

Modèles Linéaires Généralisés

Motivation:

Tarification a priori

Math-Risque Habitation

b

Objectif: Calcul d'un tarif (prime) en assurance IARD (auto, MRH, construction, ...)

En effet, la prime est calculée en fonction de variables que l'on appelle les:
VARIABLES DE TARIFICATION qui peuvent être quantitatives ou qualitatives.

Parmi ces variables, on retrouve :

* des informations sur l'assuré :

→ Si particulier : âge, CSP, ...

(⚠ La variable sexe n'est plus une variable tarifaire grâce à une directive européenne)

→ Si entreprise : secteur d'activité, nb de salariés, ...

* des informations sur le bien assuré :

→ En assurance auto : puissance du véhicule, type, age...

→ En MRH : surface du logement, nb de pièces, ...

→ En perte d'exploitation : CA de l'entreprise, ...

* des informations géographiques :

- densité de population dans la commune, revenu moyen ...

On note N la Va mb de sinistres sur une période de temps étant en général l'année puisque la plupart des contrats sont annuels.

On note Y_i la Va "Coût du $i^{\text{ème}}$ sinistre" durant la période d'exposition au risque (i.e les indemnités versées par l'assureur à l'assuré).

On note la **CHARGE TOTALE** par police :

$$S = \begin{cases} \sum_{i=1}^N Y_i & \text{si } N > 0 \\ 0 & \text{si } N = 0 \end{cases}$$

La **PRIME PURE** est donnée par :

$$\boxed{E[S] = E[Y]E[N]}$$

dès lors que les Y_i sont iid et indépendantes de N .

On cherche un ensemble de variables explicatives $X = (X_1, \dots, X_p)$ © Théo Jalabert

$$\mathbb{E}[S|X] = \mathbb{E}[Y|X]\mathbb{E}[N|X].$$

Pour estimer ces espérances, on utilisera les MLG mais d'autres approches (non-paramétriques) existent comme les arbres et les forêts ou les réseaux de neurones par exemple.

Afin de constituer un tarif, on passe par les étapes suivantes :

* Construction de la base de données

* Distribution Sinistres attribuables / graves (pour cela on fixe, grâce à la théorie des valeurs extrêmes par exemple, un seuil de grave et on tente à part les graves).

* Choix des variables tarifaires qui sont la plupart du temps, qualitative ce qui permet à la compagnie d'assurance de "segmenter" son portefeuille, c'est à dire de constituer des classes homogènes d'assurés.

Si par exemple, on travaille avec l'âge en assurance auto, on constituera des classes d'âge.

Ex : 18-30, 30-40, 40-50, 50-60, 60+

classe de référence ↗
ou ↗
NIVEAU DE
RÉFÉRENCE ↗
Il s'agit de la modalité la + représentée dans le portefeuille.

La compagnie calcule une prime pour l'assuré de référence et utilise ensuite des coefficients de majoration ou de minoration de la prime pour les assurés présentant des caractéristiques différentes.

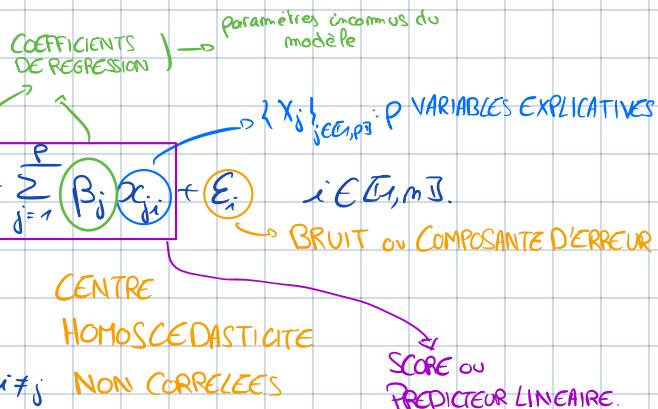
* Modélisation de l'effet des caractéristiques des individus (représentés par les modalités des variables tarifaires) sur la variable à expliquer (dans notre cas, la fréquence (N) et la sévérité (Y)). Ceci nous donne une prime pure ($\mathbb{E}[S] = \mathbb{E}[N]\mathbb{E}[Y]$).

* Passage du tarif pur au tarif technique (en rajoutant un changement de sécurité, les frais de gestion et les taxes) puis au tarif commercial (on modifie le tarif technique en fonction de la politique commerciale de la compagnie).

MODÈLE DE RÉGRESSION LINÉAIRE - RAPPEL

Ce modèle prend la forme suivante : $Y_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \epsilon_i$

VARIABLE A EXPLIQUER →



avec les hypothèses : 1) $\mathbb{E}[\epsilon_i] = 0 \forall i$

2) $\text{Var}(\epsilon_i) = \sigma^2 \forall i$

3) $\text{Cov}(\epsilon_i, \epsilon_j) = 0 \forall i \neq j$ NON CORRÉLÉES

erreurs centrées, homoscélastiques et indépendantes.

On peut écrire le modèle sous forme matricielle : $Y = X\beta + \varepsilon$

© Théo Jalabert

avec $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix}$ VECTEUR A EXPLIQUER

$\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$ VECTEUR DES COEFFICIENTS DE REGRESSION.

$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_m \end{pmatrix}$ VECTEUR DES ERREURS.

$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mp} \end{pmatrix}$ MATRICE DU PLAN D'EXPÉRIENCE

1 bis) $\varepsilon \sim N_m(0, \sigma^2 I_m)$ \Downarrow
 $Y \sim N_m(X\beta, \sigma^2 I_m)$

avec les hypothèses :

1) $E[\varepsilon] = 0 \Rightarrow E[Y] = X\beta$
 2) $\varepsilon \sim N_m(0, \sigma^2 I_m) \Rightarrow E[\varepsilon] = \varepsilon = \sigma^2 I_m$

3) X déterministe.

4) $\text{rg}(X) = p+1 < m$. (La matrice X est de rang plein).

Grâce à la méthode MCO (MOINDRES CARRES ORDINAIRES) on obtient une estimation du vecteur β :

$\hat{\beta}_{MCO} = (X'X)^{-1}X'Y = \hat{\beta}_{MLV}$

$(Y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2$

$$\begin{aligned} E[\hat{\beta}] &= E[(X'X)^{-1}X'Y] = (X'X)^{-1}X'E[Y] \\ &= (X'X)^{-1}X'X\beta = \beta \end{aligned}$$

Donc $\hat{\beta}$ est un estimateur sans biais de β .

On pose $(X'X)^{-1}X' = A \Rightarrow \hat{\beta} = AY$, alors $E[\hat{\beta}] = E[AY] = A E[Y] = A \sigma^2 I_m A' = \sigma^2 (X'X)^{-1}X'X(X'X)^{-1}$

On peut définir maintenant le VECTEUR DES VALEURS AJUSTÉES par le modèle.

$\hat{Y} = X\hat{\beta} = \underbrace{X(X'X)^{-1}X'}_H Y = HY$

où H est une matrice carrée $m \times m$ appelée HAT MATRIX (MATRICE CHAPEAU).

géométriquement, il s'agit de la matrice de projection orthogonale du vecteur Y dans le sous-espace de \mathbb{R}^m , Vect(X), engendré par les vecteurs colonnes de la matrice X .

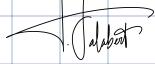
On peut aussi définir le VECTEUR DES RESIDUS, $\hat{\varepsilon} = Y - \hat{Y} = Y - X(X'X)^{-1}X'Y$
 $= [I - \underbrace{X(X'X)^{-1}X'}_H] Y$

$= \underbrace{[I - H]}_M Y$

Avec M la matrice de projection orthogonale de Y sur $\text{Vect}(X)^\perp$, l'espace orthogonal à Vect(X), appelé l'ESPACE DES RESIDUS. M est une matrice idempotente tg $M^m = M \quad \forall m \geq 1$.

On a aussi que $\text{rg}(M) < m$ (donc M n'est pas inversible) et $\text{rg}(M) = \text{Tr}(M)$.

$\hat{\varepsilon} = MY = M(X\beta + \varepsilon) = \cancel{MX\beta}^= 0 + ME = ME$

$\mathbb{E}[\hat{\epsilon}] = \mathbb{E}[M\epsilon] = M\mathbb{E}[\epsilon] = 0$ Les résidus, comme les erreurs, sont centrés. © Théo Jalabert 

$$\hat{\Sigma}_{\hat{\epsilon}} = \hat{\Sigma}_{ME} = M \hat{\Sigma}_{\epsilon} M^T = M \frac{\sigma^2}{\sigma^2 I_m} M = \sigma^2 M^2 = \sigma^2 M = \sigma^2 (I_m - H) \neq \hat{\Sigma}_{\epsilon} = \sigma^2 I_m$$

Donc les résidus $\hat{\epsilon}$, à la différence des erreurs ϵ , ne sont ni homoskedastiques ni corrélés. Il faudra construire d'autres types de résidu !

Un autre paramètre inconnu qu'il va falloir estimer est la variance de l'erreur σ^2 .

