Modélisation Statistique
Chapitre 2: Régression linéaire multiple

Mohamed Essaied Hamrita

ISMAI, Université Kairouan. Tunisie
mhamrita@gmail.com
http://hamrita.e-monsite.com/

Avril 2016

Plan du chapitre

Présentation du modèle

Le **modèle linéaire multiple** permet d'expliquer une variable endogène (à expliquée ou dépendante) en fonction de plusieurs variables exogènes (explicatives).

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i, \qquad i = 1, \ldots, n \tag{1}$$

avec:
y: la variable endogène, elle est déterminée par le modèle.
x_k: les variables explicatives supposée exogène.
ε: le terme d'erreur qui est aléatoire.
n: Le nombre d'observations.
β_0 est la constante et β_i $(i = 1, \ldots, k)$ mesurent les variations de y suite à une variation unitaire de $x_k i$, toutes choses restant égales par ailleurs.

Écriture matricielle

L'équation (1) peut être reproduite sous forme matricielle comme suit:

$$\underset{n\times 1}{Y} = \underset{n\times(k+1)}{X}\,\underset{(k+1)\times 1}{\beta} + \underset{n\times 1}{\varepsilon} \tag{2}$$

avec

$$\underset{n\times 1}{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix}, \quad \underset{n\times(k+1)}{X} = [s \;\; x] = \begin{bmatrix} 1 & x_{11} & \dots & x_{k1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x1i & \dots & x_{ki} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & \dots & x_{kn} \end{bmatrix}$$

$$\underset{(k+1)\times 1}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \underset{n\times 1}{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Exemples

Exemple 1: La relation entre le niveau de production (en logarithme) d'une entreprise en fonction du capital utilisé (en logarithme) et nombre d'heures de travail (en logarithme)

$$y_t = \beta_0 + \beta_1 k_i + \beta_2 l_i + \varepsilon_i$$

Dans ce cas, les paramètres β_i $(i = 1, 2)$ représentent les élasticités partielles de la production par rapport au capital et au travail respectivement.

Exemple 2: La relation d'une fonction de demande d'un bien en fonction de son prix et du prix d'un bien sustituant

$$D_t = \beta_0 + \beta_1 P_t + \beta_2 PS_t + \varepsilon$$

Hypothèses sur le modèle

On rappelle les hypothèses du modèles, nécessaires afin de compléter la spécification.

Les hypothèses:

H1: L'espérance des aléas est nulle, $E(\varepsilon) = 0$.

H2: Les alés sont homoscédastiques (de variance constante) et non auto-corrélés (leur covariance est nulle).

$$V(\varepsilon) = E(\varepsilon'\varepsilon) = \sigma^2 I$$

H3: X est une matrice de nombres fixes et données supposées certaines (exogènes).

H4: $\text{rang}(X) = k + 1 \leq n$ La matrice X est de rang complet. Selon cette hypothèse, aucun vecteur de la matrice X ne peut s'écrire comme combinaison linéaire des autres vecteurs. Les vecteurs de la matrice X sont, donc, linéairement indépendants.

H5: $cov(X, \varepsilon) = E(X\varepsilon) = 0$

Estimateurs des MCO

L'estimateur des MCO est déduit de la minimisation de la somme des carrés des erreurs (alés).

$$\min_{\beta} \sum_{i=1}^{n} \varepsilon_i^2 = \min_{\beta} \varepsilon' \varepsilon \qquad (3)$$

qui s'écrit sous la forme quadratique suivante:

$$S(\beta) = (Y - X\beta)'(Y - X\beta) = Y'Y - \beta'X'Y - Y'X\beta + \beta'X'X\beta.$$

La dérivée matricielle de la forme quadratique $S(\beta)$ par rapport au vecteur β conduit à la condition suivante:

$$-2X'Y + 2X'X\beta = 0 \Longleftrightarrow \widehat{\beta} = (X'X)^{-1}X'Y$$

La dérivée de second ordre s'écrit:

$$\frac{\partial^2 S}{\partial \beta \partial \beta'} = 2(X'X).$$

Cette dernière matrice étant définie positive, on vérifie que la condition suffisante est respectéé et donc il s'agit bien d'un minimum.

