

On reprend les définitions du DFFITS

$$\text{DFFITS}_i = |t_i^*| \sqrt{\frac{h_{ii}}{1-h_{ii}}} \sim t_{m-p-2}$$

Si H (la matrice de projection) est équilibrée alors on devrait avoir $h_{ii} \approx \frac{p+1}{m}$ et, du coup,

$$\frac{h_{ii}}{1-h_{ii}} \approx \frac{p+1}{m-p-1}$$

Au final, l'i-ième observation attire contre attention si

$$|\text{DFFITS}_i| > \sqrt{\frac{p+1}{m-p-1}} t_{m-p-2; 1-\frac{\alpha}{2}}$$

▷ quantile
 d'ordre $1 - \frac{\alpha}{2}$
 d'une loi
 de Student
 avec $m-p-2$ d.d.e.

Rappelez vous que le modèle linéaire s'écrit

$$Y = X\beta + \varepsilon$$

avec les hypothèses :

- 1) $\varepsilon \sim N(0, \sigma^2 I_m)$
- 2) X déterministe
- 3) $\tau(X) = p+1 < \infty$

Nous allons nous concentrer sur l'hypothèse $\mathbb{E}_\varepsilon = \sigma^2 I_m$ ce qui revient à dire que les erreurs sont homoskedastiques et indépendantes.

En pratique, nous pouvons avoir :

- * heteroscedasticité : \mathbb{E}_ε n'est pas diagonale mais tous les éléments sur la diagonale sont différents des tous les autres
- * autocorrelation : \mathbb{E}_ε n'est plus diagonale.

On note la "nouvelle" matrice de variance et covariance $\mathbb{E}_\varepsilon = \Sigma V$

avec $\lambda > 0$ et V une matrice symétrique définie positive et de rang m . Quel est l'impact de cette modification de \mathbb{E}_ε sur les propriétés des estimateurs des MC ?

L'estimateur des MC de β s'écrit toujours $\hat{\beta} = (X'X)^{-1}X'y$ et le reste sans biais :

$$E(\hat{\beta}) = (X'X)^{-1}X'\underbrace{E(y)}_{x\beta} = \beta$$

cous

$$\Rightarrow = \mathbb{E}_\varepsilon = \lambda^2 V$$

$$\begin{aligned} \text{Var}_{\hat{\beta}} &= (X'X)^{-1} X' \underbrace{\mathbb{E}_y}_{\lambda^2 V} X (X'X)^{-1} = \\ &= (X'X)^{-1} X' \lambda^2 V X (X'X)^{-1} \\ &\neq \lambda^2 (X'X)^{-1} \end{aligned}$$

qui dépend de V ; du coup, l'estimateur des MC n'est plus BLUE (i.e. le meilleur estimateur linéaire sans biais).

En conclusion, en cas d'erreur

hétéroscedastique, les paramètres du modèle doivent être estimés par la MÉTHODE DES MOINDRES CARRES PONDÉRÉS (MCP) qui consiste à transformer les variables du modèle (grâce à des pondérations) afin de se ramener à un modèle avec des erreurs homoskedastiques.

Si les erreurs sont moins corrélées, alors la méthode d'estimation à utiliser est la MÉTHODE DES MOINDRES CARRES GÉNÉRALISÉS (l'idée est la même que pour les MCP).

En cas de hétéroscéasticité et/ou asymétrie des résidus du modèle, une possibilité alternative serait de Transformer la variable à expliquer y grâce à la TRANSFORMÉE DE BOX-COX définie par:

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \quad y > 0 \\ \ln y & \lambda = 0 \end{cases}$$

© Théo Jalabert

5

Il s'agit d'une Transformation
non linéaire qui dépend d'un
paramètre (inconnu) à qui il va
falloir estimer grâce à la loi de
vraisemblance profilée (voir exercice
en TD).

Remarque: Il se peut qu'il n'existe pas
(pour nos données) une Transformation
de la variable y qui rende les résidus
symétriques et homoskedastiques. Une
certaine forme des MCP pourraient
mieux fonctionner !

On s'intéresse à présent à l'hypothèse
 $\pi(x) = p+1 < \infty$

Si cette hypothèse n'est pas vérifiée,
alors le déterminant de $(x'x)$ est
nul et $(x'x)$ n'est plus inversible
ce qui pose problème pour l'estimation
de β (on rappelle que $\hat{\beta} = (x'x)^{-1}x'y$).
Il faut éviter d'inclure dans le
modèle des variables explicatives très

fortement corrélées entre elles (multicollinearité) puisque celles auraient un impact sur $\hat{\beta} = \sigma^2(X'X)^{-1}$ et donc sur la précision des estimations obtenues (la variance explosive dès que $|X'X|$ est proche de zéro et donc les estimations sont moins précises). De plus, on a un effet sur le test de significativité de Student : les β_j vont avoir l'variance à ne pas être significativement $\neq 0$.

En effet,

$$t = \frac{\hat{\beta}_j}{\text{Var}(\hat{\beta}_j)}$$

mais la $\text{Var}(\hat{\beta}_j)$ devient très grande à cause de la multicollinearité et on se déplace vers la région d'acceptation du test. De plus, la présence de multicollinearité, de petites variations sur les données provoquent de fortes

Variations des estimations.

© Théo Jalabert

7

Comment détecter la multicollinearité?

- * Simple examen de la corrélation des corrélations entre variables explicatives qui permettrait de détecter des corrélations dangereuses de variables deux à deux (on ne détecte pas pour contre des relations complexes parmi plusieurs variables explicatives)
- * p-value du test de significativité très différente selon les modèles utilisés
- * on peut calculer le FACTEUR D'INFLATION DE LA VARIANCE (VIF)
VIF \rightarrow VARIANCE INFLATION FACTOR

$$VIF_j = \frac{1}{1 - R_j^2} \quad j = 1, \dots, p$$

où R_j^2 est le coefficient de détermination (ou R^2) de la régression de la variable X_j sur les autres variables

explicatives du modèle.

