathcal{M}_s$ (36 paramètres). Valeur p : 0,345. On n'a pas de raison de rejeter les hypothèses du modèle.
`Différence 58.230-34.59 = 23.64 à 3 degrés de liberté.`	On adopte donc le modèle $\mathcal{M}$ (4 paramètres) et dans le cadre de ce modèle on teste de $\mathcal{M}_o$ (36 paramètres). Valeur p : 0,0000297. On en déduit qu'au moins certains des facteurs invoqués pour expliquer l'incidence du cancer ne sont pertinents.

Remarque à propos de l'exemple *Les paramètres qui reflètent les différences entre les centres ont été considérés ici comme des paramètres fixes. Ce qui est raisonnable si l'intérêt porte particulièrement sur les trois centres impliqués, et ce qu'on veut savoir, c'est s'il y a des différences entre les trois centres. Mais il est probable que la question est plutôt « Est-ce que le taux de décès varie selon le centre où le traitement a eu lieu ? », la question portant sur tous les centres possibles. Dans ce cas, il faut envisager une modèle à effets aléatoires* (*non traité dans ces notes*). ■

8.1.3 Détails techniques

Quelques propriétés des fonctions de vraisemblance

Nous énonçons ci-dessous un certain nombre de propriétés des fonctions de vraisemblance, propriétés qu'on peut établir sous certaines conditions de régularité. Soit $L(\boldsymbol{\theta};\boldsymbol{y})$ la fonction de vraisemblance d'un échantillon $\boldsymbol{y} = (y_1; \ldots ; y_k)$:

$$L(\boldsymbol{\theta} ; \boldsymbol{y}) = \prod_{i=1}^{k} f_i(y_i;\boldsymbol{\theta}) .$$

On maximise normalement son logarithme $\log L = \sum_{i=1}^{k} \ln f_i(y_i;\boldsymbol{\theta})$ afin déterminer les estimateurs du maximum de vraisemblance. Donc soit $\boldsymbol{h}(\boldsymbol{\theta})$ le vecteur des dérivées partielles de $\log L$ par rapport aux paramètres et $\hat{\boldsymbol{\theta}}$ la valeur de $\boldsymbol{\theta}$ qui annule $\boldsymbol{h}(\boldsymbol{\theta})$ et maximise $\boldsymbol{h}$:

$$\boldsymbol{h}(\boldsymbol{\theta}) = \frac{\partial \ln L(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}} ; \quad \boldsymbol{h}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$$

On a

$$\mathrm{E}[\boldsymbol{h}(\boldsymbol{\theta})] = \mathbf{0}$$

Soit $\mathbf{H}$ la matrice des dérivées secondes de $\log L$

$$\mathbf{H}(\boldsymbol{\theta}) = \frac{\partial^2 \ln L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

La matrice d'information est

$$\mathcal{I}(\boldsymbol{\theta}) = -\mathrm{E}[\mathbf{H}(\boldsymbol{\theta})] = -\mathrm{E}\left[\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]$$

$\mathcal{I}(\boldsymbol{\theta})$ fournit une approximation de la matrice de covariance de l'estimateur $\hat{\boldsymbol{\theta}}$:

$$\mathbf{V}(\hat{\boldsymbol{\theta}}) \approx \mathcal{I}^{-1}(\boldsymbol{\theta})$$

On peut montrer que

$$\mathcal{I}(\boldsymbol{\theta}) = \mathrm{E}[\boldsymbol{h}(\boldsymbol{\theta})\boldsymbol{h}'(\boldsymbol{\theta})]$$

On a également ceci:

$$\mathbf{V}[\boldsymbol{h}(\boldsymbol{\theta})] = \mathrm{E}[\boldsymbol{h}(\boldsymbol{\theta})\boldsymbol{h}'(\boldsymbol{\theta})] = \mathcal{I}(\boldsymbol{\theta})$$

Méthodes itératives de solutions

Newton-Raphson

$\boldsymbol{h}(\hat{\boldsymbol{\theta}}) \approx \boldsymbol{h}(\boldsymbol{\theta}_o) + \mathbf{H}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)$ donne la procédure itérative suivante: $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mathbf{H}^{-1}(\boldsymbol{\theta}_t)\boldsymbol{h}(\boldsymbol{\theta}_t)$

Méthode des scores

On remplace $\mathbf{H}$ par son espérance $\boldsymbol{\theta}_{m+1} = \boldsymbol{\theta}_m + \mathcal{I}^{-1}(\boldsymbol{\theta}_m)\boldsymbol{h}(\boldsymbol{\theta}_m)$

Tests d'hypothèse

Statistique pour tester l'hypothèse

$$\mathrm{H_o} : \boldsymbol{\theta} = \boldsymbol{\theta}_o$$

La statistique de Wald

Puisque, asymptotiquement, $\hat{\boldsymbol{\theta}} \sim \mathcal{N}\left(\boldsymbol{\theta}; \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta})\right)$, la statistique $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)$ est à peu près de loi χ^2, ce qui demeure vrai lorsqu'on remplace $\boldsymbol{\mathcal{I}}^1(\boldsymbol{\theta})$ par $\boldsymbol{\mathfrak{I}}^{-1}(\hat{\boldsymbol{\theta}})$ si l'échantillon est grand et on peut se servir alors de la statistique $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)$.

