Régression de Poisson

ZHANG Mudong & LI Siheng & HU Chenyang

21 Mars, 2013

Plan

- Composantes des modèles
- Estimation
- Qualité d'ajustement et Tests
- Exemples
- Conclusion

Introduction de modèle linéaire généralisé

- La relation d'une variable par rapport à une ou plusieurs autres
- Regrouper tous les modèles linéaire
- Nelder et Wedderburn(1972)
- Logiciel:SAS(*genmod*) et R(*glm*)

Composantes des modèles

Distribution

On suppose que l'échantillon statistique est constitué de n variables aléatoires $\{Y_i; i = 1, \ldots, n\}$ indépendantes admettant des distributions issues d'une structure exponentielle. Cela signifie que les lois de ces variables sont dominées par une même mesure dite de référence et que la famille de leurs densités par rapport à cette mesure se met sous la forme:

$$f(y_i, \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - v(\theta_i)}{u(\phi)} + w(y_i, \phi)\right\}$$

Le paramètre θ_i est appelé *paramètre naturel* de la famille exponentielle.

Pour certaines lois, la fonction u est de la forme :

$$u(\phi) = \frac{\phi}{\omega_i}$$

où les poids ω_i sont les poids connus des observations, fixés ici à 1 pour simplifie; ϕ est appelé alors paramétre de dispersion

L'expression de la structure exponentielle se met alors sous *la forme canonique* en posant :

$$Q(\theta) = \frac{\theta}{\phi}$$

$$a(\theta) = exp\left\{-\frac{v(\theta)}{\phi}\right\}$$

$$b(y) = exp\left\{w(y, \phi)\right\}$$

on obtient

$$f(y_i, \theta_i) = a(\theta_i)b(y_i)exp\left\{y_i Q(\theta_i)\right\}$$

Prédicteur linéaire

Des variables explicatives construisent la matrice $\mathbf{X} = (x_1, \ldots, x_p)^T$, soit β un vecteur de p paramètres. Alors le prédicteur linéaire, composante déterministe du modèle est le vecteur à n composantes :

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

Lien

Une *relation fonctionnelle* entre la composante aléatoire et le prédicteur linéaire. Soit $\{\mu_i = E(Y_i); i = 1, \ldots, n\}$, on pose

$$\eta_i = g(\mu_i) \quad i = 1, \ldots, n$$

où g, appelée *fonction lien*. Ceci revient donc à écrire un modèle dans lequel une *fonction de la moyenne* appartient au sous-espace engendré par les variables explicatives :

$$g(\mu_i) = x_i'\beta = \theta_i \quad i = 1, \ldots, n$$

Loi de Bernouilli

Considérons n variables aléatoires binaires indépendantes Z_i de probabilité de succès i et donc d'espérance $E(Z_i) = \pi_i$. Les fonctions de densité de ces variables sont éléments de la famille :

$$f(z_i, \pi_i) = \pi_i^{z_i}(1-\pi_i)^{1-z_i} = (1-\pi_i)exp\left\{z_i ln\frac{\pi_i}{1-\pi_i}\right\}$$

qui est la forme canonique d'une structure exponentielle de paramètre naturel

$$\theta_i = ln\frac{\pi_i}{1-\pi_i}$$

Cette relation définit la fonction *logit* pour fonction lien canonique associée à ce modèle.

Loi de Poisson

On considère n variables indépendantes Y_i de loi de Poisson de paramètre $\mu_i = E(Y_i)$. Les Y_i sont par exemple les effectifs d'une table de contingence. Ces variables admettent pour densités :

$$f(y_i, \mu_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} = exp\{-\mu_i\} \frac{1}{y_i!} exp\{y_i \ln \mu_i\}$$

qui sont issues d'une structure exponentielle et, mises sous la forme canonique, de paramètre naturel

$$\theta_i = ln\mu_i$$

définissant comme fonction lien canonique le logarithme pour ce modèle.