Dans la situation idéale, VIF = 1 (car $R_j^2 = 0$), la variable X_j est considérée responsable du problème de multicollinearité quand le VIF est "grand". Cet indicateur s'appelle "facteur d'inflation de la variance" car il peut être vu comme le rapport entre $\text{Var}(\hat{\beta}_j)$ dans le modèle que l'on estime et la $\text{Var}(\hat{\beta}_j)$ d'un modèle optimisé avec des variables explicatives orthogonales entre elles. En effet,

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\underbrace{\text{Den}(X_j)(1 - R_j^2)}_{\rightarrow \text{numérateur de la Var}(X_j)}}$$

et, en cas d'absence de multicollinearité, comme $R_j^2 = 0$, on a que $\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\text{Den}(X_j)}$ donc effectivement le VIF nous donne une idée de l'inflation de la $\text{Var}(\hat{\beta}_j)$.



due à la présence de multicollinearité = 9
Té.

Remarque: si R_j^2 est élevé mais en même temps $\text{Dev}(x_j)$ est élevée, la $\text{Var}(\hat{\beta}_j)$ n'explose pas ce qui veut dire que la multicollinearité peut intéresser l'estimation de tous les paramètres ou seulement de quelques paramètres.

* on peut calculer l'indice de conditionnement:

$$K = \frac{\lambda_1}{\lambda_p} = \frac{\max\{\lambda_j\}}{\min\{\lambda_j\}}$$

avec $\lambda_1, \dots, \lambda_p$ les valeurs propres de la matrice $X'X$ rangées par ordre décroissant.

Le déterminant de $|X'X|$ est donné par le produit des valeurs propres et donc dès que les dernières valeurs propres sont trop petites, on risque d'avoir des problèmes numériques.

En pratique, pas de soucis si $K < 100$
 et inquiétant si $K > 1000$. Cet indicateur permet d'identifier un problème de multicollinearité mais ne nous permet pas d'identifier les (ou les) variable(s) responsable(s) du problème. Qu'est-ce qu'on fait en cas de multicollinearité ?

REMEDES

- * Virez les variables explicatives responsables du problème
- * obtenez des observations supplémentaires afin d'essayer d'augmenter la $\text{Dev}(X_j)$
- * effectuer une régression sur les composantes principales i.e. remplacer les variables explicatives par leurs premières composantes principales (i.e. par de nouvelles variables qui sont des combinaisons linéaires des variables initiales de variance maximale sous une contrainte

d'orthogonalité.

→ problème d'interprétabilité
des nouvelles variables !

* régression aux MOINDRES CARRÉS PARTIELS (MCP) (PLS)

↳ PARTIAL LEAST SQUARES

Comme pour les régressions sur composantes principales, on cherche à construire de nouvelles variables explicatives, combinées linéaires des variables initiales, qui soient orthogonales entre elles et classées par ordre d'importance. À la différence de les régressions sur composantes principales, le choix des composantes est ici dicté par leur lien avec la variable à expliquer.

* régression ridge

L'idée est de travailler non pas avec $(X'X)$ mais avec $(X'X + \delta I)$ de façon à augmenter les valeurs propres et donc éviter les problèmes d'invérissage de la matrice $(X'X)$.

MODELES LINEAIRES GENERALISES

© Théo Jalabert

12

Le MLG est une généralisation du modèle linéaire qui présente quelques limites :

1) caractère gaussien de la variable à expliquer Y (on fait l'hypothèse que $\varepsilon \sim N$ ce qui implique $Y \sim N$).

En pratique, en assurance maladie par exemple, très rarement on travaille avec des variables gaussiennes. Pensez par exemple à la Tarification : on a dit que la prime pure s'écrit (dans le modèle collectif)

$$E(S) = E(N) E(X)$$

nombre moyen de sinistres le coût moyen

on travaille donc avec N (v.a. nb de sinistres) et X (v.a. coût du sinistre) qui ne suivent pas une loi gaussienne (N est plutôt Poisson)

ou BN et X plutôt Gamma ou Gaussienne inverse). Pensez aussi à un modèle qui expliquerait le comportement client où la variable à expliquer serait qualitative.

- 2) La variance de Y est constante et ne peut pas dépendre des variables explicatives

Le MLG est un modèle expliqué/expliquantes qui présente trois éléments :

- 1) $Y = (Y_1, \dots, Y_n)^T \rightarrow \text{vecteur à expliquer}$

Les v.a. Y_i sont indépendantes et leur loi appartient à la famille exponentielle, i.e. leur loi peut s'écrire sous la forme suivante :

$$f(y; \theta, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{\phi} + c(y, \phi) \right\}$$

avec

θ : paramètre réel appelé
PARAMETRE CANONIQUE ou
paramètre de la moyenne

ϕ : paramètre réel appelé PARAMETRE
DE DISPERSION

a: l'est définie sur les réels et nulle

b: l'est définie sur les réels et deux fois dérivable

c: l'est définie sur \mathbb{R}^2

La famille exponentielle inclut la plupart des lois usuelles : gaussienne, Gamma, Poisson, Bernoulli, ...

