Rappels algèbre linéaire

$$(\mathbf{A}+\mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$$
$$(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top$$
$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^\top = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$
$$\text{tr}(\mathbf{A}) = \sum_{i=1}^{m} a_{i,i}$$

Rang : Nombre de colonnes *(ou lignes)* linéairement indépendantes.

symétrique : Lorsqu'une matrice carrée $\mathbf{A} = \mathbf{A}^\top$.

Idempotente : Lorsqu'une matrice carrée $\mathbf{A} = \mathbf{AA}$.

Dérivées où

$\mathbf{a} = (a_1, \ldots, a_p)^\top$

$\mathbf{b} = (b_1, \ldots, b_p)^\top$

$\mathbf{A}_{p\times p}$ est symétrique

$f(\mathbf{b})$ est dérivable du vecteur $\mathbf{b}$

$$\frac{\partial}{\partial \mathbf{b}}\mathbf{b}^\top\mathbf{a} = \mathbf{a}$$
$$\frac{\partial}{\partial \mathbf{b}}\mathbf{b}^\top\mathbf{Ab} = 2\mathbf{Ab}$$
$$\frac{\partial}{\partial \mathbf{b}} f(\mathbf{b})^\top \mathbf{A} f(\mathbf{b}) = 2\left(\frac{\partial}{\partial \mathbf{b}} f(\mathbf{b})\right)^\top \mathbf{A} f(\mathbf{b})$$

2 Régression linéaire simple

Postulats

- $\mathbf{H}_1$ Linéarité : $\mathrm{E}[\varepsilon_i] = 0$
- $\mathbf{H}_2$ Homoscédasticité : $\mathrm{Var}(\varepsilon_i) = \sigma^2$
- $\mathbf{H}_3$ Indépendance : $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$
- $\mathbf{H}_4$ Normalité : $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

Modèle

Y_i : Valeurs observées d'une variable aléatoire.

x_i : Valeurs **connues**

β_i : Paramètres **fixés** mais inconnus à ***estimer***

ϵ_i : Réalisations inconnues d'une variable aléatoire.

$$\mathrm{E}[Y_i|x_i] = \beta_0 + \beta_1 x_i$$
$$\mathrm{Var}(Y_i|x_i) = \sigma^2$$
$$Y_i|x_i \overset{\mathbf{H}_4}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

Estimation des paramètres

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i Y_i - \bar{Y}\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_i} = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{XY}}{S_{XX}}$$

S_{XY} : La somme des produits croisés corrigée.
S_{XX} : La somme des carrés corrigée.
β_0 : On peut interpréter β_0 comme la vidange pour tout biais dont le modèle ne tient pas compte.
On peut visualiser en imaginant le droite de régression monter ou descendre jusqu'à un point où la moyenne des résidus est de zéro.

Estimation de σ^2

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n}$$
$$s^2 = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n-p'} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2} = \text{MSE}$$
$$\text{où } \frac{(n-2)s^2}{\sigma^2} \sim \chi^2_{(n-2)}$$

où $p' = 2$ en régression linéaire simple.

Propriété des estimateurs

$\mathrm{E}[\hat{\beta}_j]$	$\mathrm{V}(\hat{\beta}_j)$	Sous l'hypothèse de normalité
β_0	$\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$	$\hat{\beta}_0 \overset{H_4}{\sim} \mathcal{N}\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right]\right)$
β_1	$\frac{\sigma^2}{S_{xx}}$	$\hat{\beta}_1 \overset{H_4}{\sim} \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_{XX}}\right)$

Tests d'hypothèse sur les paramètres

Hypothèses	t_{obs}	C
$H_0 : \hat{\beta} = \theta_0$ $H_1 : \hat{\beta} \neq \theta_0$	$\frac{\hat{\beta}-\theta_0}{\sqrt{\widehat{V(\hat{\beta})}}} \overset{H_1}{\sim} T_{(n-2)}$	$\lvert t_{obs}\rvert > \left\lvert t_{(n-2),\alpha/2}\right\rvert$

$\therefore$ rejete H_0 si $|t_{obs}| > \left|t_{(n-2),\alpha/2}\right|$.

Intervalle de confiance

Pour les paramètres $\widehat{\beta}_0$ et $\widehat{\beta}_1$

$$\left[\widehat{\beta}_0 \pm t_{(n-2),\alpha/2} s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\right]$$
$$\left[\hat{\beta}_1 \pm t_{(n-2),\alpha/2}\frac{s}{\sqrt{S_{xx}}}\right]$$

Prévisions

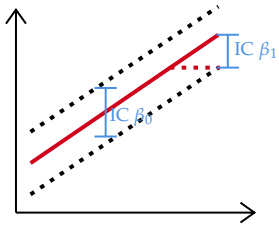
2 types de prévisions possibles pour une valeur x_0 donnée

1. Prévoir la valeur moyenne
   $$\mathrm{E}[Y_0|x_0] = \beta_0 + \beta_1 x_0$$
2. Prévoir la 'vraie' valeur de Y_0
   $$Y_0 = \beta_0 + \beta_1 x_0 + \epsilon$$

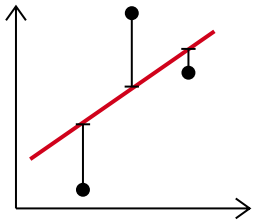
$\because \mathrm{E}[\epsilon] = 0 \therefore \widehat{\mathrm{E}[Y|x_0]} = \widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$

2 sources d'erreur dans nos prévisions

1. **Parameter risk** pour $\mathrm{E}[Y|x_0]$ et Y_0.
   alias l'incertitude liée à l'estimation des paramètres β_0 et β_1.



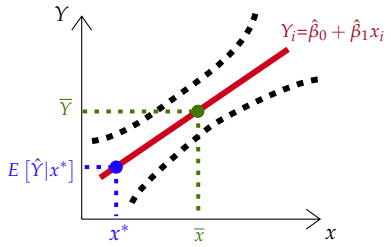
2. **Process risk** pour Y_0.
   alias ϵ qui est la fluctuation des valeurs de la variable endogène Y autour de sa moyenne.



Intervalles de confiance de niveau $1-\kappa$

$$\mathrm{E}[Y|x_0] \in \left[\widehat{Y}_0 \pm t_{(n-2),\alpha/2}\sqrt{s^2\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{XX}}\right)}\right]$$

On peut voir cet I.C. ci-dessous.
Plus x^* s'éloigne de $\bar{x}$, plus l'incertitude augmente et l'I.C. est large.
On voit alors que les limites de l'intervalle sont des *hyperboles* centrées en $(\bar{x},\bar{Y})$.
De plus, on peut voir qu'on tient compte uniquement du **parameter risk** et non le **process risk**.



L'I.C. pour la prévision tient compte du **process risk** ***en plus*** du **parameter risk**.

$$Y_0 \in \left[\widehat{Y}_0 \pm t_{(n-2),\alpha/2}\sqrt{s^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{XX}}\right)}\right]$$

Analyse de la variance (ANOVA)

Pour déterminer la proportion de la variabilité de Y **expliquée par le modèle**

Source	dl	SS	MS	F
Model	p	$\sum_{i=1}^n(\hat{Y}_i-\bar{Y})^2$ **(SSR)**	SSR/dl_1 **(MSR)**	$\frac{MSR}{MSE}$
Residual error	$n-p'$	$\sum_{i=1}^n(Y_i-\hat{Y}_i)^2$ **(SSE)**	SSE/dl_2 **(MSE** $=s^2$**)**	
Total	$n-1$	$\sum_{i=1}^n(Y_i-\bar{Y})^2$ **(SST)**		

Où :

- p : Nombre de variables explicatives dans le modèle.
- p' : Nombre de variables estimées dans le modèle.