Estimateurs des MCO

Estimation de σ^2: Par analogie avec la régression simple, les résidus après estimation sont

$$e = Y - X\widehat{\beta}$$

et la somme des résidus carrés est

$$SCR = e'e$$

On en déduit:

$$\widehat{\sigma}^2 = \frac{SCR}{n-(k+1)} = \frac{e'e}{n-(k+1)}$$

avec $k+1$ est le nombre de paramètres à estimer dans le modèle.
Le nombre $n-k-1$ est appelé le degré de liberté.

Propriétés des estimateurs

Théorème de Gauss-Markov

Sous les hypothèse H1, H2, H3 et H4, l'estimateur des MCO est BLUE (Best Linear Unbiaised Estimator), ie; meilleur estimateur linéaire sans biais.

Démonstration

- Linéarité:

$$\begin{aligned} \widehat{\beta} &= (X'X)^{-1}X'Y = AY = A(X\beta + \varepsilon) \\ &= \beta + (X'X)^{-1}X'\varepsilon = \beta + A\varepsilon \end{aligned}$$

- $E(\widehat{\beta}) = \beta$:

$$E\left(\widehat{\beta}\right) = \beta + (X'X)^{-1}X'E(\varepsilon) = \beta$$

car $E(\varepsilon) = 0$.

Demonstration

- $V(\widehat{\beta}) = \sigma^2 (X'X)^{-1}$

$$V(\widehat{\beta}) = E\left[(\widehat{\beta} - \beta)(\widehat{\beta} - \beta)'\right]$$

Or, $(\widehat{\beta} - \beta) = (X'X)^{-1} X'\varepsilon$ et $(\widehat{\beta} - \beta)' = \varepsilon' X (X'X)^{-1}$ puisque $(X'X)^{-1}$ est symétrique.
D'où $(\widehat{\beta} - \beta)(\widehat{\beta} - \beta)' = (X'X)^{-1} X'\varepsilon'\varepsilon X (X'X)^{-1}$, donc
$V(\widehat{\beta}) = (X'X)^{-1} X' E(\varepsilon'\varepsilon) X (X'X)^{-1}$ avec $E(\varepsilon'\varepsilon) = \sigma^2 I$, donc

$$\boxed{V(\widehat{\beta}) = \sigma^2 (X'X)^{-1}}$$

Estimation de σ^2

σ^2 est généralement inconnu, donc il faut l'estimer. On peut démontrer qu'un estimateur sans biais de σ^2 est donné par

$$\widehat{\sigma}^2 = \frac{e'e}{n-k-1} = \frac{SCR}{n-k-1}$$

avec $n-k-1$ est le degré de liberté qui est égale à nombre d'observations - nombre de paramètres à estimer.

Estimation de σ^2

σ^2 est généralement inconnu, donc il faut l'estimer. On peut démontrer qu'un estimateur sans biais de σ^2 est donné par

$$\widehat{\sigma}^2 = \frac{e'e}{n-k-1} = \frac{SCR}{n-k-1}$$

avec $n-k-1$ est le degré de liberté qui est égale à nombre d'observations - nombre de paramètres à estimer.

On déduit, donc, un estimateur sans biais de $V(\widehat{\beta})$.

$$\widehat{V}(\widehat{\beta}) = \widehat{\sigma}^2 (X'X)^{-1}$$

Décomposition de la variance

L'équation d'analyse de la variance

$$\underbrace{(Y-\bar{y})'(Y-\bar{y})}_{SCT} = \underbrace{(\hat{Y}-\bar{y})'(\hat{Y}-\bar{y})}_{SCE} + \underbrace{e'e}_{SCR} \tag{4}$$

