1 Régression linéaire simple

Postulats

- $\mathbf{H_1}$ Linéarité : $E[\varepsilon_i] = 0$
- $\mathbf{H_2}$ Homoscédasticité : $Var(\varepsilon_i) = \sigma^2$
- $\mathbf{H_3}$ Indépendance : $Cov(\varepsilon_i, \varepsilon_j) = 0$
- $\mathbf{H_4}$ 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

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

Pour la prévision de Y_0

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

$$\hat{Y_0} \pm t_{n-2}\left(\frac{\alpha}{2}\right)\sqrt{s^2\left(1 + \frac{1}{m} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right)}$$

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)		

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

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

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

Intervalle de confiance sur les paramètres

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

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

Estimation de σ^2

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

Test d'hypothèse sur un paramètre du modèle

On rejète $H_0 : \beta_j = 0$ si

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

Propriétés de la droite de régression

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

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

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

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 $E\left[\mathbf{Y}|\mathbf{X}^*\right]$

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

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

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

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 $\mathbf{X}^* = \begin{bmatrix} \frac{x_1 - \bar{x}_1}{s_1} & ... & \frac{x_p - \bar{x}_p}{s_p} \end{bmatrix}$
- 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*.

Solution

On retire les variables ayant un VIF élevé (une à la fois)

Analyse des résidus et validation du modèle

Vérification de la linéarité

- On trace les graphiques à variable ajoutée ($\hat{\varepsilon}_{\mathbf{Y}|\mathbf{X}_{-\mathrm{j}}}$ en fonction de $\hat{\varepsilon}_{x_j|\mathbf{X}_{-\mathrm{j}}}$).
- 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.