Régression de Poisson

Modèle de comptage

Ricco Rakotomalala

Université Lumière Lyon 2

Plan

1. Problématique – Modélisation – Estimation
2. Qualité de l'ajustement
3. Inférence statistique
4. Interprétation des coefficients
5. Sélection de variables
6. Etude des résidus – Points atypiques et influents
7. Surdispersion – Problème et solutions
8. Conclusion
9. Références

Principe et estimation des paramètres

RÉGRESSION DE POISSON

Principe de la Régression de Poisson

La régression de Poisson est un modèle de prédiction qui s'applique lorsque la variable cible Y est une variable de comptage (nombre d'apparition d'un évènement durant un laps de temps).

Distribution de la loi de Poisson

$$P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!} \quad (y = 0, 1, 2, \ldots)$$

λ est le paramètre de la loi de Poisson, où E[Y] = λ et V[Y] = λ

Objectif de la Régression de Poisson

Modéliser ce paramètre λ à l'aide des covariables (X_1, X_2, …, X_J) avec

$$\ln \lambda = a_0 + a_1 X_1 + \cdots + a_j X_j$$

De fait, $E[Y/X] = V[Y/X] = \exp(a_0 + a_1 X_1 + \cdots + a_j X_j)$

Données et notations

Nombre d'infidélités d'une personne (sur 1 année)

Affairs	Constante	Gender	YearsMarried	RatingMarriage
0	1	0	15	4
0	1	0	7	4
0	1	1	4	3
0	1	1	10	4
0	1	0	1.5	5
0	1	0	0.125	3
1	1	0	15	2
1	1	1	4	4
2	1	1	15	4
3	1	0	10	4
3	1	0	15	4
7	1	0	10	3
7	1	1	10	3
7	1	0	15	2
7	1	1	10	4
12	1	1	15	2
12	1	1	10	2
1	1	1	1.5	4
1	1	1	10	5
1	1	0	1.5	5

Gender : 1, homme ; 0, femme

Nombre d'années de mariage

Satisfaction dans l'union (1: faible ; 5 : très satisfait.e)

$n = 20$ observations, $p = 3$ descripteurs

$X = (1, X_1, X_2, ..., X_p)$, matrice de description

Attention au rôle de la constante

X_i description de l'individu n°i

Ex. $X_1 = (1, 0, 15, 4)$

Y_i valeur de la cible pour l'individu n°i

Ex. $Y_8 = 1$

Notation matricielle : $a_0 + a_1 X_{i,1} + ... = X_i a$

On souhaite modéliser le nombre d'infidélités (Y = Affairs) d'une personne sur l'année passée.

Fonction de vraisemblance

Puisque

$$P[Y = yi/Xi, a] = \frac{e^{-\exp(X_i a)}[\exp(X_i a)]^{y_i}}{y_i!}$$

On peut écrire la fonction de vraisemblance

$$L(a) = \prod_{i=1}^{n} \frac{e^{-\exp(X_i a)}[\exp(X_i a)]^{y_i}}{y_i!}$$

Et la log vraisemblance, que l'on doit maximiser en fonction de a

$$LL(a) = \sum_{i=1}^{n} y_i \times X_i a - \exp(X_i a) - \ln y_i!$$

LL1 LL2

Ne dépend pas de a, on peut laisser de côté pour le calcul de â.

Exemple sous Excel avec le ‘’solveur’’

Cellules variables

Coefficients	1.84038	0.75084	0.07634	-0.59515		
Affairs	Constante	Gender	YearsMarried	RatingMarriage	LL1	LL2
0	1	0	15	4	0.0000	1.8309
0	1	0	7	4	0.0000	0.9941
0	1	1	4	3	0.0000	3.0377
0	1	1	10	4	0.0000	2.6485
0	1	0	1.5	5	0.0000	0.3603
0	1	0	0.125	3	0.0000	1.0666
1	1	0	15	2	1.7951	6.0201
1	1	1	4	4	0.5160	1.6753
2	1	1	15	4	2.7113	3.8793
3	1	0	10	4	0.6694	1.2500
3	1	0	15	4	1.8144	1.8309
7	1	0	10	3	5.7280	2.2666
7	1	1	10	3	10.9839	4.8024
7	1	0	15	2	12.5658	6.0201
7	1	1	10	4	6.8178	2.6485
12	1	1	15	2	30.5514	12.7553
12	1	1	10	2	25.9713	8.7083
1	1	1	1.5	4	0.3251	1.3842
1	1	1	10	5	0.3788	1.4606
1	1	0	1.5	5	-1.0209	0.3603

SUM	99.8074	65

LL(a)	**34.8074**

Cellule cible à maximiser

Estimation en pratique 1 – Algorithme ''IRLS''

IRLS : Iterated Reweighted Least Squares. S'appuyer sur la régression pondérée. Fixer une valeur de départ, itérer jusqu'à convergence (déviance – à voir plus loin – stable, ou vecteur des coefficients stable). On parle aussi de **Fisher Scoring**.

