Université de Strasbourg
M1 - Magistère
Statistique - études de cas

Ségolen Geffray
geffray@math.unistra.fr
Année 2016/2017

Sujet 5: Etude du comportement de l'estimateur des coefficients d'un modèle de régression de Poisson et des résidus de Pearson studentisés

---

• Introduction à la régression de Poisson:
De manière générale, le but de tout modèle de régression est de spécifier une relation (de nature stochastique) entre une variable Y appelée variable à expliquer et un certain nombre de variables explicatives $X^{(1)}, ..., X^{(p)}$. Dans toute la suite, on supposera que les variables explicatives sont de loi continue.

Notons $(Y_i, X_i^{(1)}, ..., X_i^{(p)})$ pour $i = 1, ..., n$, les observations correspondant à n individus que l'on suppose distribuées comme $(Y, X^{(1)}, ..., X^{(p)})$. Le modèle de régression de Poisson s'écrit sous la forme suivante:

1. **indépendance des individus**: les vecteurs $(Y_i, X_i^{(1)}, ..., X_i^{(p)})$ sont mutuellement indépendants,

2. **hypothèse distributionnelle**: conditionnellement à $X_i^{(1)}, ..., X_i^{(p)}$, la loi de la variable réponse Y_i est la loi de Poisson.

3. **linéarité en les paramètres**: les covariables influent sur la loi conditionnelle des Y_i au travers d'un prédicteur linéaire

$$\eta_i := \beta_0 + \sum_{k=1}^{p} \beta_k X_i^{(k)} .$$

   On supposera toujours que le modèle est identifiable, à savoir ici:

$$\beta := (\beta_0, ..., \beta_p)^t = (\beta'_0, ..., \beta'_p)^t := \beta' \quad \Longrightarrow \quad \mathbb{E}_\beta[Y_i | X_i^{(1)}, ..., X_i^{(p)}] = \mathbb{E}_{\beta'}[Y_i | X_i^{(1)}, ..., X_i^{(p)}].$$

4. **fonction de lien**: le prédicteur linéaire η_i influe l'espérance conditionnelle de Y_i notée

$$\mu_i := \mathbb{E}[Y_i | X_i^{(1)}, ..., X_i^{(p)}]$$

   via une fonction de lien g monotone (souvent croissante), inversible et deux fois différentiable, choisie par l'utilisateur, de sorte que

$$g(\mu_i) = \eta_i .$$

   Il convient de prêter attention à l'ensemble de définition de g. En effet, dans le cas de la loi de Poisson, les variables Y_i prennent des valeurs positives. Comme $\mu_i = g^{-1}(\eta_i)$, il est alors judicieux de s'assurer que $g^{-1}(\eta_i) > 0$ quelque soit la valeur de η_i. Ainsi, la fonction de lien presque toujours utilisée est la fonction logarithme.

• **Estimation de l'effet des variables explicatives**:
Dans le cadre du modèle de régression présenté ci-dessus, on se demande si les variables explicatives introduites dans le modèle ont un effet ou non sur la variable à expliquer. Cet effet est quantifié par le coefficient de régression correspondant.

Dans un modèle de régression de Poisson, la fonction de lien et la loi conditionnelle des Y_i étant choisies par l'utilisateur, la seule inconnue à estimer à partir des observations $(y_i, x_i^{(1)}, ..., x_i^{(p)})_{i=1,...,n}$ de $(Y_i, X_i^{(1)}, ..., X_i^{(p)})_{i=1,...,n}$ est le vecteur des paramètres $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$. Notons $\widehat{\beta}$ l'estimateur de β obtenu par la méthode du maximum de vraisemblance. Cet estimateur est solution des équations du score que l'on obtient en annulant le vecteur gradient de la log-vraisemblance de $(Y_1, ..., Y_n)$ en $(y_1, ..., y_n)$ conditionnellement à $(X_i^{(1)}, ..., X_i^{(p)}) = (x_i^{(1)}, ..., x_i^{(p)})$ au point β. Les équations du score ainsi obtenues ne sont pas résolubles analytiquement. L'algorithme dit du *Fisher scoring*, une variante de l'algorithme de Newton-Raphson, en fournit une résolution numérique.
Sous des hypothèses assez techniques, on peut montrer que le vecteur $\widehat{\beta}$ existe avec une probabilité qui tend vers 1, est consistent pour β et suit approximativement la loi $\mathcal{N}_{p+1}\left(\beta, (\mathbb{X}^t.A(\beta).\mathbb{X})^{-1}\right)$ pour n assez grand, avec $A(\beta) = \text{diag}\,(\mu_i)_{i=1,...,n}$.