On peut vérifier que:

$$SCE = \hat{\beta}'X'Y - T\overline{Y}^2$$

$$SCT = Y'Y - T\overline{Y}^2$$

On en déduit que:

$$SCR = Y'Y - \hat{\beta}'X'Y$$

Qualité d'ajustement

Le coefficient de détermination R^2 est définit par:

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

Remarques

- La SCR diminue toujours avec l'ajout de variables supplémentaires même si celle-ci n'a aucun pouvoir explicative.
- Comme SCR diminue, R^2 augmente, donc on peut faire tendre R^2 vers 1 en ajoutant de maximum de variables.
- R^2 doit être interpreter avec prudence et dans tous les cas on lui préfère R^2 ajusté,

$$\overline{R^2} = 1 - \frac{n-1}{n-k-1}(1 - R^2)$$

Exemple 1

Les données présentées dans le tableau ci-dessous concernent 9 entreprises de l'industrie chimique. Nous cherchons à établir une relation entre la production y_i, les heures de travail l_i et le capital utilisé k_i.

$$y = \begin{pmatrix} 60 \\ 120 \\ 190 \\ 250 \\ 300 \\ 360 \\ 380 \\ 430 \\ 440 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 1100 & 300 \\ 1 & 1200 & 400 \\ 1 & 1430 & 420 \\ 1 & 1500 & 400 \\ 1 & 1520 & 510 \\ 1 & 1620 & 590 \\ 1 & 1800 & 600 \\ 1 & 1820 & 630 \\ 1 & 1800 & 610 \end{pmatrix}$$

1) Estimer par MCO les paramètres de l'équation
$y_i = \beta_0 + \beta_1 l_i + \beta_2 k_i + \varepsilon_i$.
2) Donner une estimation de la matrice des variances-covariances des paramètres estimés.

Exemple 1

1) Il s'agit de calculer le vecteur $\widehat{\beta} = (X'X)^{-1}X'Y$.

$$X'X = \begin{pmatrix} n & \sum l_i & \sum k_i \\ \sum l_i & \sum l_i^2 & \sum l_i k_i \\ \sum k_i & \sum k_i l_i & \sum k_i^2 \end{pmatrix} = \begin{pmatrix} 9 & 13790 & 4460 \\ 13790 & 21672100 & 7066200 \\ 4460 & 7066200 & 2323600 \end{pmatrix}$$

$$(X'X)^{-1} = \begin{pmatrix} 6.304 & -0.007 & 0.011 \\ & 0.000015 & -0.000031 \\ & & 0.000072 \end{pmatrix} \quad \text{et} \quad X'Y = \begin{pmatrix} \sum y_i \\ \sum l_i y_i \\ \sum k_i y_i \end{pmatrix}$$

$$\widehat{\beta} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \begin{pmatrix} -437.71 \\ 0.336 \\ 0.41 \end{pmatrix}$$

Donc, le modèle estimé est donné par:

$$\widehat{y}_i = -437.71 + 0.336 l_i + 0.41 k_i$$

Exemple 1

2) Afin de donner une estimation de la matrice des variances-covariance, on commence par estimer la variance des erreurs

$$\widehat{\sigma}^2 = \frac{SCR}{n-3} = \frac{Y'Y - \widehat{\beta}X'Y}{6} = \frac{3194}{6} = 532$$

On peut alors calculer

$$\begin{aligned}
\widehat{V}(\widehat{\beta}) &= \widehat{\sigma}^2 (X'X)^{-1} = 532 \begin{bmatrix} 6.30477 & -0.007800 & 0.011620 \\ -0.007800 & 0.000015 & -0.000031 \\ 0.011620 & -0.000031 & 0.000072 \end{bmatrix} \\
&= \begin{bmatrix} 3355.997857 & -4.152117707 & 6.18520559 \\ -4.152118 & 0.008039731 & -0.01647956 \\ 6.185206 & -0.016479558 & 0.03847225 \end{bmatrix}
\end{aligned}$$