Coefficient estimé à l'étape t : $\hat{a}^{(t)} = (X'WX)^{-1}X'Wz$

W : matrice diagonale où

$$w_{ii} = \hat{\lambda}_i = \exp\left[X_i \hat{a}^{(t-1)}\right]$$

Espérance estimée de la cible pour l'individu n°i

$$z = \ln \hat{\lambda}_i + \frac{y_i - \hat{\lambda}_i}{\hat{\lambda}_i}$$

Attention inversion de matrice, problème potentiel en cas de colinéarité.

Estimation en pratique 2 – Algorithme ''Newton-Raphson''

Algorithme d'optimisation itératif également, s'appuie sur deux informations supplémentaires : vecteur gradient, matrice hessienne

Passage de l'étape *t* à *t+1* : $a^{(t+1)} = a^{(t)} - H^{-1} \times g$

H est la matrice Hessienne (courbure locale de la fonction objectif)

g est le vecteur gradient (pente locale de la fonction objectif)

$$H(j_1, j_2) = \sum_{i=1}^{n} x_{i,j_1} x_{i,j_2} \hat{\lambda}_i \Rightarrow H = X'WX$$

$$g_j = \sum_{i=1}^{n} (y_i - \hat{\lambda}_i) x_{i,j} \Rightarrow g = X'(y - \hat{\lambda})$$

Son inverse correspond à la matrice de variance covariance des coefficients

$$\hat{V}(\hat{a}) = H^{-1}$$

Attention colinéarité !

A l'optimum (lorsque la solution â a été trouvée), $g = 0$

Différentes approches possibles pour définir la convergence.

Estimation – Un exemple sous R

```r
#importation des données
library(xlsx)
D <- read.xlsx("affairs_for_R.xlsx",sheetIndex=1)

#régression
reg <- glm(Affairs ~ Gender+YearsMarried+RatingMarriage, data = D, family = "poisson")

#esp. de Y estimé - Lambda^
print(reg$fitted.values)

##          1          2          3          4          5          6
##  1.8309084  0.9941470  3.0377357  2.6484522  0.3602838  1.0665972
##          7          8          9         10         11         12
##  6.0201366  1.6752519  3.8792892  1.2499902  1.8309084  2.2666083
##         13         14         15         16         17         18
##  4.8024406  6.0201366  2.6484522 12.7553352  8.7082695  1.3842040
##         19         20
##  1.4605697  0.3602838

sreg <- summary(reg)
print(sreg)

##
## Call:
## glm(formula = Affairs ~ Gender + YearsMarried + RatingMarriage,
##     family = "poisson", data = D)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5397  -1.4227  -0.3740   0.8893   2.5140
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.84038    0.73660   2.498  0.01247 *
## Gender          0.75084    0.26759   2.806  0.00502 **
## YearsMarried    0.07634    0.03418   2.233  0.02554 *
## RatingMarriage -0.59515    0.14437  -4.123 3.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 90.977  on 19  degrees of freedom
## Residual deviance: 44.588  on 16  degrees of freedom
## AIC: 95.089
##
## Number of Fisher Scoring iterations: 5

#mat. var.covar des coefs.
print(sreg$cov.unscaled)

##                (Intercept)       Gender YearsMarried RatingMarriage
## (Intercept)     0.54257513 -0.074329634 -0.020630975   -0.084548820
## Gender         -0.07432963  0.071603746  0.001815600    0.002104339
## YearsMarried   -0.02063097  0.001815600  0.001168516    0.002088234
## RatingMarriage -0.08454882  0.002104339  0.002088234    0.020841480

#ecart-type des coefs.
print(sqrt(diag(sreg$cov.unscaled)))

##    (Intercept)         Gender   YearsMarried RatingMarriage
##     0.73659699     0.26758876     0.03418356     0.14436579
```

$$\hat{\lambda}_i = \exp(X_i \hat{a})$$

Le même qu'avec le ''solveur'' d'Excel.

$$\hat{a} = \begin{pmatrix} 1.84038 \\ 0.75084 \\ 0.07634 \\ -0.59515 \end{pmatrix}$$

$$\hat{\sigma}_{\hat{a}} = \begin{pmatrix} 0.73660 \\ 0.26759 \\ 0.03418 \\ 0.14437 \end{pmatrix}$$

$$\hat{V}(\hat{a})$$

Goodness-of-fit

QUALITÉ DE L'AJUSTEMENT

Statistique de Pearson

Ajustement du modèle $\exp(X_i\hat{a})$

Statistique de Pearson

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

Suit une loi du χ^2 à (n – J – 1) degrés de liberté.