2) pour chaque Y_i , $i=1, \dots, n$, on connaît la valeur d'un p-uplet (X_{1i}, \dots, X_{pi}) de variables décrivant Y_i .

$$3) g_m \begin{pmatrix} E(Y_1) \\ \vdots \\ E(Y_n) \end{pmatrix} = \beta_0 + \beta_1 \begin{pmatrix} X_{11} \\ \vdots \\ X_{1m} \end{pmatrix} + \dots + \beta_p \begin{pmatrix} X_{p1} \\ \vdots \\ X_{pm} \end{pmatrix}$$

on a donc aussi

$$\boxed{g(\underbrace{\epsilon(y_i)}_{\mu_i}) = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi}}$$
$$= \underline{\underline{x_i^T \beta}}$$

Le MLG s'écrit

$$g(\underbrace{\mathbb{E}(Y_i)}_{\mu_i}) = \mathbf{x}_i^T \boldsymbol{\beta}$$

γ_i : SCORE

OU
PREDICTEUR
LINEAIRE

Si Y appartiennent à la famille exponentielle, alors

$$\text{Var}(Y) = \alpha(\phi) b''(\theta)$$

$$\alpha = V(\mu)$$

FONCTION
VARIANCE

Loi	$V(\mu)$
NORMALE	1
POISSON	μ
GAMMA	μ^2
BINOMIALE	$\mu(1-\mu)$
GAUSSIENNE INVERSE	μ^3

La fonction variance joue un rôle important : en effet, la variance

de Y varie avec la moyenne (2)
 (Y n'est donc pas homoskedastique comme dans le modèle gaussien) qui peut varier avec les variables explicatives du modèle.

Remarque : le modèle gaussien est un cas particulier du MLG que l'on obtient lorsqu'on choisit la loi gaussienne pour Y et la loi identité pour f_m .

Concernant la loi $\alpha(\phi)$, elle est souvent remplacée par ϕ/w_i avec w_i un réel strictement positif qui est commun et qui représente le poids attribué à l'ième observation.

Exemple : pondération décroissante avec le temps lorsqu'on utilise des données sur une période très longue.

Les constructions d'un MLG passe par différentes étapes:

- * recueillir des observations y_1, \dots, y_n de la variable Y et les valeurs correspondantes des variables explicatives
- * choisir une loi pour la variable Y à expliquer (dans la famille exponentielle) → choix dicté par la nature de la variable Y
- * choisir une fonction: on peut choisir la fonction "canonique" mais on n'est pas obligé ! Souvent en Tarification on choisit le log.
- * estimer le modèle (donc β et ϕ)
- * valider le modèle (indicateurs de la qualité, analyse des résidus, ...)
- * utiliser le modèle pour faire de la prediction → Tarification

N.B. Le fait de choisir la forme canonique simplifie l'estimation des paramètres inconnus.

$$\hat{\mu}_i = g^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \dots + \hat{\beta}_p x_{pi})$$

VALEUR AJUSTEE par le modèle

$$\eta_i = x_i^\top \beta \Rightarrow \text{du coup} \quad \frac{\partial \eta_i}{\partial \beta_j} = x_{ji}$$

$\downarrow (1 \ x_{1i} \dots x_{ji} \dots x_{pi})$

On a aussi

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{g'(\mu_i)}$$

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

On rappelle que le Hessian est la matrice des dérivées seconde

$$[H]_{ij} = \left[\frac{\partial^2 \mathcal{L}(\beta)}{\partial \beta_i \partial \beta_j} \right]$$

La matrice d'information est donnée par $I = X'W X$

de forme générale

$$[I]_{jk} = -E\left[\frac{\partial^2 \mathcal{L}(\beta)}{\partial \beta_j \partial \beta_k}\right] =$$

$$= - \sum_{i=1}^m \frac{x_{ji} x_{ki}}{\text{Var}(y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$$

et où W est une matrice diagonale de pondération

$$[W]_{ii} = \frac{1}{\text{Var}(y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$$

Lorsqu'on utilise la formule carabinique, l'écriture des équations de vraisemblance se simplifie et on peut constater que le Hessian est égal à la matrice d'information de Fisher et que les deux méthodes

méthodiques (Newton-Raphson et score de Fisher) convergent.

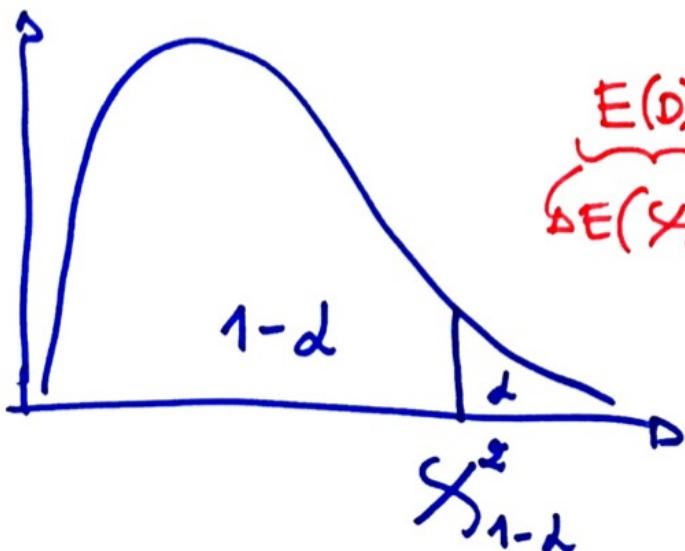
(6)

En ce cas, les équations de Valsenblanc s'écrivent (suite 10):

$$\sum_{i=1}^m \omega_i \underbrace{(y_i - \hat{\mu}_i)}_{\text{résidus du modèle}} x_{ji} = 0 \quad j=0, \dots, p$$

Cette expression traduit l'orthogonalité entre les variables explicatives du modèle et les résidus (semblable à celle obtenue dans le modèle gaussien)

$$\sum_{i=1}^m x_i \hat{\epsilon}_i = 0$$



$E(D) \leftarrow \frac{D}{m-p-1} \approx 1$

$\Delta E(x_{m-p-1}^2)$ l' ajustement du modèle est plutôt bon

Remarque: en pratique, le modèle saturé n'a pas trop d'intérêt en stat: en général, pour obtenir des estimations fiables, il faut une taille d'échantillon, m , beaucoup plus grande que le nb de paramètres à estimer (et donc $m \neq p+1$) et aussi souvent on cherche à trouver l'information et pas à reproduire les observations (et dans le modèle saturé $\hat{\mu}_i = y_i \forall i$)

Remarque: la deviance n'est pas considérée comme un bon indicateur de la qualité d'ajustement en cas de loi de Bernoulli, $y_i \sim B(q_i)$ puisqu'on peut montrer que, en ce cas, elle ne dépend que des valeurs ajustées \hat{q}_i et pas des observations y_i .