Exemple 1 (Avec R)

```
y=c(60, 120, 190, 250, 300, 360, 380, 430, 440)
l=c(1100, 1200, 1430, 1500, 1520, 1620, 1800, 1820, 1800)
k=c(300, 400, 420, 400, 510, 590, 600, 630, 610)
reg=lm(y~l+k)
ss=summary(reg)
ss$coefficients

                Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) -437.7136328 57.93097494 -7.555779 0.0002789705
l              0.3365303  0.08966455  3.753215 0.0094738487
k              0.4100157  0.19614345  2.090387 0.0815546320

vcov(reg)

            (Intercept)            l           k
(Intercept) 3355.997857 -4.152117707  6.18520559
l             -4.152118  0.008039731 -0.01647956
k              6.185206 -0.016479558  0.03847225
```

Modèle centré

Pour simplifier les calculs, on peut estimer un modèle où on élimine la constante qu'on appelle **modèle centré**.

$$\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i \\
\overline{y} &= \beta_0 + \beta_1 \overline{x}_1 + \ldots + \beta_k \overline{x}_k + \overline{\varepsilon} \\
y_i - \overline{y} &= \beta_1 (x_{1i} - \overline{x}_1) + \ldots + \beta_k (x_{ki} - \overline{x}_k + (\varepsilon_i - \overline{\varepsilon})) \\
y_i^c &= \beta_1 x_{1i}^c + \ldots + \beta_k x_{ki}^c + \varepsilon_i^c
\end{aligned}$$

Soit, sous la forme matricielle $Y^c = X^c \beta + \varepsilon^c$.
Le terme d'erreur ε^c vérifie les mêmes hypothèses que ε. En applicant les MCO sur ce modèle, on obtient:

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

Le paramètre $\widehat{\beta}_0$ peut être déduit comme suit:

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}_1 - \ldots - \widehat{\beta}_k \overline{x}_k$$

Distribution des estimateurs

Comme dans le cas du modèle linéaire simple, on introduit une hypothèse supplémentaire
H5: Les aléas sont distribué identiquement et indépendamment selon une loi normale

$$\varepsilon \overset{iid}{\sim} N(0, \sigma^2 I)$$

puisque la variable endogène Y est une fonction linéaire de ε, donc est aussi normale

$$Y \sim N(X\beta, \sigma^2 I)$$

Les paramètres estimés sont linéaires par rapport à Y. On en déduit

$$\boxed{\widehat{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1})}$$

Ce résultat implique que pour un paramètre β_i, $i = 0, 1, \ldots, k$

$$\frac{\widehat{\beta}_i - \beta_i}{\sqrt{V(\widehat{\beta}_i)}} \sim N(0, 1)$$

Or, la variance $V(\widehat{\beta}_i)$ n'est pas connue et elle est esitmée,

Distribution des estimateurs

donc

$$\frac{\widehat{\beta}_i - \beta_i}{\sqrt{\widehat{V}(\widehat{\beta}_i)}} \sim St(n-k-1)$$

où n est la taille de l'échantillon et k désigne le nombre de **variables** utilisées dans le modèle.
L'estimateur de la variance des erreurs a pour distribution

$$(n-k-1)\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-k-1)$$

Test de significativité individuelle

Il s'agit de tester si les paramètres estimés du modèle sont statistiquement significatifs, ie;

$$H_0 : \ \beta_i = 0 \quad i = 0, 1, \ldots, k$$

$$H_1 : \ \beta_i \neq 0$$

On a $\hat{t}_i = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_{\hat{\beta}_i}} \overset{iid}{\sim} St(n-k-1)$, donc

Si $|\hat{t}_i| > t_{(1-\frac{\alpha}{2}, n-k-1)}$, on décide de rejeter l'hypothèse nulle. Dans ce cas, on conclut que le paramètre β_i est statistiquement significatif au seuil α %.