SSR : Quantifie la variabilité des prévisions $\widehat{Y}_i$ expliquée par le modèle car elles ne sont pas tous égales à la moyenne $\bar{Y}_i$.
SSE : Quantifie la variabilité des $Y_i-\widehat{Y}_i$ ***pas*** expliquée par le modèle car il n'explique pas parfaitement Y_i.

Coefficient de détermination

Représente la proportion de la variation de la variable endogène Y qui est expliquée par le modèle.

$$R^2=\frac{SSR}{SST}=1-\frac{SSE}{SST}$$

Test F de Fisher pour la validité globale de la régression

On rejette $H_0:\beta_1=\beta_2=...=\beta_p=0$ si

$$F_{obs}=\frac{MSR}{MSE}\geq F_{p,n-p'}(1-\alpha)$$

À noter qu'on peut réécrire $F_{\text{obs}}=\frac{1-R^2}{R^2}$

Distribution d'un résidu ε

$E[\widehat{\epsilon}_i]$	0	$\mathcal{H}_1$
$V(\widehat{\epsilon}_i)$	$\sigma^2(1-h_{ii})$	$\mathcal{H}_2$
$Cov(\widehat{\epsilon}_i,\widehat{\epsilon}_j)$	$-\sigma^2\left(\frac{1}{n}+\frac{(x_i-\bar{x})(x_j-\bar{x})}{S_{xx}}\right)$	$\mathcal{H}_3$
$\widehat{\epsilon}_i$	$\sim\mathcal{N}(0,\sigma^2(1-h_{ii}))$	$\mathcal{H}_4$

où $h_{ii}=\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{XX}}$.
On peut interpréter h_{ii} comme étant la *"proportion"* de la variabilité du modèle σ^2 *"expliquée"* par la i$^{\text{e}}$ observation.
Pour mieux le voir : $h_{ii}=\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum_{j=1}^n(x_j-\bar{x})^2}$.

Donc, h_{ii} est une des n observations et $(x_i-\bar{x})^2$ est une des n erreurs au carré ; en multipliant par σ^2 on obtient la *"proportion"* de la variabilité du résidu attribué au i$^{\text{e}}$ résidu $\hat{\epsilon}_i$.

Vérification des postulats

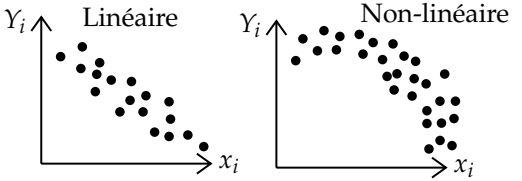
Les résidus studentisés sont définis par

$$r_i=\frac{\hat{\varepsilon}_i}{\sqrt{s^2(1-h_{ii})}}$$

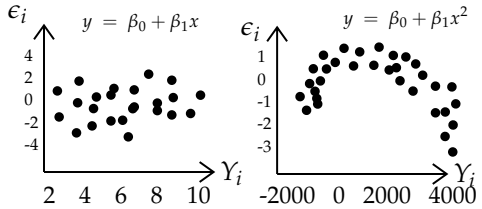
Linéarité

Utilise 3 types de graphiques *(en nuage de points)*.

› graphique $Y_i|x_i$: On veut que l'allure de la relation ait l'air linéaire.



› graphique $\hat{\varepsilon}_i|\hat{Y}_i$: On veut que les points soient centrés verticalement en 0.
Il ne devrait pas avoir de tendances discernables du nuage de points *(il devrait avoir l'air aléatoire)*.

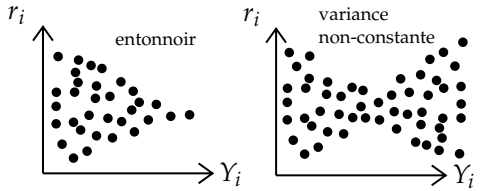


› graphique $\hat{\varepsilon}_i|x_i$: Le graphique et l'interprétation est la même que $\hat{\varepsilon}_i|\hat{Y}_i$.

Homoscédasticité

Signifie que la variance des résidus n'est pas constante. Si un modèle de régression linéaire simple est approprié, les r_j sont issus d'une distribution normale centrée réduite et devraient généralement se situer entre -3 et 3.

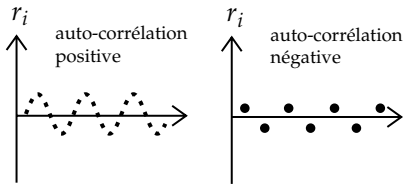
› Graphique $r_i|\hat{Y}_i$: La dispersion des résidus doit être constante, pas de forme d'entonnoir ou de résisus absolus supérieurs à 3 *(données aberrantes ou manque de normalité)*.



Indépendance

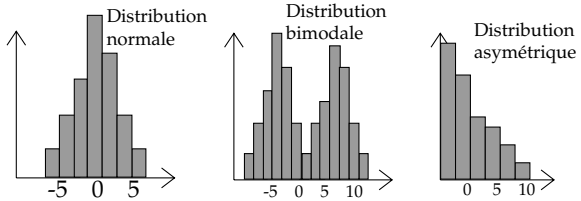
Puisque c'est difficile à tester, on peut faire la graphique des résidus en fonction du numéro d'observation pour des données chronologiques.

› Graphique $r_i|i$: Si il y a un *pattern*, présence d'auto-corrélation (le postulat H_3 n'est donc pas respecté).

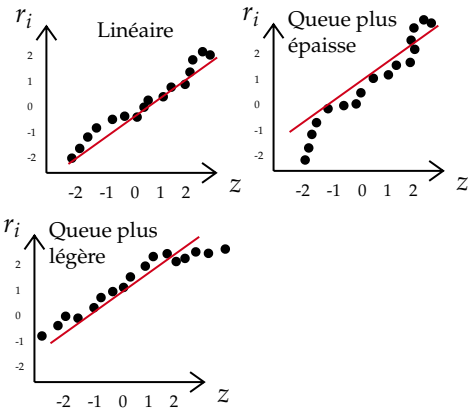


Normalité

› Histogramme des r_i.



› Q-Q Plot Normal : les résidus du modèle doivent suivre la droite des quantiles normaux théoriques.



Transformation des données

1. $V(\epsilon_i) \propto \mathrm{E}[Y_i]$ et les données de type Poisson.
   $g(Y) = \sqrt{Y}$
2. $V(\epsilon_i) \propto (\mathrm{E}[Y_i])^2$ avec la situation la plus efficace étant si Y possède une très grande étendue.
   $g(Y) = log(Y)$
3. $V(\epsilon_i) \propto (\mathrm{E}[Y_i])^4$.
   $g(Y) = 1/Y$
4. $V(\epsilon_i) \propto \mathrm{E}[Y_i](1 - \mathrm{E}[Y_i])$, $Y \in [0,1]$ et $Y \sim$ Bern.
   $g(Y) = \arcsin(\sqrt{Y})$

Prévision sous transformation

$\mathrm{E}[g(Y)] \neq g(\mathrm{E}[Y])$

Soit une transformation $g(Y) \in [a,b]$

En appliquant la transformation inverse, on peut trouver un intervalle pour $Y \in [g^{-1}(a), g^{-1}(b)]$.

Alias, le théorème de la fonction quantile pour une fonction monotone :

$$\Pr(a \leq g(Y) \leq b) = \Pr(g^{-1}(a) \leq Y \leq g^{-1}(b))$$

3 Régression linéaire multiple

Le modèle et ses propriétés

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times p'}\boldsymbol{\beta}_{p'\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$$
$$\mathrm{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$$
$$\mathrm{V}(\mathbf{Y}) = \sigma^2\mathbf{I}_{n\times n}$$
$$Y \overset{H_4}{\sim} \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_{n\times n})$$

Matrice d'incidence

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{n\times 1} = \underbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} & x_{np} \end{pmatrix}}_{n\times p'} \underbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}}_{p'\times 1} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}}_{n\times 1}$$

Types de variables exogènes possible :

- **Dichotomiques** e.g. : présence / absence.
- **Discrètes** e.g. : une classe de revenu.
- **Continues** e.g. : le poids d'une personne.
- **Qualitatives** e.g. : le sexe.