L'estimation a) en slide 15
Nécessité de la relation

(8)

$$D = \frac{D^*}{\phi} \sim \chi_{m-p-1}^2$$

H_0 : modèle restreint

H_1 : modèle complexe

$$H_0: \beta = \beta_0 = (\beta_0, \beta_1, \dots, \beta_q)'$$

$$H_1: \beta = \beta_1 = (\beta_0, \beta_1, \dots, \beta_p)'$$

avec $q < p < m$

(se renvoient à tester la "mutualité" simultanée de $\beta_{q+1}, \dots, \beta_p$)

$$\Delta = D_0 - D_1 = 2(\ell_m \hat{\mathcal{L}}_{\beta_1}^*(y) - \ell_m \hat{\mathcal{L}}_{\beta_0}^*(y))$$

deviance du modèle sous H_0

deviance du modèle sous H_1

$$= 2(\ell_m \hat{\mathcal{L}}_{SAT} - \ell_m \hat{\mathcal{L}}_{\beta_0}^*)$$

$$= 2(\ell_m \hat{\mathcal{L}}_{SAT} - \ell_m \hat{\mathcal{L}}_{\beta_1}^*)$$

C'est un test du rapport de vraisemblance avec

$$\lambda = \frac{\mathcal{L}_{\hat{\beta}_1}}{\mathcal{L}_{\hat{\beta}_0}}$$

$\mathcal{L}_{\hat{\beta}_0}$

vraisemblance du modèle restreint

On regarde si le fait de rajouter des variables explicatives au modèle (sous H_1) fait significativement diminuer la deviance.

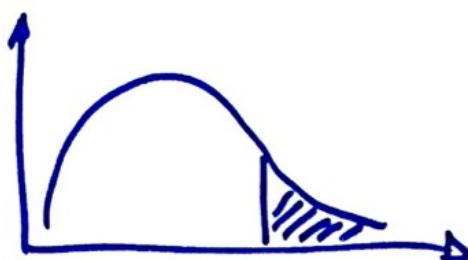
$\hat{\beta}$: EMV de β dans le modèle sous H_1
(modèle complet)

$\tilde{\beta}$: EMV de β dans le modèle sous H_0
(modèle restreint)

Le Test du rapport vraisemblance (slide 18) est utilisé par SAS pour les analyses de Type 1 et 3.

Au slide 18, on peut noter $\hat{\lambda} = \mathcal{L}_{\hat{\beta}_1}$
et $\tilde{\lambda} = \mathcal{L}_{\hat{\beta}_0}$.

q: comb de contraintes sous H_0



la region de rejet du Test

$$H_0: C\beta = r$$

Le Test de Wald sera utilisé pour tester la significativité de chacune des variables explicatives du modèle

$$H_0: \beta_J = 0$$

La statistique du Test de significativité de Wald s'écrit:

$$\frac{\hat{\beta}_J^2}{\Phi \Psi_J} \sim \chi^2_1$$

D = \text{Var}(\hat{\beta}_J)

En slide 21, $\hat{\beta}$ est l'EMV sous la contrainte $C\beta = r$

On avait écrit les équations de vraisemblance commune :

$$\sum_{i=1}^n \frac{w_i}{\phi} (y_i - \hat{\mu}_i) \frac{x_{ij}}{b''(\theta_i) g'(\mu_i)} = 0$$

La motivation de la statistique du score est que si $C\beta = r$ (sous H_0) alors $\ell'(\tilde{\beta})$ ne devrait pas être trop différent de zéro.

* dans le cas gaussien,
on avait écrit

$$H = X(X'X)^{-1}X' \quad (\text{slide 22})$$

$$\hat{\mu}_i = g^{-1}(x_i' \hat{\beta}) \quad \begin{array}{l} \text{valeurs ajustées} \\ \text{par le modèle} \end{array}$$

$$\text{Var}(z_i) = \text{Var}(Y_i) = V(\mu_i) \frac{\phi}{w_i} \quad \begin{array}{l} \phi \rightarrow \text{paramètre de dispersion} \\ w_i \rightarrow \text{poids connus} \\ a(\phi) \end{array}$$

\hookrightarrow donc variance pas constante !

$\begin{array}{l} \text{fondation} \\ \text{variance} \\ = b''(\phi) \end{array}$

Le nom du résidu de Pearson provient du fait que z_i^P peut être vu comme la racine carrée de la contribution de l'i-ème observation à la statistique de Pearson.

On représente les résidus de Pearson et on juge de la validité du modèle.

en fonction de la borne
répartition des résidus autour de
l'axe des abscisses avec une forme
cylindrique (et non pas "entourée"
par exemple).

On peut représenter les résidus de
deviance en fonction de i ce qui permet
d'identifier les observations ayant
un "grand" résidu ; en effet, si
 π_i^D est grand alors cela veut dire
que l'i-ème observation contribue
"Trop" à la construction de la
deviance D. Il reste à définir ce que
veut dire "Trop". On se rappelle
que $D \sim \chi_{n-p-1}^2$ et que $E(D) = n-p-1$.
On s'attend alors à ce que chaque
observation contribue approximativement
 $\frac{n-p-1}{n} \approx 1$ à la deviance.

(3)

Donc $|r_i^D| \gg 1$ est un indicateur du fait que l'observation i est en train de contribuer au manque d'ajustement du modèle. Cela peut signifier une mauvaise spécification du modèle ou une erreur dans les données.

* slide 25

Exercice : montrer que pour une loi gaussienne inverse avec $V(\mu) = \mu^3$, les résidus d'Anscombe s'écrivent :

$$\frac{\text{Res } y_i - \text{Res } \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

(solution sur slide 26)

© Théo Jalabert 
Selon le Type de variable Y a' expliquer,
on Tombe sur des modèles différents: (4)

1) MODELES POUR DONNEES DE COMPTAGE

(Exemple: nb de sinistres ou alors nb de
deces dans une étude de
mortalité, ...)