Test de significativité globale

Il s'agit de tester si toutes les variables explicatives n'ont pas d'effet sur l'ajustement

$$\begin{aligned} H_0 : &\quad \beta_1 = \beta_2 = \ldots = \beta_k = 0 \\ H_1 : &\quad \exists \text{ au moin un } \beta_i \neq 0 \, (i = 1, \ldots, k) \end{aligned}$$

Pour effectuer ce test, on utilise la statistique de Ficher

$$F = \frac{n-k-1}{k} \frac{R^2}{1-R^2} = \frac{n-k-1}{k} \frac{SCR_c - SCR}{SCR} \sim F(k, n-k-1)$$

où SCR_c est la somme des carrées résiduelle sous H_0.
On rejette H_0 si $\widehat{F} > F_\alpha(k, n-k-1) \Longrightarrow$ le modèle est globalement linéaire, c-à-d, il existe au moins une variable explicative ayant un pouvoir explicatif significatif.

Combinaison linéaire

Il s'agit de tester

$$
\begin{aligned}
H_0 : &\qquad r\beta = q \\
H_1 : &\qquad r\beta \neq q
\end{aligned}
$$

où r est un vecteur ligne de longueur k et q est un scalaire. Sous H_0, on a

$$t = \frac{r\beta - q}{\sqrt{r\widehat{V}(\widehat{\beta})r'}} \sim St(n-k-1)$$

Combinaison linéaire

Il s'agit de tester

$$\begin{aligned} H_0 : &\quad r\beta = q \\ H_1 : &\quad r\beta \neq q \end{aligned}$$

où r est un vecteur ligne de longueur k et q est un scalaire. Sous H_0, on a

$$t = \frac{r\beta - q}{\sqrt{r\widehat{V}(\widehat{\beta})r'}} \sim St(n-k-1)$$

Exemple

Reprenons l'exemple 1 et testons si les rendements d'échelles sont constants au seuil $\alpha = 5\%$.

Combinaison linéaire

Il s'agit de tester

$$\begin{aligned} H_0 : &\qquad r\beta = q \\ H_1 : &\qquad r\beta \neq q \end{aligned}$$

où r est un vecteur ligne de longueur k et q est un scalaire. Sous H_0, on a

$$t = \frac{r\beta - q}{\sqrt{r\widehat{V}(\widehat{\beta})r'}} \sim St(n-k-1)$$

Exemple

Reprenons l'exemple 1 et testons si les rendements d'échelles sont constants au seuil $\alpha = 5\%$.

Il s'agit de test $H_0 : \beta_1 + \beta_2 = 1$, donc $r = (1, 1)$ et $q = 1$.

$$r\widehat{V}(\widehat{\beta})r' = (1 \;\; 1)\begin{pmatrix} 0.008039731 & -0.01647956 \\ -0.01647956 & 0.03847225 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = 0.01355287.$$

Combinaison linéaire

D'où

$$t = \frac{\beta_1 + \beta_2 - 1}{\sqrt{r\widehat{V}(\widehat{\beta})r'}} = \frac{0.3365 + 0.4100 - 1}{0.1164} = -2.1778$$

On a $|\widehat{t}| = 2.1778 < t_{1-\frac{\alpha}{2},6} = 2.4469$, donc on rette H_0, c-à-d, au seuil $\alpha = 5\%$, les rendements d'échelles ne sont pas constants.

Combinaison linéaire

D'où

$$t = \frac{\beta_1 + \beta_2 - 1}{\sqrt{r\widehat{V}(\widehat{\beta})r'}} = \frac{0.3365 + 0.4100 - 1}{0.1164} = -2.1778$$

On a $|\widehat{t}| = 2.1778 < t_{1-\frac{\alpha}{2},6} = 2.4469$, donc on rette H_0, c-à-d, au seuil $\alpha = 5\%$, les rendements d'échelles ne sont pas constants.