La méthode du MV nous donne :

$\hat{\sigma}^2 = \frac{\hat{\epsilon}' \hat{\epsilon}}{m}$ mais $\hat{\sigma}^2$ est un estimateur biaisé de σ^2 et, du coup, on utilise plutôt l'estimateur :

$$S^2 = \frac{\hat{\epsilon}' \hat{\epsilon}}{m-p-1} \quad \text{qui, lui, est sans biais.}$$

Un indicateur (classique) qui permet de juger de la qualité de ce modèle est le COEFFICIENT DE DETERMINATION que l'on note R^2 tel que

$$R^2 = \frac{SCE}{SCT} = \frac{\sum_{i=1}^m (\bar{Y}_i - \bar{Y})^2}{\sum_{i=1}^m (Y_i - \bar{Y})^2} \quad \begin{array}{l} \text{avec :} \\ \text{SCE: REGRESSION SUM OF SQUARES} \\ \text{SCT: SOMME DES CARRÉS EXPLIQUÉS PAR LE MODÈLE} \end{array}$$

La formule de décomposition de la variance permet d'écrire :

$$\sum_{i=1}^m (Y_i - \bar{Y})^2 = \sum_{i=1}^m (\bar{Y}_i - \bar{Y})^2 + \sum_{i=1}^m (\bar{Y}_i - Y_i)^2$$

$$SCT = SCE + SCR \quad \begin{array}{l} \text{SOMME DES} \\ \text{CARRÉS RÉSIDUELLES} \end{array} \quad \rightarrow SSE: ERROR SUM OF SQUARES.$$

On a $0 \leq R^2 \leq 1$.

$$R^2_{\text{ajusté}} = 1 - \frac{m-1}{m-p-1} \frac{\sum_{i=1}^m \hat{\epsilon}_i^2}{\sum_{i=1}^m (Y_i - \bar{Y})^2}$$

TEST D'HYPOTHÈSE DANS UN MODÈLE DE RÉGRESSION GAUSSIEN

- 1) TEST DE SIGNIFICATIVITÉ DE STUDENT
- 2) TEST GLOBAL DE FISHER
- 3) TEST ENTRE MODÈLE ENBOITES
- 4) TEST DE CONTRAINTES LINÉAIRES ENTRE PARAMÈTRES.

1) Test de Significativité de Student.

© Théo Jalabert

$$H_0: \beta_j = 0 \quad \text{vs} \quad H_1: \beta_j \neq 0.$$

On sait que $\hat{\beta}_j \sim N(\beta_j, \sigma^2(X'X)^{-1})$

et sous H_0 , $\frac{\hat{\beta}_j}{\sqrt{\sigma^2(X'X)^{-1}}} \sim N(0, 1)$

Or σ^2 est inconnu, on le remplace donc par $S^2 = \frac{\hat{\epsilon}' \hat{\epsilon}}{m-p-1}$
et on obtient la statistique du test de significativité

$$T = \frac{\hat{\beta}_j}{S\sqrt{(X'X)^{-1}_{jj}}}, \sim t_{m-p-1}$$

Pour un test bilatéral, on n'accepte pas H_0 si $|t| > t_{1-\alpha/2, m-p-1}$

quantile d'ordre $1-\alpha/2$
d'une t_{m-p-1} .

2) Test de Fisher.

$$H_0: \forall j \in \{0, \dots, p\}, \beta_j = 0 \quad \text{vs} \quad H_1: \exists j, \beta_j \neq 0$$

$$F = \frac{\left(\sum_{i=1}^m (Y_i - \bar{Y})^2 - \sum_{i=1}^m \hat{\epsilon}_i^2 \right) / p}{\sum_{i=1}^m \hat{\epsilon}_i^2 / (m-p-1)} \sim F_{p, m-p-1}$$

Sous H_0 , $\hat{Y}_i = \bar{Y}$ et donc on peut écrire:

$$F = \frac{[SCR(H_0) - SCR(H_1)] / p}{[SCR(H_1)] / (m-p-1)}$$

Il s'agit des q derniers paramètres

3) Test entre modèles emboîtés.

On souhaite tester la nullité simultanée d'un certain nombre de paramètres β_j de notre modèle $Y = X\beta + \epsilon$ avec p variables explicatives.

L'hypothèse H_0 s'écrit :

$$H_0: \beta_{p-q+1} = \dots = \beta_p = 0 \quad \text{vs} \quad H_1: \exists j \in \{p-q+1, \dots, p\}: \beta_j \neq 0$$

(modèle restreint)

(modèle complet).

On peut noter le modèle sous H_0 (i.e. le modèle privé des q dernières variables explicatives) comme $Y = X_0\beta_0 + \epsilon_0$

avec $\epsilon_0 \sim N_m(0, \sigma^2 I_m)$.

Idee: On estime les 2 modèles, celui sous H_0 et celui sous H_1 , et si \hat{Y} est "proche de" alors intuitivement on garderait le modèle sous H_0 puisque l'information apportée par les 2 modèles est la même et le modèle sous H_0 présente moins de paramètres à exprimer (principe de parcimonie).

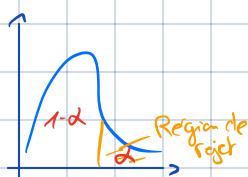
Pour quantifier le terme "proche" on utilise comme mesure de l'écart entre \hat{Y} et \bar{Y} la distance euclidienne ce qui nous amène à définir la statistique du test suivante :

$$F = \frac{\|\hat{Y} - \bar{Y}\|^2/q}{\|\bar{Y} - \hat{Y}\|^2/(m-p-1)} \sim F_{q, m-p-1} \text{ (sous } H_0)$$

SCR: Somme des carrés des résidus.

Une écriture équivalente de cette statistique du test est donnée par :

$$F = \frac{SCR(H_0) - SCR(H_1)}{SCR(H_1)} \frac{m-p-1}{q} \sim F_{q, m-p-1} \text{ (sous } H_0)$$



4) Test de contraintes linéaires entre paramètres.

Soit c un vecteur non nul de $(p+1)$ constantes réels. On souhaite tester l'hypothèse suivante :

$$H_0: c'\beta = r \quad (r \text{ connu})$$

On sait que $\hat{\beta} \sim N_{p+1}(\beta, \sigma^2(X'X)^{-1})$ et, alors, $c'\hat{\beta} \sim N(c'\beta, \sigma^2 c'(X'X)^{-1}c)$

Sous H_0 ,

$$\frac{c'\hat{\beta} - r}{\sqrt{c'(X'X)^{-1}c}} \sim t_{m-p-1}$$

avec $S^2 = \frac{\hat{E}'\hat{E}}{m-p-1}$ l'estimateur de la variance de l'erreur σ^2

Exemple d'application :

On suppose que parmi les variables explicatives de notre modèle, il y a la variable âge de l'assuré avec les classes d'âge $[18; 30[$; $[30; 40[$; $[40; 50[$; $[50; 60[$; $[60; \infty[$.

Supposé non significative

Or ici l'âge et les autres classes sont bien significatives. On va donc vouloir considérer les classes $[18; 40[$; ... ou $[18; 30[$; $[30; 50[$; ...

Pour cela on effectue le test $H_0: \beta_{[18; 30[} = \beta_{[30; 40[}$ ou $H_0: \beta_{[30; 40[} = \beta_{[40; 50[}$.

Tester $H_0: \beta_{[18; 30[} = \beta_{[30; 40[}$ revient à effectuer un test t avec $c' = (0 \ 0 \ 1 \ 0 \ 0)$ $\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$ et $r = 0$.

et sert à éventuellement regrouper des modalités.

© Théo Jalabert



Rappel:

Dans la construction d'un modèle de régression, il y a différentes étapes:

1) Définition du modèle.

2) Construction de la base de données.

3) Estimation du modèle (principalement le vecteur β et la variance de l'erreur σ^2) et tests statistiques (Student, ...)

4) Validation du modèle (étude des résidus du modèle, ...)

5) Utilisation du modèle

à des fins d'illustration d'un phénomène donné
prévision (ou prediction)

Trafic.

PRÉVISION :

Supposons que le modèle s'écrive : $y_i = \beta_0 + \beta_1 x_{1,i} + \dots + \beta_p x_{p,i} + \epsilon_i \quad i \in \{1, m\}$

Un des buts de la régression est de proposer des prévisions pour la variable Y .

Soit $x'_{m+1} = (1 \ x_{1,m+1} \ \dots \ x_{p,m+1})$ un vecteur de nouvelles valeurs pour les p variables explicatives.
On souhaite prédire y_{m+1} .

Or, $y_{m+1} = x'_{m+1} \beta + \epsilon_{m+1}$ avec $\begin{cases} E[\epsilon_{m+1}] = 0 \\ \text{Var}(\epsilon_{m+1}) = \sigma^2 \\ \text{Cov}(\epsilon_{m+1}, \epsilon_i) = 0 \quad \forall i \in \{1, m\} \end{cases}$

Grâce au modèle estimé à partir des m observations, on peut prédire y_{m+1} par $\hat{y}_{m+1} = x'_{m+1} \hat{\beta}$.

$\hat{y}_i \quad i \in \{1, m\}$.