→ REGRESSION DE POISSON

→ REGRESSION BINOMIALE NEGATIVE

→ MODELES A INFLATION DE ZEROS
(ZERO INFLATED MODELS)

ZERO INFLATED POISSON ZIP ZINB → ZERO INFLATED
NEGATIVE BINOMIAL

2) MODELES POUR VARIABLE REPONSE CATEGORIELLE

→ REPONSE BINAIRE → REGRESSION LOGISTIQUE

(Variable dichotomique) (Exemple: un individu souscrit un
contrat d'assurance ou pas)

PREPONSE

AVEC r
MODALITES ($r \geq 2$)

(GLM MULTIVARIE)

(Variable polytomique)

REPONSE ORDINALE
REPONSE NOMINALE

MODELE LOGISTIQUE
CUMULATIF
MODELE LOG-LOG
COMPLEMENTAIRE
CUMULATIF
MODELE PROBIT
CUMULATIF

→ MODELE DE REGRESSION
NOMINALE

3) MODELES POUR VARIABLE REPONSE CONTINUE

→ REGRESSION GAMMA

→ REGRESSION GAUSSIENNE INVERSE

→ REGRESSION TWEEDIE

Pour la classe 3), on peut aussi choisir d'estimer un MODÈLE LOGNORMALE sur les coûts individuels Y_i ou plutôt un modèle gaussien sur le log des coûts, $\log(Y_i)$.

On rappelle que la variance pure dans le modèle collectif s'écrit

$$E(S) = E(Y) E(N)$$

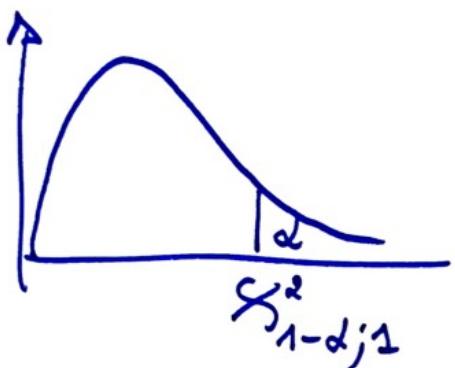
\hookrightarrow ^{nb de sinistres}
le coût du sinistre

Pour la loi de Poisson, l'espérance est égale à la variance mais ce n'est souvent pas ce qu'on observe en pratique : la variance empirique est souvent plus grande que la moyenne empirique (on parle de SURDISPERSION).

On se rappelle que la statistique du TEST de significativité de Wald est

donnée par :

$$\frac{\hat{\beta}_j^2}{\text{Var}(\hat{\beta}_j)} \sim \chi^2_{1-\alpha_j}$$



Ne pas prendre en compte la surdispersion mais admettre à l'estimer la statistique du test de significativité de Wald. Du coup, il se pourrait qu'une variable jugée pertinente dans le modèle de Poisson ne le soit plus après avoir pris en compte la surdispersion.

Pour modéliser la surdispersion, on peut utiliser des lois de Poisson composées : $Y \sim \text{Poisson}(k)$ mais k est elle aussi une v.a. continue de densité $g(k)$. A k donné, $Y \sim P(k)$ et alors :

$$f(y) = \int_0^\infty \underbrace{\frac{e^{-\lambda} \lambda^y}{y!}}_{f(y|\lambda)} g_\lambda(\lambda) d\lambda$$

Si on choisit pour g_λ la loi $I(\mu, \nu)$
 alors $f(y)$ est une BN(μ, k) 7
 avec $k = 1/\nu$.

VARIABLE OFFSET

Lorsqu'on estime un modèle pour une variable de comptage (par exemple le nb de sinistres ou de décès dans un groupe) il faut corriger par le nb d'exposés au risque.

Soit $\mu = E(Y)$, on s'intéresse à μ/n et alors le modèle s'écrit

$$g\left(\frac{\mu}{n}\right) = x'\beta$$

si g est la log, on écrit

$$\ln\left(\frac{\mu}{n}\right) = x'\beta \Rightarrow \ln\mu = \text{constante} + x'\beta$$

+
**VARIABLE
OFFSET**

Avec l'offset, Y a une espérance directement proportionnelle à x :

$$\mu = m e^{\alpha' \beta}$$

(8)

Les variables offset sont utilisées pour tenir en compte la taille du groupe ou différentes périodes d'observation.

Quand on fait de la quasi-vraisemblance, le paramètre de dispersion ϕ est estimé par $\frac{D^*}{m-p-1}$ (puisque

on ne peut pas faire de maximisation de vraisemblance car on ne fait pas d'hypothèse sur la loi de Y (aucune relation moyenne/variance décrite par une loi de la famille exponentielle correspond à ce qu'on souhaite décrire).



Nous avons vu que le modèle linéaire généralisé s'écrit :