Remarque:

La quantité $r\widehat{V}(\widehat{\beta})r'$ n'est autre que la $\widehat{V}(\widehat{\beta}_1 + \widehat{\beta}_2)$. Or

$$\widehat{V}(\widehat{\beta}_1 + \widehat{\beta}_2) = \widehat{V}(\widehat{\beta}_1) + \widehat{V}(\widehat{\beta}_2) + 2Cov(\widehat{\beta}_2, \widehat{\beta}_2)$$

D'une manière plus générale

$$V(aX + bY) = a^2V(X) + b^2V(Y) + 2abCov(X, Y)$$

Test sur q restrictions

Les tests portant sur q restrictions demandent généralement un peu plus d'effort et utilisent le test de Ficher au lieu du test t.
L'hypothèse nulle dans le cas à q restrictions linéaires s'écrit comme suit:

$$H_0 : R\beta = r$$

avec R est une matrice de dimension $(q \times (k+1))$ et r est un vecteur colonne de longueur $(k+1)$. La statistique de Ficher est donnée par:

$$\boxed{F = \frac{(R\beta - r)'\left(R(X'X)^{-1}R'\right)^{-1}(R\beta - r)}{q\widehat{\sigma}^2} = \frac{SCR_c - SCR}{q\widehat{\sigma}^2} \sim F(q, n-k-1)}$$

Si $\widehat{F} \leq F_\alpha$, on accepte l'hypothèse nulle selon laquelle toutes les restrictions contenues dans le modèle sont valides.

Test sur q restrictions

On peut démontrer que la statistique F est égale à

$$F = \frac{n-k-1}{q}\frac{R^2 - R_c^2}{1-R^2}$$

En effet

$$\begin{array}{lcl}
F & = & \dfrac{n-k-1}{q}\dfrac{SCR_c - SCR}{SCR} \\
\text{or} & & R^2 = 1 - \dfrac{SCR}{SCT} \Longrightarrow (1-R^2)SCT = SCR \\
\text{et} & & R_c^2 = 1 - \dfrac{SCR_c}{SCT} \Longrightarrow (1-R_c^2)SCT = SCR_c \\
F & = & \dfrac{n-k-1}{q}\dfrac{SCT(1-R_c^2) - SCT(1-R^2)}{SCT(1-R^2)} \\
 & = & \dfrac{n-k-1}{q}\dfrac{R^2 - R_c^2}{1-R^2}
\end{array}$$

Estimation sous contraintes

On veut estimer un modèle linéaire sous contraintes linéaires. Il s'agit de minimiser $S(\beta)$ sous la contrainte $R\beta = r$.
L'estimateur des MCO sous contrainte, noté $\widehat{\beta}_c$, vaut

$$\widehat{\beta}_c = \widehat{\beta} + (X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}\left(r - R\widehat{\beta}\right)$$

Estimation sous contraintes

On veut estimer un modèle linéaire sous contraintes linéaires. Il s'agit de minimiser $S(\beta)$ sous la contrainte $R\beta = r$.
L'estimateur des MCO sous contrainte, noté $\widehat{\beta}_c$, vaut

$$\widehat{\beta}_c = \widehat{\beta} + (X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}\left(r - R\widehat{\beta}\right)$$

La somme des carrés des résidus du modèle contraint est:

$$SCR_c = SCR + \left(\widehat{\beta}_c - \widehat{\beta}\right)'(X'X)\left(\widehat{\beta}_c - \widehat{\beta}\right)$$

La démonstration de ces deux propriétés est laissée comme exercice.

Exemple

Reprenons l'exemple 1. Par deux méthodes, tester simultanément les hypothèses $\beta_1 = 1$ et $\beta_2 = 0$ au seuil $\alpha = 5\%$.

Exemple

Reprenons l'exemple 1. Par deux méthodes, tester simultanément les hypothèses $\beta_1 = 1$ et $\beta_2 = 0$ au seuil $\alpha = 5\%$.