Si rejet de Ho (p-value < α), alors les valeurs prédites s'écartent significativement des valeurs observées. Le modèle est mal ajusté.

(Remarque : $df.residual = n – p – 1 = 16)

Notre modèle n'est pas bien ajusté.

```
#valeurs pour Stat. de Pearson
rp <- (D$Affairs-reg$fitted.values)^2 / reg$fitted.values
print(rp)

##          1          2          3          4          5          6
## 1.83090841 0.99414697 3.03773568 2.64845219 0.36028379 1.06659719
##          7          8          9         10         11         12
## 4.18624574 0.27217706 0.91040590 2.45004671 0.74650111 9.88481242
##         13         14         15         16         17         18
## 1.00558609 0.15948681 7.14982449 0.04472883 1.24427587 0.10664089
##         19         20
## 0.14523403 1.13587359

#Statistique de Pearson
khi2 <- sum(rp)
print(khi2)

## [1] 39.37996

#probabilité critique
print(pchisq(khi2,reg$df.residual,lower.tail=FALSE))

## [1] 0.0009583282
```

Statistique déviance

Statistique déviance

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

Avec : $0 \times \ln 0 = 0$

Suit une loi du χ^2 à (n – J – 1) degrés de liberté. Si rejet de Ho (p-value < α), le modèle est mal ajusté.

```
y <- D$Affairs

#valeurs pour Stat. Deviance
rd <- ifelse(D$Affairs==0,0,y*log(y/reg$fitted.values))-(y-reg$fitted.value
s)
print(rd)

##          1          2          3          4          5          6
## 1.83090841 0.99414697 3.03773568 2.64845219 0.36028379 1.06659719
##          7          8          9         10         11         12
## 3.22502661 0.15928836 0.55427966 0.87641995 0.31230855 3.15998743
##         13         14         15         16         17         18
## 0.43994193 0.07573800 2.45199550 0.02281942 0.55587220 0.05907877
##         19         20
## 0.08174312 0.38114705

#Statistique deviance
DS = 2 * sum(rd)
print(DS)

## [1] 44.58754
```

cf. summary() du modèle

Notre modèle n'est pas bien ajusté.

```
#probabilité critique
print(pchisq(DS,reg$df.residual,lower.tail=FALSE))

## [1] 0.0001605066
```

Pseudo-R²

Modèle réduit à la constante

(H0 : modèle trivial)

$$\hat{a}_0 = \ln \bar{y}$$

cf. annulation du gradient.

Statistique déviance sous H0

$$D_0 = 2\sum_{i=1}^{n} y_i \ln y_i - y_i \hat{a}_0 - \left(y_i - e^{\hat{a}_0}\right) = \cdots$$

Facile à simplifier…

Pseudo-R²

$$R^2 = \frac{D_0 - D}{D_0} = 1 - \frac{D}{D_0}$$

#1, modèle parfait

#0, pas mieux que modèle trivial

$R^2 \approx 0.51$

Pas terrible, vraiment.

```r
#estimation de la constante
a0 <- log(mean(D$Affairs))
print(a0)

## [1] 1.178655

#résidu déviance simplifié
rd0 <- ifelse(y==0,0,y *log(y)) - y * a0 - (y - exp(a0))

#Statistique deviance
D0 <- 2 * sum(rd0)
print(D0)

## [1] 90.97727

#pseudo-R²
R2 <- (D0-DS)/D0
print(R2)

## [1] 0.5099046
```

Test de significativité, intervalle de confiance

INFÉRENCE STATISTIQUE

Test du rapport de vraisemblance

Réaliser un test de significativité d'un ensemble de coefficient

$$\begin{cases} H_0 : a_{(1)} = a_{(2)} = \ldots = a_{(q)} = 0, q \leq p \\ H_1 : \exists\, j \,/\, a_{(j)} \neq 0 \end{cases}$$

La constante n'est pas intégrée dans ce type de test.

Rapport de vraisemblance : écart entre les déviances du modèle vs. du modèle réduit sous H0

$$LR = D(\text{modèle réduit}) - D(\text{modèle complet})$$

Sous H0, suit une loi du χ^2 à (q) degrés de liberté

Utilisé notamment pour tester la significativité globale du modèle

$$\begin{cases} H_0 : a_1 = a_2 = \ldots = a_p = 0 \\ H_1 : \exists\, j \,/\, a_j \neq 0 \end{cases}$$

```
#statistique du rapport de vraisemblance
LR <- D0 - DS
print(LR)

## [1] 46.38973

#probabilité critique
print(pchisq(LR,df=3,lower.tail = FALSE))

## [1] 4.686244e-10
```

Notre modèle est globalement significatif à 5%

Test de Wald

Puisque â est un estimateur du maximum de vraisemblance, il est asymptotiquement sans biais et suit une loi normale. On peut répondre aux mêmes tests avec la statistique suivante :

$$W_{(q)} = \hat{a}'_{(q)} \hat{V}^{-1}_{\hat{a}_{(q)}} \hat{a}_{(q)}$$

Sous H0, suit une loi du χ^2 à (q) degrés de liberté.

