Université de Strasbourg
M1 - Magistère
Statistique - études de cas

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

Sujet 8: Comparaison d'un modèle de régression Gamma à un modèle gaussien avec possible transformation de la réponse en termes de qualité d'ajustement et d'estimation d'effets des variables explicatives

---

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.

• **Quelques rappels sur la régression linéaire**:

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 linéaire gaussien standard s'écrit sous la forme générique suivante:

$$Y_i = \beta_0 + \sum_{k=1}^{p} \beta_k X_i^{(k)} + \varepsilon_i, \qquad \text{où} \quad \varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2). \tag{1}$$

Le modèle de régression linéaire standard repose ainsi sur les hypothèses suivantes:

($\mathbf{H_1}$) **linéarité**: l'espérance conditionnelle de la variable réponse vaut

$$\mathbb{E}[Y_i | X_i^{(1)}, ..., X_i^{(p)}] = \beta_0 + \sum_{k=1}^{p} \beta_k X_i^{(k)}.$$

Attention, bien que l'espérance conditionnelle de la variable réponse soit une combinaison affine des covariables, la linéarité s'apprécie relativement à la linéarité de $\mathbb{E}[Y_i | X_i^{(1)}, ..., X_i^{(p)}]$ en les paramètres.
On supposera toujours que le modèle est identifiable, à savoir

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

($\mathbf{H_2}$) **centrage des erreurs**: $\mathbb{E}[\varepsilon_i] = 0$ pour $i = 1, ..., n$.

($\mathbf{H_3}$) **exogénéité = non-endogénéité**: $(X_i^{(1)}, ..., X_i^{(p)})$ et ε_i sont indépendants pour $i = 1, ..., n$.

($\mathbf{H_4}$) **non-colinéarité des covariables**: les covariables sont non-corrélées entre elles, ce qui se traduit en pratique par le fait que les vecteurs $(X_1^{(k)}, ..., X_n^{(k)})$ sont non-colinéaires pour $k = 1, ..., p$. On écartera également le cas trivial où l'un de ces vecteurs serait colinéaire au vecteur de longueur n dont toutes les composantes sont égales à 1.

($\mathbf{H_5}$) **non-corrélation**: les vecteurs $(Y_i, X_i^{(1)}, ..., X_i^{(p)})$ sont non-corrélés pour $i = 1, ..., n$.

($\mathbf{H_6}$) **homoscédasticité**: les lois conditionnelles de Y_i sachant $X_i^{(1)}, ..., X_i^{(p)}$ ont même variance donc on a $\sigma_i^2 = \sigma^2$ pour $i = 1, ..., n$.

Le modèle de régression linéaire standard est gaussien lorsqu'on effectue l'hypothèse supplémentaire suivante:

($\mathbf{H_7}$) **normalité**: conditionnellement à $X_i^{(1)}, ..., X_i^{(p)}$, la variable de réponse Y_i suit la loi normale.

Notons I_n=matrice identité de taille $n \times n$ et avec

$$\mathbb{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} 1 & X_1^{(1)} & \dots & X_1^{(p)} \\ \vdots & \vdots & & \vdots \\ 1 & X_n^{(1)} & \dots & X_n^{(p)} \end{pmatrix} = \begin{pmatrix} \mathbb{X}_1 \\ \vdots \\ \mathbb{X}_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Le modèle de régression linéaire homoscédastique se réécrit matriciellement sous la forme:

$$\mathbb{Y} = \mathbb{X}.\beta + \varepsilon, \qquad \text{où} \quad \varepsilon \sim (0_n, \sigma^2 I_n) \text{ et } \varepsilon \perp\!\!\!\perp \mathbb{X}.$$

Le modèle de régression linéaire gaussien homoscédastique s'obtient sous forme matricielle lorsque l'on ajoute l'hypothèse $\varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$.

• **Estimation de l'effet des variables explicatives**:
Dans le cadre du modèle linéaire gaussien standard 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.
L'estimateur des moindres carrés ordinaires de β est $\widehat{\beta} = (\mathbb{X}^t.\mathbb{X})^{-1}.\mathbb{X}^t.\mathbb{Y}$. Sous les hypothèses $(H_1) - (H_4)$, cet estimateur est sans biais

$$\mathbb{E}[\widehat{\beta}|\mathbb{X}] = \beta\,.$$

Sous les hypothèses $(H_1) - (H_6)$, la variance de $\widehat{\beta}$ est

$$\mathrm{Var}(\widehat{\beta}|\mathbb{X}) = \sigma^2(\mathbb{X}^t.\mathbb{X})^{-1}\,.$$

Notons $\widehat{\sigma}^2$ l'estimateur de la variance de σ^2 défini par:

$$\widehat{\sigma}^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \frac{\widehat{\varepsilon}^t.\widehat{\varepsilon}}{n-(p+1)} \tag{2}$$

en notant

$$\widehat{\varepsilon}_i = Y_i - \mathbb{X}_i.\widehat{\beta}$$

et

$$\widehat{\varepsilon} = \mathbb{Y} - \mathbb{X}.\widehat{\beta}\,.$$

Sous les hypothèses (H_1)-(H_6), l'estimateur $\widehat{\sigma}^2$ est sans biais pour σ^2.

Introduisons une hypothèse supplémentaire:
($\mathbf{H_8}$): $\dfrac{\mathbb{X}^t.\mathbb{X}}{n} \xrightarrow{\mathbb{P}} Q$ lorsque $n \to \infty$ où Q est une matrice symétrique définie positive.