↳ Valeurs ajustées
par le modèle.

↳ VALEURS PREDITES
par le modèle.

Remarque : Il y a une différence entre "ajustement" et "prévision".

La valeur (ici: la $(m+1)^{\text{ème}}$ valeur) pour laquelle nous effectuons la prévision n'a pas été utilisée lors de l'estimation du modèle et donc du β .

Donc cette quantité est \neq de \hat{y}_i , la valeur ajustée, qui, elle, fait intervenir y_i dans l'estimation du β .

On souhaite évaluer l'erreur associée à cette prédition, erreur qui provient de deux sources:

- le fait qu'on ne connaît pas ϵ_{m+1} et le fait que l'on a estimé β .

L'ERREUR DE PRÉVISION est définie comme $E_{\text{me}}^P = Y_{\text{me}} - \hat{Y}_{\text{me}}^P$

© Théo Jalabert

$$\mathbb{E}[E_{\text{me}}^P] = \mathbb{E}[Y_{\text{me}} - \hat{Y}_{\text{me}}^P] = \mathbb{E}[x_{\text{me}}' \beta - x_{\text{me}}' \hat{\beta}] = 0.$$

$$\begin{aligned} \mathbb{E}[E_{\text{me}}^P] &= \mathbb{E}[x_{\text{me}}' \hat{\beta}] = x_{\text{me}}' \mathbb{E}[\hat{\beta}] = x_{\text{me}}' \beta \\ \mathbb{E}[Y_{\text{me}}] &= \mathbb{E}[x_{\text{me}}' \beta] \\ &= x_{\text{me}}' \beta + \mathbb{E}[E_{\text{me}}] \\ &= x_{\text{me}}' \beta \end{aligned}$$

dépend de E_{me}

$$\begin{aligned} \text{Var}(E_{\text{me}}^P) &= \text{Var}(Y_{\text{me}} - \hat{Y}_{\text{me}}^P) = \text{Var}(Y_{\text{me}}) + \text{Var}(\hat{Y}_{\text{me}}^P) - 2 \text{Cov}(Y_{\text{me}}, \hat{Y}_{\text{me}}^P) \\ &= \sigma^2 + x_{\text{me}}' \cancel{\sigma^2 \beta} x_{\text{me}} \\ &= \sigma^2 [1 + x_{\text{me}}' (X'X)^{-1} x_{\text{me}}] \end{aligned}$$

2 sources d'erreurs

$$\begin{aligned} \text{Cov}(Y_{\text{me}}, \hat{Y}_{\text{me}}^P) &= 0 \\ \text{Car } Y_{\text{me}} &\text{ dépend de } E_{\text{me}} \\ \text{et } \hat{Y}_{\text{me}}^P &\text{ des } E. \end{aligned}$$

Or les erreurs sont incorrélées.

Remarque:

La variance de l'erreur de prévision peut s'écrire de la façon suivante :

$$\begin{aligned} \text{Var}(E_{\text{me}}^P) &= \mathbb{E}[(E_{\text{me}}^P)^2] - \mathbb{E}[E_{\text{me}}^P]^2 \\ &= \mathbb{E}[(Y - \hat{Y}_{\text{me}}^P)^2] \end{aligned}$$

On voit donc que la variance de l'erreur de prévision est mesurée par l'**ERREUR QUADRATIQUE MOYENNE DE PRÉVISION (EQMP)**

La version empirique de cette quantité nous donne un indicateur du pouvoir prédictif du modèle estimé.

$$\text{PRESS} = \sum_{i=1}^m (E_{(i)}^P)^2 \quad \text{avec} \quad E_{(i)}^P = Y_i - \hat{Y}_i^P = Y_i - x_i' \hat{\beta}_{(i)} \quad \text{où } \hat{\beta}_{(i)} \text{ est l'estimateur de } \beta \text{ obtenu dans le modèle privé de } i^{\text{ème observation.}}$$

PREDICTED RESIDUAL SUM OF SQUARES.

Lorsqu'on compare le pouvoir prédictif de deux modèles, on peut utiliser le PRESS et choisir celui avec le PRESS le plus petit.

On a vu comment calculer la valeur prédictive par le modèle, \hat{Y}^P , mais on peut aussi calculer des **INTERVALLES DE PRÉVISION**.

On rappelle que $E \sim N_m(0, \sigma^2 I_m)$ et $E_{\text{me}} \sim N(0, \sigma^2)$ et que E et E_{me} sont indép.
Alors,

$$E_{\text{me}}^P = Y - \hat{Y}_{\text{me}}^P \sim N(0, \sigma^2 [1 + x_{\text{me}}' (X'X)^{-1} x_{\text{me}}])$$

$$\frac{Y_{\text{me}} - \hat{Y}_{\text{me}}^P}{S \sqrt{1 + x_{\text{me}}' (X'X)^{-1} x_{\text{me}}}} \sim t_{m-p-1}$$

d'où, un **INTERVALLE DE PRÉVISION** de niveau $1-\alpha$ est donné par :

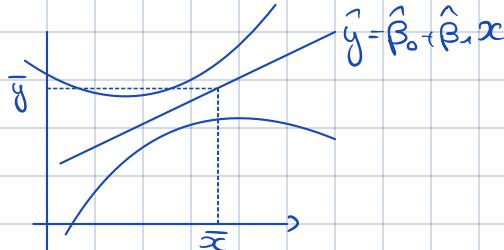
$$\hat{Y}_{\text{me}}^P \pm b_{(2, m-p-1)} S \sqrt{1 + x_{\text{me}}^T (X^T X)^{-1} x_{\text{me}}}$$

© Théo Jalabert avec $b_{(2, m-p-1)}$ le quantile d'ordre $\frac{1}{m-p-1}$ d'une Student à $m-p-1$ ddl.

Si on prend le cas de la régression simple (avec une seule variable explicative), on a que :

$$\text{Var}(\hat{Y}_{\text{me}}^P) = \sigma^2 \left(\frac{1}{m} + \frac{(x_{\text{me}} - \bar{x})^2}{\sum_{i=1}^m (x_i - \bar{x})^2} \right)$$

$$\text{Var}(\hat{\epsilon}_{\text{me}}^P) = \sigma^2 \left(1 + \frac{1}{m} + \frac{(x_{\text{me}} - \bar{x})^2}{\sum_{i=1}^m (x_i - \bar{x})^2} \right)$$



On remarque que effectuer de la prédiction pour des points qui sont éloignés du centre de gravité du nuage de points (\bar{x}, \bar{y}) , est dangereux et on risque d'obtenir des prévisions qui ne sont pas fiables.

VALIDATION DU MODÈLE DE RÉGRESSION ESTIMÉ

$$1\text{bis) } \epsilon \sim N_m(0, \sigma^2 I_m)$$

- 1) $E[\epsilon] = 0$
- 2) $\sum \epsilon = \sigma^2 I_m$
- 3) X déterministe
- 4) $\text{rg}(X) = p+1 < m$

On rappelle que le modèle s'écrit, sous forme matricielle, $Y = X\beta + \epsilon$ avec

L'**ANALYSE DES RESIDUS** du modèle constitue une partie importante de l'étape de validation du modèle.

On rappelle que les erreurs du modèle font l'objet d'un certain nb d'hypothèses qu'il va falloir vérifier.

Comme les erreurs on ne les observe pas, l'analyse est effectuée sur les RESIDUS du modèle:

$$\hat{\epsilon} = Y - \hat{Y} = Y - X\hat{\beta} = MY \quad \text{où } M = I - H$$

RESIDUS

et $H = X(X^T X)^{-1} X^T = [h_{ij}]$ est la matrice de projection ou matrice d'héritage.

On avait trouvé que $E[\hat{\epsilon}] = 0$ et $\text{Cov}[\hat{\epsilon}] = \sigma^2 M$

$$= \sigma^2 (I - H)$$

Les résidus sont donc centrés (comme les erreurs) mais ils ne sont :

- ni homoscedastiques
- ni corrélés

Car $\text{Var}(\hat{\epsilon}_i) = \sigma^2 (1 - h_{ii})$ et $\text{Cov}(\hat{\epsilon}_i, \hat{\epsilon}_j) = -\sigma^2 h_{ij}$

On a aussi que $\hat{\epsilon} \sim N_m(0, \sigma^2 M)$ (car $\hat{\epsilon} = M\epsilon$) $\Rightarrow \hat{\epsilon}$ gaussien.

déterministe

On peut définir différents types de résidus :

$$* R_i = \frac{\hat{\epsilon}_i}{\sqrt{\text{Var}(\hat{\epsilon}_i)}} = \frac{\hat{\epsilon}_i}{\sigma \sqrt{1 - h_{ii}}}$$

RESIDUS NORMALISÉS

avec ici l'élément sur la diagonale principale de la matrice de projection H . $i \in \{1, m\}$ et σ l'écart-type de l'erreur.