Sous-vecteur des paramètres à tester

Partie de la matrice de variance covariance concernée par le test.

Utilisé notamment pour tester la significativité individuelle des coefficients

$$\begin{cases} H_0 : a_j = 0 \\ H_1 : a_j \neq 0 \end{cases}$$

Statistique de test

$$z_{\hat{a}_j} = \frac{\hat{a}_j}{\hat{\sigma}_{\hat{a}_j}} \cong N(0,1)$$

$\hat{\sigma}^2_{\hat{a}_j}$ lue sur la diagonale de $\hat{V}_{\hat{a}}$

```
#summary de l'objet régression
print(summary(reg))

##
## Call:
## glm(formula = Affairs ~ Gender + YearsMarried + RatingMarriage,
##     family = "poisson", data = D)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.84038    0.73660   2.498  0.01247 *
## Gender          0.75084    0.26759   2.806  0.00502 **
## YearsMarried    0.07634    0.03418   2.233  0.02554 *
## RatingMarriage -0.59515    0.14437  -4.123 3.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Tous les coefficients semblent significatifs à 5%

Intervalle de confiance des coefficients

Puisque

$$\frac{\hat{a}_j - a_j}{\hat{\sigma}_{\hat{a}_j}} \cong N(0,1)$$

On peut en déduire l'intervalle de confiance des coefficients au niveau (1-α)

Quantile de la loi normale

$$\hat{a}_j \pm u_{1-\frac{\alpha}{2}} \times \hat{\sigma}_{\hat{a}_j}$$

$\hat{a}_j$

$\hat{\sigma}_{\hat{a}_j}$

Pour le niveau de confiance (1 - α) = 90%, on utilise le quantile à $u_{1 - \alpha/2} = u_{0.95}$

```
#récupération des coefficients dans l'objet summary.glm
coef <- sreg$coefficients[,1]

#écarts-type des coefs
se_coef <- sreg$coefficients[,2]

#quantile de la loi normale 0.95 (pour intervalle à 90%)
u <- qnorm(0.95)

#bornes basses
print("Bornes basses")

## [1] "Bornes basses"

print(coef - u * se_coef)

##    (Intercept)         Gender   YearsMarried RatingMarriage
##     0.62878379     0.31069534     0.02010836    -0.83260944

#bornes hautes
print("Bornes hautes")

## [1] "Bornes hautes"

print(coef + u * se_coef)

##    (Intercept)         Gender   YearsMarried RatingMarriage
##      3.0519723      1.1909840      0.1325623     -0.3576883

#
# utilisation de la fonction de R - normalité asymptotique
#
print(confint.default(reg,level=0.90))
##                              5 %       95 %
## (Intercept)     0.62878379  3.0519723
## Gender          0.31069534  1.1909840
## YearsMarried    0.02010836  0.1325623
## RatingMarriage -0.83260944 -0.3576883
```

Variables quantitatives, indicatrices

INTERPRÉTATION DES COEFFICIENTS

Ecart en logarithme – Variable binaire

Pour une variable binaire X_j, le coefficient $\hat{a}_j$ représente l'écart des logarithmes des nombres espérés (λ) selon l'apparition (X_j=1) ou pas (X_j=0) de la caractéristique.

$$\hat{a}_j = \ln \hat{\lambda}_{(X_j=1)} - \ln \hat{\lambda}_{(X_j=0)}$$

$$\hat{a}_j = \ln \hat{\lambda}_1 - \ln \hat{\lambda}_0$$

```
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.84038    0.73660   2.498  0.01247 *
## Gender           0.75084    0.26759   2.806  0.00502 **
## YearsMarried     0.07634    0.03418   2.233  0.02554 *
## RatingMarriage -0.59515    0.14437  -4.123 3.75e-05 ***
```

(Gender = 1, homme ; Gender = 0, femme)

A années de mariage et satisfaction égales, le logarithme du nombre moyen d'infidélité chez les hommes est supérieur de 0.75 à celui des femmes.

Puisque $\hat{a}_j = \ln \hat{\lambda}_1 - \ln \hat{\lambda}_0 = \ln \frac{\lambda_1}{\lambda_0}$

$$\Rightarrow e^{\hat{a}_j} = \frac{\lambda_1}{\lambda_0} = e^{0.75084} = 2.12$$

A années de mariage et satisfaction égales, les hommes trompent leur conjoint 2.12 fois plus souvent que les femmes.

Ecart en logarithme – Autre type de variable

Variable X_j quantitative, le coefficient $\hat{a}_j$ correspond à une différence en logarithme consécutive à l'augmentation d'une unité de X_j

$$\hat{a}_j = \ln \hat{\lambda}_{(X_j+1)} - \ln \hat{\lambda}_{(X_j)}$$