La statistique des scores

Sachant que, asymptotiquement, $\boldsymbol{h}(\boldsymbol{\theta}) \sim \mathcal{N}\left(\mathbf{0}; \boldsymbol{\mathcal{I}}(\boldsymbol{\theta})\right)$ la statistique de score est $\boldsymbol{h}'(\boldsymbol{\theta}_o)\boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta}_o)\boldsymbol{h}(\boldsymbol{\theta}_o)$ est à peu près de loi χ^2 sous H_o et peut donc servir également à tester H_o.

Le test du rapport de vraisemblances

Soit L_o la fonction de vraisemblance définie par une hypothèse H_o qui consiste en une restriction de l'espace paramétrique d'un modèle dont la fonction de vraisemblance est L. La statistique basée sur le rapport de vraisemblances L_o/L,

$$-2\ln\left(\frac{\max L_o}{\max L}\right)$$

suit asymptotiquement, sous H_o, une loi χ^2_r, où r est la réduction de la dimension de l'espace paramétrique imposée par H_o.

Application au modèle logistique

Le modèle général est

$$\eta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i'\boldsymbol{\beta}, \quad i = 1, \ldots k$$

où

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}.$$

Lorsqu'on a m_i observations qui correspondent à une même valeur de $\boldsymbol{x}_i$, le nombre de succès y_i parmi les m_i fait partie des statistiques exhaustives, qui sont

$$y_i \sim \mathcal{B}(m_i\,;\,\pi_i), \quad i = 1, \ldots, k$$

La fonction de vraisemblance est :

$$L(\boldsymbol{y}\,;\,\boldsymbol{\pi}(\boldsymbol{\beta})) = \prod_{i=1}^{k}\binom{m_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{m_i-y_i},$$

son logarithme est

$$\ell(\boldsymbol{y}\,;\,\boldsymbol{\pi}(\boldsymbol{\beta})) = \sum_{i=1}^{k}\ln\binom{m_i}{y_i} + \sum_{i=1}^{k}\left[y_i\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right) - m_i\ln\left(1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}\right)\right].$$

où $\boldsymbol{\pi}(\boldsymbol{\beta}) = (\pi_1(\boldsymbol{\beta}), \ldots, \pi_k(\boldsymbol{\beta}))$.

Le vecteur des dérivées partielles de $\ell(\boldsymbol{y}\,;\,\boldsymbol{\pi}(\boldsymbol{\beta}))$ par rapport à $\boldsymbol{\beta}$ est

$$\boldsymbol{h}(\boldsymbol{\beta}) = \frac{\partial}{\partial\boldsymbol{\beta}}\ell(\boldsymbol{y};\boldsymbol{\pi}(\boldsymbol{\beta})) = \mathbf{X}'(\boldsymbol{y}-\mathbf{M}\boldsymbol{\pi})$$

où $\mathbf{M}$ est la matrice diagonale dont les éléments sont les m_i.

La matrice des dérivées secondes est donc

$$\mathbf{H}(\boldsymbol{\beta}) = -\mathbf{X}'\mathbf{V}\mathbf{X}$$

où $\mathbf{V}$ est la matrice diagonale dont les éléments sont $m_i\pi_i(1-\pi_i)$.

La méthode de Newton-Raphson donne

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t - \mathbf{H}^{-1}(\boldsymbol{\beta}_t)\boldsymbol{h}(\boldsymbol{\beta}_t) = \boldsymbol{\beta}_t + (\mathbf{X}'\mathbf{V}_t\mathbf{X})^{-1}\mathbf{X}'(\boldsymbol{y} - \mathbf{M}\boldsymbol{\pi}_t),$$

où $\mathbf{V}_t$ est la matrice diagonale dont les éléments sont $m_i\hat{\pi}_{it}(1-\hat{\pi}_{it})$ et $\hat{\pi}_{it} = \dfrac{e^{\boldsymbol{x}_i\hat{\boldsymbol{\beta}}_t}}{1+e^{\boldsymbol{x}_i\hat{\boldsymbol{\beta}}_t}}$.

Une façon d'interpréter cette procédure est celle-ci :

On approche la fonction $\ln\dfrac{\boldsymbol{p}}{\boldsymbol{e}-\boldsymbol{p}}$ par

$$\ln\frac{\boldsymbol{p}}{\boldsymbol{e}-\boldsymbol{p}} = \ln\frac{\boldsymbol{\pi}_t}{\boldsymbol{e}-\boldsymbol{\pi}_t} + \frac{\partial}{\partial p}\ln\frac{\boldsymbol{p}}{\boldsymbol{e}-\boldsymbol{p}}\bigg|_{\boldsymbol{p}=\boldsymbol{\pi}_t} (\boldsymbol{p}-\boldsymbol{\pi}_t), \text{ où } \boldsymbol{p} = \boldsymbol{y}/\boldsymbol{m}.$$

Donc

$$\ln\frac{\boldsymbol{p}}{\boldsymbol{e}-\boldsymbol{p}} = \ln\frac{\hat{\boldsymbol{\pi}}_t}{\boldsymbol{e}-\hat{\boldsymbol{\pi}}_t} + \frac{1}{\hat{\boldsymbol{\pi}}_t(\boldsymbol{e}-\hat{\boldsymbol{\pi}}_t)}(\boldsymbol{p}-\boldsymbol{\pi}_t)$$

$$= \mathbf{X}\boldsymbol{\beta}_t + \frac{1}{\hat{\boldsymbol{\pi}}_t(\boldsymbol{e}-\hat{\boldsymbol{\pi}}_t)}(\boldsymbol{p}-\hat{\boldsymbol{\pi}}_t),$$

où $\dfrac{1}{\hat{\boldsymbol{\pi}}_t(\boldsymbol{e}-\hat{\boldsymbol{\pi}}_t)}(\boldsymbol{p}-\hat{\boldsymbol{\pi}}_t)$ serait de moyenne nulle et de matrice de covariance $\mathbf{V}_t^{-1}$ si $\boldsymbol{\pi}_t$ était égal à $\boldsymbol{\pi}$.