Paramètres du modèle

Estimation et propriétés des paramètres

$$\hat{\boldsymbol{\beta}} = \overbrace{(\mathbf{X}^\top\mathbf{X})^{-1}}^{\text{"}S_{XX}\text{"}}\overbrace{\mathbf{X}^\top\mathbf{Y}}^{\text{"}S_{XY}\text{"}}$$
$$\mathrm{E}\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta}$$
$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$$
$$\hat{\boldsymbol{\beta}} \overset{H_4}{\sim} \mathcal{N}_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$$

Intervalle de confiance sur les paramètres

$$\mathrm{V}(\beta_j) = \sigma^2 v_{jj}$$
$$\beta_j \in \left[\hat{\beta}_j \pm t_{n-p'}\left(1-\frac{\alpha}{2}\right)\sqrt{s^2 v_{jj}}\right]$$

où v_{jj} est l'élément (j,j) de la matrice $(\mathbf{X}^\top\mathbf{X})^{-1}$.

Estimation de σ^2

À noter que $s^2 \perp \boldsymbol{\beta}$ et le Biais$(s^2) = 0$.

$$s^2 = \frac{\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}}{n-p'} \qquad \text{où } \frac{\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}}{\sigma^2} \sim \chi^2_{n-p'}$$

Test d'hypothèse sur un paramètre du modèle

On rejète $H_0 : \beta_j = 0$ si

$$|t_{obs,j}| = \frac{\beta_j}{\sqrt{s^2 v_{jj}}} > t_{n-p'}\left(1-\frac{\alpha}{2}\right)$$

Propriétés de la droite de régression

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} & \hat{\varepsilon} &= \mathbf{Y}-\hat{\mathbf{Y}}\\ &= \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} & &= (\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y})\\ &= \mathbf{H}\mathbf{Y} & &= (\mathbf{I}_n - \mathbf{H})\mathbf{Y}\end{aligned}$$

où $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ est la *hat matrix*.

On a aussi que

$$\mathrm{E}\left[\hat{\mathbf{Y}}\right] = \mathbf{X}\boldsymbol{\beta} \qquad \mathrm{V}(\hat{\mathbf{Y}}) = \sigma^2\mathbf{H}$$
$$\hat{\mathbf{Y}} \overset{H_4}{\sim} N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{H})$$

Pour les résidus de la droite de régression, on a

$$\mathrm{E}\left[\hat{\boldsymbol{\varepsilon}}\right] \overset{H_1}{=} 0 \qquad \mathrm{V}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$$
$$\hat{\boldsymbol{\varepsilon}} \overset{H_4}{\sim} \mathcal{N}_n(0, \sigma^2(\mathbf{I}_n - \mathbf{H}))$$

Matrice de projection

Les matrices $\mathbf{H}$ et $\mathbf{I_n} - \mathbf{H}$ peuvent être vues commes des matrices de projection. Ces deux opérateurs possèdent plusieurs propriétés :

1. $\mathbf{H}^\top = \mathbf{H}$ (symétrie)
2. $\mathbf{HH} = \mathbf{H}$ (idempotence)
3. $\mathbf{HX} = \mathbf{X}$
4. $(\mathbf{I}_n - \mathbf{H}) = (\mathbf{I}_n - \mathbf{H})^\top$ (symétrie)
5. $(\mathbf{I}_n - \mathbf{H})(\mathbf{I}_n - \mathbf{H}) = (\mathbf{I}_n - \mathbf{H})$
6. $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = 0$
7. $(\mathbf{I}_n - \mathbf{H})\mathbf{H} = 0$

Intervalle de confiance pour la prévision

Théorème de Gauss-Markov

Selon les postulats H_1 à H_4, l'estimateur

$$\mathrm{a}^\top\hat{\boldsymbol{\beta}} = \mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$$
$$\sim \mathcal{N}(\mathbf{a}^\top\boldsymbol{\beta}, \sigma^2\mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{a})$$

est le meilleur estimateur pour $\mathbf{a}^\top\boldsymbol{\beta}$ où $\mathbf{a}^\top = \mathbf{c}^\top\boldsymbol{X}$ (**BLUE** : Best linear unbiased estimator).

I.C. pour la prévision de la valeur moyenne $\mathrm{E}[\mathbf{Y}|\mathbf{x}^*]$

$$\left[\mathbf{x}^{*\top}\hat{\boldsymbol{\beta}} \pm t_{(n-p'),\alpha/2}\sqrt{s^2\mathbf{x}^{*\top}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}^*}\right]$$

I.C. pour la valeur prédite $\hat{\mathbf{Y}}|\mathbf{x}^*$

$$\left[\mathbf{x}^{*\top}\hat{\boldsymbol{\beta}} \pm t_{(n-p'),\alpha/2}\sqrt{s^2\left(1+\mathbf{x}^{*\top}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}^*\right)}\right]$$

Analyse de la variance

Tableau ANOVA

- On utilise le même tableau ANOVA qu'en régression linéaire simple.
- $SSR_{\text{régression}} = \sum_{i=1}^{p} SSR_i$, où SSR_i représente le SSR individuel de la variable explicative i calculé par R. On peut ensuite trouver MSR et la statistique F_{obs}.
- $SSR(x)$ SSR pour le modèle incluant la variable x.

Test F pour la validité globale de la régression

Même test qu'en régression linéaire simple.

Test F partiel pour la réduction du modèle

Avec $k < p$, on va rejeter le modèle réduit :

$$H_0 : Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{ik}x_{ik} + \epsilon_i$$

Pour le modèle complet :

$$H_1 : Y_i = \beta_0 + \beta_1 x_{i1} + ...\beta_{ip}x_{ip} + \epsilon_i$$

Si

$$F_{obs} = \frac{(SSE^{(0)} - SSE^{(1)})/\Delta dl}{SSE^{(1)}/(n-p')} \geq F_{p-k,n-p'}(1-\alpha)$$

où :

$\Delta dl = p - k.$

$SSE^{(0)}$ pour le modèle réduit (H_0)

$SSE^{(1)}$ pour le modèle complet (H_1)

À noter que $\because SST^{(0)} = SST^{(1)} \therefore$

$$F_{obs} \Leftrightarrow \frac{(SSR^{(1)} - SSR^{(0)})}{\Delta dl\ MSE^{(1)}}$$

Régression avec variables qualitatives

Lorsque nous avons une variable exogène représentant des catégories, il est important de la coder sous forme d'indicatrice :

$$x_i = \begin{cases} 1, & \text{si} \dots \\ 0, & \text{sinon} \end{cases}$$

Lorsqu'il y a n $\geq$ 2 états, alors il faut créer **n - 1** variables indicatrices.
C'est **n - 1** puisque le n$^{\text{ème}}$ état est représenté lorsque toutes les autres variables indicatrices sont à 0.
Exemple avec 3 états *(rouge, bleu, vert)*.

$$x_{i1} = \begin{cases} 1, & \text{si rouge} \\ 0, & \text{sinon} \end{cases}$$

$$x_{i2} = \begin{cases} 1, & \text{si bleu} \\ 0, & \text{sinon} \end{cases}$$

$$x_{i3} = \begin{cases} 1, & \text{si vert} \\ 0, & \text{sinon} \end{cases}$$

compléter cette section

Multicollinéarité

Soit une combinaison de constantes *(pas toutes égales à 0* $d_1, \dots, d_p$.

