FOAD
COURS D' ECONOMETRIE 1
CHAPITRE 1 : Rappels de regression multiple
version du 23 mars 2013

Christine Maurel
Maître de conférences en Sciences Economiques
Université de Toulouse 1 - Capitole
Toulouse School of Economics-ARQADE

Table des matières

I Introduction

Ce premier chapitre reprend les notions essentielles d'un cours de niveau Licence troisième année, mention Econométrie. Nous reprenons les notions de base déjà étudié dans le cours "Modèle linéaire 1" du trimestre 2 de ce Master mais nous appliquons ces notions à des données économiques. Nous faisons ensuite des rappels sur les variables indicatrices [1] et nous terminons par la question de la spécification du modèle.

II Le modèle et ses hypothèses

Dans un modèle de regression multiple la variable endogène Y_i est expliquée par plusieurs variables explicatives. On note k, le nombre de variables du modèle ou encore le nombre de paramètres du modèle.

II.1 Le modèle

Le modèle de regression multiple s'écrit de la manière suivante :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_{k-1} X_{k-1i} + u_i \text{ avec } i = 1, \ldots, N \tag{1}$$

Notation et vocabulaire : β_0 est la constante
Dans ce modèle il y a k paramètres

Ecriture matricielle : $Y = X\beta + u$ avec :

Y vecteur de la var. endogène de dimension (N,1)
X matrice des var. exogènes de dimension (N,k)

β est le vecteur qui contient la liste des paramètres de dimension (k,1)
u est le vecteur des erreurs de dimension (N,1)

1. ou variable muette

II.2 Les hypothèses

1. sur les erreurs u_i :
   $H_1 : E(u/X) = 0$ [2] où u est un vecteur (N,1) et 0 est aussi un vecteur (N,1) ou encore $H_1 : E(u_i) = 0 \ \forall i = 1, \ldots, N$. C'est une hypothèse forte car elle contient en particulier $E(u_i) = 0$ mais aussi $E(X_i u_i) = 0$ [3]
   $H_2 : Var(u_i/X) = \sigma^2 \ \forall i = 1, \ldots, N$
   $H_3 : Cov(u_i, u_j/X) = 0 \ , \forall i \neq j$

   En utilisant l'écriture matricielle, les deux dernières hypothèses, H_2 et H_3, deviennent $Var(u/X) = \sigma^2 I$ où $Var(u/X)$ est une matrice (N,N) ainsi que I la matrice identité.
   $Var(u/X)$ est nommée matrice de Variance-Covariance de u : sur sa diagonale figurent les variances de chaque terme d'erreur et hors diagonale se trouvent les covariances
   La dernière hypothèse est :
   H_4 : Les erreurs u_i suivent une loi normale. Cette hypothèse sert à faire des tests (voir la section "tests").
2. sur les var. exogènes

On suppose qu'il n'existe aucune combinaison linéaire entre les variables exogènes, c'est à dire que X est de rang plein en colonne que nous écrirons dans la suite
$H_5 : Rang(X) = k$

III Estimation et propriétés des estimateurs

III.1 Estimation

Pour estimer les paramètres du modèle on utilise le critère des Moindres Carrés Ordinaires (MCO) c'est à dire que l'on minimise $\sum u_i^2$
On montre que

Si $H_5 : Rang(X) = k$ est vérifiée alors $X'X$ est une matrice qui peut être inversée. Si $X'X$ est une matrice qui peut être inversée, alors on peut calculer une estimation des paramètres de la manière suivante :

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{2}$$

EXEMPLE : Estimation d'une fonction de production Cobb-Douglas.
RAPPEL : On appelle fonction de production la relation entre la quantité de travail utilisée (le facteur travail) et le nombre de machines utilisée par exemple (facteur capital) avec la quantité produite. Cette relation peut prendre plusieurs formes ;la

2. En effet $E(u/X) = 0$ implique que $E(u_i) = 0 et E(u_i x_i) = 0$ et donc $Cov(u_i, x_i) = 0$; lecture complémentaire chapitre 4 sur la page :http ://russell.vcharite.univ-mrs.fr/EIE/

3. Les erreurs ne sont pas corrélées avec les variables explicatives ; en particulier les variables explicatives peuvent être supposées exogènes ;voir chapitre 3 de ce cours.