L'estimateur de $\boldsymbol{\beta}$ est donc

$$(\mathbf{X}'\mathbf{V}_t\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_t\left[\mathbf{X}\boldsymbol{\beta}_t + \frac{1}{\hat{\boldsymbol{\pi}}_t(\boldsymbol{e}-\hat{\boldsymbol{\pi}}_t)}(\boldsymbol{p}-\hat{\boldsymbol{\pi}}_t)\right] = \boldsymbol{\beta}_t + (\mathbf{X}'\mathbf{V}_t\mathbf{X})^{-1}\mathbf{X}'(\boldsymbol{y}-\boldsymbol{m}\hat{\boldsymbol{\pi}}_t).$$

La déviance

Soit $\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}(\hat{\boldsymbol{\beta}})$ le vecteur $\boldsymbol{\pi}$ évalué au point $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, l'estimateur du maximum de vraisemblance de $\boldsymbol{\beta}$ dans un certain modèle $\mathcal{M}$. La vraisemblance maximale est

$$L_m = L(\boldsymbol{y}\,;\,\hat{\boldsymbol{\pi}}) = \prod_{i=1}^{k}\binom{m_i}{y_i}\hat{\pi}_i^{y_i}(1-\hat{\pi}_i)^{m_i-y_i}$$

et son logarithme est

$$\ell_m = \ell(\boldsymbol{y}\,;\,\hat{\boldsymbol{\pi}}) = \sum_{i=1}^{k}\ln\binom{m_i}{y_i} + \sum_{i=1}^{k}\left[y_i\ln\hat{\pi}_i + (m_i-y_i)\ln(1-\hat{\pi}_i)\right].$$

L'estimation de $\boldsymbol{\pi}$ est soumise à des contraintes, puisque $\boldsymbol{\pi}$ est fonction d'un nombre plus restreint de paramètres $\boldsymbol{\beta}$. L'estimateur non contraint de $\boldsymbol{\pi}$ est le vecteur des fréquences observées $\boldsymbol{p} = (y_1/m_1, \ldots, y_k/m_k)$. Le logarithme de la fonction de vraisemblance en ce point est

$$\ell_s = \ell_s(\boldsymbol{y}\,;\,\boldsymbol{p}) = \sum_{i=1}^{k}\ln\binom{m_i}{y_i} + \sum_{i=1}^{k}\left[y_i\ln p_i + (m_i-y_i)\ln(1-p_i)\right].$$

$\ell(\boldsymbol{y}\,;\,\boldsymbol{p})$ est le maximum du log de la vraisemblance d'un modèle $\mathcal{M}_s$ appelé *modèle saturé*. C'est le modèle où le nombre de paramètres est égal au nombre k d'observations binomiales. Il est évident que $\ell_s = \ell(\boldsymbol{y}\,;\,\boldsymbol{p}) \geq \ell(\boldsymbol{y}\,;\,\hat{\boldsymbol{\pi}}) = \ell_m$. La différence $\ell(\boldsymbol{y}\,;\,\boldsymbol{p}) - \ell(\boldsymbol{y}\,;\,\hat{\boldsymbol{\pi}})$ est une mesure de l'effet des contraintes; elle est

d'autant plus importante que les p_i s'éloignent des $\hat{\pi}_i$. Une fonction de cette différence, appelée *déviance* et définie par

$$\textit{Déviance}\ (\mathcal{M}) = D(\boldsymbol{p}\ ;\ \hat{\boldsymbol{\pi}}) = 2[\ell(\boldsymbol{y}\ ;\ \boldsymbol{p}) - l(\boldsymbol{y}\ ;\ \hat{\boldsymbol{\pi}})] = 2\sum_{i=1}^{k}\left[y_i \ln\left(\frac{p_i}{\hat{\pi}_i}\right) + (m_i - y_i)\ln\left(\frac{1-p_i}{1-\hat{\pi}_i}\right)\right],$$

sert à tester le modèle $\mathcal{M}$ par rapport à $\mathcal{M}_s$: sous le modèle $\mathcal{M}$,

$$D(\boldsymbol{p}\ ;\ \hat{\boldsymbol{\pi}}) = -2\sum_{i=1}^{k}\left(y_i \ln\frac{p_i}{\hat{\pi}_i} + (m_i - y_i)\ln\frac{1-p_i}{1-\hat{\pi}_i}\right) \xrightarrow{L} \chi^2_{k-r}$$

r étant le nombre de paramètres dans le modèle $\mathcal{M}$.