$$g(\mu) = \mathbf{x}'\beta$$

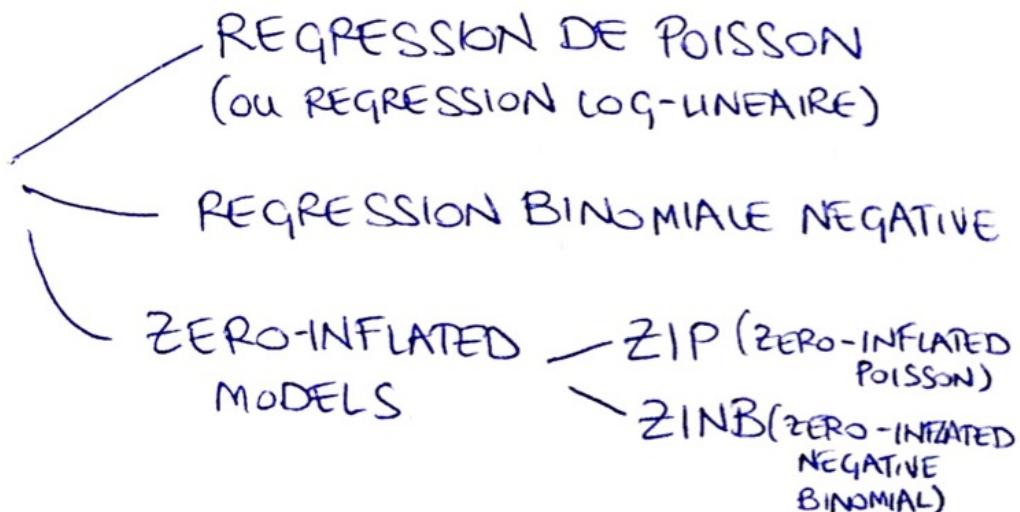
avec $\mu = E(Y)$, g la fonction lien et $\mathbf{x}'\beta$ le score du prédicteur linéaire. Or, selon le choix que l'on effectue pour la loi de la variable à expliquer Y (il faut choisir, je vous rappelle, une loi qui appartient à la famille exponentielle) et pour la fonction lien g (fonction lien canonique ou autre (souvent le log en Tarification)) on tombe sur un modèle différent.

Pour la loi de la variable à expliquer Y , le choix doit bien évidemment être fait parmi les lois de la famille exponentielle (comme déjà dit plus haut) mais il doit aussi être dicté par le type de variable (quantitative (discrete ou continue), qualitative (ordinaire ou nominale)).

Les modèles que l'on étudiera sont résumés dans le schéma ci-dessous.

1) MODELES POUR DONNEES DE COMPTAGE

(Exemple : nombre de sinistres en assurance automobile)



2) MODELES

POUR
VARIABLE
REPONSE
CATEGORIELLE
(Exemple : choix
du niveau de
garantie par
l'assuré son sacré)

VARIABLE
BINNAIRE
(ou DICHOTOMIQUE
(deux modalités))

→ REGRESSION
LOGISTIQUE

VARIABLE
POLYTOMIQUE
($K > 2$ modalités)

VARIABLE
QUALITATIVE
NOMINALE

MODELE DE
REGRESSION
NOMINALE

VARIABLE QUALITATIVE
ORDINAIRE

MODELE
LOGISTIQUE
CUMULATIF

MODELE
LOG-LOG
COMPLEMENTAIRE
CUMULATIF

MODELE
PROBIT
CUMULATIF

3) MODELES POUR

VARIABLE
REPONSE
CONTINUE

(Exemple :
montant de
saisie)

REGRESSION GAMMA

REGRESSION GAUSSIENNE INVERSE

REGRESSION TWEEDIE

Dans un souci de tarification, seulement les modèles 1) et 3) seront utilisés ; les modèles 2) pourraient être utiles par exemple pour une étude du comportement clients en assurance.

Nous allons regarder de plus près ces modèles et en étudier les spécificités.

Comme vous avez les Transports comme support, je vais juste rajouter quelques commentaires quand nécessaire.

On va commencer par les MODELES POUR DONNEES DE COMPTAGE. Quand la variable Y à expliquer est une variable de comptage (nb de sinistres en assurance auto, nb de décès dans une étude de mortalité, ...), classiquement nous avons deux possibilités : soit Y suit une loi de Poisson soit elle suit une BN. Une troisième possibilité est offerte par les modèles à inflation de zéros (ZIP et ZINB).

Dans un modèle de REGRESSION DE POISSON, on suppose que Y suit une loi de Poisson et on choisit comme fm linc soit le log (fm linc canonique pour une Poisson) soit l'identité (la fm linc log garantit des valeurs ajustées $\hat{\mu}$ positives).

L'effet du choix de la fm linc est expliqué au slide 4 par un exemple simple : un modèle avec une seule variable explicative - Avec une fm linc identité, l'effet est additif : si la variable explicative passe de x_1 à x_1+1 alors la variable à expliquer passe de $\beta_0 + \beta_1 x_1$ à $\beta_0 + \beta_1 (x_1+1) = \beta_0 + \underline{\beta_1 x_1} + \underline{\beta_1}$.

Avec une fm linc log, l'effet est multiplicatif : le passage de x_1 à x_1+1 se traduit cette fois-ci par un passage de $e^{\beta_0 + \beta_1 x_1}$ à $e^{\beta_0 + \beta_1 (x_1+1)} = e^{\beta_0 + \beta_1 x_1} e^{\beta_1}$

Le même modèle, sur slide 5, est toujours un modèle simple avec une seule variable explicative car cette fois-ci elle est qualitative avec r modalités et on choisit comme niveau de référence la modalité r . Comme elle est qualitative, on la remplace par $r-1$ variables indicatrices des niveaux.

On a parlé par le passé de segmentation, niveau de référence, modèle multiplicatif et coefficients correcteurs pour le calcul de la prime. C'est exactement ce qu'on retrouve sur slide 6 et 7 : si le but est de construire un tarif, il vaut mieux travailler avec des variables explicatives qualitatives (qui seront remplacées par des indicatrices) ce qui permettra de définir facilement un assuré de référence (celui qui présente la modalité de référence pour chacune des variables explicatives qualitatives (la modalité de référence est choisie en général comme étant la modalité la plus représentée dans l'échantillon)) et segmenter au sein du portefeuille. D'autre part, la fm log permet de définir un modèle multiplicatif : la moyenne pour l'assuré de référence est e^{β_0} ; pour les autres assurés, on obtiendra la moyenne en multipliant e^{β_0} par un certain nb de coefficients correcteurs qui correspondent aux caractéristiques qu'il présente.

