

# Régression linéaire simple

## Postulats

- H<sub>1</sub>** Linéarité :  $E[\varepsilon_i] = 0$   
**H<sub>2</sub>** Homoscédasticité :  $Var(\varepsilon_i) = \sigma^2$   
**H<sub>3</sub>** Indépendance :  $Cov(\varepsilon_i, \varepsilon_j) = 0$   
**H<sub>4</sub>** Normalité :  $\varepsilon_i \sim N(0, \sigma^2)$

## Modèle

$$E[Y_i|x_i] = \beta_0 + \beta_1 x_i$$

$$Var(Y_i|x_i) = \sigma^2$$

$$Y_i|x_i \stackrel{H_4}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$$

## 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 $\sigma^2$

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

## Propriété des estimateurs

$$E[\hat{\beta}_1] = \beta_1, \quad Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{XX}}$$

$$\hat{\beta}_1 \stackrel{H_4}{\sim} N(\beta_1, \frac{\sigma^2}{S_{XX}})$$

$$E[\hat{\beta}_0] = \beta_0, \quad Var(\hat{\beta}_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right)$$

$$\hat{\beta}_0 \stackrel{H_4}{\sim} N(\beta_0, \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right))$$

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x} \sigma^2}{S_{XX}}$$

## Tests d'hypothèse sur les paramètres

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

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

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  $H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$  si

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

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

## Distribution d'un résidu $\varepsilon$

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

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  $\hat{\varepsilon}_i|\hat{x}_i$

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 résidus 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.

## Régression linéaire multiple

### Le modèle et ses propriétés

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

### Paramètres du modèle

#### Estimation et propriétés des paramètres

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \\ E[\hat{\boldsymbol{\beta}}] &= \boldsymbol{\beta}, \text{Var}(\mathbf{Y}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} \\ \hat{\boldsymbol{\beta}} &\stackrel{H_4}{\sim} N_p(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}) \end{aligned}$$

#### Intervalle de confiance sur les paramètres

$$\begin{aligned} \text{var}[\beta_i] &= \sigma^2 v_{jj} \\ \beta_i &\in \left[ \hat{\beta}_i \pm t_{n-p'} \left(1 - \frac{\alpha}{2}\right) \sqrt{s^2 v_{jj}} \right] \\ \text{où } v_{jj} &\text{ est l'élément } (j, j) \text{ de la matrice } (\mathbf{X}^\top \mathbf{X})^{-1}. \end{aligned}$$

#### Estimation de $\sigma^2$

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

#### Test d'hypothèse sur un paramètre du modèle

On rejette  $H_0 : \beta_j = 0$  si

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

### Propriétés de la droite de régression

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

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

$$\begin{aligned} E[\hat{\mathbf{Y}}] &= \mathbf{X}\boldsymbol{\beta}, \text{Var}(\hat{\mathbf{Y}}) = \sigma^2 \mathbf{H} \\ \hat{\mathbf{Y}} &\stackrel{H_4}{\sim} N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{H}) \end{aligned}$$

Pour les résidus de la droite de régression, on a

$$\begin{aligned} E[\hat{\boldsymbol{\varepsilon}}] &\stackrel{H_1}{=} \mathbf{0}, \text{Var}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2 (\mathbf{I}_{n \times n} - \mathbf{H}) \\ \hat{\boldsymbol{\varepsilon}} &\stackrel{H_4}{\sim} N_n(\mathbf{0}, \sigma^2 (\mathbf{I}_{n \times n} - \mathbf{H})) \end{aligned}$$

### Matrice de projection

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

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

### Intervalle de confiance pour la prévision

#### Théorème de Gauss-Markov

Selon les postulats  $H_1$  à  $H_4$ , l'estimateur

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

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

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

Pour

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

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  $(\mathbf{X}^\top \mathbf{X})^{-1}$ , i.e. une petite variation de  $\mathbf{Y}$  peut changer de grandes variations en  $\hat{\boldsymbol{\beta}}$  et  $\hat{\mathbf{Y}}$ ;
- $\hat{\beta}_i$  de signes contre-intuitif;
- $\text{Var}(\hat{\beta}_i)$  et  $\text{Var}(\hat{\mathbf{Y}})$  très grandes;
- Les méthodes de sélection de variable ne concordent pas;
- Conclusions erronées sur la significativité de certains paramètres, malgré une forte corrélation avec  $\mathbf{Y}$ .

## Détection

- Si  $r_{ij}$  dans la matrice de corrélation  $\mathbf{X}^{*\top} \mathbf{X}^*$  est élevée, où  $\mathbf{X}^* = \begin{bmatrix} x_1 - \bar{x}_1 & \dots & x_p - \bar{x}_p \\ s_1 & & s_p \end{bmatrix}$

- 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  $j^{\text{e}}$  variable et les  $(j - 1)$  autres variables exogènes en *input*.

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

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

## 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 ( $\hat{\varepsilon}_{Y|X_{-j}}$  en fonction de  $\hat{\varepsilon}_{x_j|X_{-j}}$ ).
- Ces graphiques doivent normalement donner une droite de pente  $\beta_j$ .
  - 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

- Mesure de la qualité global du modèle :

$$R_2 = \frac{SSR}{SST}$$

Si on ajoute un 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 trouver les variables exogènes.

- Coefficient de détermination ajusté :

$$R_a^2 = \frac{MSE}{MST}$$

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 à prédire de nouvelles données.

### Principe de la validation croisée

- Enlever la  $i^{\text{e}}$  observation du jeu de données.
- Estimer les paramètres du modèle à partir des  $n - 1$  données restante.
- Prédire  $Y_i$  à partir de  $x_i$  et du modèle obtenue en 2, noté  $\hat{Y}_{i,-i}$
- Répéter les étapes 1-3 pour chaque  $i, i = 1, \dots, n$
- Calculer la somme des carrées des erreurs de prévision  $PRESS = \sum_{i=1}^n (Y_i - \hat{Y}_{i,-i})^2$

On cherche à 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 calculé  $n$  régression :

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

### Échantillon de test et validation croisée par $k$ ensemble

- Enlever le  $k^{\text{e}}$  ensemble de de jeu de donner.
- Estimer les paramètres du modèle à partir des  $k - 1$  Échantillon restant.
- Prédire les observations du  $k^{\text{e}}$  ensemble ( $\hat{Y}_{i,-k}$ ) et calculer  $MSEP_k = \frac{1}{n_k} \sum_{i \in \text{group } k} (Y_i - \hat{Y}_{i,-k})^2$
- Répéter les étapes 1 à 3 pour chaque  $k, k = 1, \dots, k$
- Calculer la moyenne 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 à AIC, mais la pénalité des paramètres dépend de la grandeur de l'échantillon. On cherche à minimiser ces 2 critères.

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

## Méthode algorithmiques