plus simple est la fonction de production de type Cobb-Douglas que nous allons estimer. Une fonction de production décrit donc de manière mathématique la technologie utilisée par l'entreprise. Après avoir estimé cette fonction de production en appliquant les formules des MCO du cours avec SAS, nous verrons quelles sont les utilisations économiques que nous pouvons faire de ces résultats.
FICHIER DE DONNEES :cobb1 (ce fichier contient aussi les variables en log dont nous aurons besoin un peu plus loin)
Pour 27 entreprises de la branche "Industrie des métaux", le fichier SAS cobb contient les variables suivantes :
L_i la quantité de facteur travail utilisé par l'entreprise i
K_i la quantité de facteur capital utilisé par l'entreprise i
Y_i la valeur ajoutée de l'entreprise i
Source : "Econométrie", W. H. Greene, Ed Pearson Education, 2005, Cinquième édition
Le fichier de données contient des observations individuelles ; on parle parfois dans ce cas de coupe transversale (en anglais cross section).
Pour la définition des variables, l'auteur renvoie aux articles initiaux où ces variables ont été construites ; étant donné que ces articles sont assez anciens je n'ai pas plus d'information à fournir sur la construction de ces 3 variables ; je peux tout de même indiquer que la construction de la variable "Capital" est toujours assez délicate ; en pratique il faut tester plusieurs définitions c'est à dire plusieurs variables pour mesurer le capital ; pour le facteur "Travail", on peut aussi choisir le nombre d'heures travaillées ou le rapport entre la masse salariale totale et le salaire moyen. Il y a toujours en pratique plusieurs manières de calculer l'équivalent empirique d'un concept économique.
Revenons à notre exemple. On fait l'hypothèse que la fonction de production de ces 27 entreprises est de type Cobb-Douglas c'est à dire qu'elle s'écrit de la manière suivante :
$Y_i = AL_i^{\beta_1} K_i^{\beta_2}$
Remarque : dans les cours de microéconomie la fonction Cobb-Douglas est souvent écrite de la manière suivante : $Y_i = AL_i^{\beta_1} K_i^{1-\beta_1}$. Pour trouver cette formulation il faut imposer la contrainte $\beta_1 + \beta_2 = 1$ ou encore $\beta_2 = 1 - \beta_1$ à l' écriture plus générale de ce cours. Nous proposons donc ici de commencer par une fonction plus générale puis de tester l'hypothèse $\beta_1 + \beta_2 = 1$ un peu plus loin.
La fonction Cobb-Douglas n'est pas linéaire dans les paramètres et il est habituel de transformer les variables avec la fonction logarithme (par défaut toujours népérien dans ce cours) de la manière suivante :
$Ln(Y_i) = Ln(A) + \beta_1 Ln(L_i) + \beta_2 Ln(K_i) + u_i$
Nous en profitons pour ajouter le terme d'erreur mais on aurait pu aussi l'introduire dans la fonction Cobb-Douglas sous la forme e^{u_i}.
Posons $Ln(A) = \beta_0$ et nous obtenons :

$$Ln(Y_i) = \beta_0 + \beta_1 Ln(L_i) + \beta_2 Ln(K_i) + u_i \tag{3}$$

Nous avons renommé $Ln(A)$ en β_0 pour avoir les mêmes notations que dans le cours. Etant donné que ce modèle est linéaire dans les paramètres (grâce aux log), on peut

appliquer la formule (2) :
$\hat{\beta} = (X'X)^{-1}X'Y$

Vous pouvez remarquer que l'utilisation du log des variables n'est pas justifiée par un argument pratique, sur cet exemple, mais par un argument de théorie économique : si nous faisons l'hypothèse que la fonction de production de ces 27 entreprises est de type Cobb-Douglas alors en transformant cette fonction avec la fonction log nous obtenons un modèle linéaire dans les paramètres et nous pouvons appliquer les formules de calcul habituelles. Nous regardons ensuite si les données valident cette hypothèse sur la fonction de production Cobb-Douglas.
APPLICATION AVEC SAS [4] :
Remarque : sur mon ordinateur, j'ai crée une bibliothèque SAS, avec l'assistant de création (et non pas avec l'instruction libname) pour stocker mes fichiers de données que j'ai nommée "TPFOAD" et que j'active au démarrage ; ainsi les noms de mes fichiers seront toujours de la forme "tpfoad.nom" dans mes programmes.