Comme la variance de l'erreur n'est pas connue, en pratique, on remplace σ^2 par son estimateur sans biais $S = \sqrt{\frac{\hat{E}'\hat{E}}{m-p-1}}$ et on obtient :

$$* h_i = \frac{\hat{E}_i}{S\sqrt{1-h_{ii}}} \quad i \in \{1, m\}$$

RESIDUS STANDARDISÉS

On peut encore remplacer S par $S_{(i)}$, l'estimateur de l'écart-type de l'erreur obtenu dans le modèle privé de l' $i^{\text{ème}}$ observation, ce qui permet d'obtenir :

$$* h_i^* = \frac{\hat{E}_i}{S_{(i)}\sqrt{1-h_{ii}}} \quad i \in \{1, m\}$$

RESIDUS STUDENTISÉS. (obtenus par VALIDATION CROISEE) (CROSS VALIDATION)

Les h_i^* sont construits en deux étapes :

1) On considère l'échantillon privé de l' $i^{\text{ème}}$ observation (donc de taille $m-1$), et on estime les paramètres inconnus du modèle, β et σ^2 par $\hat{\beta}_{(i)}$ et $S_{(i)}^2$

2) On considère maintenant que l' $i^{\text{ème}}$ observation $x_i^* = [1 \ x_{i1} \ \dots \ x_{ip}]$ est une nouvelle donnée et on prévoit y_i par $\hat{y}_i = x_i^* \hat{\beta}_{(i)}$ de façon classique.

Rappelez-vous que, lorsqu'on avait parlé de prévision, on avait dit que $\frac{Y_{mi} - \hat{Y}_{mi}}{S\sqrt{1+x_{mi}^*(X^*X)^{-1}x_{mi}}} \sim t_{m-p-1}$

Dans notre cas, cela se réécrit de la façon suivante :

$$h_i^* = \frac{y_i - \hat{y}_i^*}{S_{(i)}\sqrt{1+x_i^*(X_{(i)}^*X_{(i)})^{-1}x_i}} \sim t_{m-p-2} \quad \text{avec } X_{(i)} \text{ la matrice } X \text{ privée de la } i^{\text{ème}} \text{ ligne.}$$

I l y a un + car $H = -H$.

Au final, les résidus que l'on retiendra pour notre analyse, ce sont les h_i^* pour les deux raisons suivantes :

1) $h_i^* \sim t_{m-p-2}$ ce qui permet de détecter des valeurs aberrantes et, de plus, on peut montrer que

$$h_i^* = h_i \sqrt{\frac{m-p-2}{m-p-1-h_i^2}}$$

Donc h_i^* est une fonction monotone de h_i et permet de mieux détecter les valeurs aberrantes.

En effet, lorsque $h_i > 1$, $h_i^* > h_i$ car $\sqrt{\frac{m-p-2}{m-p-1-h_i^2}} > 1$.

2) h_i^* fait intervenir $S_{(i)}$ et pas S pour l'estimation de l'écart-type de l'erreur. Ceci est un avantage puisque si l' $i^{\text{ème}}$ observé est aberrant, elle aura un impact exclusivement sur le numérateur de h_i^* et pas sur le dénominateur et cela évitera une éventuelle compensation entre numérateur et dénominateur comme on peut l'avoir pour h_i .

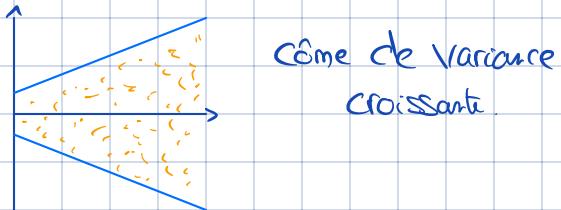
© Théo Jalabert

Une bonne analyse graphique des résidus consiste à les représenter en fonction de quantiles du théorème.



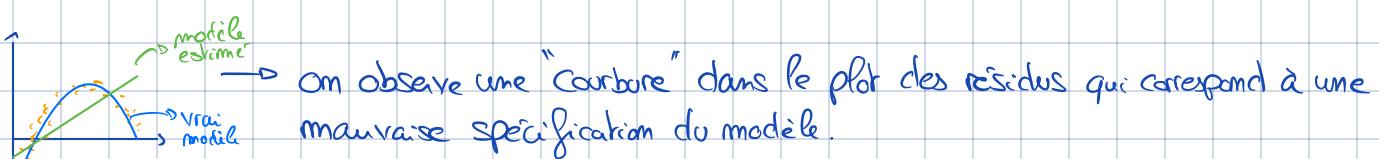
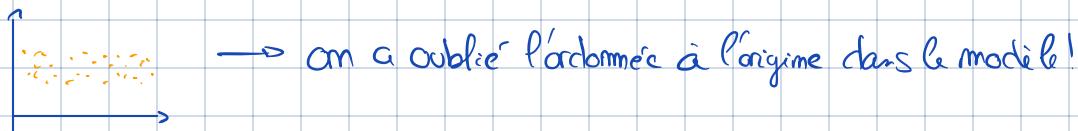
On s'attend à observer des résidus qui se distribuent autour de l'axe des abscisses (puisque ils sont centrés à zéro) de façon aléatoire (pas de "structure" dans les résidus) et on s'attend à ce que 95% de ces résidus soit à l'intérieur de ces bornes de confiance [-2, 2].

Le simple plot de résidus est aussi un moyen de détecter d'éventuels problèmes.
Par exemple:



Lorsque les hypothèses du modèle sont vérifiées, on s'attend à ce que la corrélation entre les résidus et les valeurs ajustées par le modèle soit nulle.

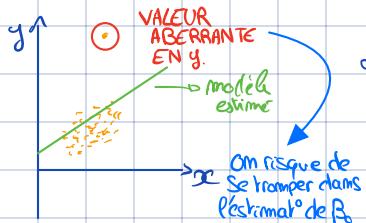
Dans le modèle de régression, on trouve $\hat{E}'\hat{Y} = 0$.



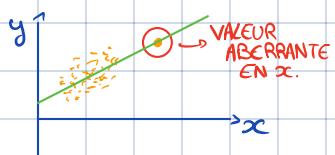
VALEURS ABERRANTES ET/OU INFLUENTES DANS UN MODÈLE LINÉAIRE

Le plot des résidus peut être utile afin de détecter des valeurs aberrantes qui peuvent ou pas constituer un problème.

On va voir quelques exemples:



La valeur aberrante en y peut être problématique puisqu'elle a comme effet d'attirer la droite de régression et alors on se retrouve probablement avec un modèle qui ne représente ni la majeure partie des données ni les valeurs aberrantes.

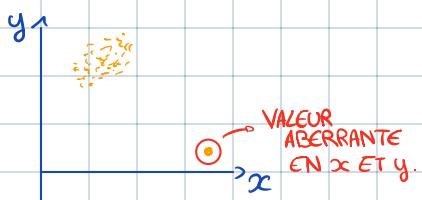


Dans cet exemple, la valeur aberrante en x ne pose pas de problème et, au contraire, a un effet "benifique" sur l'estimation du modèle car dans le modèle $y = \beta_0 + \beta_1 x + \epsilon$.

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

et que la présence de

Cette valeur aberrante fait augmenter la variance de X et diminuer la variance d'estimation (donc estimations plus fiables et précises !) © Theo Jalabert 



Dans ce cas, on risque de se tromper dans l'estimation de β_0 et β_1 .

Ici la valeur aberrante est aussi **INFLUENTE** c'ds qu'elle a une influence sur l'estimation des paramètres du modèle.
(Le modèle avec ou sans la valeur en question me donne pas les mêmes estimations).

MESURES D'INFLUENCE

Supposons que nous avons détecté des valeurs aberrantes dans notre échantillon, c'est important d'établir si ces valeurs sont influentes ou pas au moyen d'indicateurs que nous allons définir.

* MESURE D'INFLUENCE BASEE SUR LA MATRICE DE PROJECTION H.

On rappelle que $H = X(X'X)^{-1}X'$ et que $\hat{Y} = HY$.

De plus, $\hat{Y}_i = \hat{Y}_{H_i Y} = H \hat{Y}_i H' = H \sigma^2 I_m H = \sigma^2 H H = \sigma^2 H$

$\hookrightarrow H$ idempotente.

d'où $\text{Var}(\hat{Y}_i) = \sigma^2 h_{ii}$ ce qui implique $h_{ii} > 0$ (puisque la variance est toujours positive)

On rappelle aussi que $\hat{\epsilon} = MY$ et $\hat{\epsilon}_i = \hat{\epsilon}_M = \sigma^2(I - H)$ et donc $\text{Var}(\hat{\epsilon}_i) = \sigma^2(1 - h_{ii})$
 $\Rightarrow h_{ii} < 1$.

On a donc trouvé que $0 < h_{ii} < 1$.

De plus, $\sum_{i=1}^m h_{ii} = \text{Tr}(H) = \text{Tr}(X(X'X)^{-1}X') = \text{Tr}((X'X)^{-1}X'X) = p + 1$.

On a aussi que la variance des résidus estimés est d'autant plus faible que h_{ii} est grand et la valeur de h_{ii} mesure l'influence de la $i^{\text{ème}}$ observation sur \hat{y}_i (ou, autrement, le poids de la $i^{\text{ème}}$ observation dans la construction de \hat{y}_i). On va le voir tout de suite!