Multicolinéarité *exacte* Une variable est la combinaison linéaire d'une, *ou plusieurs*, autres variables.
$\sum_{j=1}^{p} d_j X_j = 0$

Multicolinéarité *au sens large* Lorsqu'une variable est *presque* la combinaison linéaire d'une, ou plusieurs, autres variables.
Plus fréquente et difficile à détecter.
$\sum_{j=1}^{p} d_j X_j \approx 0$

Problèmes potentiels

- Instabilité de $(\mathrm{X}^\top \mathrm{X})^{-1}$, i.e. une petite variation de Y peut causer de grandes variations en $\hat{\beta}$ et $\hat{Y}$.
- $\hat{\beta}_i$ de signes contre-intuitif.
  Pour exemple, le taux de chômage avec $\hat{\beta}_i > 0$ et le taux d'emploi avec $\hat{\beta}_k < 0$.
- $\mathrm{V}(\hat{\beta}_i)$ et $\mathrm{V}(\hat{Y})$ très grandes.
- Les méthodes de sélection de variable ne concordent pas.
- Conclusions erronées sur la significativité de certains paramètres, malgré une forte corrélation avec Y.

Détection

- Si r_{ij} dans la matrice des coefficients de corrélation échantillonnaux $\mathbf{X}^{*\top}\mathbf{X}^*$ est élevée, où $\forall j, k \in \{1, \dots, p\}$.

$$\mathbf{X}^{*\top}\mathbf{X}^* = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix}_{p \times p}$$

$$\mathbf{X}^* = \begin{bmatrix} \frac{x_1 - \bar{x}_1}{s_1} & \dots & \frac{x_p - \bar{x}_p}{s_p} \end{bmatrix}_{1 \times p}$$

$$r_{jk} = \sum_{i=1}^{n} \frac{(x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k)}{s_j s_k}$$

- Si le facteur d'inflation de la variance (VIF_j) est élevé où :

$$VIF_j = \frac{1}{1 - R_j^2}$$

VIF_j : Pour évaluer le degré de dépendance de chaque variable exogène sur les autres variables exogènes.
Pour la j^{e} variable exogène (x_j), on mesure le niveau de dépendance en effectuant une régression ayant x_j comme variable explicative (*endogène*) et les $(p-1)$ variables restantes comme variables explicatives (*exogènes*).
R_j^2 mesurera la proportion de la variabilité de x_j expliquée par les autres variables explicatives, s'il est élevé il semblerait avoir un problème.
Le nom "*facteur d'inflation*" vient du fait que la variance de $\hat{\beta}_j$ s'exprime en fonction du VIF :

$$\mathrm{V}(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2} VIF_j$$

Solution

- On retire les variables ayant un VIF élevé (une à la fois).
- On combine des variables exogènes redondantes

Problèmes et détails

- Généralement, on peut considérer un VIF > 10 comme un point ou considérer la présence de la multicolinéarité.
- Incapable de détecter des mutlicolinéarités de la colonne de 1.
- Incapable de cerner le nombre de *quasi* dépendances linéaires dans les données.
- Incapable de pointer une valeur précise du VIF où l'on doit vraiment commencer à s'inquiéter (10 est une valeur *ad hoc*).

Validation du modèle et des postulats

Linéarité

- On trace les **graphiques à variable ajoutée** ($\hat{\varepsilon}_{\mathbf{Y}|\mathbf{X}_{-\mathbf{j}}}$ en fonction de $\hat{\varepsilon}_{x_j|\mathbf{X}_{-\mathbf{j}}}$).
  $\hat{\varepsilon}_{\mathbf{Y}|\mathbf{X}_{-\mathbf{j}}}$ **:** Vecteur des résidus de la régression de $\mathbf{Y}$ sur toutes les variables sauf $\mathbf{x}_j$.
  $\hat{\varepsilon}_{x_j|\mathbf{X}_{-\mathbf{j}}}$ **:** Vecteur des résidus de la régression avec $\mathbf{x}_j$ comme variable endogène sur toutes les autres variables exogènes.
- Ces graphiques doivent normalement donner une droite de pente β_j.
  - Si le graphique ressemble à un graphique de résidus normaux, x_j est **inutile**.
  - Si il y a une courbe, x_j est **non-linéaire**.

Homogénéité des variances

- Graphique $r_i | \hat{Y}_i$

Normalité des erreurs

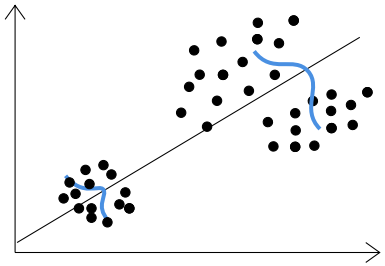
- Graphique Q-Q plot normal.

Indépendance entre les observations

- Graphique $\hat{\varepsilon}_i | i$
- Test de Durbin-Watson pour détecter la présence d'autocorrélation positive entres les résdius.
- S'il y a autocorrélation, alors $\mathrm{V}(\epsilon) \neq \sigma^2 \mathbf{I_n}$ puisque les éléments autres que la diagonale ne sont pas toutes 0.
  Toutes nos procédure (test d'hypothèses, intervalles de confiance, etc.) dépendent sur ce fait et donc des conclusions erronées pourraient être obtenues.

Hétéroscédasticité et régression pondérée

Jusqu'à dâte, les procédure sont adéquates seulement si $\epsilon_i, \ldots, \epsilon_n$ sont (iid) telle que $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I_n})$.
Que doit on faire si $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2))$?
On appelle ceci l'**hétéroscédasticité**, c'est à dire que la variance n'est pas la même pour tous les résidus.



On utilise la méthode du **moindre carré pondéré** (MCP) pour trouver $\widehat{\boldsymbol{\beta}}$ tel que :

$$\sum_{i=1}^{n} \omega_i (Y_i - \widehat{Y}_i)^2 \Leftrightarrow \widehat{\boldsymbol{\epsilon}}^\top \mathbf{V}^{-1} \widehat{\boldsymbol{\epsilon}}$$

est minimale et qu'on pose :

$$\omega_i \propto 1/\sigma_i^2$$

$$\mathbf{V} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$$

On obtient alors :

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{Y}$$

$$\mathrm{V}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1}$$

On peut faire le **test** $H_0 : \beta_j = \beta_{j,0}$ avec la statistique :

$$t = \frac{\widehat{\beta}_j - \beta_{j,0}}{\sqrt{(\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1}_{\mathbf{jj}}}}$$

Et utiliser $V = V^* \sigma^2$ en estimant σ^2 par :

$$s^2 = \frac{\sum_{i=1}^{n} \omega_i (Y_i - \widehat{Y}_i)^2}{n - p'}$$

On peut contruire une table d'**analyse de la variance** avec la décomposition :

$$\underbrace{\mathbf{Y}^\top \mathbf{V}^{-1} \mathbf{Y}}_{SST_V} = \underbrace{\widehat{\boldsymbol{\beta}} \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{Y}}_{SSR_V} + \underbrace{(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^\top \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})}_{SSE_V}$$

4 Sélection de modèle et régression régularisée

En présence de beaucoup de variable exogènes, on court le danger d'en garder trop ou pas assez

- **Trop** : On augmente inutilement la variance des estimations $(\hat{\beta})$.
- **Moins** : On simplifie notre modèle trops en augmentant inutilement le biais des estimations $(\hat{\beta})$.

ajouter information sur début du chapitre pour l'examen

Critères de comparaison classiques

- Coefficient de détermination (pour mesurer la qualité globale du modèle) :
  $$R^2 = \frac{SSR}{SST}$$
  Si on ajoute une variable exogène, il est certain que R^2 augmentera, on utilise donc ce critère pour valider si la régression est utile pour prédire Y, mais pas pour critère de sélection des variables exogènes.
- **Coeficient de détermination ajusté** :
  $$R_a^2 = \frac{SSE/p}{SST/(n-1)} = \frac{MSE}{MST}$$
  Ce critère permet de valider l'ajout de nouvelles variables exogènes.
  De plus, **il peut diminuer avec l'ajout de nouvelles variables**.

Ces deux critères sont inutiles pour comparer des modèles avec des transformations différentes et pour des modèles avec/sans ordonnée à l'origine.

Méthode basées sur la puissance de prévision

Ce critère maximise l'habileté du modèle a prédire de nouvelles données.