```
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.84038    0.73660   2.498  0.01247 *
## Gender          0.75084    0.26759   2.806  0.00502 **
## YearsMarried    0.07634    0.03418   2.233  0.02554 *
## RatingMarriage -0.59515    0.14437  -4.123 3.75e-05 ***
```

Variable X_j indicatrice d'une variable V à K modalités, le coefficient $\hat{a}_j$ correspond à une différence en logarithme de cette modalité avec la modalité de référence (celle qui a été omise)

$$\hat{a}_j = \ln \hat{\lambda}_{(V=X_j)} - \ln \hat{\lambda}_{(V=X_{réf.})}$$

Coefficients standardisés

Laquelle des deux variables à le plus d'impact ?

```
#régression avec les variables initiales
reg2 <- glm(Affairs ~ YearsMarried + RatingMarriage, data = D, family = "poisson")
print(summary(reg2))

##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     2.56902    0.68356   3.758 0.000171 ***
## YearsMarried    0.05700    0.03204   1.779 0.075259 .
## RatingMarriage -0.62071    0.14716  -4.218 2.47e-05 ***
## ---


#centrer et réduire les variables
Z <- data.frame(scale(D,center=TRUE,scale = TRUE))

#régression sur variables centrées et réduites (pas la cible !)
reg3 <- glm(D$Affairs ~ YearsMarried + RatingMarriage, data = Z, family = "poisson")
print(summary(reg3))

##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.8774     0.1595   5.500 3.81e-08 ***
## YearsMarried     0.2995     0.1683   1.779   0.0753 .
## RatingMarriage  -0.6199     0.1470  -4.218 2.47e-05 ***
## ---


#écarts-type des variables concernées
sd_vars <- sapply(D[,c('YearsMarried','RatingMarriage')],sd)
print(sd_vars)

##   YearsMarried RatingMarriage
##      5.2541141      0.9986833

#coefficients standardisés
print(reg2$coefficients[2:3]*sd_vars)

##   YearsMarried RatingMarriage
##      0.2994724     -0.6198965
```

Pour apprécier les contributions relatives des variables quantitatives dans un modèle, une stratégie simple consiste à centrer et (surtout) réduire les variables (qui sont sans unités par conséquent). On a alors les coefficients ''bêta'' (β_j) qui sont interprétables sous la forme de différence en écart-type de X_j.

Nous pouvons retrouver les coefficients (β_j) en multipliant les a_j par les écarts-type des variables σ_j

La satisfaction dans le mariage influe comparativement plus que les années de mariage sur le nombre des infidélités !

Optimisation des critères AIC et BIC

SÉLECTION DE VARIABLES

Intérêt de la sélection de variables

La sélection de variables – ne conserver que les variables explicatives pertinentes – est importante pour (1) mieux situer les phénomènes de causalité (interprétation du modèle) ; (2) assurer la robustesse du modèle (principe de parcimonie, rasoir d'Occam).

L'automatiser devient indispensable dès lors que le nombre de variables candidates augmente, le nombre de combinaisons devient important. On adopte souvent des démarches pas à pas, ascendantes (forward), descendantes (backward) ou bidirectionnelles.

Les approches basées sur des tests statistiques (test de Wald) constituent une réponse possible, mais avec le danger des faux positifs induits par les comparaisons multiples (en multipliant les tests, on augmente les chances d'intégrer à tort une variable, il faut restreindre le risque α).

Une alternative est de s'appuyer sur un critère qui effectue un arbitrage entre la qualité de l'ajustement (la log-vraisemblance) et la complexité du modèle (nombre de paramètres), les critères AIC (Akaike) ou BIC (Schwarz) sont souvent utilisés dans ce contexte.

Critère Akaike

Formule LL(a) en page 6

$$AIC(a) = -2 \times LL(a) + 2 \times (p+1)$$

Critère Schwarz

$$BIC(a) = -2 \times LL(a) + \ln(n) \times (p+1)$$

Plus restrictif dès que $\ln(n) > 2$ c.-à-d. conduit à sélectionner moins de variables.

Sélection – Un exemple

```
#régression avec l'ensemble des variables
#disponibles dans la base initiale
reg.all <- glm(Affairs ~ ., data = D, family = "poisson")
print(summary(reg.all))

## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -5.34374    2.74755  -1.945  0.05179 .
## Gender          0.32785    0.64691   0.507  0.61230
## Age            -0.05048    0.03226  -1.565  0.11766
## YearsMarried    0.19393    0.08137   2.383  0.01716 *
## Children        0.68167    0.66429   1.026  0.30482
## Religiousness  -0.05947    0.13429  -0.443  0.65786
## Education       0.66216    0.15818   4.186 2.84e-05 ***
## Occupation     -0.80025    0.26715  -2.996  0.00274 **
## RatingMarriage -0.68795    0.15141  -4.544 5.53e-06 ***
## ---
```

Régression avec l'ensemble des variables disponibles.