Avant d'estimer la fonction Cobb-Douglas par MCO, nous allons tout d'abord procéder à une étude exploratoire très simple des données.
Commençons par calculer les moyennes des 3 variables du modèle : LNY, LNL et LNK. Le programme est :

```
data tpfoad.cobb1 ;set tpfoad.cobb ;
LNY=Log(Y) ;LNL=log(L) ;LNK=log(K) ;run ;
proc means data=tpfoad.cobb ;
var LNY LNL LNK ;run ;
```

Remarquer que j'évite d' utiliser le même nom quand je crée un nouveau fichier car il faut revenir en arrière si je me trompe.
Remarquer de plus que SAS ne fait pas de différence entre les majuscules et les minuscules mais que par convention on écrit les facteurs de production et la quantité produite en majuscules.

Le listing est le suivant :

```
 La procédure MEANS

Variable    Nb         Moyenne      Écart-type         Minimum         Maximum
-------------------------------------------------------------------------------
LNY         27       7.4436313       0.7611529       6.3849414       9.1951425
LNL         27       5.7636521       0.6562399       4.9199809       7.3555325
LNK         27       7.4459224       0.9684820       5.6347539       9.5460659
-------------------------------------------------------------------------------
```

4. L'application commence par des statistiques descriptives ;je n'ai pas indiqué le code R car il a été vu dans le cours de "logiciels statistiques".

Dans un premier temps, les résultats de la procédure “means" servent à identifier des erreurs de saisie : par exemple la valeur minimale de certaines variables ne doit pas être négative. Sur cet exemple , et avec des variables transformées , la procédure "means" n'a pas d'intérêt pratique particulier à mon avis.
Calculons ensuite la matrice de corrélation (de Pearson) avec le programme suivant :

```
proc corr data=tpfoad.cobb1 ;var LNY LNL LNK ;run ;
```

Les résultats sont les suivants :

```
  Coefficients de corrélation de Pearson, N = 27
  Prob > |r| under H0: Rho=0

            LNY           LNL           LNK

LNY        1.00000       0.94753       0.94312
                          <.0001        <.0001

LNL        0.94753       1.00000       0.89456
            <.0001                      <.0001

LNK        0.94312       0.89456       1.00000
            <.0001        <.0001
```

Commentaires :
Remarquons tout d'abord que toutes les corrélations sont significatives, c'est à dire significativement différentes de 0. En effet en dessous de chaque coefficient de corrélation SAS donne une probabilité de dépasser la statistique de test de l'hypothèse H_0 la corrélation est nulle entre les deux variables correspondantes [5]. Si cette probabilité est inférieure à 5% on dit que la corrélation est significative à 5%. Si cette probabilité est inférieure à 1%, elle est significative à 1%. En général par abus de langage il n'y a pas de vocabulaire particulier si cette probabilité est inférieure à 1% ; nous reviendrons plus longuement sur les tests dans un prochain paragraphe.
Après avoir établi la liste des corrélations significatives [6] nous pouvons commenter les valeurs des corrélations de la manière suivante :
Etant donné que les 3 variables n'ont pas le même statut en économétrie, c'est à dire qu'il y a une variable expliquée ou variable endogène, ici LNY la valeur ajoutée ou la production par abus de langage, et des variables explicatives ou exogènes, ici LNL et LNK, il y a deux types de commentaires : le premier type concerne la corrélation de LNY avec les deux variables explicatives LNL et LNK ;nous constatons que ces deux corrélations sont très proches de 1 et positives ; il existe donc une lien positif très important entre la valeur ajoutée en log d'une entreprise et la quantité

5. pour la statistique de test voir http ://www.math-info.univ-paris5.fr/smel/cours/ts/node15.html

6. nous commentons seulement les corrélations significatives.

en log des deux facteurs qu'elle utilise [7]. Le second type de commentaire concerne la corrélation entre les variables explicatives ici entre LNL et LNK ;nous constatons qu'il y a aussi un lien positif élevé (0.89456) entre les deux variables explicatives de notre modèle. Je reviendrai sur ce commentaire dans la conclusion de ce paragraphe "Estimation".
Terminons cette partie analyse exploratoire simple par des graphiques, toujours utiles. Utilisons la procédure Gplot , avec des graphiques beaucoup plus agréables que la proc plot :

```
proc gplot data=tpfoad.cobb1 ;plot LNY*LNL ;run ;
```