Principe de la validation croisée

Algorithme de validation croisée classique

1. Pour $i = 1, ..., n$,
   1.1 Enlever la i^{e} observation du jeu de données.
   1.2 Estimer les paramètres du modèle à partir des $n - 1$ données restante.
   1.3 Prédire Y_i à partir de x_i et du modèle obtenu en 2, noté $\hat{Y}_{i,-i}$
2. Calculer la somme des carrés des erreurs de prévision $PRESS = \sum_{i=1}^{n} (Y_i - \hat{Y}_{i,-i})^2$

On cherche a minimiser le **PRESS** ou, à maximiser le **coefficient de détermination de prévision** :

$$R_p^2 = 1 - \frac{PRESS}{SST} = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_{i,-i})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

Les résidus PRESS

Il est possible de trouver la statistique PRESS sans devoir calculer n régressions :

$$PRESS = \sum_{i=1}^{n} \left(\frac{\hat{\varepsilon}_i}{1 - h_{ii}} \right)^2$$

Le $\text{i}^{\text{ème}}$ résidu PRESS est le $\text{i}^{\text{ème}}$ terme de la sommation.

$$\hat{\varepsilon}_{i,-i} = \frac{\hat{\varepsilon}_i}{1 - h_{ii}}$$

Utile dans le cas où il n'aurait qu'un seul résidu mauvais qui biaise la sommation.

Échantillion de test et validation croisée par k ensemble

Le **dernier recourt** après avoir essayé tous les autres tests.
Lorsqu'on a un très grand jeu de données, on peut les séparer en deux :

Training set : Utilisé pour ajuster le modèle et faire la sélection de variables explicatives.

Test set : Échantillon sur lequel on effectue des prévisions.

La puissance de prévision est donc calculé avec :

$$\textbf{MSEP} = \frac{1}{n_{\text{test}}} \sum_{i \in \text{test}} (Y_i - \hat{Y}_i)^2$$

Algorithme de validation croisée par K ensembles

1. Pour $k = 1, ..., K$,
   1.1 Enlever le k^{e} ensemble du jeu de donnée.
   1.2 Estimer les paramètres du modèle à partir des données des $k-1$ échantillons restants.
   1.3 Prédire les observations du k^{e} ensemble $(\hat{Y}_{i,-k})$ et calculer
   $$MSEP_k = \frac{1}{n_k} \sum_{i \in group\, k} (Y_i - \hat{Y}_{i,-k})^2$$
2. Calculer la moyenne des sommes des carrés des erreurs de prévision $\frac{1}{k}\sum_{k=1}^{k} MSEP_k$

On choisit le modèle qui minimise $\frac{1}{k}\sum_{k=1}^{k} MSEP_k$

Le C_p de Mallows

$$\begin{aligned} C_p &= p' + \frac{(s_p^2 - \widehat{\sigma}^2)(n - p')}{\widehat{\sigma}^2} \\ &= \frac{SSE}{\widehat{\sigma}^2} + 2p' - n \end{aligned}$$

On cherche le modèle pour lequel $C_p \approx p'$

Critère d'information d'akaike et critère bayésien de Schwarz

› Ce critère est le plus utilisé dans la pratique et permet d'évaluer la qualité de l'ajustement d'un modèle.
$$AIC = n \cdot \ln\left(\frac{SSE}{n}\right) + 2p'$$
AIC prend en compte à la fois la qualité des prédictions du modèle et sa complexité.

› BIC est similaire a AIC, mais la pénalité des paramètres dépend de la taille de l'échantillon.
$$BIC = n \cdot \ln\left(\frac{SSE}{n}\right) + \ln(n)p'$$

› On cherche donc à minimiser ces 2 critères.

Méthode algorithmiques

Méthode d'inclusion forward

1. On commence avec le modèle le plus simple (i.e. $\hat{Y}_i = \beta_0$)
2. On essaie d'ajouter la variable qui, en l'incluant dans le modèle, permet de réduire le plus le SSE du modèle.
3. On valide si la variable diminue de façon significative les résidus avec un test F, où
   $$F_{obs} = \frac{SSE_{\text{petit modèle}} - SSE_{\text{grand modèle}}}{SSE_{\text{grand modèle}}/(n-p')}$$
   On ajoute la variable au modèle si
   $$F_{obs} > F_{1,n-p'}(1-\alpha)$$
4. On répète jusqu'à ce qu'aucune variable ne vaille la peine d'être ajoutée.

Méthode d'exclusion backward

1. On débute avec le modèle complet
2. On veut enlever la variable exogène qui, en l'excluant du modèle, permet de minimiser l'augmentation du SSE de la régression.
3. Même test F qu'à l'étape 3 de la méthode *forward*, sauf qu'on enlève la variable seulement si
   $$F_{obs} < F_{1,n-p'}(1-\alpha)$$
4. On répète jusqu'à ce qu'aucune variable ne vaille la peine d'être enlevée.

Méthode pas à pas (step-wise)

1. On débute avec la méthode d'inclusion
2. Après l'ajout d'une variable au modèle, on effectue la méthode d'exclusion pour les variables qui sont actuellement dans le modèle (on remet constamment le modèle en question).

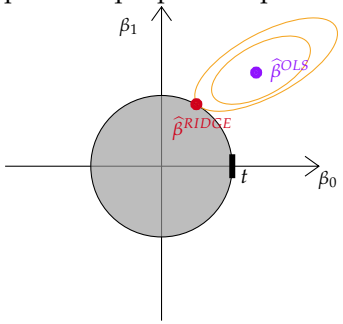
Graphique des variables ajoutées

Régression Ridge

Deux normes sont utilisées en régression, les normes L2 *(régression Ridge)* et L1 *(régression Lasso)*.

$$\|\boldsymbol{\beta}\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2} \qquad \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$$

La **régression Ridge** est habituellement utilisée lorsque nous avons une variance très élevée avec l'estimateur BLUE.
Souvent, ceci est lorsqu'il y a beaucoup de variables explicatives qui sont **possiblement corrélés**.
La méthode consiste à minimiser l'estimateur en acceptant un certain biais *(budget total)* t.
Donc, à minimiser la variance de l'estimateur sous la contrainte que $\|\boldsymbol{\beta}\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2} \leq t$.
Cette condition représente un cercle dont on prends le point le plus près de l'estimateur OLS $\widehat{\beta}$; c'est-à-dire, qu'on prends le point qui est *on the "**ridge**" of the circle*.
Les lignes oranges représentent différentes valeurs possible que peuvent prendre $\widehat{\beta}$.s



On veut minimiser l'équation suivante :

$$S^R(\beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

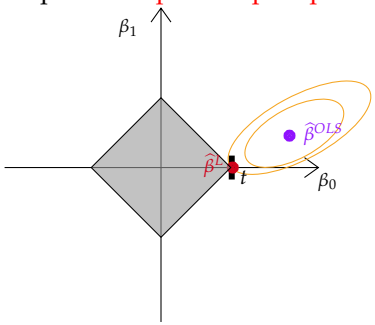
Et on trouve que

$$\hat{\boldsymbol{\beta}}^R = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p \right)^{-1} \mathbf{X}^\top \mathbf{Y}$$

On choisit la valeur optimale pour le coefficient de régularisation λ avec une validation croisée.

Régression Lasso (*Least Absolute Shrinkage and Selection Operator*)

La **régression Lasso** est habituellement utilisée pour la sélection de variables.
Souvent, notre estimateur va tomber sur un des coins du diamant, ce qui implique qu'un paramètre serait à 0.
Donc, nous pouvons mettre à 0 des paramètres contrairement à la régression **Ridge** où les paramètres peuvent être près de 0 mais très rarement égale à 0.
La méthode minimise la variance de l'estimateur sous la contrainte que $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j| \leq t$.
Cette condition représente le diamant de l'image dont on prends le point le plus près de l'estimateur OLS $\widehat{\beta}$.



