Économétrie

Régression Linéaire

$Y = X\beta + \varepsilon$
n nombre d'observation
k nombre de coefficients i.e. $\beta = (\beta_0 \quad \cdots \quad \beta_{k-1})'$

Termes d'erreur $\varepsilon_i = Y_i - X_i\beta$
Représente la variation entre la valeur observée pour un individu (Y_i), de la valeur supposée du modèle ($X_i\beta$)
MCO, hypothèses :
- les termes ϵ_t sont iid et suivent une loi normale, en général $N(0,1)$
- ils sont également indépendants des variables explicatives X
- $E(\epsilon) = 0$ et $\text{Var}(\epsilon)$ est une constante

Estimation des coefficients par MCO
- $\beta = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}$ i.e. $y_i = a_0 + a_1 x_i + \varepsilon_i$

$$\boxed{\hat{a}_1 = \frac{\sum_{t=1}^{12}(y_t - \overline{y})(x_t - \overline{x})}{\sum_{t=1}^{12}(x_t - \overline{x})^2} \quad \hat{a}_0 = \overline{y} - \hat{a}_1\overline{x}}$$

- Sous forme matricielle

$$\boxed{\hat{\beta} = (X'X)^{-1}X'Y}$$

Résidus $e_i = y_i - \hat{y}_i$
Estimation de la *variance résiduelle*

$$\hat{\sigma}_\epsilon{}^2 = \frac{\sum_{i=1}^{n} e_i{}^2}{n-k}$$

Analyse de la variance

Source	Somme des carrés	Degrés de liberté	Variance
Modèle	$SCE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	$k-1$	$\frac{SCE}{k-1}$
Résidus	$SCR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i{}^2$	$n-k$	$\frac{SCR}{k-1} = \hat{\sigma}_\epsilon{}^2$
Total	$SCT = \sum_{i=1}^{n}(y_i - \bar{y})^2 = SCE + SCR$	$n-1$	$\frac{SCT}{n-1}$

Coefficient de détermination et de détermination ajustée
$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$
Le coefficient de détermination non ajusté (R^2) ne prend pas en compte le nombre de variables explicatives et augmente ainsi mécaniquement avec celles-ci.
$\bar{R}^2 = 1 - \frac{SCR/n-k}{SCT/n-1}$
Le coefficient de détermination ajusté ($\bar{R}^2$) prend en compte le nombre de variables explicatives et permet ainsi la comparaison de modèle de complexités différentes.
Si on veut comparer un modèle améliorer avec l'ajout de nouvelles variables, il ne suffit pas de regarder le $\bar{R}^2$, il faut également tester la significativité des nouvelles variables.

Interprétation : le modèle permet d'expliquer [R^2 ou $\bar{R}^2$]% des variations de y ➔ si le ratio est faible (élevé), le modèle est de mauvaise (bonne) qualité.
Si il y a qu'une seule variable dans le modèle, alors cette variable permet d'expliquer [R^2 ou $\bar{R}^2$]% des variations de y.

Variance des estimateurs

- $\beta = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}$ i.e. $y_i = a_0 + a_1 x_i + \varepsilon_i$

$$\hat{\sigma}_{\hat{a}_0}{}^2 = \hat{\sigma}_\epsilon{}^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right) \quad \hat{\sigma}_{\hat{a}_1}{}^2 = \frac{\hat{\sigma}_\epsilon{}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

- Matrice de variance-covariance des coefficients

$$\widehat{\Omega}_{\widehat{\beta}} = \hat{\sigma}_\epsilon{}^2 (X'X)^{-1}$$

Intervalle de confiance des coefficients

À $(1-\alpha)$%,

$$\boxed{\beta_i = \left[\hat{\beta}_i \pm t_{(n-k)}^{\alpha/2} \times \hat{\sigma}_{\widehat{\beta}_i}\right]}$$

Interprétation

- si 0 appartient à l'IC, la variable x^i n'est à priori pas significative au seuil de α%
- si 0 n'appartient pas à l'IC, la variable x^i est significative au seuil de α%

Variables qualitatives

Lorsqu'on cherche à introduire une variable explicative de nature qualitative avec m modalités dans le modèle, il nous faut introduire uniquement $(m-1)$ variables dichotomique correspondant à $(m-1)$ modalités choisies, par soucis de colinéarité. On appelle groupe de référence celui correspondant à la modalité non introduite, et les interprétations se font par rapport au groupe de référence. Afin d'introduire de telles variables dichotomiques, il faut s'assurer que chacune des modalités comprend un effectif suffisant, soit 5% de l'effectif total.