```
#Sélection backward - critère BIC (k = ln(n))
library(MASS)
reg.sel <- stepAIC(reg.all,direction="backward",k=log(nrow(D)))

## Start:  AIC=85.47
## Affairs ~ Gender + Age + YearsMarried + Children + Religiousness +
##     Education + Occupation + RatingMarriage
##
##                 Df Deviance     AIC
## - Religiousness  1   16.204  82.671
## - Gender         1   16.270  82.737
## - Children       1   17.082  83.548
## - Age            1   18.491  84.958
## <none>               16.008  85.471
##
## Step:  AIC=82.67
## Affairs ~ Gender + Age + YearsMarried + Children + Education +
##     Occupation + RatingMarriage
##
##                 Df Deviance     AIC
## - Gender         1   16.331  79.802
## - Children       1   17.731  81.202
## - Age            1   18.887  82.358
## <none>               16.204  82.671
##
## Step:  AIC=79.8
## Affairs ~ Age + YearsMarried + Children + Education + Occupation +
##     RatingMarriage
##
##                 Df Deviance     AIC
## - Age            1   19.019  79.495
## <none>               16.331  79.802
##
## Step:  AIC=79.49
## Affairs ~ YearsMarried + Children + Education + Occupation +
##     RatingMarriage
##
##                 Df Deviance     AIC
## <none>               19.019  79.495

print(summary(reg.sel))

##
## Call:
## glm(formula = Affairs ~ YearsMarried + Children + Education +
##     Occupation + RatingMarriage, family = "poisson", data = D)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -6.11705    2.14347  -2.854  0.00432 **
## YearsMarried    0.11287    0.05339   2.114  0.03451 *
## Children        0.94685    0.42396   2.233  0.02553 *
## Education       0.61117    0.13832   4.418 9.94e-06 ***
## Occupation     -0.66775    0.21523  -3.102  0.00192 **
## RatingMarriage -0.68984    0.14717  -4.687 2.77e-06 ***
## ---
```

Valeur de départ du BIC (avec toutes les variables).

Le retrait de ''Religiousness'' entraîne la plus forte baisse du BIC

Aucun retrait de variable n'induit une réduction du BIC

Modèle final.

Etude des résidus.

RÉSIDUS – POINTS ATYPIQUES ET INFLUENTS

Résidus

L'étude des résidus permet de diagnostiquer la régression, pour détecter les régularités (problème de spécification), ou identifier les points isolés (atypiques ou mal modélisés).

Résidus bruts. Ecart entre endogène observée et la prédiction de E(Y) = λ du modèle.

$$r_i = y_i - \hat{\lambda}_i \qquad \text{où } \hat{\lambda}_i = \exp(X_i \hat{a})$$

Résidus de Pearson. Ecart normalisé par l'écart-type. En effet, E(Y) = V(Y) = λ pour la loi de Poisson

$$rp_i = \frac{r_i}{\sqrt{\hat{\lambda}_i}}$$

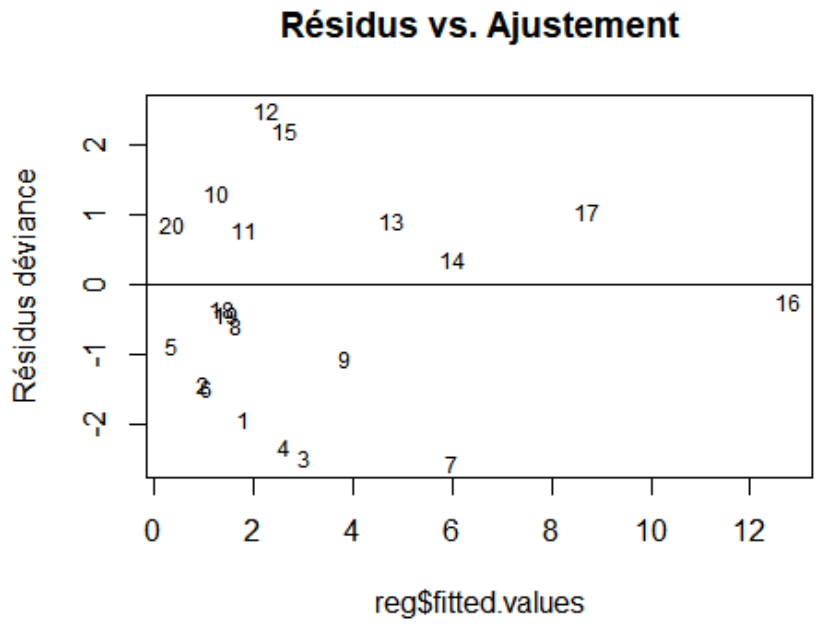
Résidus déviance. Qui est une composante (pour l'individu n°i) de la Statistique Déviance utilisée pour évaluer la modélisation (Goodness-of-fit)

$$rd_i = sgn(r_i) \times \sqrt{2 \times \left[y_i \ln \frac{y_i}{\hat{\lambda}_i} - (r_i) \right]}$$

Résidus - Illustration