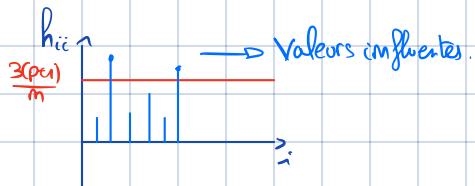
La matrice H est une matrice de projection, symétrique et idempotente. La $i^{\text{ème}}$ composante de $\hat{Y} = HY$ s'écrit:

$$\hat{Y}_i = \sum_{j=1}^m h_{ij} \hat{Y}_j = h_{ii} \hat{Y}_i + \sum_{\substack{j=1 \\ j \neq i}}^m h_{ij} \hat{Y}_j \quad i \in \{1, m\}$$

\hookrightarrow représente le rôle joué par la $i^{\text{ème}}$ observation dans la détermination de \hat{Y}_i .

Si $h_{ii} = 1$ alors \hat{Y}_i est déterminé par la seule donnée y_i alors que si $h_{ii} = 0$ alors l' $i^{\text{ème}}$ observation n'a aucune influence sur \hat{Y}_i .

Or, on voudrait que toutes les observations participent de la même manière à la construction des \hat{y}_i . Comme $\sum_{i=1}^m h_{ii} = p+1$, la moyenne des h_{ii} vaut $(p+1)/m$ ainsi on considère que la i ème valeur est influente si $h_{ii} > 2 \frac{(p+1)}{m}$ (ou $h_{ii} > \frac{3(p+1)}{m}$).



Un autre indicateur qui permet de détecter une observation influente est la **DISTANCE DE COOK**

L'idée est de "mesurer" l'influence de l' i ème observation sur l'estimation de β en comparant l'estimateur de β obtenu dans le modèle avec les m observations, $\hat{\beta}$, à l'estimateur de β obtenu dans le modèle privé de l' i ème observation, $\hat{\beta}_{(i)}$.

Intuitivement, si l'écart est "grand" alors l' i ème observation doit être considérée comme influente.

$\hat{\beta} \in \mathbb{R}^{p+1}$ et, en général, une distance basée sur un produit scalaire s'écrit:

$$d(\hat{\beta}_{(i)}, \hat{\beta}) = (\hat{\beta}_{(i)} - \hat{\beta})' Q (\hat{\beta}_{(i)} - \hat{\beta})$$

où $Q \in \mathbb{S}_{p+1}^+$ (\mathbb{R})

Sous l'hypothèse de normalité et lorsque σ^2 est inconnue,
la **REGION DE CONFIANCE** du vecteur β de niveau α s'écrit:

$$RC_\alpha(\beta) = \left\{ \beta \in \mathbb{R}^{p+1} \mid \frac{1}{(p+1)\hat{S}^2} (\hat{\beta} - \beta)' (X'X) (\hat{\beta} - \beta) \leq F_{p+1, m-p-1}^{1-\alpha} \right\}$$

avec $F_{p+1, m-p-1}^{1-\alpha}$ le quantile d'ordre $1-\alpha$
d'une $\mathcal{F}(p+1, m-p-1)$

Cette inégalité définit un ellipsoïde centré en $\hat{\beta}$ et l'influence de l' i ème observation peut être mesurée par le décentrage de cet ellipsoïde quand on supprime l' i ème observat° de l'échantillon.

Au final, la **distance de Cook** est définie comme:

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})' (X'X) (\hat{\beta}_{(i)} - \hat{\beta})}{S^2(p+1)} \quad i \in \{1, m\}.$$

On compare D_i aux quantiles de la loi de Fisher $\mathcal{F}(p+1, m-p-1)$ et on considère que l' i ème observat° est influente au sens de Cook si D_i dépasse le quantile.

On pourrait se dire que pour calculer D_i il faut estimer m modèles mais en réalité on peut utiliser l'expression suivante

$$D_i = \frac{h_{ii}}{(p+1)(1-h_{ii})} f_i^2 \quad \text{avec } h_{ii} \text{ l'}i\text{ème élément sur la diagonale principale de la matrice de projection } H.$$

De cette écriture, on voit que D_i mesure à la fois le caractère aberrant de l' $i^{\text{ème}}$ observat° (quand h_i est élevé) et le caractère influent (quand $\frac{h_{ii}}{1-h_{ii}}$ est élevé).

Un autre type d'indicateur d'influence est le DFFITS (ou ECART de WELSH-KU) défini comme:

$$\text{DFFITS}_i = |h_i^*| \sqrt{\frac{h_{ii}}{1-h_{ii}}} \quad \text{où } h_i^* \text{ est le résidu studentisé obtenu en remplaçant } S_i^2 \text{ (inconnu) par } S_{(i)}^2$$

Si la matrice H est "équilibrée", on devrait avoir $h_{ii} \approx \frac{p+1}{m}$ et donc $\frac{h_{ii}}{1-h_{ii}} \approx \frac{p+1}{m-p-1}$; donc,

Si $|DFFITS_i| > \sqrt{\frac{p+1}{m-p-1}} \sqrt{\frac{1-\alpha/2}{m-p-2}}$ l' $i^{\text{ème}}$ observat° risque d'être influente.

Le but, maintenant, est d'aller regarder de plus près certaines hypothèses du modèle.
 On va commencer par l'hypothèse :

$$\mathbb{E}\epsilon = \mathbb{E}(\epsilon\epsilon') = \sigma^2 I_m \quad \text{donc erreurs homoscedastiques et corrélées.}$$

En pratique, on peut rencontrer :

- * erreurs hétéroscedastiques
- * erreurs hétéroscedastiques et corrélées.

On suppose que la "nouvelle" matrice de variance et covariance des erreurs s'écrit :

$$\mathbb{E}\epsilon = \Sigma \quad \text{avec } \Sigma \text{ une matrice symétrique et définie positive et } \Sigma \neq 0.$$

On se demande quel va être l'impact de cette modification sur les estimations des paramètres du modèle.
 L'estimateur des MCO de β sera toujours $\hat{\beta} = (X'X)^{-1}X'Y$ et reste sans biais ($\mathbb{E}(\hat{\beta}) = \beta$) mais

$$\mathbb{E}\hat{\beta} = \mathbb{E}_{(X'X)^{-1}X'Y} = (X'X)^{-1}X'\mathbb{E}_Y X(X'X)^{-1} = (X'X)^{-1}X'\Sigma X(X'X)^{-1} = \Sigma(X'X)^{-1}X'VX(X'X)^{-1} \neq \Sigma(X'X)^{-1}$$

et donc $\hat{\beta}$ n'est plus un estimateur BLUE i.e. le meilleur estimateur linéaire sans biais.

La conclusion est que, en cas d'hétéroscedasticité, il faudrait ne pas utiliser les MCO mais utiliser, pour estimer β , la méthode des MOINDRES CARRÉS PONDÉRÉS (MCP). Si en plus, on a un problème d'erreurs corrélées, alors on utilise la méthode des MC généralisés (MCG).

Parfois, pour résoudre un problème d'asymétrie des résidus ou d'hétéroscedasticité, une simple transformation de la variable Y suffit (sans passer par les MCP ou les MCG) et la transformation

qui va être utilisée est celle dictée par la **TRANSFORMEE DE BOX-COX** qui est une transformation linéaire qui s'écrira comme :

© Théo Jalabert

T. Jalabert

$$Y^{(\lambda)} = \begin{cases} \frac{Y^{\lambda}-1}{\lambda} & \text{si } \lambda \neq 0, Y > 0 \\ \ln Y & \text{si } \lambda = 0 \end{cases}$$

Selon la valeur du paramètre λ , la transformation à appliquer à la variable Y est différente. λ doit être estimé à partir des données de l'échantillon en maximisant la fonction de vraisemblance PROFILEE (voir TD).

On va maintenant s'intéresser à l'hypothèse $rg(X) = p$ où p est le nb de variables explicatives du modèle. Si le $rg(X)$ n'est pas plein ou si certaines variables explicatives sont très fortement corrélées entre elles, on parle de **MULTICOLINÉARITÉ** et cela représente un problème car, en présence de multicolinéarité dans le modèle, on va avoir du mal à inverser la matrice $(X'X)$ et donc à définir un estimateur $\hat{\beta}$ de β .

De plus, la $\text{Var}(\hat{\beta})$ explose (puisque $\text{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}$ et, en présence de multicolinéarité, le déterminant de $X'X$ tend vers zéro) et donc les estimations ne sont pas robustes ni précises.

Cela a aussi un effet sur la significativité des variables. On se rappelle que la stat du test de significativité ($H_0: \beta_j = 0$) est : $\frac{\hat{\beta}_j}{\sqrt{\text{Var}(\hat{\beta}_j)}}$

La $\text{Var}(\hat{\beta}_j)$ augmente, la valeur de la stat du test de significativité diminue et on se déplace donc vers la région d'acceptation du test (du coup on a tendance à dire que certaines variables explicatives ne sont pas significatives alors que probablement en l'absence de multicolinéarité elles auraient été significatives).