Estimation

L'estimation des paramètres β_j est calculée en maximisant la log-vraisemblance du modèle linéaire généralisé. Pour n observations supposées indépendantes et en tenant compte que dépend θ de β, la log-vraisemblance s'écrit

$$\mathfrak{L}(\beta) = \sum_{i=1}^{n} \ln f(y_i; \theta_i, \phi) = \sum_{i=1}^{n} l(\theta_i; \phi; y_i)$$

Calculons

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}$$

$$\frac{\partial l_i}{\partial \theta_i} = \left[y_i - v'(\theta_i)\right] / u(\phi) = (y_i - \mu_i)/u(\phi)$$

$$\frac{\partial \mu_i}{\partial \theta_i} = v''(\theta_i) = Var(Y_i)/\mu(\phi)$$

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij} \quad \text{car} \quad \eta_i = x_i'\beta$$

$$\frac{\partial \mu_i}{\partial \eta_i} \quad \text{depend de la fonction lien} \quad \eta_i = g(\mu_i)$$

Les équations de la vraisemblance sont :

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\text{Var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0 \qquad j = 1, \ldots, p$$

Qualité d'ajustement

Il s'agit d'évaluer la qualité d'ajustement du modèle sur la base des différences entre observations et estimations. Plusieurs critères sont proposés.

- Déviance
- Test de Pearson

Déviance

Le modèle estimé est comparé avec le modèle dit *saturé*.

Cette comparaison est basée sur l'expression de la déviance D des log-vraisemblances $\mathcal{L}$ et $\mathcal{L}_{sat}$:

$$D = -2(\mathcal{L} - \mathcal{L}_{sat})$$

qui est le logarithme du carré du rapport des vraisemblances. D suit une loi du χ^2 à $n - p$ degrés de liberté ce qui permet de construire un test de rejet ou d'acceptation du modèle.

Test de Pearson

Un test du χ^2 est également utilisé pour comparer les valeurs observées y_i à leur prévision par le modèle. La statistique du test est définie par

$$X^2 = \sum_{i=1}^{I} \frac{(y_i - \hat{\mu}_i)^2}{\widehat{var}(\hat{\mu}_i)}$$

(μ_i est remplacé par $n_i\pi_i$ dans le cas binomial) et on montre qu'elle admet asymptotiquement la même loi que la déviance.

le modèle peut être jugé satisfaisant pour un rapport D/ddl plus petit que 1.

Tests

Rapport de vraisemblance

Le rapport de vraisemblance ou la différence de déviance est une évaluation de l'apport des variables explicatives supplémentaires dans l'ajustement du modèle. La différence des déviances entre deux modèles emboîtés respectivement à q_1 et $q_2 (q_2 > q_1)$ variables explicatives

$$D_2 - D_1 = 2(\mathcal{L}_1 - \mathcal{L}_{sat}) - 2(\mathcal{L}_2 - \mathcal{L}_{sat}) = 2(\mathcal{L}_1 - \mathcal{L}_2)$$

suit approximativement une loi du χ^2 à $(q_2 - q_1)$ degrés de liberté pour les lois à 1 paramètre (binomial, Poisson) et une loi de Fisher pour les lois à deux paramètres (gaussienne).

Test de Wald

Ce test est basé sur la forme quadratique faisant intervenir la matrice de covariance des paramètres, l'inverse de la matrice d'information observée $(X'WX)^{-1}$.

Diagnostics

Résidus

Pearson

Les résidus obtenus en comparant valeurs observées y_i et valeurs prédites $\hat{y}_i$ sont pondérés par leur précision estimée par l'écart-type : s_i de $\hat{y}_i$. Ceci définit les résidus de Pearson :

$$r_{p_i} = \frac{y_i - \hat{y}_i}{s_i}$$

dont la somme des carrés conduit à la statistique du même nom.