Variables croisées

Afin de tester l'effet d'une variable sur une autre, on peut introduire une variable croisée correspondant à l'intersection des deux variables précédentes :

$$var\ croisée = var1 * var2$$

Interprétation : les variables croisées d'interprète de manière conjointe avec les variables non croisées.

Variables quadratiques

Afin de capter des relations non linéaires et en particulier des relations quadratiques entre y et la variable explicative x^i, on intègre une nouvelle variable $(x^i)^2$.

Interprétation des résultats

- Pour une variable quantitative x^i
    - La variable x^i est statistiquement significative (test de Student/IC/p-value)

Toutes choses étant égales par ailleurs, une augmentation de 1 unité de x^i entraine en moyenne une variation $\hat{\beta}_i$ de y.

Toutes choses étant égales par ailleurs, une augmentation de 1 unité de x^i entraine une variation comprise entre [borne inférieure de l'IC] et [borne supérieure de l'IC] de y avec une certitude de $(1-\alpha)$%.

- La variable x^i n'est pas statistiquement significative (test de Student/p-value)

Toutes choses étant égales par ailleurs, la variable x^i n'a pas d'effet significatif sur y.

- Pour une variable qualitative
  - La modalité x^i est statistiquement significative (test de Student/IC/p-value)

Toutes choses étant égales par ailleurs, la modalité entraîne une variation $\hat{\beta}_i$ de y par rapport au groupe de référence.

- La modalité x^i n'est pas statistiquement significative (test de Student/p-value)

Toutes choses étant égales par ailleurs, il n'y a pas de différence en y pour la modalité x^i par rapport au groupe de référence ➔ cette modalité fait partie du groupe de référence également.

- Les modalités x^i et x^j sont statistiquement significatives et différentes (Test de contraintes linéaire de Fisher/signes des coefficients opposés)

Toutes choses étant égale par ailleurs la modalité x^i entraîne une variation $\hat{\beta}_i - \hat{\beta}_j$ de y par rapport à la modalité x^j.

- Pour une variable dichotomique croisée x^i
  - La variable est significative (test de Student/IC/p-value)

Toute choses étant égales par ailleurs, pour un individu vérifiant $var1$ et $var2$, y varie de $\hat{\beta}_i$ en moyenne, par rapport à un individu satisfaisant $var1$ mais pas $var2$.

Prévisions

$\hat{Y} = X\hat{\beta}$ où $\hat{\beta}$ correspond à tous les coefficients estimés, même ceux non significatifs.
Interprétation : si les caractéristiques X sont observées, Y aura une valeur moyenne de $\hat{Y}$.
Intervalle de confiance à $(1-\alpha)$% :

$$y_t = \hat{y}_t \pm t_{(n-k)}^{\alpha/2} \times \hat{\sigma}_\varepsilon$$

Interprétation : si les caractéristiques X sont observées, on peut affirmer avec une certitude de $(1-\alpha)$% que y aura une valeur comprise entre [borne inférieure de l'IC] et [borne supérieure de l'IC].