Afin de détecter un problème de multicolinéarité, on peut :

- * aller regarder une simple matrice des corrélations linéaires entre variables explicatives qui va permettre de détecter des relations entre variables 2 par 2.
- * Calculer le **FACTEUR D'INFLATION DE LA VARIANCE** ou **VIF** (VARIANCE INFLATION FACTOR) qui est défini comme :

$$\text{VIF}_j = \frac{1}{1 - R_j^2} \quad j \in [1, p] \quad \text{avec } R_j^2 \text{ le coeff de déterm. du modèle } X_j \text{ contre tous les autres.}$$

variables explicatives du modèle ($X_j = \beta_0 + \beta_1 X_1 + \dots + \beta_{j-1} X_{j-1} + \beta_j X_j + \dots + \beta_p X_p + \epsilon$).
Dans la situation idéale, $\text{VIF}_j = 1$ (puisque $R_j^2 = 0$).

Cet indicateur est communément appelé facteur d'inflation de la variance puisqu'il peut s'écrire comme le rapport entre la $\text{Var}(\hat{\beta}_j)$ dans le modèle que l'on estime et la $\text{Var}(\hat{\beta}_j)$ dans un

modèle où les variables explicatives sont orthogonales entre elles.

© Théo Jalabert



Donc, c'� nous dit de combien la $\text{Var}(\hat{\beta}_j)$ augmente par rapport à la situation "idéale".

En effet, on peut écrire $\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\text{Dev}(X_j)(1-R_j^2)}$ où $\text{Dev}(X_j)$ est le numérateur de la $\text{Var}(X_j)$.

Si R_j^2 est élevé mais qu'en même temps $\text{Dev}(X_j)$ est élevée alors la $\text{Var}(\hat{\beta}_j)$ n'explique pas donc la multicollinearité n'est pas toujours un problème et ça peut entraîner l'estimation de tous les β ou alors seulement de quelques paramètres.

* Calculer l'**INDICE DE CONDITIONNEMENT** qui est défini comme

$$K = \frac{\max\{\lambda_j\}_1^p}{\min\{\lambda_j\}_1^p} \quad \text{où } \lambda_1, \dots, \lambda_p \text{ sont les valeurs propres de la matrice } X'X.$$

Or, $|X'X| = \prod_{j=1}^p \lambda_j$ et des problèmes numériques ou des variances excessives apparaissent quand les dernières valeurs propres (ordonnées en ordre décroissant) sont trop petites. En pratique, si $K < 100$, on considère qu'il n'y a pas de pb de multicollinearité, alors qu'il y a un pb important si $K > 1000$.

REMEDES AU PROBLEME DE MULTICOLINÉARITÉ

- 1) éliminer a priori (avant d'estimer le modèle) une ou plusieurs variables explicatives.
- 2) obtenir des observations supplémentaires pour les variables explicatives afin d'augmenter $\text{Dev}(X_j)$.
- 3) effectuer une **REGRESSION SUR COMPOSANTES PRINCIPALES (PCR - PRINCIPAL COMPONENT REGRESSION)** qui consiste à remplacer les variables explicatives de départ par de nouvelles variables obtenues comme combinaison linéaire des variables de départ et qui sont orthogonales et de variance maximale. → Problème d'interprétabilité des nouvelles variables!
- 4) **REGRESSION PLS (PARTIAL LEAST SQUARES)**: Comme au point 3), la régression est effectuée sur de nouvelles variables, combinaison linéaire des variables initiales, qui sont orthogonales entre elles et classées par ordre d'importance (en prenant en compte leur covariance avec Y).
- 5) **REGRESSION RIDGE**: En présence d'un problème de multicollinearité, le déterminant de la matrice $(X'X)$ est proche de 0, ce qui veut dire aussi que une ou des valeurs propres de cette matrice sont égales à 0 ou proches de 0. On note $\{\lambda_j\}_{j=1}^p$ les valeurs propres de $(X'X)$; l'idée est d'agir sur les valeurs propres de la matrice afin d'éviter que le déterminant soit proche de 0. Un résultat d'algèbre linéaire nous dit que les matrices $(X'X)$ et $(X'X + \delta I_p)$,

avec δ une constante positive, on a les mêmes vecteurs propres mais des valeurs propres différentes : $\{\lambda_j\}$ et $\{\lambda_j + \delta\}$ respectivement.

© Théo Jalabert

On pourrait donc penser à remplacer la matrice $(X'X)$ en $\hat{\beta}$ par la matrice $(X'X + \delta I_p)$, ce qui permettrait d'augmenter toutes les valeurs propres, y compris celles proches de zéro, et obtenir donc un coefficient $\hat{\beta}$ unique et stable.

Cette méthode, connue comme **régression ridge**, consiste donc à estimer β par

$$\hat{\beta}_R(\delta) = (X'X + \delta I)^{-1}X'Y \quad (\text{rappelons-nous qu'on aurait } \hat{\beta} = (X'X)^{-1}X'Y)$$

avec δ une constante positive à déterminer.

Si $\delta \rightarrow 0$, alors $\hat{\beta}_R \rightarrow \hat{\beta}$; si par contre $\delta \rightarrow +\infty$ alors $\hat{\beta}_R(\delta) \rightarrow 0$.

Que peut-on dire sur les propriétés de $\hat{\beta}_R$?

$$\begin{aligned} * E[\hat{\beta}_R] &= (X'X + \delta I)^{-1}X'E[Y] = (X'X + \delta I)^{-1}X'X\beta \\ &\quad = X\beta \\ &= (X'X + \delta I)^{-1}(X'X + \delta I - \delta I)\beta \\ &= \beta - \delta(X'X + \delta I)^{-1}\beta \neq \beta \end{aligned}$$

donc $\hat{\beta}_R$ est un estimateur biaisé de β (à la différence de $\hat{\beta}$).

$$\begin{aligned} * \text{Var}_{\hat{\beta}_R} &= \text{Var}_{(X'X + \delta I)^{-1}X'Y} = (X'X + \delta I)^{-1}X'\text{Var}_Y X(X'X + \delta I)^{-1} \\ &\quad \text{X déterministe} \\ &\quad \text{Y v.a.} \\ &= \delta^2 (X'X + \delta I)^{-1}X'X(X'X + \delta I)^{-1} \neq \text{Var}_{\hat{\beta}} = \sigma^2 (X'X)^{-1} \end{aligned}$$

$$A^{-1} = \frac{1}{\det A} \dots$$

quand $\det A \neq 0$ alors A^{-1}

$\text{Var}_{\hat{\beta}_R}$ fait intervenir $(X'X + \delta I)$ au lieu de $(X'X)$, ce qui permet de diminuer la variance d'estimation puisque les valeurs propres de $(X'X + \delta I)$ sont plus élevées que celle de $(X'X)$. La valeur de la constante δ peut être choisie par **VALIDATION CROISEE**.

→ SUIVRE AVEC LE POLycopié.

Le **MODÈLE LINÉAIRE GÉNÉRALISÉ (MLG)** est une généralisation du modèle visant à faire face à certains limites du modèle linéaire, à savoir :

1) Caractère gaussien de la variable à expliquer.

En effet en tarification, par exemple, on calculera la prime pure en utilisant un modèle coût moyen-fréquence dans lequel on cherche à expliquer soit le montant du sinistre, soit le nombre de sinistres et aucune variable n'est gaussienne (on observe une loi asymétrique pour le montant et le nb de sinistres est plutôt modélisé par une variable de comptage).

De plus, le modèle linéaire ne permet pas d'expliquer une variable qualitative, ce qui pourrait être utile, par exemple, lorsqu'on souhaite expliquer le choix par un assuré d'un certain niveau de

2) Dans le modèle linéaire, on fait l'**hypothèse de variance constante** (homosérialité)
i.e. $V(Y_i) = V(E_i) = \sigma^2 \quad \forall i \in \{1, \dots, n\}$. Cette hypothèse n'est pas toujours réaliste puisque ça revient à dire que seulement la moyenne de Y varie avec les variables explicatives (et pas la variance) alors qu'en pratique, on observe souvent un phénomène d'inflation de la variance (donc en général elle n'est pas constante).

3) Il n'est pas possible de construire un modèle à coeff corrélés (quoiqu'on peut estimer un modèle gaussien sur le log de la variable à expliquer (si elle est positive)).

$$\log Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + E \quad \text{avec } \log Y \sim N(\cdot, \cdot)$$

Slide 2

$\theta \in \mathbb{R}$: paramètre canonique (ou paramètre de la moyenne).

$\phi \in \mathbb{R}$: paramètre de dispersion

a: fonction définie sur les réels et non nulle

b: $\frac{d}{dx} a(x) = -\frac{1}{a(x)}$ et 2 fois dérivable (Δ^2)

c: $\mathbb{R}^p \rightarrow \mathbb{R}^2$

Font partie de la famille exponentielle la majorité partie des lois usuelles:

gaussiennes, gamma, Poisson, Bernoulli, Binomiale ...

Ce qui traduit une richesse de lois possibles pour la variable à expliquer Y qu'on n'avait pas dans le modèle gaussien.

