2 Régression linéaire simple

Postulats

H$_1$ Linéarité : $\mathrm{E}[\varepsilon_i] = 0$

H$_2$ Homoscédasticité : $\mathrm{Var}(\varepsilon_i) = \sigma^2$

H$_3$ Indépendance : $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$

H$_4$ Normalité : $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

Modèle

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

Estimation des paramètres

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

Estimation de σ^2

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

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

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

Intervalle de confiance

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

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

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
   $Y_0 = \beta_0 + \beta_1 x_0 + \epsilon$

$\because \mathrm{E}[\epsilon] = 0 \therefore \widehat{\mathrm{E}[Y|x_o]} = \widehat{Y}_o = \beta_0 + \beta_1 x_0$

2 sources d'erreur dans nos prévisions

1. **Parameter risk** pour $\mathrm{E}[Y|x_o]$ et Y_0.
   alias incertitude liée à l'estimation des paramètres β_0 & β_1.
2. **Process risk** pour Y_0.
   alias fluctuation des valeurs de la variable réponse autour de sa moyenne ϵ.

Intervalles de confiance de niveau $1 - \kappa$

$$\mathrm{E}[Y|x_o] : \left[\widehat{Y}_0 \pm t_{(n-2),\frac{k}{2}}\sqrt{s^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right)}\right]$$
$$Y_0 : \left[\widehat{Y}_0 \pm t_{(n-2),\frac{k}{2}}\sqrt{s^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right)}\right]$$

Analyse de la variance (ANOVA)

Pour déterminer la proportion de la variabilité de Y est expliquée par le modèle
C'est-à-dire, que ça explique la variabilité des Y_i à la moyenne $\bar{Y}$.

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 est le nombre de variables explicatives dans le modèle.
Où p' est le nombre de variables estiméees 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 totale dans Y qui est expliquée par x

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

Test F de Fisher pour la validité globale de la régression

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

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

où p est le nombre de variables explicatives dans le modèle (régression linéaire simple, $p = 1$ et $p' = p+1$).
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
$V(\widehat{\epsilon}_i)$	$\sigma^2(1-h_{ii})$
$Cov(\widehat{\epsilon}_i, \widehat{\epsilon}_j)$	$-\sigma^2\left(\frac{1}{n} + \frac{(x_i-\bar{x})(x_j-\bar{x})}{S_{xx}}\right)$

où $h_{ii} = \frac{1}{n} + \frac{(\bar{x}-x_i)^2}{S_{XX}}$.

Vérification des postulats

Les résidus studentisés sont définis par

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

Linéarité

- graphique $Y_i|x_i$
- graphique $\hat{\varepsilon}_i|\hat{Y}_i$
- graphique $\hat{\varepsilon}_i|x_i$

Les deux derniers graphique doivent être centrés à 0 et d'allure aléatoire.

Homoscédasticité

- 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.

Indépendance

- 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.
   $g(Y) = \arcsin(\sqrt{Y})$

3 Régression linéaire multiple

Le modèle et ses propriétés

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

Paramètres du modèle

Estimation et propriétés des paramètres

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

Intervalle de confiance sur les paramètres

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

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

Estimation de σ^2

$$\hat{\sigma}^2 = s^2 = \frac{\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}}{n-p'}$$

Il peut être démontré que cette estimateur est sans biais et indépendant de $\hat{\boldsymbol{\beta}}$

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

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

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

On a aussi que

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

Pour les résidus de la droite de régression, on a

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

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

$$\mathrm{a}^\top\hat{\boldsymbol{\beta}} = \mathbf{a}^\top(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top\mathbf{Y}$$

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

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

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

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

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

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}.

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

$$H_0 : Y_i = \beta_0 + \beta_1 x_{i1} + ...\beta_{ik} \quad \text{(modèle réduit)}$$

Pour

$$H_1 : Y_i = \beta_0 + \beta_1 x_{i1} + ...\beta_{ip} \quad \text{(modèle complet)}$$

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

Multicollinéarité

Problèmes potentiels

- Instabilité de $(\mathrm{X}^\top\mathrm{X})^{-1}$, i.e. une petite variation de Y peut changer de grandes variations en $\hat{\beta}$ et $\hat{Y}$;
- $\hat{\beta}_i$ de signes contre-intuitif ;
- $Var(\hat{\beta}_i)$ et $Var(\hat{Y})$ 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 de corrélation $\mathbf{X}^{*\top}\mathbf{X}^*$ est élevée, où $\mathbf{X}^* = \begin{bmatrix} \frac{x_1-\bar{x}_1}{s_1} & ... & \frac{x_p-\bar{x}_p}{s_p} \end{bmatrix}_{1\times p}$
- Si le facteur d'influence de la variance (VIF_j) est élevé, où

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

avec R_j^2 le coefficient de détermination de la régression ayant comme variable réponse le j^{e} variable et les $(j-1)$ autres variables exogènes en *input*.