Changement dans le groupe de référence

On suppose que :

- le groupe de référence initial correspond à la modalité g_1 et que le nouveau groupe de référence correspond à la modalité g_2 pour une variable qualitative
- les autres modalités de la variable qualitative sont notées m_i
- Les autres variables sont notées j

Les nouveaux coefficients se calculent de la manière suivante :

$\beta'_0 = \beta_0 + \beta_{g_2}$: constante pour le nouveau groupe de référence

$\beta'_{g_1} = -\beta_{g_2}$: coefficient associé à l'ancien groupe de référence

$\beta'_{m_i} = \beta_{m_i} - \beta_{g_2}$: coefficient associé à la modalité m_i

$\beta'_j = \beta_j$: coefficient associé à la variable j

Si le nouveau groupe de référence était significatif par rapport aux autres modalités dans l'ancien modèle – par exemple g_2 significatifs et tous les m_i non significatifs –, alors les autres modalités sont forcément significatives dans le nouveau modèle. Si ce n'était pas le cas, on ne peut rien avancer quand à la significativité des modalités dans le nouveau modèle. Si les variables j étaient significatives dans l'ancien modèle elles le sont également dans le nouveau et inversement.

Spécification du modèle

Lorsque l'on demande la spécification d'un modèle, il s'agit d'écrire :

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-1} x_{k-1}$$

Classement de l'effet des variables

Afin de classer les variables en fonction de leur effet, il faut prendre en compte uniquement les coefficients des variables significatives et vérifier qu'ils sont différents (test de contraintes linéaire de Fisher/coefficients de signes opposés).

Les tests

<u>Test de significativité globale</u> : Test de Fisher
- Hypothèses : $H_0 : \beta_1 = \cdots = \beta_{k-1} = 0 \quad H_1 : \exists \beta_i \neq 0, i = 1, \cdots, k-1$
- Statistique de test :
$$F^* = \frac{R^2}{1-R^2}\frac{n-k}{k-1} \sim F(k-1, n-k)$$
- Règle de décision : rejet de H_0 au seuil de $\alpha\%$ si $F^* > F^{\alpha}_{(k-1,n-k)}$

Interprétation : le modèle est/n'est pas globalement significatif au seuil de $\alpha\%$.

<u>Test de significativité d'un coefficient</u> : Test de Student
- Hypothèses : $H_0 : \beta_i = 0 \quad H_1 : \beta_i \neq 0,$
- Statistique de test :
$$t^* = \frac{\hat{\beta}_i}{\hat{\sigma}_{\hat{\beta}_i}} \sim t(n-k)$$
- Règle de décision : rejet de H_0 au seuil de $\alpha\%$ si $|t^*| > t^{\alpha/2}_{(n-k)}$

Interprétation : la variable x^i est/n'est pas statistiquement significative au seuil de $\alpha\%$
➔ l'issu du test conditionne l'interprétation de l'effet de x^i sur y.

<u>Test de contraintes</u> : Test de contraintes linéaires de Fisher
- Hypothèses : H_0: contraintes vérifiés $\quad H_1$: une des contraintes n'est pas vérifiée
- Statistique de test :
  1) Estimer le modèle général et le modèle contraint
$$F^* = \frac{SCR_c - SCR_g}{c}\frac{n-k}{SCR_g} \sim F(c, n-k)$$
  où SCR_g (SCR_c) correspond au SCR du modèle général (avec contraintes).
  2) Sans passer par l'estimation des deux modèles – Stata
$$F^* = \frac{\left(R\hat{\beta} - q\right)'[R(X'X)^{-1}R']^{-1}\left(R\hat{\beta} - q\right)/c}{s^2}$$
     où $R\beta = q$
     avec
     - R une matrice de c lignes exprimant les contraintes : pour chaque contrainte – qui correspond à une ligne – 1 sur la colonne correspondant à la variable pour égaliser le coefficient à une valeur ou 1 et -1 sur les colonnes correspondantes aux variables pour effectuer une égalité
     - q un vecteur de longueur c exprimant les égalités : pour chaque contrainte – qui correspond à une ligne – la valeur correspondant pour une égalité d'un coefficient de variable, 0 si deux coefficients sont égalisés.
- Règle de décision : rejet de H_0 au seuil de $\alpha\%$ si $F^* > F^{\alpha}_{(c,n-k)}$

Interprétations : si H_0 est rejeté alors les contraintes émises en H_0 sont fausses au seuil de $\alpha\%$.
- ➢ Pour une variable qualitative, on peut se servir de ce test afin de tester sa pertinence dans le modèle en émettant en H_0 l'hypothèse que tous les coefficients associés aux variables dichotomiques correspondant aux modalités sont nuls.
  ➔ si H_0 est rejeté alors la variable qualitative est pertinente.