L'approche de $D(\boldsymbol{y}\ ;\hat{\boldsymbol{\pi}})$ vers la loi χ^2, cependant, n'est vraie que lorsque les m_i sont grands. En pratique, ce sont surtout des *différences* de déviances qu'on emploie pour tester des hypothèses dans le cadre d'un modèle $\mathcal{M}$. Considérons une hypothèse H_o emboîtée dans le modèle $\mathcal{M}$. Elle définit un certain sous-modèle $\mathcal{M}_o$ qui impose de nouvelles contraintes sur les paramètres. La déviance de ce sous-modèle mesure la distance qui l'éloigne du modèle saturé, ce qui n'est pas normalement ce qu'on cherche à mesurer ; dans le cadre d'un modèle donné, c'est l'écart entre $\mathcal{M}_o$ et $\mathcal{M}$ qui est pertinent, et donc on calcule la différence entre les deux déviances. Si $\hat{\boldsymbol{\beta}}_o$ est l'estimateur du maximum de vraisemblance dans $\mathcal{M}_o$, et $\hat{\boldsymbol{\pi}}_o = \boldsymbol{\pi}(\hat{\boldsymbol{\beta}}_o)$, alors la différence de déviances

$$\begin{aligned} G^2 &= \textit{Déviance}\ (\mathcal{M}_o)\ - \textit{Déviance}\ (\mathcal{M}) \\ &= D(\boldsymbol{p}\ ;\ \hat{\boldsymbol{\pi}}_o) - D(\boldsymbol{p}\ ;\ \hat{\boldsymbol{\pi}}) \\ &= 2[l(\boldsymbol{y}\ ;\ \hat{\boldsymbol{\pi}}) - l(\boldsymbol{y}\ ;\ \hat{\boldsymbol{\pi}}_o)] \\ &= 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}_i}{\hat{\pi}_{oi}}\right) + (m_i - y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-\hat{\pi}_{oi}}\right)\right] \end{aligned}$$

est égale à $-2\log\lambda$, où λ est le rapport des vraisemblances maximales :

$$\lambda = \frac{\prod_{i=1}^{k}\binom{m_i}{y_i}\hat{\pi}_{oi}^{\,y_i}(1-\hat{\pi}_{oi})^{m_i-y_i}}{\prod_{i=1}^{k}\binom{m_i}{y_i}\hat{\pi}_i^{\,y_i}(1-\hat{\pi}_i)^{m_i-y_i}} = \frac{\prod_{i=1}^{k}\hat{\pi}_{oi}^{\,y_i}(1-\hat{\pi}_{oi})^{m_i-y_i}}{\prod_{i=1}^{k}\hat{\pi}_i^{\,y_i}(1-\hat{\pi}_i)^{m_i-y_i}} = \prod_{i=1}^{k}\left(\frac{\hat{\pi}_{oi}}{\hat{\pi}_i}\right)^{y_i}\left(\frac{1-\hat{\pi}_{oi}}{1-\hat{\pi}_i}\right)^{m_i-y_i}$$

Lorsque H_o est vraie,

$$G^2 \xrightarrow{\mathcal{L}} \chi^2_\nu$$

le nombre de degrés de liberté ν étant la différence entre le nombre de paramètres dans le modèle et le nombre de paramètres sous H_o.

Dans le modèle logistique à une variable, l'hypothèse nulle est que $\eta_i = \beta_o$, et donc chaque p_i est estimé par p, la fréquence globales des incidents cardio-vasculaires. Alors

$$L = \prod_{i=1}^{k}\left(\frac{p}{\hat{\pi}_i}\right)^{y_i}\left(\frac{1-p}{1-\hat{\pi}_i}\right)^{m_i-y_i},$$

et

$$G^2 = 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}_i}{p}\right) + (n_i - y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-p}\right)\right]$$

Le khi-deux de Pearson

La statistique de Pearson est une autre mesure définie par

$$\chi^2 = \sum_{i=1}^{k}\frac{(y_i - m_i\hat{\pi}_i)^2}{\hat{\pi}_i(m_i - m_i\hat{\pi}_i)} = \sum_{i=1}^{k}\frac{m_i(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)}$$

où les $\hat{\pi}_i$ sont les estimateurs des p_i sous le modèle et les p_i sont les estimations dans le modèle saturé (les fréquences observées). On peut montrer aisément que

$$\chi^2 = \sum_{i=1}^{k}\frac{(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i} + \sum_{i=1}^{k}\frac{(y_i - m_i\hat{\pi}_i)^2}{m_i(1-\hat{\pi}_i)} = \sum_{i=1}^{k}\frac{(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i} + \sum_{i=1}^{k}\frac{\left[(m_i - y_i) - (m_i - m_i\hat{\pi}_i)\right]^2}{m_i(1-\hat{\pi}_i)}.$$

la statistique χ^2 habituelle. Cette statistique est asymptotiquement équivalente à la déviance du modèle. Lorsque les n_i sont grands, χ^2 suit une loi χ^2_ν, où ν est le nombre de restrictions imposées par le modèle sur les k paramètres du modèle saturé :

$$\nu = k - r,$$

r étant le nombre de paramètres du modèle. Dans l'exemple traité dans cette section, le nombre de valeurs distinctes de la variable *âge* est k = 43. Le modèle de régression ayant deux paramètres, r = 1 et ν = 41. La valeur de χ^2 est 21,304, pas trop éloignée de G^2 = 23,754.

8.2 Modèle de Poisson

Les observations $y_1, y_2, \ldots, y_n$ sont indépendantes, de loi de Poisson de paramètres $\lambda_1, \lambda_2, \ldots, \lambda_n$ respectivement, et

$$\eta_i = \ln \lambda_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots, \beta_p x_{ip}$$