- La variance de $\hat{\beta}_j$ s'exprime en fonction du VIF comme suit :

$$\mathrm{Var}\left(\hat{\beta}_j\right) = \frac{\sigma^2}{(\mathbf{X}^{*\top}\mathbf{X}^*)_{\mathbf{jj}}}VIF_j$$

Solution

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

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 $\hat{\varepsilon}_{x_j|\mathbf{X}_{-\mathbf{j}}}$).
- 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 $r_i|\hat{Y}_i$

Indépendance entre les observations

› Graphique $\hat{\varepsilon}_i|i$

› Test de Durbin-Watson (pas à l'examen)

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 augmente inutilement le biais des estimations$(\hat{\beta})$

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.

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

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 :

$$R_p^2 = 1 - \frac{PRESS}{SST}$$

Les résidus PRESS

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

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

Échantillion de test et validation croisée par k ensemble

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

$$C_p = p' + \frac{(s_p^2 - \hat{\sigma^2})(n-p')}{\hat{\sigma^2}} = \frac{SSE}{\hat{\sigma^2}} + 2p' - n$$

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. On cherche à minimiser ces 2 critères.

$$BIC = n \cdot \ln\left(\frac{SSE}{n}\right) + \ln(n)p'$$

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).

Régression Ridge

- Les coefficients de la régularisation sont réduits (*shrinked*) car on applique une pénalité sur leur taille totale avec la norme $\ell_2 = \sqrt{\sum_{i=1}^{p} \beta_j^2}$
- On veut minimiser l'équation suivante :
  $$R^{Ridge}(\beta) = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x^{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$
  Et on trouve que
  $$\hat{\boldsymbol{\beta}}^{Ridge} = \left(\mathbf{X}^\top Y + \lambda \mathbf{I}_{p\times p} \right)^{-1} \mathbf{X}^\top \mathbf{Y}$$
- Cette méthode est très utile **Lorsqu'il y a beaucoup de variables explicatives**. On choisit la valeur optimale pour le coefficient de régularisation λ avec une validation croisée.
- Si la valeur de λ augmente, le modèle perd en flexibilité et donc la variance des estimateurs diminue. Par contre, le biais augmente.
- Le modèle de régression Ridge est plus difficile à interpréter, car plusieurs coefficients des paramètres peuvent être près de 0.

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

- Très similaire à la régression Ridge, sauf qu'on utilise la norme ℓ_1 pour appliquer une contrainte à l'équation à minimiser : $\ell_1 = \sum_{j=1}^{p} |\beta_j|$
- L'équation à minimiser est donc
  $$S^{Lasso}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$
- La différence avec Ridge est que les paramètres peuvent être égaux à zéro **(il y a donc une sélection des variables)**.

5 Modèles linéaires généralisés (GLM)

Famille exponentielle linéaire

Définition

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

- On peut exprimer la fonction de densité (ou masse) de probabilité comme
  $$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y;\phi) \right)$$
  où θ est le paramètre canonique et ϕ est le paramètre de dispersion.
- la fonction c ne dépend pas du paramètre θ.
- Le support de Y ne dépend pas des paramètres θ ou ϕ.

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

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

Lemme de la Log-vraisemblance

Soit $\ell(\theta,\phi;Y) = \mathrm{L}(\theta,\phi;Y)$ la log-vraisemblance. Alors,

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

et

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

Fonction de lien

Soit $\eta = \mathbf{X}\boldsymbol{\beta}$. La fonction de lien est la transformation qu'on applique à η afin de limiter le support de Y.