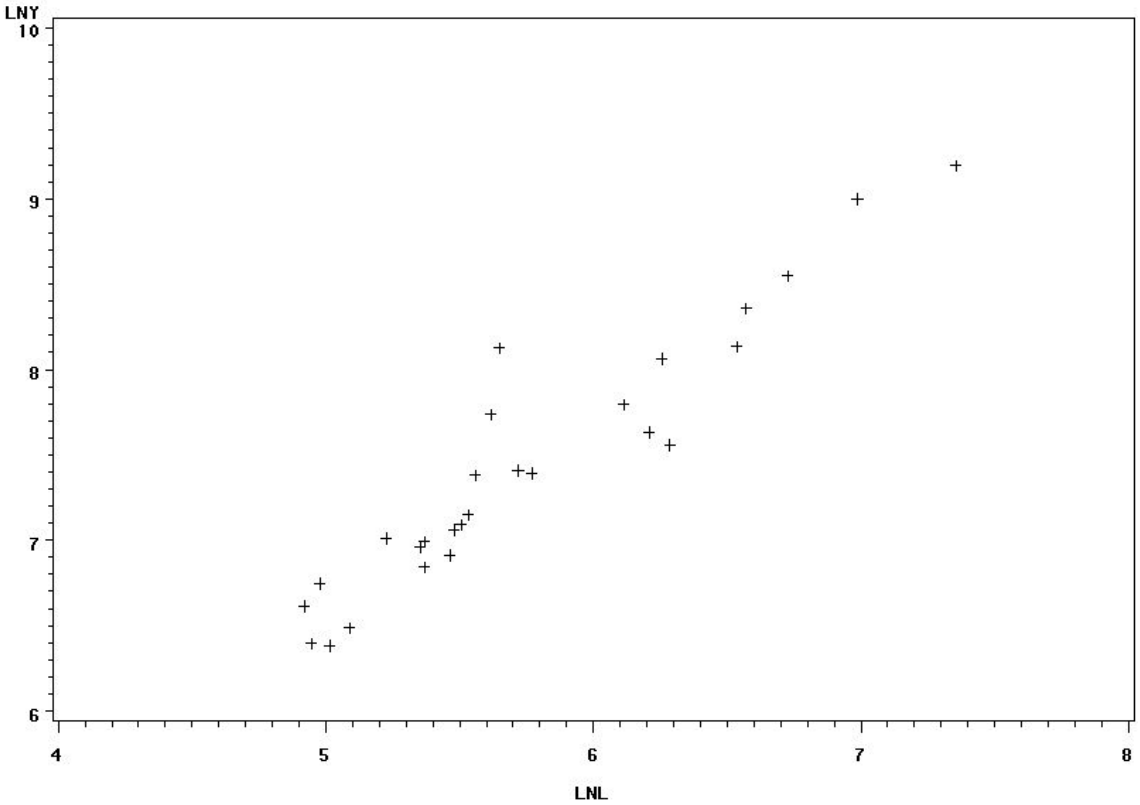

FIGURE 1 – LNY et LNL

Commentaire : La valeur élevée et positive de la corrélation entre ces deux variables est confirmée par le graphique **??** car les points sont effectivement répartis autour d'une droite [8].

7. je ne reviens pas sur le fait qu'une corrélation entre deux variables peut être élevée de manière trompeuse à cause d'une troisième variable qui a un effet sur les deux premières. Ce problème est corrigé quand on effectue une régression multiple

8. Le lecteur peut comparer l'allure de ce nuage de points avec celui où les variables ne sont pas transformées en log.

Nous obtenons le même commentaire sur le graphique **??** avec le second facteur de production dont le programme est :

```
proc gplot data=tpfoad.cobb1 ;plot LNY*LNK ;run ;
```