On veut minimiser l'équation suivante :

$$S^L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

À noter qu'un méthode plus efficace que Ridge et Lasso est de minimiser sous la contrainte que $\sum_{j=1}^{p} \mathbf{1}_{\{\beta_j \neq 0\}} \leq t$.

5 Modèles linéaires généralisés (GLM)

Famille exponentielle linéaire

Définition

Une loi de probabilité fait partie de la famille exponentielle linéaire si :

- Sa fonction de densité *(ou de masse)* de probabilité peut être exprimée comme :

Densité de la famille exponentielle

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

où

θ : paramètre canonique

ϕ : paramètre de dispersion

- La fonction c ne dépend pas du paramètre θ.
- Le support de Y ne dépend pas de θ puisqu'il ne peut pas varier.

note : Si ϕ est ***inconnu***, alors le modèle fait partie de la **famille exponentielle *de dispersion*** et le support de Y ne peut pas dépendre de ϕ.

note : Si chaque donnée à un ϕ_i différent, on réécrit la fonction de dispersion comme $\alpha_i = \phi/\omega_i \ \forall i \in \{1,\ldots,n\}$.
Ceci est la **famille exponentielle *pondérée***.

Propriétés

Soit $\mu = \dot{b}(\theta) = \frac{\partial}{\partial\theta} b(\theta)$ et $V(\mu) = \ddot{b}(\theta) = \frac{\partial^2}{\partial\theta^2} b(\theta)$.
Alors, si Y fait partie de la famille exponentielle linéaire, on peut exprimer l'espérance et la variance comme

$$\mathrm{E}[Y] = \dot{b}(\theta) = \mu$$
$$\mathrm{V}(Y) = a(\phi)\ddot{b}(\theta) = a(\phi)V(\mu)$$

Lemme de la Log-vraisemblance

Soit $\ell(\theta,\phi;y) = \ln f(y;\theta,\phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y;\phi)$.
Alors, sous certains conditions de régularité,

$$\mathrm{E}\left[\frac{\partial}{\partial\theta}\ell(\theta,\phi;Y)\right] = 0$$

Et l'information de Fisher peut être réécrite comme :

$$\mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\ell(\theta,\phi;Y)\right)^2\right] = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta,\phi;Y)\right]$$

note : Ces résultats sont utilisés pour obtenir les formules de la variance et l'espérance.

Fonctions de lien

Le but est de relier l'espérance de Y aux variables exogènes au lieu de Y elle-même ; soit :

$\mathbf{x}_{1\times p'}$: une ligne de la matrice d'incidence ; et

$\boldsymbol{\beta}_{p'\times 1}$: paramètre sur lequel $\mathrm{E}[Y|\mathbf{x}]$ dépend.

Alors,

$$\mathrm{E}[Y|\mathbf{x}] = \mu(\mathbf{x};\boldsymbol{\beta}) = \mu(\mathbf{x}\boldsymbol{\beta}) = \mu(\eta)$$

où η est le **prédicteur linéaire** :

$$\eta = \mathbf{x}\boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{p} x_j\beta_j$$

En *régression linéaire*, $\mu(\mathbf{x};\boldsymbol{\beta}) = \eta$ sans contraintes sur les valeurs que peut prendre μ.
En *régression linéaire **généralisée***, certaines variables réponses ont des contraintes *(e.g. Bernoulli où $0 \le \pi \le 1$)*.
La **fonction de lien** est donc la transformation appliquée à η afin de restreindre le support de μ.

$$g(\mu_i) = \eta_i \qquad \text{ou} \qquad \mu_i = g^{-1}(\eta_i)$$

Lien	η	μ
identité	$\eta = \mu$	$\mu = \eta$
logit	$\eta = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{e^\eta}{1+e^\eta}$
probit	$\eta = \Phi^{-1}(\mu)$	$\mu = \Phi(\eta)$
log-log complémentaire	$\eta = \ln(-\ln(1-\mu))$	$\mu = 1 - e^{-e^\eta}$
log	$\eta = \ln\mu$	$\mu = e^\eta$
inverse	$\eta = 1/\mu$	$\mu = 1/\eta$
canonique	$\eta = \theta$	

Estimation des paramètres

- Estimer $\hat{\beta}$ avec la méthode de l'**EMV** ***(MLE)***.
- L'EMV est **convergent**,
  $$\hat{\boldsymbol{\beta}} \underset{n\to\infty}{\longrightarrow} \boldsymbol{\beta}$$
- L'estimateur est **asymptotiquement normale**,
  $$\hat{\boldsymbol{\beta}} \underset{n\to\infty}{\sim} \mathcal{N}\left(\boldsymbol{\beta}, \frac{\mathcal{I}(\boldsymbol{\beta})^{-1}}{n}\right)$$
  où $\mathcal{I}(\boldsymbol{\beta})_{p'\times p'}$ est la matrice d'information Fisher :
  $$\mathcal{I}(\boldsymbol{\beta}) = \mathrm{E}\left[\dot{\ell}(\boldsymbol{\beta};Y_1,...,Y_n)\dot{\ell}(\boldsymbol{\beta};Y_1,...,Y_n)^\top\right]$$
  $$\Leftrightarrow -\mathrm{E}\left[\ddot{\ell}(\boldsymbol{\beta};Y_1,...,Y_n)\right]$$
- On peut estimer la matrice d'information de Fisher avec **l'information observée** lorsque β est inconnu :

$$\mathcal{I}(\hat{\boldsymbol{\beta}}) = -\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\beta}}\ell(\boldsymbol{\beta};y_i)\left\{\frac{\partial}{\partial\boldsymbol{\beta}}\dot{\ell}(\boldsymbol{\beta};Y_1,...,Y_n)\right\}^\top\Bigg|_{\hat{\boldsymbol{\beta}}}$$
$$\Leftrightarrow -\sum_{i=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\beta}^2}\ell(\boldsymbol{\beta};y_i)\Bigg|_{\hat{\boldsymbol{\beta}}}$$

Algorithme de Newton-Raphson

Afin d'évaluer $\hat{\boldsymbol{\beta}}$, on utilise l'approximation de Taylor du premier ordre avec cet algorithme itératif :

Newton Raphson

(1) Choisir des valeurs de départ pour le vecteur $\hat{\boldsymbol{\beta}}^{H_0}$

(2) Pour $k = 1, 2, ...$

(2.1)
$$\hat{\boldsymbol{\beta}}^{(k)} = \hat{\boldsymbol{\beta}}^{(k-1)} + \left\{-\ddot{\ell}(\hat{\boldsymbol{\beta}})^{(k-1)}\right\}^{-1}\dot{\ell}(\hat{\boldsymbol{\beta}})^{(k-1)}$$

(2.2) Si $|\dot{\ell}(\hat{\boldsymbol{\beta}})^{(k)}| < \varepsilon$, on converge vers les paramètres optimaux pour le modèle et on arrête.

(2.3) Répéter les étapes (2.1) et (2.2) jusqu'à une convergence.

Méthode du score de Fisher

Cette variante de l'algorithme de Newton-Raphson remplace $-\ddot{\ell}(\boldsymbol{\beta})$ par son espérance $-\mathrm{E}\left[\ddot{\ell}(\boldsymbol{\beta})\right] = \mathcal{I}(\boldsymbol{\beta})$ à l'étape (2.1) ; l'information de Fisher $\mathcal{I}(\boldsymbol{\beta})$ doit alors être approximée.
La distinction est que l'algorithme de Newton-Raphson est basé sur la **matrice d'information observée** $-\ddot{\ell}(\boldsymbol{\beta})$ alors que l'algorithme de Fisher-Scoring est basé sur la **matrice d'information *attendue*** Fisher $\mathrm{E}\left[-\ddot{\ell}(\boldsymbol{\beta})\right]$.
Ceci veut alors dire que si le lien canonique est utilisé alors les deux méthodes sont équivalents.