Méthode 1:

Il s'agit de tester $\beta_1 = 1$ et $\beta_2 = 0$, c-à-d

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \Longrightarrow R\beta = r$$

Donc, on a $q = 2$ restrictions. La statistique F sous H_0 est:

$$\begin{aligned} F &= \frac{(R\widehat{\beta} - r)'\left(R(X'X)^{-1}R'\right)^{-1}(R\widehat{\beta} - r)}{q\widehat{\sigma}^2} \\ &= \frac{(R\widehat{\beta} - r)'\left(R\widehat{V}(\widehat{\beta})R'\right)^{-1}(R\widehat{\beta} - r)}{q} \end{aligned}$$

avec $R\widehat{\beta} - r = \begin{bmatrix} -0.6634 \\ 0.4100 \end{bmatrix}$

Exemple

et $\left(R\widehat{V}(\widehat{\beta})R'\right)^{-1} = 10^2 \begin{bmatrix} 0.804 & -1.648 \\ -1.648 & 3.847 \end{bmatrix}^{-1} = \begin{bmatrix} 1019.653 & 436.767 \\ 436.767 & 213.0817 \end{bmatrix}$

Donc

$$\begin{aligned} F &= \frac{1}{2}(R\widehat{\beta} - r)'\left(R\widehat{V}(\widehat{\beta})R'\right)^{-1}(R\widehat{\beta} - r) \\ &= \frac{1}{2}[-0.6634 \;\; 0.4100] \begin{bmatrix} 1019.653 & 436.767 \\ 436.767 & 213.0817 \end{bmatrix} \begin{bmatrix} -0.6634 \\ 0.4100 \end{bmatrix} \\ &= \frac{247.0346}{2} = 123.5173 \end{aligned}$$

et comme $\widehat{F} > F_{0.05}(2,6) = 5.143$, on doit rejeter l'hypothèse nulle en faveur de l'hypothèse altenative.

Exemple

Méthode 2:

On utilise les variations résiduelles des modèles contraint et non contraint, pour calculer la statistique F comme suit:

$$F = \frac{6}{2}\frac{SCR_c - SCR}{SCR}$$

SCR_c est la somme des carrés des résidus sous H_0. Le modèle contraint est

$$y_i = \beta_0 + l_i + \varepsilon_i \Longrightarrow y_i - l_i = \beta_0 + \varepsilon_i$$

On vérifie bien que $\widehat{\beta}_0 = \overline{y} - \overline{l}$, soit $\widehat{\beta_0} = -1251.11$. Par suite, la valeur de SCR_c devient:

$$\begin{aligned} SCR_c &= \sum \left(y_i - l_i - \widehat{\beta}_0\right)^2 = \sum \left((y_i - \overline{y}) - (l_i - \overline{l})\right)^2 \\ &= S_{yy} + S_{ll} - 2S_{yl} = 147888.9 + 542755.6 - 2 \times 277977.8 \\ &= 134688.9 \end{aligned}$$

On peut vérifier que $SCR = 3193.767$

Exemple

D'où, on tire la statistique estimée $\widehat{F}$

$$\widehat{F} = \frac{6}{2}\frac{134688.9 - 3193.767}{3193.767} = 123.5173$$

D'une manière identique, on remarque que $\widehat{F} > \mathrm{F}_{\alpha}(2,6)$. On constate qu'on obtient les mêmes conclusions, puisque les deux procédures sont, en fait, équivalentes.

Test de Chow

Ce type de test permet de tester s'il y a un **changement de régime** d'une période à une autre ou s'il ya **une différence entre deux groupes**. Pour ce faire, on divise les observations en deux sous périodes.

$$H_0 : \quad y_t = \beta_0 + \beta_1 x_{1t} + \ldots + \beta_k x_{kt} + \varepsilon_t,\ t = 1, \ldots, T$$

$$H_1 : \quad \begin{cases} y_t = \beta_0^1 + \beta_1^1 x_{1t} + \ldots + \beta_k^1 x_{kt} + \varepsilon_{1t} & \text{si } t \leq t_0 \\ y_t = \beta_0^2 + \beta_1^2 x_{1t} + \ldots + \beta_k^2 x_{kt} + \varepsilon_{2t} & \text{si } t > t_0 \end{cases}$$

On estime le modèle par les MCO d'une part sur l'ensemble de la période (modèle contraint) et d'une part sur les deux sous-périodes $1, \ldots, t_0$ et $t_0 + 1, \ldots, T$. Puis on déduit les sommes carrés résiduelles de chaque modèle, afin de calculer la statistique F comme suit:

$$F = \frac{T - 2(k+1)}{k+1} \frac{SCR_c - SCR}{SCR} \overset{H_0}{\sim} F(k+1, T - 2(k+1))$$

avec SCR_c est la somme carrés des résidus du modèle contraint et $SCR = SCR_1 + SCR_2$.