On passe maintenant au MODÈLE DE RÉGRESSION BINOMIALE NÉGATIVE - Le problème avec la Poisson est que l'espérance est égale à la variance alors que souvent, en pratique, on observe une variance empirique supérieure à la variance théorique (on parle de phénomène de surdispersion). Si on ne prend pas en compte la surdispersion, on risque de sous-estimer la variance des estimateurs ce qui entraîne des intervalles de confiance trop étroits et aussi une surestimation de la valeur de la statistique du test du χ^2 de significativité (je vous rappelle qu'elle est donnée par $\hat{\beta}_j^2 / \text{Var}(\hat{\beta}_j)$) → on se déplace donc vers la région de rejet du test ce qui veut dire qu'on juge significative une variable qui ne l'est probablement pas. En cas de surdispersion, il faut donc pas estimer un modèle de Poisson mais plutôt un modèle avec $Y \sim BN$.

Exercice: montrer que la loi binomiale négative est un mélange de Gamma-Poisson (en fait $Y \sim P(\lambda)$ mais λ est une V.A. qui suit une loi Gamma (cela correspond au cas de populations différentes dans lesquelles la Poisson change de paramètres)) -

Dans le slide 8, il y a un $\text{term}(x)$ qui apparaît dans l'écriture du modèle - Ce $\text{term}(x)$ est ce qu'on appelle une VARIABLE OFFSET.

Soit Y une variable de comptage et $\mu = E(Y)$. On va s'intéresser pas au nombre cours mais s'intéresser à un taux d'occurrence μ/m ; du coup:

fréquence $g(\frac{\mu}{m}) = \alpha^{\beta}$ → on divise par le nb d'exposés au risque par exemple

et si g est la f_m lsm, alors

$$\text{lsm}(\mu) = \underbrace{\text{lsm}(\alpha)}_{\text{VARIABLE}} + \alpha^{\beta}$$

Grâce à l'offset, on aura $\mu = m e^{\alpha^{\beta}}$

Exemple: on s'intéresse au nb moyen de décès dans un groupe; alors, il va falloir corriger par le nb m d'exposés au risque - Au final, le nb moyen de décès sera directement proportionnel à m (on a écrit plus haut $\mu = m e^{\alpha^{\beta}}$).

La variable offset est comme une nouvelle variable dans le modèle mais son coefficient vaut 1 (pas besoin de l'estimer!).

En cas de surdispersion, une alternative à la loi Binomiale Négative est représentée par la QUASI-VRAISEMBLANCE (seules 9 et 10).

Pour une loi de la famille exponentielle,

$\text{Var}(Y) = \phi V(\mu)$ et, plus spécifiquement, pour une Poisson, $\phi = 1$ et $V(\mu) = \mu$; du coup $\text{Var}(Y) = \mu = E(Y)$.

On voudrait plutôt une relation moyenne-variance du type $\text{Var}(Y) = \phi \mu$ avec $\phi > 1$ mais cela me

correspond à aucune loi de la famille exponentielle et du coup les paramètres ne peuvent pas être estimés par un MU comme on l'a fait vu. La solution est donnée par la quasi-vraisemblance définie en slide 9. En slide 10, vous avez l'exemple de la quasi-Poisson (elle prend ce nom du fait que sa variance est celle d'une Poisson, $V(\mu) = \mu$). Avec une quasi-Poisson, $\hat{\beta}$ est le même que l'on aurait obtenu si on avait estimé une régression de Poisson (et qu'on n'aurait pas pris en compte la surdispersion) mais les écart-types sont "gommés" ce qui permet de résoudre les problèmes listés en slide 8.

Ce n'est pas consigné sur les slides, mais parmi les modèles pour données de comptage, on retrouve les MODELES A INFLATION DE ZEROS.

Ces modèles sont utilisés lorsque on observe un très important de zeros. En ce cas, on assume que ces zeros proviennent de deux processus différents : un processus qui génère des zeros "structuels" et un processus qui génère des mb de sinistres aléatoires. Ces zeros "structuels" peuvent être liés, par exemple, à un système bonus-malus ; l'assuré ne déclare pas le sinistre si le montant n'est pas important afin d'éviter une augmentation de prime l'année suivante.

Un modèle à inflation de zéros est un mélange entre une masse en 0 et un modèle classique de comptage (Poisson ou BN).

© Théo Jalabert

Pour modéliser la probabilité de ne pas déclarer un sinistre (Surpoids en 0), on considère un modèle logistique

$$\pi_i = \frac{e^{x_i^\top \beta}}{1 + e^{x_i^\top \beta}}$$

Pour le modèle de comptage, sait $p_i(k)$ la probabilité que l'i-eme individu ait k sinistres, alors

$$p(N_i=k) = \begin{cases} \pi_i + (1-\pi_i)p_i(0) & \text{si } k=0 \\ (1-\pi_i)p_i(k) & \text{si } k=1, 2, \dots \end{cases}$$

Pour le ZIP (ZERO-INFLATED POISSON) on a

$$p_i(0) = e^{-\lambda_i} \quad \text{et} \quad p_i(k) = e^{-\lambda_i} \frac{\lambda_i^k}{k!}$$