- Lorsqu'on ajoute plusieurs nouvelles variables à un modèle et qu'on veut tester si elles sont significatives, ou alors lorsqu'on veut tester simultanément la significativité de plusieurs variables, on se sert de ce test en émettant en H_0 l'hypothèse que tous les coefficients correspondant aux variables mentionnées sont égaux à 0.

Test de stabilité des paramètres : Test de Chow

- Hypothèses : $H_0: \beta^T = \beta^A = \beta^B \quad H_1: \beta^A \neq \beta^B$
  où $\beta^T, \beta^A, \beta^B$ correspondent aux coefficient du modèle : comportant toutes les variables, pour la population avec la modalité A, pour la population avec la modalité B, respectivement.
- Statistique de test :
$$F^* = \frac{\left(SCR_T - (SCR_A + SCR_B)\right)/k}{(SCR_A + SCR_B)/(n-2k)} \sim F(k, n-2k)$$
  Ici, k correspond aux modèles A et B et n au modèle T.
- Règle de décision : rejet de H_0 au seuil de $\alpha\%$ si $F^* > F^{\alpha}_{(k,n-2k)}$

Interprétations : si H_0 est rejeté alors les variables explicatives on des effets différents selon les modalités A et B.

Démonstrations

$$\hat{a}_1 = \frac{\sum_{t=1}^{12}(y_t - \overline{y})(x_t - \overline{x})}{\sum_{t=1}^{12}(x_t - \overline{x})^2} \quad \hat{a}_0 = \overline{y} - \hat{a}_1\overline{x}$$

Notons $Q = \sum_{i=1}^{n} {e_i}^2$. Le problème consiste à minimiser cette valeur.
Nous avons $e_i = y_i - \hat{y}_i$.
Ainsi, $Q = \sum_{i=1}^{n} {y_i}^2 - 2a_0 \sum_{i=1}^{n} y_i - 2a_1 \sum_{i=1}^{n} y_i x_i + n \cdot {a_0}^2 + 2a_0 a_1 \sum_{i=1}^{n} x_i + {a_1}^2 \sum_{i=1}^{n} {x_i}^2$

Équations normales : $\begin{cases} \frac{d}{da_0} Q = -\sum_{i=1}^{n} y_i + n \cdot a_0 + a_1 \sum_{i=1}^{n} x_i \\ \frac{d}{da_1} Q = -\sum_{i=1}^{n} y_i x_i + a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} {x_i}^2 \end{cases}$

Or, minimiser Q revient à résoudre $\begin{cases} \frac{d}{da_0} Q = 0 \\ \frac{d}{da_1} Q = 0 \end{cases} \Rightarrow \begin{cases} \hat{a}_1 = \frac{\sum_{t=1}^{12}(y_t - \overline{y})(x_t - \overline{x})}{\sum_{t=1}^{12}(x_t - \overline{x})^2} \\ \hat{a}_0 = \overline{y} - \hat{a}_1\overline{x} \end{cases}$

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Notons $Q = \epsilon'\epsilon$. Le problème consiste à minimiser cette valeur.
Nous avons $\epsilon = Y - X\beta$.
Ainsi, $Q = (Y - X\beta)'(Y - X\beta) = YY' - \beta'X'Y - Y'X\beta + \beta'X'X\beta$
Remarquons $[(Y')_{1\times n}(X)_{n\times k}(\beta)_{k\times 1}]_{1\times 1} \Rightarrow Y'X\beta = \beta'X'Y$
Donc $Q = YY' - 2\beta'X'Y + \beta'X'X\beta$
Équation normale : $\frac{\partial Q}{\partial \beta} = -2X'Y + 2X'X\beta$
Or, minimiser Q revient à résoudre $\frac{\partial Q}{\partial \beta} = 0 \Rightarrow X'X\beta = X'Y \Rightarrow \hat{\beta} = (X'X)^{-1}X'Y$

$$\beta_i = \left[\hat{\beta}_i \pm t_{(n-k)}^{\alpha/2} \times \hat{\sigma}_{\hat{\beta}_i}\right]$$

$\hat{\beta}_i$ est une somme de variable aléatoire suivant une loi normale donc $\hat{\beta}_i$ suit également une loi normale d'espérance β_i et variance estimée à $\hat{\sigma}_{\hat{\beta}_i}^2$. Or, $\hat{\sigma}_{\hat{\beta}_i}^2$ est une Chi-deux avec $n - k$ degrés de liberté. Ainsi, $\frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_{\hat{\beta}_i}}$ suit une loi de Student avec $n - k$ degrés de liberté.