Lien log $\eta = \ln\mu \leftrightarrow \mu = e^\eta$

Lien logistique $\eta = \ln\left(\frac{\mu}{1-\mu}\right) \leftrightarrow \mu = \frac{e^\eta}{1+e^\eta}$

Lien probit $\eta = \Phi^{-1}(\mu) \leftrightarrow \mu = \Phi(\eta)$

Lien log-log complémentaire $\eta = \ln(-\ln(1-\mu)) \leftrightarrow \mu = 1 - e^{-e^\eta}$

Lien canonique $\eta = \theta$

Estimation des paramètres

- On estime $\hat{\beta}$ avec la méthode du maximum de vraisemblance (EMV ou *MLE* en anglais)
- L'EMV est cohérent, i.e.
  $$\hat{\boldsymbol{\beta}} \underset{n\to\infty}{\longrightarrow} \boldsymbol{\beta}$$
- L'estimateur a une normalité asymptotique, i.e. lorsque $n \to \infty$,
  $$\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \frac{\mathcal{I}(\boldsymbol{\beta})^{-1}}{n} \right)$$

où $\mathcal{I}(\boldsymbol{\beta})_{(p'\times p')}$ est la matrice d'information de Fisher :

$$\begin{aligned}\mathcal{I}(\boldsymbol{\beta}) &= \mathrm{E}\left[\dot{\ell}(\boldsymbol{\beta}; Y_1, ..., Y_n)\dot{\ell}(\boldsymbol{\beta}; Y_1, ..., Y_n)^\top\right] \\ &= -\mathrm{E}\left[\ddot{\ell}(\boldsymbol{\beta}; Y_1, ..., Y_n)\right]\end{aligned}$$

› On peut estimer la matrice d'information de Fisher avec **l'information observée** :

$$\mathcal{I}(\hat{\boldsymbol{\beta}}) = -\sum_{i=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\beta}^2}\ell(\boldsymbol{\beta}; Y_i)\Big|_{\hat{\boldsymbol{\beta}}}$$

Algorithme de Newton-Raphson

L'objectif est de trouver $\hat{\boldsymbol{\beta}}$ qui maximise $\ell(\hat{\boldsymbol{\beta}})$, ce qui revient à trouver $\dot{\ell}(\hat{\boldsymbol{\beta}}) = 0$. On utilise l'approximation de Taylor de premier ordre dans l'algorithme :

(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 méthode est la même que l'algorithme de Newton-Raphson, à l'exception qu'on remplace $\ddot{\ell}(\hat{\boldsymbol{\beta}})$ par $-\mathrm{E}\left[\ddot{\ell}(\hat{\boldsymbol{\beta}})\right]$ à l'étape (2.1)

Construction d'IC sur les paramètres

› Lorsqu'on prédit des données, on peut aussi créer un I.C de confiance pour le prédicteur linéaire η_i. Par les propriétés du maximum de vraisemblance, quand $n \to \infty$, on a que $\hat{\boldsymbol{\beta}}$ est asymptotiquement normal. Alors, puisque η est une combinaison linéaire de v.a. *approximativement* normales, alors

$$\eta_i \approx \mathcal{N}\left(\eta_i, \widehat{\mathrm{Var}}(\hat{\eta}_i)\right)$$

› Et on a que (dans le cas simple où le modèle est $\beta_0 + \beta_1 x_{i1}$),

$$\begin{aligned}\mathrm{Var}\left(\hat{\eta}_i\right) &= \mathrm{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1}\right) \\ &= \mathrm{Var}\left(\hat{\beta}_0\right) + x_{i1}^2\mathrm{Var}\left(\hat{\beta}_1\right) \\ &\quad + 2x_{i1}\mathrm{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right)\end{aligned}$$

› Dans le cas multivarié, on a

$$\mathrm{Var}\left(\hat{\eta}_i\right) = \mathbf{X}^\top\mathcal{I}(\hat{\boldsymbol{\beta}})^{-1}\mathbf{X}$$

› L'intervalle de confiance pour η_i est

$$\hat{\eta}_i \pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{\eta}_i)}$$

› Un intervalle de confiance (non-centré) pour μ_i, en utilisant la fonction de lien inverse $g^{-1}(\eta)$ serait

$$\mu_i \in \left[g^{-1}\left(\hat{\eta}_i^{(L)}\right), g^{-1}\left(\hat{\eta}_i^{(U)}\right)\right]$$

› En utilisant la méthode Delta, on obtient un I.C qui est centré pour μ_i, on a

$$\mu_i \in z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}(\hat{\mu}_i)}$$

où

$$\mathrm{Var}\left(\hat{\mu}_i\right) = \left(\frac{\partial}{\partial\eta_i}g^{-1}(\eta_i)\Big|_{\eta_i=\hat{\eta}_I}\right)^2\mathrm{Var}\left(\hat{\eta}_i\right)$$