Le modèle ZINB (ZERO-INFLATED NEGATIVE BINOMIAL) est un mélange d'une Bernoulli (pour les zéros structurels) et une BN pour les zéros aléatoires. On écrit

$$\begin{aligned} \pi + (1-\pi) NB(0, r, p) & \quad \text{si } k=0 \\ (1-\pi) NB(k, r, p) & \quad \text{si } k>0 \end{aligned}$$

avec $NB(k, r, p) = \binom{k+r-1}{k} p^k (1-p)^r$

Nous passerons maintenant aux MODELES
POUR VARIABLE REPONSE QUALITATIVE (ou
categorical). La variable Y à expliquer est qualitative.
Tunc si elle présente deux modalités (par exemple
survenance d'un accident, facilité pour une entre-
prise, ...) ou k modalités (avec $k > 2$) et alors
les modalités sont ordonnées ou pas (on parle de
variable polytomique alors que pour la variable binale
on parle aussi de variable dichotomique) - On
distingue donc plusieurs modèles listés en slide 11.
Comme mentionné en slide 14, pour la régression
logistique ($Y \sim B(\pi)$ et non binomial logistique), on
n'utilise pas la deviance coronaire comme mesure de la qualité
d'ajustement du modèle mais plutôt le χ^2 de
Pearson - En effet, on juge de la qualité du modèle
en comparant les valeurs ajustées par le modèle
aux valeurs observées mais cela n'est pas possible
pour la Bernoulli puisque la deviance dépend
uniquement de $\hat{\pi}$; (les valeurs observées n'apparaissent
pas dans la formule pour D en slide 14).
Quand la variable à expliquer présente r modalités
(avec $r > 2$), qu'elle soit ordinaire ou multivariée,
on parle de GLM multivarié (slide 15) - Nous ne
Traiterons pas cela en détail mais je vous
donne une référence - En fait, nous avons déjà
Traité des variables qualitatives à plusieurs modalités
mais elles ont toujours figuré parmi les variables

© Théo Jalabert 

explicatives du modèle. Maintenant c'est la variable à expliquer qui est qualitative mais nous procérons de la même façon, i.e. nous la décomposons par un tableau d'indicatrices égale au nb de modalités moins 1. Nous écrivons donc

$$Y = (Y_1, \dots, Y_{z-1})'$$

avec Y_j ($j=1, \dots, z-1$) la variable indicatrice qui vaut 1 si l'individu présente la modalité j (0 sinon).

Après avoir défini quelques généralités en slide 15, nous distinguons le cas où la variable Y est qualitative nominale (pas d'ordre entre les modalités) ou ordinaire.

En slide 16, nous trouvons l'écriture du modèle de régression nominale. Rappelez-vous que en cas de variable binaire, le modèle était

$$\ln\left(\frac{\pi}{1-\pi}\right) = x'\beta$$

avec $\pi = P(Y=1)$ et qu'on avait appelé le rapport $\pi/(1-\pi)$ la côte. Ici (slide 16) nous avons

$$\ln\left(\frac{\pi_j}{\pi_z}\right) = \theta_j + x'\beta_j$$

où $\pi_j = P(Y=j)$ donc la probabilité d'observer la j -ème modalité de la variable à expliquer et $\pi_z = P(Y=z)$ est la probabilité d'observer la modalité de référence. Du coup, il s'agit exactement d'une côte (comme pour le modèle logistique)

© Theo Jalabert
6

mais la côte est écrite pour le modèle^j pour rapport au niveau de référence π .

On estime $n-1$ modèles : un pour chacune des probabilités de la variable à expliquer (notez que les paramètres θ et β dépendent de j).

Un exemple où on estimerait un modèle de régression multinomiale est celui où on cherche à expliquer le choix du moyen de transport (voiture, métro, bus, vélo, etc) par un individu (il n'y a pas d'ordre entre les différents moyens de transport).

Un exemple pour une variable réponse séquentielle (slide 17) serait celui où on s'intéresse au degré de gravité d'un événement (par exemple une fracture ou un accident). On alors, comme marqué sur le slide 17, la réponse à la question "combien évaluez-vous votre santé ?"; les modèles pourraient être "Très bien", "bien", "moyens" (pour simplifier!).

En ce cas, ce n'est plus $\pi_j = P(Y=j)$ que l'on cherche à modéliser mais les probabilités cumulées

$$\gamma_j = P(Y \leq j)$$

(on exploite le fait que les modèles sont ordonnés). C'est le choix de la loi de ε (l'erreur du modèle) qui nous permettra de distinguer trois modèles (ceux listés en slide 19).

En slide 20, remarquez que, comme pour le modèle

de régression multinomiale (slide 16) on modélise le cas
et on se retrouve à estimer $r-1$ modèles mais
cette fois-ci seulement θ_j dépend de γ (pas β)
et les $r-1$ équations sont parallèles.

On revient sur cette question en slide 23 où on dit
que cette hypothèse (β est le même pour toutes les
modélisations de la variable Y) est forte et que
probablement on devrait plutôt écrire:

$$\ln \frac{\pi_j}{1-\pi_j} = \theta_j + \underbrace{\gamma' \beta}_{\beta \text{ aussi dépendrait de } \gamma} \quad j=1, \dots, r-1$$

Le problème est qu'un modèle comme celui-ci
aurait à estimer beaucoup de paramètres
(on en rajoute $r-1$) donc il faudrait l'utiliser
que quand c'est vraiment nécessaire (autrement
il faudrait utiliser β (et pas β_j)). Pour cela,
on présente le Test en slide 24 et le modèle
en slide 25.

Il nous reste à traiter les MODELES POUR VARIABLE
RESPONSE CONTINUE (slide 26). Il s'agit de
construire un modèle pour le coût de sinistre
par exemple. Cette variable étant non-négative
et asymétrique à droite, le modèle gaussien
n'est pas utile mais une possibilité serait
de Transformer la variable Y pour la rendre
symétrique (avec un log par exemple) et

consiste à estimer un modèle gaussien (sur la transformée).
Autrement, on choisit comme loi pour \mathbf{Y} la Gamma, la Gaussienne (inverse ou la Tweedie).
Entre la Gaussienne inverse et la Gamma, on préfère la Gaussienne inverse au cas d'asymétrie.
Celle plus courante. La Tweedie est en général utilisée lorsque l'on souhaite modéliser directement la charge Totale (et donc ne pas estimer deux modèles (un pour le nb de sinistres et un pour le montant de sinistres) parce que par exemple on ne dispose pas des coûts individuels mais seulement de la charge globale sur l'assuré par police.

ET voilà, c'était le dernier cours de GLM ! Je vous enverrai dès que possible des notes manuscrites aussi pour la 1ere partie des slides (celles du 24 mars). Il faut que je les écrive !!