$\mathbb{P}(l \leq \beta_i \leq u) = 1 - \alpha \Leftrightarrow \mathbb{P}\left(\frac{\hat{\beta}_i - l}{\hat{\sigma}_{\hat{\beta}_i}} \geq t_{(n-k)} \geq \frac{\hat{\beta}_i - u}{\hat{\sigma}_{\hat{\beta}_i}}\right) = 1 - \alpha$

$\Leftrightarrow \mathbb{P}\left(\frac{\hat{\beta}_i - l}{\hat{\sigma}_{\hat{\beta}_i}} \geq t_{(n-k)}\right) = \alpha/2$ par symétrie de la loi de Student avec $\frac{\hat{\beta}_i - l}{\hat{\sigma}_{\hat{\beta}_i}} = -\frac{\hat{\beta}_i - u}{\hat{\sigma}_{\hat{\beta}_i}}$

$\Leftrightarrow \mathbb{P}\left(\frac{\hat{\beta}_i - l}{\hat{\sigma}_{\hat{\beta}_i}} \geq t_{(n-k)}\right) \Leftrightarrow \frac{\hat{\beta}_i - l}{\hat{\sigma}_{\hat{\beta}_i}} = t_{(n-k)}^{\alpha/2} \Leftrightarrow \beta_i = \left[\hat{\beta}_i \pm t_{(n-k)}^{\alpha/2} \times \hat{\sigma}_{\hat{\beta}_i}\right]$

Modèles à choix discret

$Y_i = \begin{cases} 1 & \text{si } Y_i^* > 0 \\ 0 & \text{si } Y_i^* \leq 0 \end{cases}$ avec $Y_i^* = X_i\beta + \varepsilon_i$ appelée variable latente

$$\boxed{p_i = \mathbb{P}(Y_i = 1) = F(X_i\beta)}$$

où F correspond à la FdR d'une loi normale ou logistique.
La FdR utilisée correspond du modèle (logit ou probit) qui lui-même dépend de la distribution des termes d'erreur.

Réalisation du modèle

1- Récolter les données
2- Recoder les données (variables qualitatives)
3- Estimer le modèle (Maximum de vraisemblance)
4- S'assurer de la qualité d'ajustement du modèle (matrice de confusion et taux)
5- Regarder les résultats (significativité et interprétation des variables)

Loi normale

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

Loi loglogistique

$$F(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}$$

Vraisemblance du modèle

$$L(Y_i, X_i\beta) = \prod_{i=1}^{n} [p_i]^{y_i}[1-p_i]^{1-y_i}$$

Estimation

Coefficients β par $\hat{\beta}$ Estimateurs du Maximum de Vraisemblance
p_i par $\hat{p}_i = \mathbb{P}(\hat{Y}_i = 1) = F(X_i\hat{\beta})$

Effet marginal

- Variable quantitative x^j

L'effet marginal sur p_i de l'augmentation d'une unité de la variable x^j est donné par

$$\frac{\partial \mathbb{P}(Y_i = 1)}{\partial x_i^j} = \beta_j \times f(X_i\beta)$$

- Variable dichotomique x^j

L'effet marginal sur p_i pour une variable dichotomique x^j est donné par

$$\mathbb{P}(Y_i = 1 | x^j = 1) - \mathbb{P}(Y_i = 1 | x^j = 0)$$

Élasticité

L'élasticité de $\hat{p}_i$ par rapport à une augmentation de 1% de la variable x^j est donnée par :

$$\frac{\partial \mathbb{P}(Y_i = 1)}{\partial x_i^j} \times \frac{x_i^j}{\mathbb{P}(\hat{Y}_i = 1)} = \beta_j \times f(X_i\beta) \times \frac{x_i^j}{\mathbb{P}(\hat{Y}_i = 1)}$$

Interprétation : une augmentation de 1% de la variable x^j entraîne une augmentation de … % de p_i.