Sous les hypothèses (H_1)-(H_6) et (H_8), la convergence $\widehat{\beta} \xrightarrow{\mathbb{P}} \beta$ a lieu lorsque $n \to \infty$.

Sous les hypothèses $(H_1) - (H_7)$, la loi de l'estimateur $\widehat{\beta}$ de β est

$$\widehat{\beta}_{EMV} \sim \mathcal{N}_{p+1}(\beta, \sigma^2(\mathbb{X}^t.\mathbb{X})^{-1}),$$

et, pour $j = 0, 1, ..., p$, on a

$$\frac{\left(\widehat{\beta}^{(j)} - \beta^{(j)}\right)}{\sqrt{\widehat{\sigma}^2[(\mathbb{X}^t.\mathbb{X})^{-1}]_{j,j}}} \sim T(n - (p+1)).$$

• **Résidus standardisés dans le cadre du modèle linéaire gaussien standard**:
Considérons les résidus (bruts) que l'on a défini pour $i = 1, ..., n$ par $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$. Lorsque la valeur $|\widehat{\varepsilon}_i|$ est "anormalement" élevée, cela indique que le $i^{\text{ème}}$ réponse a été mal reconstituée par le modèle. Afin de pouvoir trancher si une observation est "anormalement" élevée, il faut être en mesure de pouvoir comparer des choses comparables. Or, bien que l'on ait travaillé sous l'hypothèse d'homoscédasticité qui stipule que $\text{Var}(\varepsilon_i) = \sigma^2$ pour $i = 1, ..., n$, il n'en va pas de même pour les résidus bruts. On a en effet établi que $\text{Var}(\widehat{\varepsilon}_i) = \sigma^2(1 - h_i)$ pour $i = 1, ..., n$. Pour rendre les amplitudes comparables entre les différents résidus, ie afin d'obtenir des résidus de variances égales, on normalise chacun des résidus (bruts) par son écart-type estimé pour ne plus s'intéresser qu'aux résidus normalisés. Les résidus standardisés sont des résidus normalisés définis de la façon suivante:

$$\widehat{\varepsilon}_i^{\text{Stand}} = \frac{\widehat{\varepsilon}_i}{\sqrt{\widehat{\sigma}^2(1 - h_i)}}$$

où $\widehat{\sigma}^2$ est l'estimateur sans biais de σ^2 défini en (2).
On peut montrer que la distribution des résidus standardisés est la loi de Student à $n - (p+1)$ degrés de liberté, notée $T(n - (p+1))$. Lorsque le modèle linéaire gaussien s'ajuste bien aux données, on s'attend donc à ce que 95% des résidus standardisés se trouvent entre les quantiles d'ordre respectif 2.5% et 97.5% de la loi $T(n - (p+1))$.

Lorsque la réponse est à valeurs positives et que les hypothèses d'homoscédasticité et de normalité ne sont pas satisfaites par les données, la transformation de Box-Cox de la variable réponse permet parfois d'atteindre l'homoscédasticité et de se rapprocher de la normalité. Pour une variable Y à valeurs positives, la transformation de Box-Cox est définie par

$$Y_{BC} = \begin{cases} \dfrac{Y^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log(Y), & \lambda = 0. \end{cases}$$

• **Introduction à la régression Gamma**:

Notons à nouveau $(Y_i, X_i^{(1)}, ..., X_i^{(p)})$ 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 Gamma d'espérance notée μ_i.

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 Gamma, 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 le plus souvent 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 $A(\beta) = \text{diag}\left(\frac{1}{\mu_i^2 g'(\mu_i)^2}\right)_{i=1,...,n}$.

• **Ajustement du modèle de régression de Poisson 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 $A(\widehat{\beta}) = \text{diag}\left(\frac{1}{\widehat{\mu}_i^2 g'(\widehat{\mu}_i)^2}\right)_{i=1,...,n}$. Si le modèle 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.

• Ajustement du modèle gaussien au moyen du logiciel `R`:
Le logiciel `R` permet d'ajuster simplement un modèle de régression 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 linéaire est alors ajusté au moyen de l'instruction suivante (où $p = 3$)

```
mylm<- lm(y~x1+x2+x3)
```

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

```
summary(mylm)
```

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

```
hatvalues(mylm)
```

La fonction `rstandard` fournit les résidus standardisés à partir d'un modèle ajusté avec la fonction `lm` du package `stats`, chargé par défaut:

```
rstandard(mylm)
```

La fonction `boxcox` du package `MASS` à charger au moyen de l'instruction

```
library(MASS)
```

permet de déterminer une valeur optimale du paramètre λ de la transformation de Box-Cox. La fonction `bcPower` du package `car` à charger au moyen de l'instruction

```
library(car)
```

• Ajustement du modèle de régression Gamma au moyen du logiciel `R`:
Le logiciel `R` permet d'ajuster simplement un modèle de régression Gamma 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 Gamma avec fonction de lien logarithme est ajusté au moyen de l'instruction suivante (où $p = 3$)

```
myglm<- glm(y~x1+x2+x3,family=Gamma(link=''log''))
```

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.
Comparer de manière empirique à partir de données simulées telles que la réponse est discrète à valeurs positives

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

dans le cas de l'ajustement

- d'un modèle de régression gaussien pour Y,
- d'un modèle de régression gaussien pour $\sqrt{Y}$,
- d'un modèle de régression de Poisson pour Y.

Vous veillerez à faire varier

- la taille de l'échantillon n simulé,
- la valeur moyenne attendue pour la réponse.