• **Ajustement du modèle aux données**:
L'expression des résidus de Pearson (internement) studentisés dans le cas d'une régression de Poisson est:

$$\widehat{\varepsilon}_i^{PS} = \frac{(Y_i - \widehat{\mu}_i)}{\sqrt{\widehat{\mu}_i(1-h_i)}}$$

où $\log(\widehat{\mu}_i) = \mathbb{X}_i.\widehat{\beta}$ avec $\mathbb{X}_i = (1, X_i^{(1)}, ..., X_i^{(p)})$, et où h_i est le levier de l'observation i, à savoir h_i est le $i^{\text{ème}}$ coefficient diagonal de la matrice des leviers H. La matrice des leviers est définie dans le cas d'une régression de Poisson par:

$$H = A(\widehat{\beta})^{1/2}.\mathbb{X}.(\mathbb{X}^t.A(\widehat{\beta}).\mathbb{X})^{-1}.\mathbb{X}^t.A(\widehat{\beta})^{1/2}$$

avec $A(\widehat{\beta}) = \text{diag}\,(\widehat{\mu}_i)_{i=1,...,n}$ Si le modèle de régression de Poisson est adéquat, pour n assez grand, les résidus de Pearson studentisés sont "approximativement" centrés et leur distribution s'approche "raisonnablement" de l'homoscédasticité et de la symétrie.

• **Implémentation au moyen du logiciel** `R`:
Le logiciel `R` permet d'ajuster simplement un modèle de régression de Poisson à des données. Pour ajuster un modèle linéaire sur un échantillon $(Y_i, X_i^{(1)}, ..., X_i^{(p)})_{i=1,...,n}$, on stockera les n valeurs des Y_i dans un vecteur `y`, puis, pour $k = 1, \ldots, p$, on stockera les n valeurs des $X_i^{(k)}$ dans un vecteur `xk`. Le modèle de régression de Poisson avec fonction de lien logarithme est ajusté au moyen de l'instruction suivante (où $p = 3$)

```
myglm<- glm(y~x1+x2+x3,family=poisson)
```

et le résultat est stocké dans l'objet `myglm`. L'instruction suivante permet d'accéder à l'ensemble des quantités calculées:

`summary(myglm)`

Sont notamment calculées, les estimations des β_j pour $j = 0, \ldots, p$ et les estimations de leurs écarts-types. L'objet `summary(myglm)` est une liste dont on peut récupérer chacune des composantes qui nous intéresse. Pour cela, l'instruction `str(summary(myglm))` permet de voir le nom et la structure des différentes composantes de cette liste.

La fonction `residuals` fournit différents types de résidus à partir d'un modèle ajusté avec la fonction `glm` du package `stats`, chargé par défaut. L'instruction

`residuals(myglm,type=''pearson'')`

fournit les résidus de Pearson sans la normalisation par $\sqrt{1-h_i}$. On peut effectuer la normalisation en récupérant les valeurs des leviers au moyen de l'instruction

`hatvalues(myglm)`

Exercice 1.
Etudier de manière empirique à partir de données simulées

- la distribution de l'estimateur des coefficients de régression,
- le comportement des résidus de Pearson studentisés,

en fonction de

- l'inclusion lors de l'ajustement d'une variable explicative non-influente en réalité,
- l'omission lors de l'ajustement d'une variable explicative influente en réalité,
- la taille de l'échantillon n simulé,
- le nombre moyen de comptages attendu pour la réponse.