La fonction de vraisemblance est $L = \prod_{i=1}^{n}\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} = \prod_{i=1}^{n}\frac{e^{-e^{\boldsymbol{\beta}'\mathbf{x}_i}}e^{-\boldsymbol{\beta}'\mathbf{x}_i y_i}}{y_i!} = \frac{e^{-\sum_{i=1}^{n}e^{\boldsymbol{\beta}'\mathbf{x}_i}}e^{-\sum_{i=1}^{n}\boldsymbol{\beta}'\mathbf{x}_i y_i}}{y_i!}$,

$$\ln L = -\sum_{i=1}^{n}e^{\boldsymbol{\beta}'\mathbf{x}_i} + \sum_{i=1}^{n}\boldsymbol{\beta}'\mathbf{x}_i y_i - \sum_{i=1}^{n}\ln y_i!$$

$$\boldsymbol{h}(\boldsymbol{\beta}) = \frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \mathbf{X}'(\mathbf{y} - \boldsymbol{\lambda}), \text{ où } \boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_n]'.$$

$$\mathbf{H}(\boldsymbol{\beta}) = \left[\frac{\partial^2 \ln L}{\partial\beta_j \partial\beta_k}\right] = -\mathbf{X}'\boldsymbol{\Lambda}\mathbf{X}, \text{ où } \boldsymbol{\Lambda} = \begin{bmatrix}\lambda_1 & \cdots & 0\\ \vdots & \ddots & 0\\ 0 & \cdots & \lambda_n\end{bmatrix}.$$

Le paramètre $\boldsymbol{\beta}$ est estimé par la méthode du maximum de vraisemblance, la valeur maximale étant déterminé par des méthodes itératives basées sur $\boldsymbol{h}$ et $\mathbf{H}$.

Exemple *Cancer et usine de traitement de déchets nucléaires*
Le tableau en annexe présente le nombre de cas de cancer dans 94 localités. L'analyse a pour but de déterminer si le taux de cancer est fonction de la distance x entre la localité et une usine de traitement de déchets nucléaires. Si y_i est le nombre de cas de cancer dans la localité i et x_i la distance par rapport à l'usine, alors le modèle est

$$y_i \sim \text{Poisson}(\lambda_i)$$

et le lien entre λ_i et x_i est

$$\ln \lambda_i = \beta_o + \beta_1 x_i$$

La commande permettant d'obtenir une estimation des paramètres avec **R** est la suivante :

```
> glm(Cancers~Distance,family=poisson)
(Intercept)     Distance
   0.186865    -0.006138
Degrees of Freedom: 93 Total (i.e. Null);  92 Residual
Null Deviance:      149.5
Residual Deviance: 146.6        AIC: 262.4
```

On estime donc que

$$\ln \lambda_i = 0{,}186865 - 0{,}006138 x_i$$

$\hat{\beta}_1$ est négatif, comme il se doit. Nous devons maintenant déterminer si on peut affirmer avec confiance que $\beta_1 < 0$.

```
> summary(glm(Cancers~Distance,poisson))
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.186865   0.188728   0.990   0.3221
Distance    -0.006138   0.003667  -1.674   0.0941 .
    Null deviance: 149.48  on 93  degrees of freedom
Residual deviance: 146.64  on 92  degrees of freedom
```

La valeur *p* n'est ni trop grande ni trop petite. On pourrait à la rigueur affirmer l'existence d'une relation, en admettant toutefois que notre niveau de confiance est plutôt faible.

Exemple *Cholestérol et maladies cardiovasculaires*
Dans cet exemple, les variables exogènes sont dichotomiques : le taux de cholestérol (normal ou excessif) et la maladie cardiovasculaire (présente ou absente). C'est un exemple simple qui peut facilement être traité par des moyens élémentaires, mais nous le présentons ici pour illustrer un modèle appelé *modèle loglinéaire*, utile dans l'analyse de tableaux de contingence comme le suivant :

		Maladie cardiovasculaire		
		Présente	Absente	
Cholestérol	< 260	$y_{11} = 51$	$y_{12} = 992$	$y_{1\cdot} = 1043$
	≥ 260	$y_{21} = 41$	$y_{22} = 245$	$y_{2\cdot} = 286$
		$y_{\cdot 1} = 92$	$y_{\cdot 2} = 1237$	$y_{\cdot\cdot} = 1329$

À moins que les effectifs des marges aient été fixés d'avance, on peut considérer les quatre observations y_{ij} comme des variables indépendantes. On les suppose distribués selon une loi de Poisson :

$$y_{ij} \sim \text{Poisson}(\lambda_{ij}), \quad i = 1,2 \ ; j = 1, 2.$$

L'approche vise à traiter ce problème d'une manière analogue à une analyse de variance à deux facteurs. On suppose que

$$\ln \lambda_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}.$$

On a donc les paramètres suivants :

		Maladie cardiovasculaire		
		Présente	Absente	
Cholestérol	< 260	$\lambda_{11} = e^{\mu+\alpha_1+\beta_1+\gamma_{11}}$	$\lambda_{12} = e^{\mu+\alpha_1+\beta_2+\gamma_{12}}$	$\lambda_{1\cdot} = e^{\mu+\alpha_1}[e^{\beta_1+\gamma_{11}} + e^{\beta_2+\gamma_{12}}]$
	≥ 260	$\lambda_{21} = e^{\mu+\alpha_2+\beta_1+\gamma_{21}}$	$\lambda_{22} = e^{\mu+\alpha_2+\beta_2+\gamma_{22}}$	$\lambda_{1\cdot} = e^{\mu+\alpha_2}[e^{\beta_1+\gamma_{21}} + e^{\beta_2+\gamma_{22}}]$
		$\lambda_{\cdot 1} = e^{\mu+\beta_1}[e^{\alpha_1+\gamma_{11}} + e^{\alpha_2+\gamma_{21}}]$	$\lambda_{\cdot 2} = e^{\mu+\beta_2}[e^{\alpha_1+\gamma_{12}} + e^{\alpha_2+\gamma_{22}}]$	