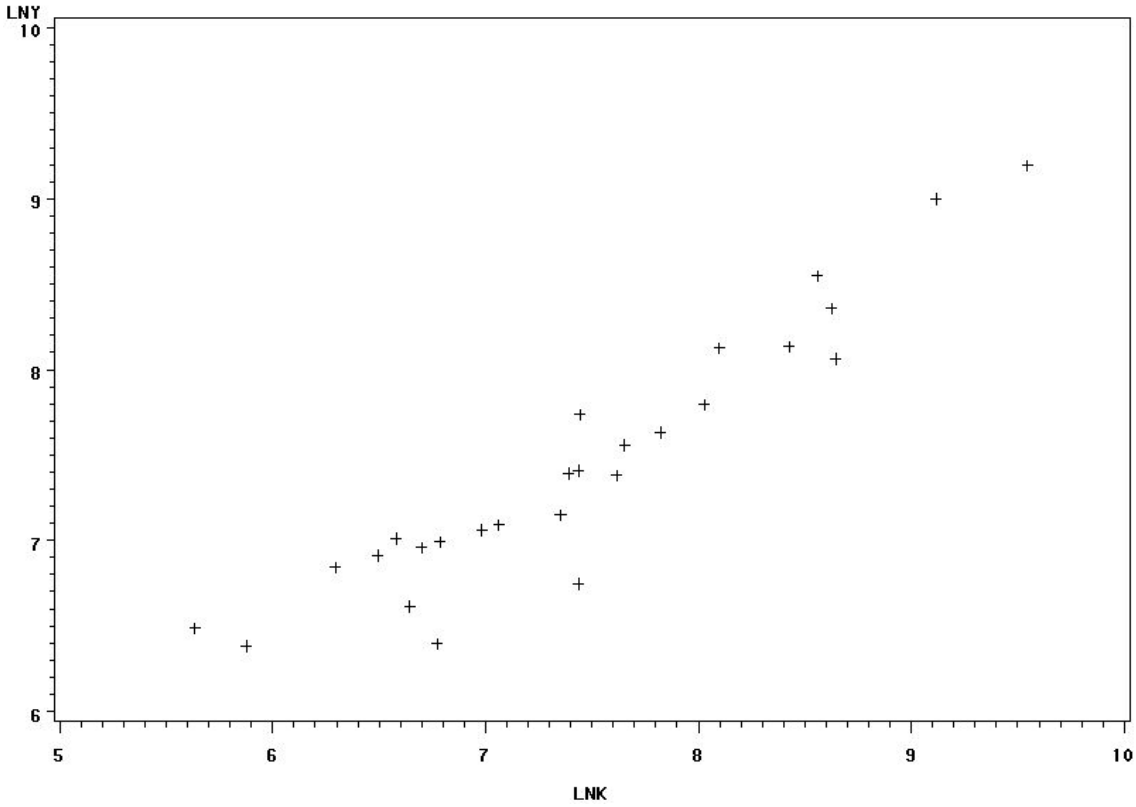

FIGURE 2 – LNY et LNK

Nous allons maintenant estimer par MCO cette fonction de production Cobb-Douglas ; il suffit d'utiliser la "proc reg" pour estimer le modèle par MCO. Le programme SAS est très simple :

```
proc reg data =tpfoad.cobb1 ;
model LNY=LNL LNK ;run ;
```

Le logiciel R ne fournit pas la même présentation des résultats mais on retrouve les mêmes informations (j'ajouterai des résultats un peu plus loin). Le programme est le suivant :

```
library(AER)
mod1 <- lm(LNY~LNL+LNK, data =cobb1)
summary(mod1)
```

Les résultats sont les suivants :

```
The REG Procedure
```

Model: MODEL1
Dependent Variable: LNY

Number of Observations Read	27
Number of Observations Used	27

Analyse de variance

Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	2	14.21156	7.10578	200.25	<.0001
Error	24	0.85163	0.03548		
Corrected Total	26	15.06320			

Root MSE	0.18837	R-Square	0.9435
Dependent Mean	7.44363	Adj R-Sq	0.9388
Coeff Var	2.53067		

Résultats estimés des paramètres

Variable	DF	Résultat estimé des paramètres	Erreur std	Valeur du test t	Pr > |t|
Intercept	1	1.17064	0.32678	3.58	0.0015
LNL	1	0.60300	0.12595	4.79	<.0001
LNK	1	0.37571	0.08535	4.40	0.0002

Il suffit de lire les valeurs estimés des paramètres de la manière suivante :
$\hat{\beta}_0 = 1.17$
$\hat{\beta}_1 = 0.60$
$\hat{\beta}_2 = 0.37$
La transformation par la fonction Log présentent deux intérêts :
Le premier est mathématique et a déjà été signalé : le modèle économétrique devient linéaire (dans les paramètres) et donc on peut l'estimer avec les formules précédentes des MCO. Le second intérêt est économique. Les paramètres estimés s'interprètent directement comme des élasticités , notion présentée dans tous les cours théoriques de microéconomie [9] et que nous allons brièvement rappeler ici.
Définition de l'élasticité de la production par rapport au travail :

$$\text{élasticité de Y par rapport à L } = \frac{\frac{dY}{Y}}{\frac{dL}{L}} = \frac{dLnY}{dLnL}$$

Les paramètres β_1 et β_2 s'interprètent donc directement comme des élasticités car les variables sont spécifiées en Log. Ainsi β_1 est l' élasticité de la production par rapport au facteur travail et β_2 s'interprète comme l' élasticité de la production par rapport au facteur capital.