Tests d'hypothèses

La principale différence dans les tests d'hypothèses pour les GLMs est que les données ne sont pas nécessairement normales, mais font plutôt parti de la famille exponentielle. Alors, on ne peut pas utiliser la distribution Student.

Statistique de Wald

Pour tester :

$$\mathrm{H}_0 : \beta_j = 0 \qquad \text{vs} \qquad \mathrm{H}_1 : \beta_j \neq 0$$

On teste la statistique :

$$Z = \frac{\hat{\beta}_j}{\sqrt{\widehat{\mathrm{V}(\hat{\beta}_j)}}} \underset{\substack{\mathrm{H}_0\\ n \text{ grand}}}{\approx} \mathcal{N}(0,1)$$

On rejète donc H_0 si $Z > z_{1-\frac{\alpha}{2}}$.

Note : On obtient $\widehat{\mathrm{V}(\hat{\beta}_j)}$ sur les éléments de la diagonale de $\{\mathcal{I}(\hat{\beta})\}^{-1}/n$.

Test du rapport de vraisemblance

Pour tester :

$$\mathrm{H}_0 : \boldsymbol{\beta} \in \mathrm{B}_0 \qquad \text{vs} \qquad \mathrm{H}_1 : \boldsymbol{\beta} \in \mathrm{B}_0^{\mathrm{C}}$$

où $\mathrm{B}_0 \subset \mathbb{R}^{p'}$ est un sous-espace de l'espace des paramètres.

On teste la statistique :

$$\lambda(y) = \frac{\sup_{\boldsymbol{\beta}\in B_0} \mathcal{L}_n(\boldsymbol{\beta})}{\sup_{\boldsymbol{\beta}\in\mathbb{R}^{p'}} \mathcal{L}_n(\boldsymbol{\beta})} = \frac{\mathcal{L}_n\left(\hat{\boldsymbol{\beta}}^{(H_0)}\right)}{\mathcal{L}_n(\hat{\boldsymbol{\beta}})}$$

À noter que $\lambda(y) \leq 1$ toujours puisque, sous H_0, il y a moins de variables explicatives.
On rejete l'hypothèse nulle H_0 lorsque $\lambda(y) \leq$ valeur critique.
Si H_0 spécifie

- ***tous*** les paramètres du modèle, on a :
  $$-2\ln\lambda(Y) \underset{H_0}{\approx} \chi^2_{p'}$$
- ***partiellement*** les paramètres du modèle, on a :
  $$-2\ln\lambda(Y) \underset{H_0}{\approx} \chi^2_{k_2-k_1}$$
  où
  k_1 : Nombre de paramètres non-spécifiés dans H_0 ;
  k_2 : Nombre de paramètres non-spécifiés dans H_1 ($k_2 > k_1$).

Note : Avec le **TRV**, on peut seulement comparer des modèles qui sont **emboités**. C'est-à-dire, que tous les paramètres sous H_0 doivent être inclus sous H_1.

Intervalles de confiances

- Le **pivot approximatif** permet d'obtenir un IC approximatif sur un paramètre β_j où $j = 0,\ldots,p$:
  $$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\widehat{V(\hat{\beta}_j)}}} \underset{n\to\infty}{\approx} \mathcal{N}(0,1)$$
  Donc, les estimateurs $\hat{\boldsymbol{\beta}}$ sont asymptotiquement normalement distribués.
- Puisque le prédicteur linéaire $\hat{\eta} = \boldsymbol{x}\hat{\boldsymbol{\beta}}$ est une combinaison linéaire de ces estimateurs, on peut également y trouver un IC approximatif.
  De plus, on peut déduire un IC approximatif ***pas*** *centré autour de la prévision* pour $\mu(\eta)$ si la fonction de lien est monotone.
  Pour en avoir un centré autour de la prévision, il faut faut définir le pivot approximatif avec la méthode de Delta *(basé sur l'expansion Taylor)*.

Méthode Delta

Soit la suite de v.a. $X_1,\ldots,X_n$ telle que
$\sqrt{n}(X_n - \theta) \rightsquigarrow \mathcal{N}(0,\sigma^2)$
Alors, pour toute fonction ψ dérivable et telle que $\psi'(\theta) \neq 0$:
$\sqrt{n}(\psi(X_n) - \psi(\theta)) \rightsquigarrow \mathcal{N}(0,\{\psi(\theta)\}^2)$

Adéquation du modèle

Statistique X^2 de Pearson

$$X^2 = \sum_{i=1}^{n}\left(\frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}\right)^2 \underset{\text{n grand}}{\approx} \chi^2_{n-p'}$$

Où $X^2 \leq \chi^2_{n-p',1-\frac{\alpha}{2}}$ si le modèle est adéquat.

Si ϕ est inconnu, on peut l'estimer avec $\hat{\phi} = \frac{X^2}{n-p'}$.

Déviance

Types de modèles où $i = 1,\ldots,n$:

nul : Il n'y a pas de relation entre Y_i et x_i, et les observations sont identiquement distribuées :
$\hat{\mu}_i = \hat{\mu}_0 \qquad \theta = \tilde{\theta}_{n\times 1}$

complet : Chaque prévision est exactement égale à la donnée :
$\hat{\mu}_i = y_i \qquad \theta = \tilde{\theta}_{n\times 1}$

avec une fonction de lien : où g^{-1} est l'inverse de la fonction de lien.
$\hat{\mu}_i = g^{-1}(\boldsymbol{x}_i\hat{\beta}) \qquad \theta = \hat{\theta}_{n\times 1}$

Déviance : Différence de l'ajustement entre le modèle complet et le modèle choisi.

$$\begin{aligned} D(y;\hat{\mu}) &= 2\big(\ell(\tilde{\theta}) - \ell(\hat{\theta})\big) \\ &= 2\sum_{i=1}^{n}\omega_i\Big(y_i(\tilde{\theta} - \hat{\theta}) - [b(\tilde{\theta}) - b(\hat{\theta})]\Big) \end{aligned}$$

Déviance *réduite* :

$$D(y;\hat{\mu})^* = \frac{D(y;\hat{\mu})}{\phi}$$

Selon le **TRV**,

$$2\big(\ell(\tilde{\theta}) - \ell(\hat{\theta})\big) \Leftrightarrow 2\ln\left(\frac{\mathcal{L}_n(\tilde{\boldsymbol{\beta}})}{\mathcal{L}_n(\hat{\boldsymbol{\beta}})}\right) \Leftrightarrow \ln\left(\frac{\mathcal{L}_n(\tilde{\boldsymbol{\beta}})}{\mathcal{L}_n(\hat{\boldsymbol{\beta}})}\right)^2 \sim \chi^2_{n-p'}$$

si le modèle de $\hat{\theta}$ est adéquat
De plus, $\mathrm{E}\left[\frac{D(y;\hat{\mu})}{\phi}\right] \approx n - p'$; alors si ϕ est inconnu, on peut l'estimer par $\hat{\phi} = \frac{D(y;\hat{\mu})}{n-p'}$.

Comparaison de modèles

Les **critères classiques AIC** et **BIC** peuvent être utilisés pour comparer des modèles.

Analyse de la déviance

Lorsque ϕ est connu, et que les deux modèles m_A et m_B sont emboités ($p_B > p_A$), alors si m_A est une bonne simplification du m_B :

$$\frac{D(y;\hat{\mu}_A) - D(y;\hat{\mu}_B)}{\phi} = 2\big(\ell(\hat{\theta}_B) - \ell(\hat{\theta}_A)\big) \sim \chi^2_{p_B-p_A}$$

- Il est certain que la déviance va augmenter en diminuant le nombre de paramètres.
- On veut valider si la déviance augmente *significativement* au point de ne pas pouvoir simplifier m_B.
- On rejète H_0 que m_A est une bonne simplification de m_B si la différence de la déviance réduite est supérieure à $\chi^2_{p_B-p_A,1-\frac{\alpha}{2}}$.