Les paramètres doivent être soumis à des contraintes telles celles introduites en analyse de variance. Les contraintes les plus communes sont

$$\sum_{i=1}^{a}\alpha_i = \sum_{j=1}^{b}\beta_j = 0,\ \sum_{i=1}^{a}\gamma_{ij} = 0 \text{ pour tout } j;\ \sum_{j=1}^{b}\gamma_{ij} = 09 \text{ pour tout } i.$$

Lorsqu'on tient compte de ces contraintes on obtient ceci :

		Maladie cardiovasculaire		
		Présente	Absente	
Cholestérol	< 260	$\lambda_{11} = e^{\mu+\alpha_1+\beta_1+\gamma_{11}}$	$\lambda_{12} = e^{\mu+\alpha_1-\beta_1-\gamma_{11}}$	$\lambda_{1.} = e^{\mu+\alpha_1}[e^{\beta_1+\gamma_{11}} + e^{-\beta_1-\gamma_{11}}]$
	≥ 260	$\lambda_{21} = e^{\mu-\alpha_1+\beta_1-\gamma_{11}}$	$\lambda_{22} = e^{\mu-\alpha_1-\beta_1+\gamma_{11}}$	$\lambda_{1.} = e^{\mu-\alpha_1}[e^{\beta_1-\gamma_{11}} + e^{-\beta_1+\gamma_{11}}]$
		$\lambda_{.1} = e^{\mu+\beta_1}[e^{\alpha_1+\gamma_{11}} + e^{-\alpha_1-\gamma_{11}}]$	$\lambda_{.2} = e^{\mu-\beta_1}[e^{\alpha_1-\gamma_{11}} + e^{-\alpha_1+\gamma_{11}}]$	

Les γ_{ij}, qui désignent les interactions dans un modèle ANOVA, sont ici les paramètres de dépendance entre le taux de cholestérol et la maladie cardiovasculaire. Nous allons montrer que l'hypothèse d'indépendance, en fonction des paramètres, s'exprime comme

$$H_o : \gamma_{11} = 0.$$

Conditionnellement, étant donné $y_{1.}$ et $y_{2.}$ variables y_{11} et de y_{21} sont de loi binomiale de paramètres $(\pi_1 ; y_{1.})$ et $(\pi_2 ; y_{2.})$ (voir l'exercice 8.17), avec $\pi_1 = \lambda_{11}/\lambda_{1.}$ et $\pi_2 = \lambda_{21}/\lambda_{2.}$. L'hypothèse d'indépendance affirme que $\pi_1 = \pi_2$, ce qui est équivalent à

$$H_o : \frac{\lambda_{11}\lambda_{22}}{\lambda_{12}\lambda_{21}} = 1.$$

Compte tenu des contraintes,

$$\frac{\lambda_{11}\lambda_{22}}{\lambda_{12}\lambda_{21}} = e^{4\gamma_{11}}$$

d'où l'hypothèse d'indépendance est équivalente à $\gamma_{11} = 0$, ce qui équivaut à « tous les γ_{ij} sont nuls ».

Les modèle qui inclut tous les paramètres est un modèle saturé.

On présente à R les valeurs de trois variables, y, mcv (pour « maladie cardiovasculaire »), et cholest :

y	mcv	cholest
51	1	1
992	0	1
41	1	0
245	0	0

Voici comment construire le modèle saturé :

```
> summary(glm(y~mcv*cholest,family=poisson))
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.50126    0.06389  86.108   <2e-16 ***
mcv         -1.78769    0.16874 -10.595   <2e-16 ***
cholest      1.39846    0.07134  19.602   <2e-16 ***
mcv:cholest -1.18021    0.22156  -5.327    1e-07 ***
    Null deviance: 1.6582e+03  on 3  degrees of freedom
Residual deviance: 2.2204e-15  on 0  degrees of freedom
AIC: 35.406
```

Modèle additif, équivalent au modèle d'indépendance :

```
> summary(glm(y~mcv+cholest,family=poisson))
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.58425    0.05960   93.69   <2e-16 ***
mcv         -2.59866    0.10806  -24.05   <2e-16 ***
cholest      1.29386    0.06675   19.38   <2e-16 ***
    Null deviance: 1658.18  on 3  degrees of freedom
Residual deviance:   26.43  on 1  degrees of freedom
AIC: 59.836
```

Dans le modèle additif, estimations des moyennes

```
> mod2$fitted.values
        1         2         3         4
 72.20166 970.79834  19.79834 266.20166
```

Estimation des λ sous le modèle additif :

		Maladie cardiovasculaire		
		Présente	Absente	
Cholestérol	< 260	72,20	970,80	1043
	≥ 260	19,80	266,20	286
		92	1237	1329

8.3 Surdispersion dans le modèle de Poisson

Soit $y_i \sim \mathcal{P}(\lambda_i)$, $\lambda_i = \lambda(\boldsymbol{x}_i) = e^{\boldsymbol{\beta}'\boldsymbol{x}_i}$. Alors, $E(y_i) = \lambda_i$ et, théoriquement, $\text{Var}(y_i) = \lambda_i$. Or il arrive souvent que $\text{Var}(y_i) > \lambda_i$, un phénomène connue sous le nom de *surdispersion*. Nous décrivons ici une façon relativement simple d'estimer la variance de y_i sous l'hypothèse que $\text{Var}(y_i) = \phi\lambda_i$, où $\phi = 1+\alpha$, appelé paramètre de surdispersion, est un paramètre à estimer. Dans ces conditions,

$$\text{Var}\left(\frac{y_i - \lambda_i}{\sqrt{\lambda_i}}\right) = \frac{\text{Var}(y_i)}{\lambda_i} = \frac{\phi\lambda_i}{\lambda_i} = \phi .$$

