

---

**Sujet 6: Etude du comportement de l'estimateur des coefficients d'un modèle de régression de Bernoulli, des résidus de Pearson studentisés et des résidus quantiles randomisés**

---

• **Introduction à la régression de Bernoulli:**

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  $X^{(1)}, \dots, X^{(p)}$ . Dans toute la suite, on supposera que les variables explicatives sont de loi continue.

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

1. **indépendance des individus:** les vecteurs  $(Y_i, X_i^{(1)}, \dots, X_i^{(p)})$  sont mutuellement indépendants,
2. **hypothèse distributionnelle:** conditionnellement à  $X_i^{(1)}, \dots, X_i^{(p)}$ , la loi de la variable réponse  $Y_i$  est la loi de Bernoulli.
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:

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

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

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

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 Bernoulli, les variables  $Y_i$  prennent les valeurs '0' ou '1'. Comme  $\mu_i = g^{-1}(\eta_i)$ , il est alors judicieux de s'assurer que  $g^{-1}(\eta_i) \in ]0, 1[$  quelque soit la valeur de  $\eta_i$ . Ainsi, la fonction de lien la plus utilisée est la fonction  $\text{logit}(\cdot) = \log\left(\frac{\cdot}{1-\cdot}\right)$ . Les fonctions de lien  $\text{probit}(\cdot) = \phi^{-1}(\cdot)$  en notant  $\phi$  la fonction de répartition de la loi  $\mathcal{N}(0, 1)$  et  $\log(-\log(1-\cdot))$  sont deux autres choix possibles.

On dit que le modèle est bien spécifié lorsqu'il correspond effectivement au mécanisme ayant servi à générer les données. Dans le cas contraire, on dit qu'il est mal spécifié.

• **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  $(y_i, x_i^{(1)}, \dots, x_i^{(p)})_{i=1, \dots, n}$

de  $(Y_i, X_i^{(1)}, \dots, X_i^{(p)})_{i=1, \dots, n}$  est le vecteur des paramètres  $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$ . Notons  $\hat{\beta}_n$  l'estimateur

de  $\beta$  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  $(Y_1, \dots, Y_n)$  en  $(y_1, \dots, y_n)$  conditionnellement à  $(X_i^{(1)}, \dots, X_i^{(p)}) = (x_i^{(1)}, \dots, x_i^{(p)})$  au point  $\beta$ . 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  $\hat{\beta}$  existe avec une probabilité qui tend vers 1, est consistant pour  $\beta$  et suit approximativement la loi  $\mathcal{N}_{p+1}(\beta, (\mathbb{X}^t \cdot A(\beta) \cdot \mathbb{X})^{-1})$

pour  $n$  assez grand avec  $A(\beta) = \text{diag} \left( \frac{1}{\mu_i(1 - \mu_i)g'(\mu_i)^2} \right)_{i=1, \dots, n}$ .

• **Ajustement du modèle aux données:**

Sont présentés ci-dessous deux types de résidus dans le cas d'une régression de Bernoulli.

L'expression des résidus de Pearson (internement) studentisés dans le cas d'une régression de Bernoulli est:

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

où  $g(\hat{\mu}_i) = \mathbb{X}_i \cdot \hat{\beta}$  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 Bernoulli par:

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

avec  $A(\hat{\beta}) = \text{diag} \left( \frac{1}{\hat{\mu}_i(1 - \hat{\mu}_i)g'(\hat{\mu}_i)^2} \right)_{i=1, \dots, n}$ .

Si le modèle de régression de Bernoulli 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.

Notons  $F_{\mu_i}$  la fonction de répartition conditionnelle de la loi conditionnelle de la réponse  $\mathcal{L}(Y_i | X_i^{(1)}, \dots, X_i^{(p)})$ , qui est ici la loi de Bernoulli de paramètre  $\mu_i$ . Notons  $\phi$  la fonction de répartition de la  $\mathcal{N}(0, 1)$ . Les résidus quantiles randomisés sont définis dans le cas d'un modèle de régression de Bernoulli par

$$\hat{\varepsilon}_i^R = \phi^{-1}(U_i)$$

où  $U_i \sim \mathcal{U}(a_i, b_i)$  avec  $a_i = \lim_{y \uparrow y_i} F_{\hat{\mu}_i}(y)$  et  $b_i = F_{\hat{\mu}_i}(y_i)$ .

Les résidus quantiles randomisés sont censés approcher “raisonnablement” la normalité.

### • Implémentation au moyen du logiciel R:

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

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

et le résultat est stocké dans l’objet `myglm`. Le modèle de régression de Bernoulli avec fonction de lien probit est ajusté au moyen de l’instruction suivante (où  $p = 3$ )

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

tandis que l’ajustement d’un modèle de régression de Bernoulli avec fonction de lien  $\log(-\log(1-))$  est ajusté au moyen de l’instruction suivante (où  $p = 3$ )

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

L’instruction suivante permet d’accéder à l’ensemble des quantités calculées:

```
summary(myglm)
```

Sont notamment calculées, les estimations des  $\beta_j$  pour  $j = 0, \dots, p$  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`. 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)
```

La fonction `qresiduals` du package `statmod` fournit les résidus quantiles randomisés.

### 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,
- le comportement des résidus quantiles randomisés,

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,
- la fonction de lien choisie.