Analyse des résidus

3 types de résidus pour vérifier l'ajustement d'un GLM :

Résidus de Pearson

Si l'ajustement est adéquat, ils ont une moyenne de 0 et une variance constante ; cependant, la distribution des résidus peut être asymétrique.

Résidus de Pearson

$$r_{P_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

Résidus d'Anscombe

On applique une transformation $A(Y)$ à Y afin de minimiser l'asymétrie de la distribution de $A(Y)$ (alors, on veut qu'elle soit approximativement normale).

$$A(t) = \int_{-\infty}^{t} \mathrm{V}^{-1/3}(s)ds$$

Cette transformation donne alors un résidu de :

Résidus d'Anscombe

$$r_{A_i} = \frac{A(y_i) - A(\hat{\mu}_i)}{\dot{A}(\hat{\mu}_i)\sqrt{\mathrm{V}(\hat{\mu}_i)}}$$

Résidus de déviance

Lorsque la déviance est définie comme :

$$D(y;\hat{\theta}) = \sum_{i=1}^{n} d_i$$

La contribution de la donnée i dans la déviance est :

$$d_i = 2\omega_i(y_i(\tilde{\theta}_i - \hat{\theta}_i) - [b(\tilde{\theta}_i) - b(\hat{\theta}_i)])$$

Le résidu est alors :

Résidus de déviance

$$r_{D_i} = \text{signe}(y_i - \hat{\mu}_i)\sqrt{d_i}$$

où

$$\text{signe}(t) = \begin{cases} 1, & t > 0 \\ -1, & t < 0 \end{cases}$$

Note : On ajoute le signe à la définition puisque la racine carrée sortira uniquement des résultats positifs alors que pour être utile à une analyse, ils ne peuvent pas tous être positifs.

Régularisation

On peut également ajouter une pénalité de type Ridge ou Lasso *(sélection de variables)* comme en modèles linéaires.

6 Modélisation de données de comptage

Terme *offset*

On veut souvent modéliser le taux de réclamation, cela se fait avec un terme *offset* t_i qui représente l'exposition au risque (i.e. le nombre d'années qu'on a assuré la personne) :

$$\ln\left(\frac{\mu_i}{t_i}\right) = \boldsymbol{x}_i\boldsymbol{\beta}$$

$$\ln(\mu_i) = \boldsymbol{x}_i\boldsymbol{\beta} + \ln(t_i)$$

$$\mu_i = t_i e^{\eta_i}$$

le terme *offset* peut être vu comme une variable explicative additionnelle (où le coefficient est toujours 1)

Notation pour les interactions

Lorsqu'on utilise des variables catégoriques qui ont plusieurs niveaux, on peut utiliser une notation abrégée. Prenons un modèle quelquonque `A * B` avec la variable `A` qui a I = 3 niveaux et `B` qui a J = 2 niveaux. Alors, on aurait

$$\ln(\mu_{i,j}) = \alpha + \beta_i^A + \beta_j^B + \gamma_{i,j} \quad i = 1,2,3 \text{ et } j = 1,2$$

0ù on impose les contraintes telles que $\beta_1^A = \beta_1^B = 0$ et $\gamma_{1,j} = \gamma_{j,1} = 0$.

Approximation de la Binomiale par une Poisson

Si la variable qu'on veut modéliser obéit à une $Bin(m,\pi)$ avec m grand et π petit, alors on peut l'approximer avec une loi de Poisson en prenant le modèle

$$\ln(\mu_i) = \ln(m_i) + \ln(\pi_i)$$

où $\ln(m_i)$ est un terme *offset*

Tableau de contingence

Lorsque toutes les variables sont des catégorielles, on peut créer un tableau de contingence, où on veut modéliser le nombre dans chaque case avec un GLM Poisson.

On a 3 modèles dans les tableaux de contingence (illustré avec des modèles simples qui ont les variables explicatives `A`, `B` et `C` avec J,K et L niveaux :

- Modèle d'indépendance : $A + B + C$
- Modèle d'indépendance partielle (celui qu'on veut tester) :
  $A + B * C$
- Modèle d'indépendance conditionnelle (aussi appelé le *modèle saturé* [1] :
  $A * B * C$

On peut alors tester l'indépendance de certaines variables en faisant une **Analyse de la déviance** (section 5).

Cote

La cote de A est définie par

$$\text{Cote}(A) = \frac{\Pr(A)}{\Pr(\overline{A})} = \frac{\Pr(A)}{1 - \Pr(A)}$$

Sousdispersion et susdispersion

Avec le modèle Poisson, on suppose que $\mathrm{E}[Y_i|x_i] = \mathrm{Var}(Y_i|x_i)$. Toutefois, les données peuvent être **sous-dispersées** si

$$\mathrm{E}[Y_i|x_i] > \mathrm{Var}(Y_i|x_i)$$

On détecte aussi la sous-dispersion si $D(y;\hat{\mu})/dl < 0.6$ ou $X^2 < 0.6$. On peut régler les problèmes de sous-dispersion en utilisant une distribution binomiale. Les

1. Ce modèle est celui qui prédit le mieux, mais n'est d'aucune utilité car il a autant de paramètres qu'on a d'observations. On essaie donc de voir si le modèle d'indépendance partielle est une bonne simplification.

données peuvent être **surdispersées** si

$$\mathrm{E}[Y_i|x_i] < \mathrm{Var}(Y_i|x_i)$$

On le détecte lorsque $D(y;\hat{\mu})/dl > 1.7$ ou $X^2 > 1.7$

Binomiale négative

Lorsque les données sont surdispersées, on peut utiliser la distribution binomiale négative dans notre modélisation. Soit $Y|Z = z \sim Pois(\mu z)$ et $Z \sim \Gamma(\theta_z, \theta_z)$, alors $\mathrm{E}[Y] = \mu$ et $\mathrm{Var}(Y) = \mu + \frac{\mu^2}{\theta_z}$ et on a que $Y \sim BinNeg(\mu, \theta_z)$ telle que

$$f_Y(y) = \frac{\Gamma(\theta_z + y)}{\Gamma(\theta_z)y!}\left(\frac{\mu}{\mu+\theta_z}\right)^y\left(\frac{\theta_z}{\mu+\theta_z}\right)^{\theta_z}$$

Lorsque $\theta_z \to \infty$, on retombe sur le modèle Poisson. On peut faire un TRV pour valider si le modèle Poisson est une bonne simplification du modèle binomiale négative :

$$\Pr\left(2\left(\ell^{Pois}(\hat{\boldsymbol{\beta}}) - \ell^{NB}(\hat{\boldsymbol{\beta}})\right) > x\right) = \frac{1}{2}\Pr\left(\chi^2_{(1)} > x\right)$$

Modèle Poisson gonflée à zéro

Lorsqu'on a une masse de probabilité à zéro plus importante à 0, on peut utiliser la loi de Poisson *gonflée à zéro*, en modélisant à la fois la probabilité π_i que la fréquence soit égale à zéro (avec un modèle binomial logistique) et λ_i la fréquence avec un modèle Poisson avec fonction de lien log.

7 Modélisation de données binomiales

7.1 Cas Bernouilli

Tableau de mauvaise classification

	Prédiction $\hat{Y}_i$	
Vrai Y_i	0	1
0	a	b
1	c	d

En forçant $\hat{Y}_i$ tel que

$$\hat{Y}_i = \begin{cases} 0 & , \hat{\pi}_i < \tau \\ 1 & , \hat{\pi}_i \geq \tau \end{cases}$$

On peut calculer la statistique de **sensitivité** (i.e. le taux de bonne classificiation des vrais 1) et de **spécificité** (i.e. le taux de bonne classification des vrais 0) :

$$\text{Sensitivité} = \alpha(\tau) = \frac{d}{c+d}$$

$$\text{Spécificité} = \beta(\tau) = \frac{a}{a+b}$$