Donc $E\left(\frac{(y_i - \lambda_i)^2}{\lambda_i}\right) = \phi$, et on estime ϕ par

$$\hat{\phi} = \frac{1}{n-q}\sum_{i=1}^{n}\frac{(y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i}$$

8.4 Modèle multinomial

Soit $\boldsymbol{x} \sim \mathcal{M}(n\,;\boldsymbol{\pi})$, où $\pi = \boldsymbol{\pi}(\boldsymbol{\theta})$ est de *dimension k* mais $\boldsymbol{\theta}$ est de dimension $q < k$. Soit $\boldsymbol{p} = (1/n)\boldsymbol{x}$. La fonction de vraisemblance est

$$L(\boldsymbol{\theta}) = \frac{n!}{\prod_{i=1}^{k}(np_i)!}\prod_{i=1}^{k}\pi_i^{np_i}$$

et

$$\ell(\boldsymbol{\theta}) = \ln L = \ln\left(\frac{n!}{\prod_{i=1}^{k}(np)_i!}\right) + \sum_{i=1}^{k} np_i \ln(\pi_i)$$

Soit $\boldsymbol{h}(\boldsymbol{\theta}) = \frac{\partial\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$. La u^e composante de $\boldsymbol{h}(\boldsymbol{\theta})$ est

$$\boldsymbol{h}_u(\boldsymbol{p};\boldsymbol{\theta}) = \sum_{i=1}^{k}\frac{np_i}{\pi_i}\frac{\partial\pi_i}{\partial\theta_u} .$$

On peut écrire $\boldsymbol{h}(\boldsymbol{\theta}) = n\mathbf{Z}'\mathbf{D}^{-1}\boldsymbol{p}$, où $\mathbf{D} = \text{diag}(\pi_i)$ et $\mathbf{Z}$ est une matrice dont la $(i\,,\,u)^{\,e}$ composante est $\frac{\partial\pi_i}{\partial\theta_u}$:

$$\mathbf{Z} = \left(\frac{\partial\pi_i}{\partial\theta_u}\right) = \left(\frac{\partial\boldsymbol{\pi}}{\partial\boldsymbol{\theta}}\right)$$

L'estimateur du maximum de vraisemblance est la solution $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{g}(\boldsymbol{p};\boldsymbol{\theta}) = 0$.

Par la méthode de Newton-Raphson, la suite des approximations successives devrait être

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \left(\frac{\partial \mathbf{g}(\boldsymbol{p};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^{-1} \boldsymbol{g}(\boldsymbol{p};\boldsymbol{\theta}),$$

mais la méthode dite de *scoring* remplace la matrice $\frac{\partial \mathbf{g}(\boldsymbol{p};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ par son espérance. L'élément (u, v) de $\frac{\partial \mathbf{g}(\boldsymbol{p};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ est

$$\frac{\partial \mathbf{g}_u(\boldsymbol{p};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_v} = \sum_{i=1}^{k} np_i \left\{ \frac{1}{\pi_i} \frac{\partial^2 \pi_i}{\partial \theta_u \partial \theta_v} - \frac{1}{\pi_i^2} \frac{\partial \pi_i}{\partial \theta_u} \frac{\partial \pi_i}{\partial \theta_v} \right\},$$

et son espérance est

$$\begin{aligned} \mathrm{E}\left(\frac{\partial \mathbf{g}_u(\boldsymbol{p};\boldsymbol{\theta})}{\partial \theta_v}\right) &= \mathrm{E}\left(\sum_{i=1}^{k} np_i \left\{ \frac{1}{\pi_i} \frac{\partial^2 \pi_i}{\partial \theta_u \partial \theta_v} \right\}\right) - \mathrm{E}\left(\sum_{i=1}^{k} np_i \left\{ \frac{1}{\pi_i^2} \frac{\partial \pi_i}{\partial \theta_u} \frac{\partial \pi_i}{\partial \theta_v} \right\}\right) \\ &= n \sum_{i=1}^{k} \frac{\partial^2 \pi_i}{\partial \theta_u \partial \theta_v} - n \sum_{i=1}^{k} \frac{1}{\pi_i} \frac{\partial \pi_i}{\partial \theta_u} \frac{\partial \pi_i}{\partial \theta_v}. \end{aligned}$$

Le premier terme est nul, car $\sum_{i=1}^{k} \pi_i = 1 \Rightarrow \sum_{i=1}^{k} \frac{\partial \pi_i}{\partial \theta_u} = 0 \Rightarrow \sum_{i=1}^{k} \frac{1}{\pi_i} \frac{\partial \pi_i}{\partial \theta_u} \frac{\partial \pi_i}{\partial \theta_v} = \frac{\partial}{\partial \theta_v}\left(\sum_{i=1}^{k} \frac{\partial \pi_i}{\partial \theta_u}\right) = 0.$

Écrivant le deuxième terme sous forme matricielle, nous avons

$$\mathrm{E}\left(\frac{\partial \mathbf{g}(\boldsymbol{p};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) = -n\mathbf{Z}'\mathbf{D}^{-1}\mathbf{Z}$$

Donc la solution s'obtient pas la suite d'approximations successives :

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + (n\mathbf{Z}'\mathbf{D}^{-1}\mathbf{Z})^{-1}\boldsymbol{g}(\boldsymbol{p};\boldsymbol{\theta}) = \boldsymbol{\theta}_t + (\mathbf{Z}'\mathbf{D}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{D}^{-1}\boldsymbol{p}.$$

La matrice $\mathbf{Z}$ et le vecteur $\mathbf{Z}'\mathbf{D}^{-1}\boldsymbol{p}$ sont évalués à $\boldsymbol{\theta}_t$ et $\mathbf{D}$ est évaluée à $\boldsymbol{\pi}_t = \boldsymbol{\pi}(\boldsymbol{\theta}_t)$.