Déviance

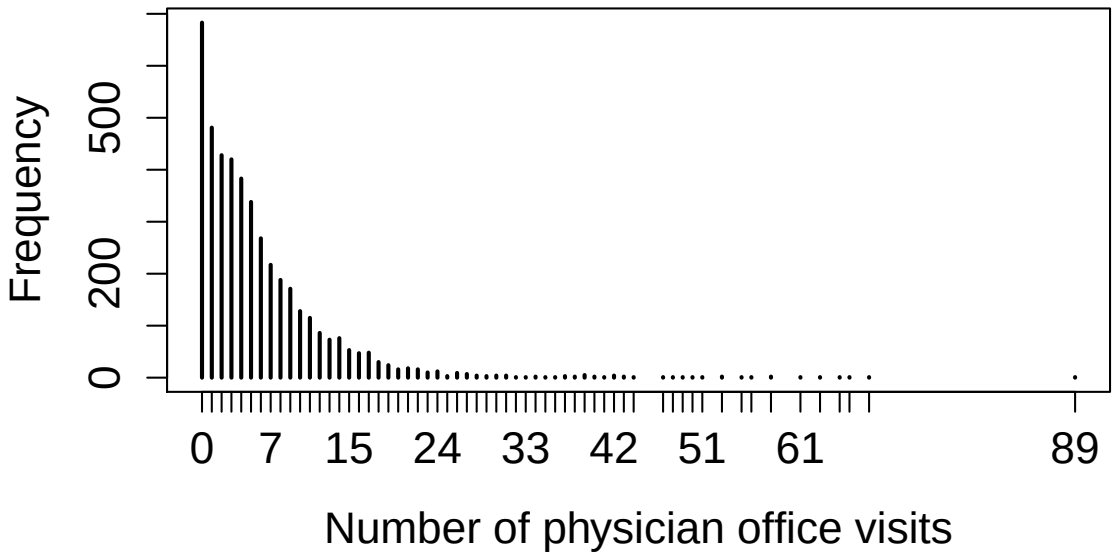
Ces résidus mesurent la contribution de chaque observation à la déviance du modèle par rapport au modèle saturé. Des versions standardisées et studentisées en sont définies comme pour ceux de Pearson.

Exemples

Deb et Trivedi (1997) analysent des données sur 4406 personnes, âgées de 66 ans et plus, qui sont couverts par Medicare, un programme public d'assurance. L'objectif est de modéliser la demande de soins médical, comme capturée par le nombre de bureaux des médecins et les consultations externes des hôpitaux par les covariables disponibles pour les patients.

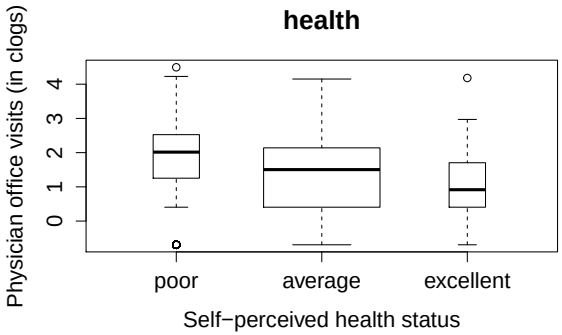
Explication des données

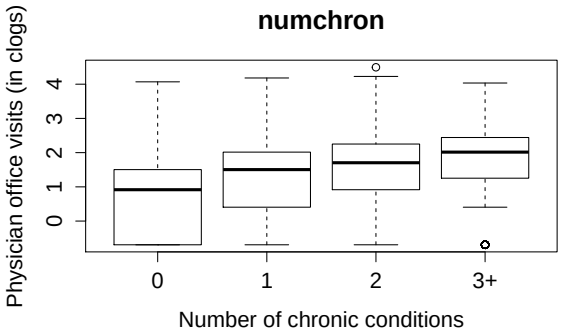
- **ofp**: Le nombre des visites chez le bureau de médecin, variable dépendante.
- **hosp**: Nombre de séjours à l'hôpital
- **health**: État de santé auto-évaluation
- **numchron**: Nombre de maladies chroniques
- **gendre**: Sexe
- **school**: Nombre des années d'études
- **privins**: Indicateur de l'assurance privée