N.B: pour les lois qui ont un seul paramètre comme par exemple la loi de Poisson, on pose $\phi = 1$ (ϕ étant le paramètre de dispersion de la famille exponentielle).

Pour les autres lois (à 2 paramètres), ϕ n'est pas connu, en général, il est estimé à l'aide des données.

Le MODELE LINÉAIRE GÉNÉRALISÉ s'écrit:

$$* g_m \left(\frac{\mathbb{E}[Y]}{\mu} \right) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

$$* g_m \begin{pmatrix} \mathbb{E}[Y_1] \\ \vdots \\ \mathbb{E}[Y_m] \end{pmatrix} = \beta_0 + \beta_1 \begin{pmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1m} \end{pmatrix} + \dots + \beta_p \begin{pmatrix} X_{p1} \\ X_{p2} \\ \vdots \\ X_{pm} \end{pmatrix}$$

$$* g \left(\frac{\mathbb{E}[Y_i]}{\mu_i} \right) = \beta_0 + \beta_1 X_{1i} + \dots + \beta_p X_{pi} = x_i' \beta \text{ où } \begin{cases} x_i' = (X_{1i}, \dots, X_{pi}) \\ \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \end{cases}$$

On a 4 écritures \neq pour le même modèle.

Écriture rapide du MLG

$$g(\mu_i) = \underbrace{x_i' \beta}_{\eta_i} \quad \Rightarrow \eta_i: \text{SCORE CL des variables explicatives}$$

On reprend la 1^{ère} égalité :

$$g_m(\mathbb{E}[Y]) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

Annotations sur l'équation :

- g_m (FONCTION LIEN)
- $\mathbb{E}[Y]$ (VARIABLE À EXPLIQUER)
- $\beta_0, \beta_1, \dots, \beta_p$ (COEFFICIENTS (ou PARAMÈTRES) du MLG)
- P VARIABLES EXPLICATIVES

N.B.: Le modèle linéaire est un cas particulier du MLG que l'on obtient en choisissant la loi Gaussienne pour Y et la fonction lien identité.

Etapes d'un MLG :

- ① Recueillir des observations à la fois pour la variable Y à expliquer (dans le cas de la tarification et donc du modèle fréquence - coût moyen, il s'agira du nombre de sinistres et du montant des sinistres) et pour les variables explicatives $X_j, j \in [1, p]$ de façon à constituer un échantillon de taille n .
- ② Choisir une loi pour la va Y . La loi doit être choisie dans la famille exponentielle et le choix dépendra du type de variable à expliquer. Si, par exemple, Y est un nb de sinistres, on choisira la loi de Poisson ou la loi Binomiale.
- ③ Choisir la fonction lien g . On peut choisir la fonction canonique mais cela n'est pas obligatoire ! En tarification, par exemple, on préfère utiliser la fonction lien Log qui donne un modèle multiplicatif et permet d'obtenir facilement les "coefficients correcteurs" utiles pour le calcul de la prime.
- ④ Estimer la loi du modèle spécifié auparavant, grâce à l'échantillon de données de l'étape 1. Cela revient à estimer les $\beta_j, j \in [0, p]$ et si inconnu, le paramètre de dispersion ϕ .
- ⑤ Valider le modèle en passant par des tests statistiques, l'analyse des résidus du modèle et des indicateurs de la qualité d'ajustement.
- ⑥ Utiliser le modèle pour faire de la prédiction ou, dans notre cas, de la tarification.

Slide 3:

$$\mathbb{E}[Y] = b(\phi) = \frac{\partial b(\phi)}{\partial \phi}$$

$$V(Y) = a(\phi) \underbrace{b''(\phi)}_{V(\mu)} = a(\phi) V(\mu)$$

V(\mu): FONCTION VARIANCE

Dans le modèle linéaire, on a donc que $V(Y_i) = \sigma^2$ $Y = X\beta + \epsilon$
 $\mathbb{E}[Y] = X\beta$

LOIS	FONCTION VARIANCE $V(\mu)$
NORMALE $\mathcal{N}(\mu, \sigma^2)$	1
POISSON $P(\mu)$	μ
GAMMA $\Gamma(\alpha, \beta)$	μ^2
BINOMIALE $B(n, p)$	$\mu(1-\mu)$
GAUSSIENNE INVERSE	μ^3

Rq: Dans le MLG, mis à part le cas de la loi Gaussienne, la variance est toujours fonction de la moyenne, μ , qui varie avec les variables explicatives du modèle. La variance n'est donc pas constante comme c'était le cas pour le modèle linéaire.

De plus, cela nous permet de voir le MLG comme un modèle pas seulement pour la moyenne de Y mais aussi pour la variance (puisque la variance varie avec μ qui varie avec les variables explicatives du modèle).

Souvent, en pratique, on remplace $a(\phi)$ par $\frac{\phi}{w_i}$ avec w_i des poids communs, attribués à chacune des m observations.

Par exemple, on peut imaginer une pondération croissante avec le temps qui consiste à attribuer un poids plus faible aux observations plus éloignées dans le temps si la période est très longue. Donc on aura:

$$V(Y_i) = \frac{\phi}{w_i} V(\mu).$$

L'utilisation des poids, w_i , peut aussi exprimer la possibilité pour Y_i de représenter une observation agrégée. Par exemple, si Y_i est la moyenne de m observations, alors $w_i = m$.

Pour la Poisson, comme $\phi=1$, on a $V(Y_i) = V(\mu)$

Slide 9:

* Le Hessian est la matrice des dérivées secondes: $[H]_{ij} = \frac{\partial^2 \mathcal{L}(y_i; \beta)}{\partial \beta_i \partial \beta_j}$

* La matrice d'information est la matrice $I = X'W X$ de forme général:

$$[I]_{jk} = -\mathbb{E}\left[\frac{\partial^2 \mathcal{L}(\beta)}{\partial \beta_j \partial \beta_k}\right] = -\sum_{i=1}^m \frac{x_{ji} x_{ki}}{\text{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

et où W est la matrice diagonale de pondération $[W]_{ii} = \frac{1}{\text{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$

$$\text{et } X = \begin{pmatrix} 1 & x_{11} & \dots & x_{p1} \\ 1 & x_{21} & \dots & x_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & \dots & x_{pm} \end{pmatrix}$$

Remarque : Dans les équations de vraisemblance pour β (slide 9), le ϕ n'apparaît plus (ϕ est une constante) ce qui veut dire qu'on peut estimer β sans se soucier de ϕ . © Théo Jalabert

Slide 10:

La fonct° lien canonique associée à la structure exponentielle est tq :

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

Le MLG est $\underbrace{g(\mu_i)}_{\text{L} = \eta_i} = x_i' \beta = \eta_i$

$L = \eta_i$ si g fonct° lien canonique

Les équations de vraisemblance "simplifiées" s'écrivent :

$$\sum_{i=1}^n \omega_i (y_i - \hat{\mu}_i) x_{ji} = 0 \quad j = 0, \dots, p$$

\hookrightarrow c'est un résidu !

Cette relation traduit l'orthogonalité entre les résidus du modèle et les variables explicatives (intuitivement il n'y a plus rien dans les variables explicatives qui puisse apporter de l'information sur les résidus). On a la même relation dans le modèle gaussien :

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

Remarque : quand on utilise la fonct° lien canonique, comme les termes $\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_k}$ ne dépendent plus de y_i , on peut montrer que le Hessian et la matrice d'informat° coïncident et donc les 2 matrices numériques coïncident aussi.

Slide 12:

En réalité, en pratique on n'estimerait jamais un modèle saturé puisque, pour obtenir des estimations robustes, on demande à travailler avec une taille d'échantillon n beaucoup plus grande que le nb de paramètres à estimer dans le modèle.

De plus, en général, par un modèle statistique on cherche à résumer l'information plutôt qu'à la reproduire exactement.

Slide 13:

$$D = 2 \ln(\lambda) = 2(l_{\text{SAT}} - l)$$

avec l_{SAT} et l les fonct° de log-vraisemblance du modèle saturé et du modèle estimé, respectivement.

Si $\frac{D}{m-p-1} \approx 1$, alors ajustement bon.

© Théo Jalabert

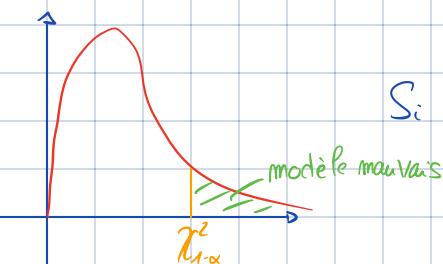
$$\hookrightarrow \mathbb{E}[\chi^2_{m-p-1}]$$



Remarque : Dans le cas où on extrime un MLG avec $Y_i \sim B(q_i)$, la déviance n'est pas un bon indicateur de la qualité d'ajustement du modèle. On verra en TD qu'elle dépend uniquement des valeurs ajustées \hat{q}_i des q_i et pas de y_i .