Matrice de covariance

La matrice $[n\mathbf{Z}'\mathbf{D}^{-1}\mathbf{Z}]^{-1}$ tend en probabilité vers la matrice de covariance $\mathrm{V}(\hat{\boldsymbol{\theta}})$.

Esquisse de démonstration :

$$\mathbf{0} = \mathbf{g}(\boldsymbol{p};\ \hat{\boldsymbol{\theta}}) = \mathbf{g}(\boldsymbol{p};\boldsymbol{\theta}) + \left[\frac{\partial \mathbf{g}(\boldsymbol{p};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right](\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

$$\Rightarrow (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = -\left[\frac{\partial \mathbf{g}(\boldsymbol{p};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^{-1} \mathbf{g}(\boldsymbol{p};\boldsymbol{\theta}) \approx -\left[-n\mathbf{Z}'\mathbf{D}^{-1}\mathbf{Z}\right]^{-1}\left(n\mathbf{Z}'\mathbf{D}^{-1}\boldsymbol{p}\right) = \left[\mathbf{Z}'\mathbf{D}^{-1}\mathbf{Z}\right]^{-1}\left(\mathbf{Z}'\mathbf{D}^{-1}\boldsymbol{p}\right)$$

Puisque $\mathbf{Z}'\mathbf{D}^{-1}\boldsymbol{\pi} = 0$, on a aussi

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \left[\mathbf{Z}'\mathbf{D}^{-1}\mathbf{Z}\right]^{-1}\left(\mathbf{Z}'\mathbf{D}^{-1}(\boldsymbol{p} - \boldsymbol{\pi}\right)$$

Par conséquent,

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}) = \left[\mathbf{Z}'\mathbf{D}^{-1}\mathbf{Z}\right]^{-1}\mathbf{Z}'\mathbf{D}^{-1}\left[(1/n)(\mathbf{D} - \boldsymbol{\pi}\boldsymbol{\pi}')\right]\mathbf{D}^{-1}\mathbf{Z}\left[\mathbf{Z}'\mathbf{D}^{-1}\mathbf{Z}\right]^{-1} = \left[n\mathbf{Z}'\mathbf{D}^{-1}\mathbf{Z}\right]^{-1}$$

Ceci découle aussi d'un résultat général selon lequel l'inverse de la matrice $\mathrm{E}\left(-\frac{\partial^2 \log(L)}{\partial \boldsymbol{\theta}^2}\right)$ est une estimation de la variance d'un estimateur du maximum de vraisemblance. Or

$$\mathrm{E}\left(-\frac{\partial^2 \log(L)}{\partial \boldsymbol{\theta}^2}\right) = \mathrm{E}\left(-\frac{\partial \boldsymbol{h}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) = n\mathbf{Z}'\mathbf{D}^{-1}\mathbf{Z}.$$

On estime $\mathrm{V}(\hat{\boldsymbol{\theta}})$ en évaluant $n\mathbf{Z}'\mathbf{D}^{-1}\mathbf{Z}$ à $\hat{\boldsymbol{\theta}}$, la limite de la suite $\{\boldsymbol{\theta}_t\}$.

Annexe

Nombre de cas de cancer dans 94 localités et distance par rapport à une usine de traitement de déchets nucléaires

Cancers	Distance	Cancers	Distance	Cancers	Distance	Cancers	Distance	Cancers	Distance
0	0,4798	0	17,1569	0	41,9548	1	56,5481	1	79,6524
0	1,3869	0	17,8122	2	42,4602	0	58,7671	0	79,8780
6	2,2000	1	18,2744	2	42,8748	2	61,9104	0	79,9097
0	2,4225	0	18,8351	4	43,1508	0	64,0548	4	80,0000
1	2,5755	3	21,6664	0	43,2731	1	65,5515	1	86,6747
0	2,7669	2	22,4411	3	43,3566	1	66,1883	1	86,9465
0	3,2124	2	22,4810	0	43,6408	0	66,5539	0	88,7387
4	5,1150	1	22,7810	0	45,2371	0	69,7822	0	91,5030
0	5,9737	0	23,3224	1	47,0691	0	70,3247	1	92,4188
3	6,9036	0	23,7612	0	47,4623	0	70,5756	1	92,5825
0	7,0254	0	24,5598	3	47,6960	1	71,1742	2	94,0659
3	7,1303	3	26,2318	0	48,3813	2	71,2182	3	94,2659
2	7,4881	1	26,5255	2	51,6088	1	71,7595	1	94,4987
0	11,4695	1	27,5936	0	51,7080	0	72,7350	0	94,7842
0	12,8757	0	32,9644	0	51,7127	0	72,9987	0	94,9556
0	14,0739	0	34,5129	1	53,1356	0	73,0988	0	95,6401
1	15,5794	2	35,0853	1	53,8110	1	73,6102	0	97,0189
2	15,6532	0	40,9007	0	54,5641	0	73,7653	1	99,1136
1	15,8407	0	41,1479	2	55,1587	0	79,4807		