9. et très utilisée par les économistes

INTERPRETATION EN PRATIQUE : l'élasticité de la production par rapport au travail représente la variation en pourcentage de la production dûe à une variation de 1% du facteur travail. Sur notre échantillon cette élasticité est égale à 0.6 ; on peut donc faire le commentaire suivant :
Quand la quantité de travail augmente de 1 %, alors la quantité produite augmente de 0,6 %
Nous obtenons une interprétation pratique similaire pour l'élasticité de la production par rapport au capital : Quand la quantité de capital augmente de 1 %, alors la quantité produite augmente de 0,37 %
AVANTAGE DES ELASTICITES :
Contrairement aux paramètres estimés, les élasticités ne dépendent pas des unités de mesure des variables (elles sont interprétées en pourcentage) , on peut donc les comparer. Sur cet échantillon la quantité produite (je devrais dire "la valeur ajoutée mais par abus de langage je dis plutôt "la quantité produite") est plus sensible à une augmentation du facteur travail qu'à une augmentation du facteur capital car l'élasticté du travail est supérieure à celle du capital pour cette branche à condition que l'on rejette l'hypothèse d'égalité des deux élasticités. Nous effectuerons ce test dans le paragraphe "Tests" (voir plus loin).
Avec les paramètres estimés nous pouvons calculer deux variables pour chaque individu :
- la valeur estimée de la variable endogène c'est à dire ici la valeur estimée de la valeur ajoutée que nous noterons dans un modèle théoique général $\hat{Y}_i$
- le résidu noté $\hat{u}_i$ qui est égal à $Y_i - \hat{Y}_i$
Remarquons qu'à partir de la définition du résidu nous avons $Y_i = \hat{Y}_i + \hat{u}_i$ qui représente une décomposition en deux parties de la variable endogène observée. Pour l'instant nous n'utiliserons pas ces deux variables $\hat{Y}_i$ et $\hat{u}_i$.

III.2 Propriétés des estimateurs

On démontre que les propiétés des estimateurs des MCO sont les suivantes :

1. On montre que la matrice de variance-covariance de $\hat{\beta}$, notée $Var\hat{\beta}$, est donnée par la formule suivante (voir Annexe) :

$$Var\hat{\beta} = \sigma^2(X'X)^{-1}$$

Etant donnée que la variance des erreurs, σ^2 est inconnue, nous calculons une estimation de cette variance avec la formule suivante :

$$\hat{\sigma^2} = \frac{\sum_{i=1}^{N} \hat{u}_i^2}{N-k}$$

Vocabulaire : on appelle Somme des Carrés des Résidus, notée SCR, la quantité $\sum_{i=1}^{N} \hat{u}_i^2$. Ainsi , $\hat{\sigma^2} = \frac{SCR}{N-k}$

$$\Rightarrow \hat{Var\beta} = \hat{\sigma}^2(X'X)^{-1} \quad (4)$$

APPLICATION AVEC SAS :
Pour obtenir la matrice de variance-covariance des estimateurs avec SAS, il faut exécuter le programme suivant :

```
proc reg data=tpfoad.cobb1 outest=var covout ;
model LNY=LNL LNK ;
run ;quit ;
```

CONSEIL : je ne détaille pas l'option "outest=" ni "covout" mais il faut que vous cherchiez dans la documentation de SAS sur la proc reg, les options que j'utilise si vous ne les connaissez pas.
Le fichier Work.var contient la matrice de variance-covariance des estimateurs mais aussi d'autres informations. Pour afficher seulement cette matrice, je vous propose le programme :

```
data varcov ;set var ;if _type_='COV' ;
keep _name_ Intercept LNL LNK ;run ;
```

puis ouvrir ce fichier varcov dans la librairie Work ou faire un proc print. Nous obtenons :

NAME	Intercept	LNL	LNK
Intercept	0.10679	-0.019835	0.001188850
LNL	-0.01984	0.015864	-.009616201
LNK	0.00119	-0.009616	0.007283931

Ainsi, et avec nos notations , $\hat{Var}\hat{\beta}_0 = 0.10679$, $\hat{Var}\hat{\beta}_1 = 0.015864$ et $\hat{Var}\hat{\beta}_2 = 0.007283931$. Hors diagonale se trouvent les covariances estimées.
Le programme R est le suivant :

```
covb <- vcov(mod1)
print(covb)
```

Nous utiliserons cette matrice dans le paragraphe Test.