Interprétation
La valeur des coefficients estimés n'est pas interprétable directement. Seuls les signes peuvent être interprétés.

- Variable quantitative
    - Pour une variable x^i significative avec un coefficient positif (négatif)

Toutes choses étant égales par ailleurs, la probabilité de y_i est plus élevée (plus faible) au fur et à mesure que x^i augmente.

-
    - Pour une variable x^i non significative

Toutes choses étant égales par ailleurs, la probabilité de y_i n'est pas impactée par x^i.
* Pour gagner du temps, regrouper l'interprétation de toutes les variables non significatives.

- Variable qualitative
    - Pour une variable x^i significative avec un coefficient positif (négatif)

Toutes choses étant égales par ailleurs, la probabilité de y_i est plus élevée (plus faible) lorsque x^i est vérifiée.

-
    - Pour une variable x^i non significative

Toutes choses étant égales par ailleurs, la probabilité de y_i est la même pour la modalité de x^i et le groupe de référence.

-
    - Pour deux variables significatives x^i et x^j

Si leur coefficient estimé sont de signe opposé, alors on peut interpréter l'effet de l'une par rapport à l'autre.
À l'inverse, si leurs signes ne sont pas opposés, on ne peut pas faire d'interprétation sur l'effet d'une variable par rapport à l'autre.

Variables au carré
Si le carré de la variable est significatif et le coefficient associé est de même signe que celui de la variable simple, alors l'effet est amplifié au fur et à mesure que la variable augmente. À l'inverse, s'il est significatif mais de signe opposé, alors l'effet est réduit au fur et à mesure que la variable augmente.

Qualité d'ajustement du modèle
Matrice de confusion :

- pour chaque individu de l'échantillon, on calcule $\mathbb{P}(\hat{Y}_i = 1)$
- on classe les individus selon deux catégories : $\mathbb{P}(\hat{Y}_i = 1) > 0.5$: "$Y = 1$" ; $\mathbb{P}(\hat{Y}_i = 1) < 0.5$: "$Y = 0$"

Prédictions \ Observée	$Y = 1$	$Y = 0$	
$Y = 1$	Bonnes prédictions	Mauvaises prédictions	
$Y = 0$	Mauvaises prédictions	Bonnes prédictions	
			Nb d'individus

- calculer le taux de mauvaises/bonnes prédictions
- Si le taux de mauvaises (bonnes) prédictions est inférieur (supérieur) à 0.5, le modèle est de bonne qualité

Parfois le taux est bon de manière général mais pas dans une catégorie spécifique : vérifier le taux de la catégorie, par exemple "$Y = 1$".

Tests

Tests de significativité globale / Tests de significativité des variables

- Tes de Wald (le plus courant)
- Test du score
- Test du rapport de vraisemblance : $\chi^2 < 0.05$

Test de contraintes linéaires

- Test de Wald

➢ Tester si deux coefficients sont égaux

Démonstrations

$$\boxed{p_i = \mathbb{P}(Y_i = 1) = F(X_i\beta)}$$

$\varepsilon_1, \cdots, \varepsilon_n$ sont iid et suivent une loi normale ou logistique.

$$p_i = \mathbb{P}(Y_i = 1) = \mathbb{P}(Y_i^* > 0) = \mathbb{P}(X_i\beta + \varepsilon_i > 0) = \mathbb{P}(\varepsilon_i > -X_i\beta) = 1 - \mathbb{P}(\varepsilon_i < -X_i\beta)$$
$$= 1 - F(-X_i\beta) = F(X_i\beta)$$

où F correspond à la FdR d'une loi normale ou logistique en focntion de la distribution des termes d'erreur.
La dernière égalité est vraie car la fd des lois normale et logistique sont symétriques.