```
#résidus déviance
rd <- residuals(reg)

#graphique
plot(reg$fitted.values,rd,main="Résidus vs. Ajustement",ylab="Résidus déviance",type="n")
text(reg$fitted.values,rd,label=1:20,cex=0.75)
abline(h=0)
```



Comme la statistique déviance

$$D = \sum_{i=1}^{n} rd_i^2$$

On peut identifier les observations *à contre-courant* dans la modélisation, qui pèsent négativement dans l'ajustement (rd_i^2 élevés par rapport aux autres)

Ex. n°7, une femme avec bcp d'années de mariage (years = 15), pas très heureux (rating = 2), et pourtant relativement fidèle (affairs = 1). Mais qu'est-ce qu'elle attend ? $\hat{\lambda}_6 = 6.02$

Ex. n°15, un homme avec quelques années de mariage (years = 10), assez heureux (rating = 4), et pourtant chaud lapin (affairs = 7). On s'y attendait, mais pas à ce point-là : $\hat{\lambda}_{15} = 2.64$

Levier

Le levier d'un individu h_i indique son éloignement dans l'espace de représentation, mais aussi son influence globale dans la régression. Elle est lue dans la diagonale de la matrice :

$$H = W^{1/2}X'(X'WX)^{-1}XW^{1/2}$$

On peut aussi la calculer directement pour l'individu n°i :

$$h_i = \hat{\lambda}_i X_i (X'WX)^{-1} X_i'$$

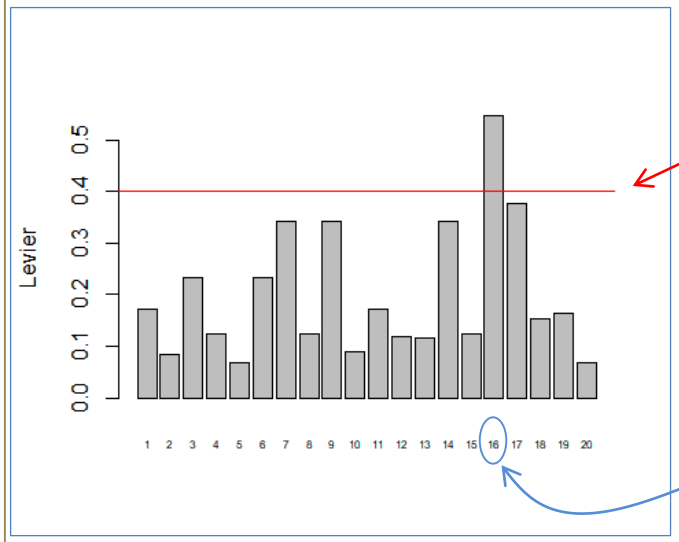
Où $X_i = (1, X_{i,1}, X_{i,2}, \ldots, X_{i,p})$

```
#seuil du levier
seuil_levier <- 2*(3+1)/20
print(seuil_levier)

## [1] 0.4

#Levier
levier <- influence.measures(reg)$infmat[,'hat']