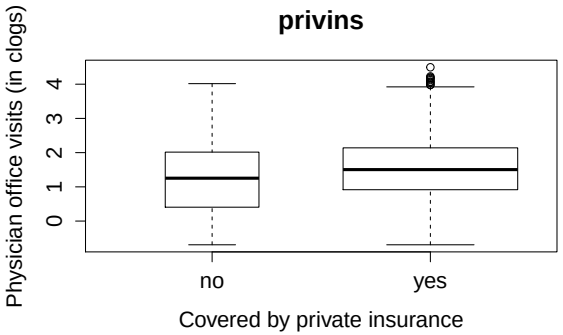


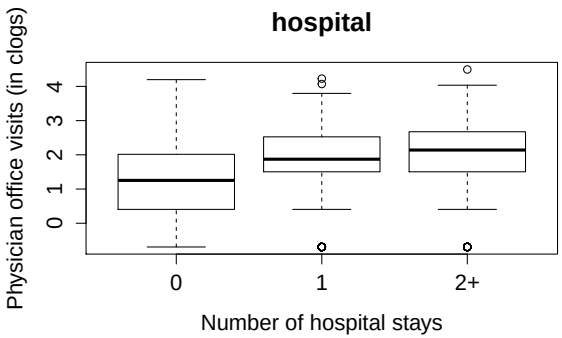
L'histogramme montre que la distribution marginale présente à la fois des variations considérables et un assez grand nombre de zéros.

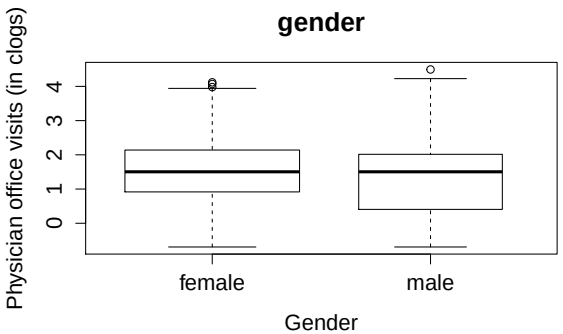
Régression de Poisson

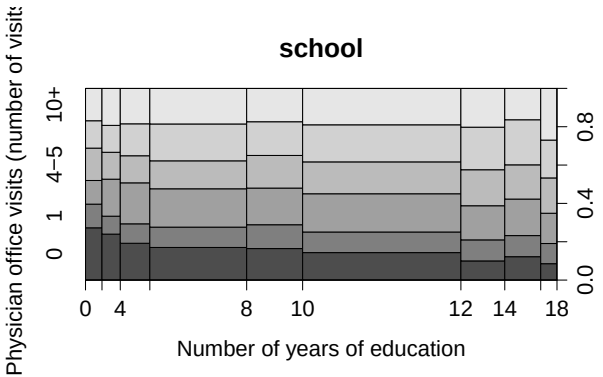












Régression de Poisson

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4055  -1.9962  -0.6737   0.7049  16.3620
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.028874   0.023785  43.258   <2e-16 ***
hosp             0.164797   0.005997  27.478   <2e-16 ***
healthpoor       0.248307   0.017845  13.915   <2e-16 ***
healthexcellent -0.361993   0.030304 -11.945   <2e-16 ***
numchron         0.146639   0.004580  32.020   <2e-16 ***
gendermale      -0.112320   0.012945  -8.677   <2e-16 ***
school           0.026143   0.001843  14.182   <2e-16 ***
privinsyes       0.201687   0.016860  11.963   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 26943  on 4405  degrees of freedom
Residual deviance: 23168  on 4398  degrees of freedom
AIC: 35959

Number of Fisher Scoring iterations: 5
```

Conclusion

- **Binômiale**: la variable binômiale $\mathfrak{B}(m, p)$ est la somme de m Bernoullis $\mathfrak{B}(1, p)$ indépendants.

  Exemple: Taille d'un sous-échantillon dans un échantillon de taille donnée (nombre de sujets réagissant favorablement).

- **Poisson** Une somme de variables de Poisson indépendantes est encore de Poisson.

  Exemple: Nombres d'événements (ex: en pannes) arrivant sur une durée donné. Comptes dans une table de contingence.