Exemple

Nous souhaitons expliquer le salaire, Y des individus à partir de leur niveau d'études, X. Une qualification d'autant plus élevée devrait induire une rémunération plus élevée. Après cette première étape, nous souhaitons savoir si la relation est la même chez les hommes et chez les femmes.

Nous disposons de $n = 40$ observations, dont $n_1 = 20$ hommes et $n_2 = 20$ femmes. Tester, au seuil $\alpha = 5\%$, si la relation est la même chez les hommes et chez les femmes sachant que:

$$\begin{aligned}
y &= -902.231 + 267.024x, \quad SCR_c = 60775962.6 \\
y_H &= -413.655 + 261.071x_H, \quad SCR_H = 36995693.5 \\
y_F &= -230.105 + 178.472x_F \quad SCR_F = 16223705.4
\end{aligned}$$

Exemple

Nous souhaitons expliquer le salaire, Y des individus à partir de leur niveau d'études, X. Une qualification d'autant plus élevée devrait induire une rémunération plus élevée. Après cette première étape, nous souhaitons savoir si la relation est la même chez les hommes et chez les femmes.

Nous disposons de $n = 40$ observations, dont $n_1 = 20$ hommes et $n_2 = 20$ femmes. Tester, au seuil $\alpha = 5\%$, si la relation est la même chez les hommes et chez les femmes sachant que:

$$\begin{aligned}
y &= -902.231 + 267.024x, \quad SCR_c = 60775962.6 \\
y_H &= -413.655 + 261.071x_H, \quad SCR_H = 36995693.5 \\
y_F &= -230.105 + 178.472x_F \quad SCR_F = 16223705.4
\end{aligned}$$

Calculons la statistique de Ficher et la comparée avec celle tabulée.

$$\widehat{F} = \frac{40 - 2 \times 2}{2} \frac{60775962.6 - 53219399.1}{53219399.1} = 2.5558$$

On remarque que $\widehat{F} < F_{0.05}(2, 36) = 3.2594$, donc on accepte l'hypothèse nulle. Au seuil $\alpha = 5\%$, la liaison entre les années d'études et le salaire **est la même** pour les hommes et les femmes.

Prévision

Soit $X'_{T+1} = (1, X_{1,T+1}, X_{2,T+1}, \ldots, X_{k,T+1})$ une nouvelle valeur et nous voulons prédire Y_{T+1}.
La valeur Y_{T+1} est estimée par: $\widehat{Y}_{T+1} = X'_{T+1}\widehat{\beta}$.
L'espérance et la variance de l'erreur de prévision $\varepsilon^p_{T+1} = Y_{T+1} - \widehat{Y}_{T+1}$ valent:

$$\begin{aligned}
E\left(\varepsilon^p_{T+1}\right) &= 0 \\
V\left(\varepsilon^p_{T+1}\right) &= V\left(X'_{T+1}(\widehat{\beta} - \beta) - \varepsilon_{T+1}\right) \\
&= X'_{T+1} V\left(\widehat{\beta} - \beta\right) + \sigma^2 \\
&= \sigma^2 \left[X'_{T+1}(X'X)^{-1} X_{T+1} + 1\right]
\end{aligned}$$

Un intervalle de confiance de niveau $(1-\alpha)$ pour Y_{T+1} est donné par

$$\left[X'_{T+1}\widehat{\beta} \pm t_{1-\alpha/2}(T+k-1)\widehat{\sigma}\sqrt{X'_{T+1}(X'X)^{-1}X_{T+1} + 1}\right]$$