#graphique
barplot(levier,ylab="Levier",cex.names=0.5)
abline(h=seuil_levier,col="red")
```



Puisque $\sum_{i=1}^{n} h_i = p + 1$, on prend empiriquement comme seuil de suspicion $2 \times \frac{p+1}{n}$ (2 fois la moyenne)

L' individu n°16 est un homme qui endure 15 années de mariage malheureux (rating = 2). **Très différent des autres individus de la base**. Pourtant il est bien modélisé (résidu faible, cf. page précédente), parce qu'il décharge à tire larigot (affairs = 12).

Problème et solutions

SURDISPERSION

Surdispersion (overdispersion)

Surdispersion ? Elle survient lorsque la variance de la variable cible est largement supérieure à sa moyenne (invalidant l'hypothèse de la loi de Poisson).

Conséquence ? Les écarts-type des coefficients sont sous-estimés, faussant les tests de significativité (les coefficients semblent tous significatifs)

Diagnostic ? Le ratio entre la statistique de Pearson (χ^2) et les degrés de liberté (n – p – 1) s'éloigne ''significativement'' de 1 (des tests plus ''formels'' existent, cf. Hilbe, section 7.4).

Solution 1. Améliorer le modèle en introduisant de nouvelles variables, ou des interactions, ou en effectuant des transformations, ou en traitant les données atypiques…

Solution 2. Corriger l'estimation de l'écart-type des coefficients (les coefficients estimés ne sont pas modifiés) avec une procédure très simple issu du ratio calculé ou en introduisant un paramètre supplémentaire dans la régression (quasi-poisson).

Traitement de la surdispersion - Exemple

```
#summary de la régression
print(summary(reg))

## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.84038    0.73660   2.498  0.01247 *
## Gender           0.75084    0.26759   2.806  0.00502 **
## YearsMarried     0.07634    0.03418   2.233  0.02554 *
## RatingMarriage -0.59515    0.14437  -4.123 3.75e-05 ***
## ---

#ratio - Statistique de Pearson / degré de liberté
ratio_surdispersion <- (khi2/reg$df.residual)
print(ratio_surdispersion)

## [1] 2.461248

#correction à introduire
se_correction <- sqrt(ratio_surdispersion)

#récupérer le tableau des coefficients
coef_table <- sreg$coefficients

#modifier les écarts-type estimés
coef_table[,2] <- sreg$coefficients[,2] * se_correction

#et les autres colonnes
coef_table[,3] <- coef_table[,1]/coef_table[,2] #z-value
coef_table[,4] <- 2*pnorm(abs(coef_table[,3]),lower.tail = FALSE)

#print du nouveau tableau des coefficients
print(round(coef_table,5))

##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.84038    1.15560 1.59257  0.11126
## Gender           0.75084    0.41980 1.78855  0.07369
## YearsMarried     0.07634    0.05363 1.42341  0.15462
## RatingMarriage -0.59515    0.22649 -2.62775  0.00860

#AUTRE APPROCHE -- modélisation avec quasi-poisson
reg.qp <- glm(Affairs ~ Gender + YearsMarried + RatingMarriage, data = D, family = "quasipoisson")
print(summary(reg.qp))

## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.84038    1.15565   1.593   0.1308
## Gender           0.75084    0.41982   1.788   0.0926 .
## YearsMarried     0.07634    0.05363   1.423   0.1738
## RatingMarriage -0.59515    0.22650  -2.628   0.0183 *
## ---
## (Dispersion parameter for quasipoisson family taken to be 2.46146)
```

La loi de distribution a changé.

A 5%, toutes les variables semblent pertinentes.

$$\tau = \frac{\chi^2}{n-p-1} = \frac{39.37996}{16} = 2.461248$$

$$\sqrt{\tau} = \sqrt{2.461} = 1.5688$$

$$\hat{\sigma}_{\hat{a}_j}^{corrigé} = \hat{\sigma}_{\hat{a}_j} \times \sqrt{\tau}$$

Conclusion : (1) La correction ad-hoc et ''quasi-poisson'' donnent des résultats très similaires ; (2) finalement, seule RatingMarriage semble pertinente dans le modèle.

''Quasi-poisson'', on modélise :

$E(Y/X) = \lambda$

$V(Y/X) = \theta \times \lambda$

θ est un paramètre supplémentaire, facteur de surdispersion, estimé à partir des données.

CONCLUSION

Régression de Poisson - Conclusion

- Adaptée dès lors que la variable cible Y représente un comptage.
- S'inscrit dans le cadre du modèle linéaire généralisé. Estimateur du maximum de vraisemblance, avec ses bonnes propriétés asymptotiques.
- On retrouve les outils habituels de la régression (interprétation des coefficients, tests de significativité, sélection de variables, étude des résidus et identification des points atypiques).
- Attention au problème de surdispersion, heureusement des solutions simples existent.
- Attention à la valeur (Y = 0). Peut-être symptomatique de deux phénomènes (0 : les personnes fidèles, 0 : des godelureaux qui n'ont pas eu l'occasion de sévir sur la période étudiée). Des solutions existent (ex. Régression Binomiale Négative, Zero-inflated Poisson,...).

RÉFÉRENCES

Références

Deux références s'imposent vraiment !

- ''STAT 501 – Regression Methods", PennState, Eberly College of Science, https://newonlinecourses.science.psu.edu/stat501/ (Poisson Regression - Lesson 15.4)
- J.M. Hilbe, ''Negative Binomial Regression", Second Edition, Cambridge University Press, 2011.