Statistique de Wald

Test d'hypothèse pour tester $H_0 : \beta_j = 0$, $H_1 : \beta_j \neq 0$. On a que

$$Z = \frac{\beta_j}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j)}} \sim \mathcal{N}(0,1)$$

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

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

Test du rapport de vraisemblance

On teste $H_0 : \beta \in \beta_0$ et $H_1 : \beta \in \beta_1$, où β_1 est le complément de l'espace β_0, qui est une sélection réduite des variables explicatives disponibles. On teste

$$\lambda(y) = \frac{\mathrm{L}\left(\hat{\beta}^{(H_0)}\right)}{\mathrm{L}(\hat{\beta})}$$

$\lambda(y)$ sera assurément plus petit que 1 (il y a moins de variables explicatives). Mais on veut tester si $\lambda(y)$ est plus petit qu'une certaine valeur critique.

› Si H_0 *spécifie tous les paramètres du modèle*, on a

$$-2\ln\lambda(y) \sim \chi^2_{p'} \quad \text{, Sous } H_0$$

› Si H_0 *spécifie partiellement les paramètres du modèle*, on a

$$-2\ln\lambda(y) \sim \chi^2_{k_2-k_1} \quad \text{, Sous } H_0$$

où k_1 est le nombre de paramètres non-spécifiés dans H_0 et k_2 le nombre de paramètres non-spécifiés dans H_1.

› Avec le TRV, on peut seulement comparer des modèles qui sont liés ($\hat{\beta}^{(H_0)}$ doit être un sous-ensemble de $\hat{\beta}$).

Adéquation du modèle

Statistiques χ^2 de Pearson

On peut valider l'adéquation du modèle avec la statistique X^2, où

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

Avec $X^2 \leq \chi^2_{n-p',1-\frac{\alpha}{2}}$ si le modèle est adéquat. Si ϕ est inconnu, on peut l'estimer avec $\hat{\phi} = \frac{X^2}{n-p'}$

Déviance

On a

$$2(\ell(\tilde{\theta}) - \ell(\hat{\theta})) \sim \chi^2_{n-p'}$$

avec $\bar{\theta}$ est le modèle nul, $\hat{\theta}$ le modèle à l'étude et $\tilde{\theta}$ le modèle complet, où $\hat{\mu}_i = y_i$. Cette expression repré-

sente la déviance $D(y;\hat{\mu})$:

$$\begin{aligned} 2(\ell(\tilde{\theta}) - \ell(\hat{\theta})) &= 2\sum_{i=1}^{n} \frac{w_i}{\phi}(y_i\tilde{\theta} - b(\tilde{\theta}) - y_i\hat{\theta} + b(\hat{\theta})) \\ &= 2\sum_{i=1}^{n} \frac{w_i}{\phi} y_i(\tilde{\theta} - \hat{\theta}) - (b(\tilde{\theta} - b(\hat{\theta})) \\ &= \frac{D(y;\hat{\mu})}{\phi} \end{aligned}$$

Si ϕ est inconnu, on peut l'estimer avec $\hat{\phi} = \frac{D(y;\hat{\mu})}{n-p'}$

Comparaison de modèles

Les critères classiques AIC et BIC peuvent être utilisés pour comparer des modèles. On peut aussi faire une analyse de la déviance

Analyse de la déviance

On compare le modèle A et le modèle B (où A est une simplification de B). Le modèle A sera une bonne simplification de B si

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

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 B. On rejète H_0 que A est une bonne simplification de B si la différence est déviance réduite est supérieure à $\chi^2_{p_B - p_A, 1-\frac{\alpha}{2}}$

Analyse des résidus

Résidus de Pearson

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

Aussi, les résidus d'Anscombe et les résidus de la déviance.

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) :

$$\begin{aligned} \ln\left(\frac{\mu_i}{t_i}\right) &= \boldsymbol{x}_i\boldsymbol{\beta} \\ \ln(\mu_i) &= \boldsymbol{x}_i\boldsymbol{\beta} + \ln(t_i) \\ \mu_i &= t_i e^{\eta_i} \end{aligned}$$

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 abbrégée. 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 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 $X^2 > 1.7$

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.

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) :

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

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