Slide 14:

$$\text{Var}(Y_i) = \frac{a(\phi)}{\phi} b''(\theta_i) \xrightarrow{\text{V}(\mu_i)} \text{FONCTION VARIANCE}$$



Si $\chi^2 > \chi^2_{1-\alpha; m-p-1}$ alors l'ajustement par le modèle estimé est mauvais.

Règle empirique: Si $\frac{\chi^2}{m-p-1} \approx 1$ alors l'ajustement par le modèle estimé est bon.

Slide 15:

ϕ est en général inconnu sauf pour les lois qui présentent un seul paramètre, comme par exemple, la loi de Poisson pour laquelle $\phi=1$ (commu).

$$\underline{\text{MVI}}: \frac{\partial \ell(y_i; \theta, \phi)}{\partial \phi} = 0 \Rightarrow \hat{\phi} \quad (\text{solution numérique}).$$

L'estimateur $\hat{\phi}$ est plutôt instable, donc peu utilisé en pratique.

L'estimateur $\hat{\phi}$ de Pearson, est basé sur un développement de Taylor à l'ordre 2 de la fonction log-vraisemblance.

Slide 16:

L'idée est de comparer 2 modèles, un modèle sous H_0 et un modèle sous H_1 en sachant que le modèle sous H_0 est plus "petit" (moins de paramètres inconnus) et surtout il est emboité dans le modèle sous H_1 .

H_0 : modèle restreint \hookrightarrow modèle obtenu en rajoutant des contraintes sur les paramètres du modèle
 H_1 : modèle complet

D_0 est la déviance du modèle sous H_0 .

© Théo Jalabert



$$D_0 = 2(l_{SAT} - l_{B_0}) \quad D_1 = 2(l_{SAT} - l_{B_1})$$

$$\Delta = 2(l_{B_1} - l_{B_0}) \quad \text{la stat du modèle.}$$

On se demande si le fait de rajouter des variables explicatives dans le modèle permet de faire "significativement" diminuer la déviance du modèle ou pas !

Slide 17.

- * H_0 : modèle restreint \rightarrow on rajoute des contraintes sur certains paramètres du modèle sous H_1 .
- * H_1 : modèle complet
- * $\tilde{\beta}$: EMV de β dans le modèle sous H_1
- * $\tilde{\beta}_0$: EMV de β dans le modèle sous H_0 .

Slide 18.

- * \tilde{L} vraisemblance du modèle sous H_1
- * \tilde{L}_0 vraisemblance du modèle sous H_0

q : nb de contraintes sur les paramètres dans le modèle sous H_0 .

NB: SAS utilise le test du rapport de vraisemblance pour effectuer les analyses de type 1 et de type 3 qui seront traitées en cours.

Slide 20.

Dans le modèle linéaire, le test de significativité est un test de Student alors que pour les modèles linéaires généralisés, le test de significativité est un test du χ^2 de Wald.

$$H_0: \beta_j = 0$$

La statistique du test s'écrit $\frac{\hat{\beta}_j^2}{\text{Var}(\hat{\beta}_j)} \stackrel{\text{sous } H_0}{\sim} \chi_1^2$

$$H_0: \beta_1 = \dots = \beta_p = 0 \quad)$$

Dans le modèle linéaire, c'était un test de Fisher.
Pour les MLG, il s'agit d'un test de Wald du χ^2 sur le modèle global.

Un autre exemple est celui où l'on souhaite grouper deux modalités (i et $j+1$) d'une variable explicative, et donc on doit tester $H_0: \beta_i = \beta_{j+1}$ © Théo Jalabert

$$C = (0 \ 0 \ \frac{1}{\delta} \ \frac{-1}{\delta} \ 0 \ \dots \ 0) \quad (\text{pour } r=0)$$

Slide 21:

Il est la fonction log de vraisemblance comme :

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

On rejette H_0 pour de grandes valeurs de la statistique du test puisque si le modèle est le modèle sous H_0 alors $\ell'(\tilde{\beta})$ devrait être proche de zéro.

Slide 22:

$$\text{Avant } H = X(X'X)^{-1}X'$$



on prend quand ça dépasse le seuil, c'est-à-dire (ici 3)

Slide 24:

$$\mu_i = g^{-1}(x_i' \beta)$$

$$g(\mu_i) = x_i' \beta$$

$$\text{Var}(r_i) = \text{Var}(Y_i) = \frac{\phi}{w_i} V(\mu_i)$$

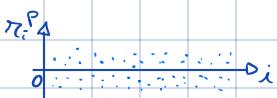
fonct^e variance

qui dépend de i donc les résidus ne sont pas homoskedastiques

r_i^P peut s'écrire comme la racine carrée de la contribution de l' i ^eme observation à la construction du χ^2 de Pearson.

indicateur de la qualité d'ajustement du modèle.

On s'attend à avoir des résidus "petits", distribués autour de zéro de façon aléatoire ; Si un résidu r_i^P est "élevé", cela voudrait dire que l' i ^eme observat^e contribue à un éventuel mauvais ajustement du modèle estimé



Si $|z_i^D|$ est "grand" alors l' $i^{\text{ème}}$ observat° risque de contribuer au mauvais ajustement du modèle estimé.

Mais qu'est-ce que ça veut dire "grand"? On rappelle que la déviance $D \sim \chi^2_{m-p-1}$, $E[D] = m-p-1$ et alors on s'attend à ce que chaque observat° contribue approximativement à hauteur de $\frac{m-p-1}{m} \approx 1$ à la déviance.

Du coup, $|z_i^D| \gg 1$ alors l' $i^{\text{ème}}$ observat° contribue au mauvais ajustement.

N.B.: Les résidus de déviance, sauf dans le cas où $Y \sim N$, sont asymétriques!

→ Hors slides à partir de maintenant.

On peut définir différents types de modèles :

(1) MODELES POUR DONNEES DE COMPTABLE

→ Exemple: On cherche à expliquer le nombre de décès dans une étude de mortalité ou alors le nombre de sinistres en assurance automobile, ...

REGRESSION DE POISSON.

REGRESSION BINOMIALE NEGATIVE

MODELES A INFILTRATION DE ZEROS
(ZEROS-INFILTRATED MODELS)

ZIP (ZERO-INFILTRATED Poisson)

ZINB (ZERO-INFILTRATED NEGATIVE BINOMIAL)

(2) MODELES POUR VARIABLES REPONSES CATEGORIELLES.

REPONSE BINAIRE → REGRESSION LOGISTIQUE

→ Exemple: Un individu souscrit à un contrat d'assurance ou non.

REPONSE avec R>2 CATEGORIES (variable polytomique) GLM MULTIVARIÉ

→ REPONSE ORDINALE (Exemple: niveau de dommage).

→ MODELE LOGISTIQUE CUMULATIF

→ MODELE LOG-LOG COMPLEMENTAIRE

→ MODELE PROBIT CUMULATIF.

→ REPONSE NOMINALE → MODELE DE REGRESSION NOMINALE

③ MODELES POUR VARIABLES REPONSES CONTINUES.

(Exemple: coût d'un sinistre).

REGRESSION GAMMA

REGRESSION GAUSSIENNE INVERSE

REGRESSION TWEEDIE

une autre possibilité est d'estimer un modèle gaussien sur le $\log(Y)$.

Slide 32:

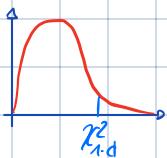
e^{β_j} : COEFFICIENT CORRECTEUR dû au fait que l'individu présente la modalité j.

Slide 35:

SURDISPERSION: la variance empirique est supérieure à la moyenne empirique.

$$H_0: \beta_j = 0$$

$$\frac{\hat{\beta}_j^2}{\text{Var}(\hat{\beta}_j)}$$



Si $\text{Var}(\hat{\beta}_j) \uparrow$, alors, toute chose étant égale par ailleurs, la valeur de la stat du test de signification de Wald \uparrow et on se déplace vers la région de rejet du test: on peut donc être amené à dire qu'une variable est significative alors que probablement, après prise en compte de la surdispersion, elle ne le serait pas!

VARIABLE OFFSET: Lorsque la variable à expliquer est une variable de comptage, comme par exemple : le nb de sinistres ou le nb de décès dans un groupe, il faut corriger par le nb d'exposés au risque.

Si $E[Y] = \mu$, alors cela revient à s'intéresser à $\frac{\mu}{m}$ (taux d'occurrence) et on écrit: $g\left(\frac{\mu}{m}\right) = x'\beta$

Si on choisit la fonction liée log, alors $\ln\left(\frac{\mu}{m}\right) = x'\beta$.

$$\Rightarrow \ln(\mu) = \ln(m) + x'\beta$$

variable offset → c'est comme si on avait une nouvelle variable dans le modèle mais pour laquelle on n'a pas de coeff à estimer.

Avec l'offset, la moyenne de Y est directement proportionnelle à m: $\mu = m e^{x'\beta}$.

L'offset sert à prendre en compte la taille du groupe ou les différentes expositions au risque

par exemple: ≠ zones géographiques

les polices dans un port peuvent avoir des exposés ≠

$$\text{Var}(Y) = \frac{a(\phi)}{\frac{\phi}{\omega}} \frac{b''(\phi)}{\sqrt{\mu}}$$