1 Régression linéaire simple

Postulats

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

Modèle

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

Estimation des paramètres

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

Estimation de σ^2

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

Propriété des estimateurs

$$E[\hat{\beta}_1] = \beta_1 \quad , Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{XX}}$$
$$\hat{\beta}_1 \overset{H_4}{\sim} N(\beta_1, \frac{\sigma^2}{S_{XX}})$$
$$E[\hat{\beta}_0] = \beta_0 \quad , Var(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)$$
$$\hat{\beta}_0 \overset{H_4}{\sim} N(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)$$
$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}\sigma^2}{S_{XX}}$$

Tests d'hypothèse sur les paramètres

$H_0 : \hat{\beta} = \theta_0$, $H_1 : \hat{\beta} \neq \theta_0$

$$t_{obs} = \frac{\hat{\beta} - \theta_0}{\sqrt{\hat{Var}(\hat{\beta})}} \sim T_{n-2}$$

On Rejette H_0 si $t_{obs} > |t_{n-2}(1 - \frac{\alpha}{2})|$

Intervalle de confiance

Pour la droite de régression ($E[Y_0|x_0]$)

Sachant que $E[Y_0|x_0] = \beta_0 + \beta_1 x_0$, on a l'IC suivant

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

Pour la prévision de Y_0

Sachant que $Y_0 = \beta_0 + \beta_1 x_0 + \varepsilon$, on a l'IC suivant

$$\hat{Y}_0 \pm t_{n-2}\left(1 - \frac{\alpha}{2}\right)\sqrt{s^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right)}$$

Analyse de la variance (ANOVA)

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.

Coefficient de détermination

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

On a aussi la relation suivante avec F_{obs} :

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - p'}{p}$$

Test F de Fisher pour la validité globale de la régression

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

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

où p est le nombre de variables explicatives dans le modèle (régression linéaire simple, $p = 1$ et $p' = p + 1$.

Distribution d'un résidu ε

$$E[\hat{\varepsilon_i}] = 0 \quad , Var(\hat{\varepsilon_i}) = \sigma^2(1 - h_{ii})$$

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'autocorrélation (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.

2 Régression linéaire multiple

Le modèle et ses propriétés

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

Paramètres du modèle

Estimation et propriétés des paramètres

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

Intervalle de confiance sur les paramètres

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

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

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

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

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

Pour les résidus de la droite de régression, on a

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

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

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

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

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

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

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 $\mathbf{X}^{*} = \left[\frac{x_1-\bar{x}_1}{s_1} \quad ... \quad \frac{x_p-\bar{x}_p}{s_p}\right]$
- 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 $\hat{\beta}_i$ s'exprime en fonction du VIF comme suit :
$$Var(\hat{\beta}_i) = \frac{\sigma^2}{(\mathbf{X}^{*\mathbf{T}}\mathbf{X}^{*})_{\mathbf{jj}}} VIF_j$$

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 ($\hat{\varepsilon}_{\mathbf{Y}|\mathbf{X}_{-j}}$ en fonction de $\hat{\varepsilon}_{x_j|\mathbf{X}_{-j}}$).
- 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)

3 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 obtenue en 2, noté $\hat{Y}_{i,-i}$
2. Calculer la somme des carrées 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 davoir 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 $k-1$ Échantillion restant.
   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 grandeur de l'échantillion. 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 variabel 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).