2. Propriétés des estimateurs :

   (a) Propriétés en échantillon fini :
   On montre que les paramètres estimés, $\hat{\beta}$, sont les meilleurs estimateurs linéaires et sans biais (MELSB ou en anglais Best Linear Unbiaised Estimator , BLUE)) si toutes les hypothèses $H1$, $H2$, $H3$ et H_5 sont vérifiées (Théorème de Gauss-Markov). [10]
   Rappel :
   - $\hat{\beta}$ est sans biais si $E(\hat{\beta}) = \beta$ (voir Annexe)
   - "meilleur" signifie que la variance de $\hat{\beta}$ est minimale ou encore que la précision de $\hat{\beta}$ est maximale.

10. Remarquons que l'hypothèse de normalité de u_i n'est pas nécessaire ici mais elle est utilisée pour l'inférence "exacte".

(b) Propriétés asymptotiques :
Les propriétés précédentes sont dites “propriétés exactes" ou “‘propriétés en échantillon fini" ;elles sont valides pour toute taille d'échantillon. Les propriétés asymptotiques ne sont pas définies pour toute taille d'échantillon mais seulement pour un taille d'échantillon qui tend vers l'infini. Un résultat important en pratique est que les statistiques de Student et de Fisher [11] sont approximativement [12] distribués selon une loi de Student ou de Fisher même si les erreurs ne sont pas des variables aléatoires normales. Avant de présenter les propriétés asymptotiques des MCO, nous donnons queslques définitions et théorèmes.

– Définitions et théorèmes :
Définition 1 : Convergence en probabilité [13] :
La suite de variables aléatoires X_N converge en probabilité vers une constante a si pour tout ϵ, $P(|X_N - a| > \epsilon) \to 0$ quand $N \to \infty$. On écrit

$$X_N \xrightarrow{P} a$$

ce qui se lit “ X_N tend en probabilité vers a.
On écrit aussi

$$Plim(X_N) = a$$

on prononce “plim".

Loi (faible) des Grands Nombres : La moyenne empirique converge vers l'Espérance :
Si X_N est une suite de variables aléatoires indépendantes admettant les mêmes moments d'ordre 1 et 2 c'est à dire $E(X_N) = m$ et $Var(X_N) = \sigma^2$, alors quand $N \to \infty$,

$$\bar{X}_N \xrightarrow{P} m$$

.

Théorème Central Limite (TCL) : si X_N est une suite de variables aléatoires indépendantes et de même loi admettant des moments d'ordre 1 et 2 noté $m = E(X_N)$ et $\sigma^2 = Var(X_N)$, alors

$$\sqrt{N}(\frac{\bar{X}_N - m}{\sigma}) \xrightarrow{\mathcal{L}} N(0,1)$$

– Les propriétés asymptotiques des paramètres estimés par MCO.
Propriété 1 : $\hat{\beta}$ est un estimateur consistent de β si $plim\hat{\beta} = \beta$.
Propriété 2 : La Distribution asymptotique de $\hat{\beta}$ est une loi Normale :

11. que nous présenterons dans le paragraphe “Tests"
12. c'est à dire quand la taille d'échantillon est grande
13. Il est très souvent plus facile d'établir une autre forme de convergence, la convergence en Moyenne Quadratique pour obtenir la convergence en probalité mais cela dépasse le cadre de ce cours

Si les u_i sont indépendants et identiquement distribués avec une espérance nulle et une variance finie σ^2 et si de plus les X se comportent bien[14] alors, et si N tend vers l'infini,

$$\sqrt{N}(\hat{\beta} - \beta)/X \xrightarrow{\mathcal{L}} N\left(0, \sigma^2(X'X)^{-1}\right)$$

$$\hat{\beta}_N/X \xrightarrow{\mathcal{L}} N\left(\beta, \hat{\sigma}^2(X'X)^{-1}\right)$$

Notons que la normalité asymptotique n'est pas obtenue en faisant l'hypothèse de normalité des erreurs. Cette hypothèse n'est pas nécessaire. La normalité asympotitique est une conséquence du Théorème Central Limite.

En conclusion si les hypothèses des MCO sont vérifiées (H1,H2,H3 et H5) alors l'estimateur des MCO est BLUE en échantillon fini et consistant et asypmtotiquement normal en échantillon infini.
En pratique, nous utiliserons les propriétés asymptotiques des estimateurs quand ces estimateurs ne sont pas, soit sans biais, soit de variance minimale c'est à dire quand les propriétés en échantillon fini ne sont pas vérifiées. Dans ces deux cas, les seules propriétés disponibles (ou connues) sont des propriétés asymptoqtiques. Dans ce premier chapitre les estimateurs possèdent toutes les propriétés en échantillon fini. Ce ne sera plus le cas dans les deux prochains chapitres.

III.3 Conclusion : la multicolinéarité

Dans la partie exploratoire de notre exemple nous avons trouvé : corr(LNL,LNK)= 0.89456. Quand deux variables explicatives (au moins) ont une corrélation élevée (positive ou négative) , on se heurte à un problème de multicolinéarité (ou de colinéarité dans certains manuels). On parle de multicolinéarité exacte quand la corrélation est égale à 1 et de multicolinéarité approchée quand la corrélation est "proche" de 1. En cas de multicolinéarité exacte, on ne peut pas calculer les estimateurs des paramètres car le déterminant de la matrice $X'X$ est nul. Nous verrons ce cas un peu plus loin dans le paragraphe sur les variables indicatrices. Dans le cas où la corrélation n'est pas égale à 1 mais proche de 1, on peut calculer les paramètres mais on se heurte à un "problème" de multicolinéarité approchée. Sur ce sujet je vous conseille le livre "Econométrie" de Damodar N. Gujarati, Bernard Bernier. Vous pouvez en trouver un extrait sur books.google.fr . Lire le paragraphe sur la multicolinéarité (paragraphe 10.4). Je partage totalement le point de vue de ces auteurs que je vais essayer de résumer.
Sur l'exemple de la fonction de production Cobb-Douglas, les deux facteurs de production sont très corrélés et nous sommes donc en présence de multicolinéarité.

14. en particulier si la matrice $\frac{X'X}{N}$ tend vers une matrice définie positive

Comme l'indiquent les auteurs "même si la multicolinéarité est forte, ... , les estimateurs MCO conservent encore les propriétés" vues dans le paragraphe précédent : ils sont toujours les meilleurs estimateurs linéaires et sans biais. "Le seul effet de la multicolinéarité est de rendre difficile l'obtention d'estimations des coefficients ayant de faibles écarts types. Mais disposer d'un petit nombre d'observations a le même effet,...".
Les deux auteurs indiquent les moyens de détection [15] de la multicolinéarité dans la paragraphe 10.7 puis les remèdes dans le paragraphe 10.8. Je ne vais pas présenter les moyens de détection car le seul remède possible sur notre échantillon c'est de ne rien faire. Il est clair qu'il n'y a pas d'autres solutions que d'introduire les deux facteurs de production : la théorie économique nous indique que la quantité produite dépend (au moins) de la quantité de travail et de capital utilisée par l'entreprise. Supprimer un de deux facteurs revient à se heurter à un biais des variables omises [16] car les deux facteurs sont significatifs. On parle aussi de biais de spécification c'est à dire que si on supprime un facteur de production pour éliminer la multicolinéarité alors le modèle est mal spécifié car il manque une variable pour expliquer la quantité produite. Or les économètres préferent toujours un estimateur avec la plus petite variance parmi les estimateurs sans biais. Le premier critère de choix est donc un estimateur sans biais. Ceci se comprend facilement si on se souvient des commentaires économiques des paramètres estimés que nous avons commencé à faire avec la notion d'élasticité. En résumé il faut faire un arbitrage entre biais de variables omises et multicoliéarité et dans ce cas les économètres évitent toujours les paramètres biaisés.

IV Tableau d'analyse de la variance

1. PRESENTATION DU TABLEAU :
   Dans le listing de SAS précédent (résultat SAS numéro 1), SAS a affiché le tableau "Analyse de variance" que nous reproduisons une nouvelle fois ci-après et que nous allons commenter rapidement dans ce paragraphe.

Analyse de variance

Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	2	14.21156	7.10578	200.25	<.0001
Error	24	0.85163	0.03548		
Corrected Total	26	15.06320			

   Définitions : On appelle Somme des Carrés Expliquée , notée SCE, la quantité $\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2$ et Somme des Carrés Totale la quantité $\sum_{i=1}^{N}(Y_i - \bar{Y})^2$, notée SCT.
   On retrouve les trois Sommes des Carrés (nous avons déjà défini SCR) dans

15. on peut penser à un problème de multicolinéarité quand le R^2 est élevé et que peu ou pas de variable sont significatives.
16. voir l'annexe du chapitre 1

la seconde colonne du tableau d'analyse de la variance de SAS. Ainsi $SCE = 14.21$, $SCR = 0.85$ et $SCT = 15.06$
Dans la première colonne se trouvent les degrés de liberté : $k-1 = 2$,$N-k = 24$ et $N - 1 = 26$
Dans la colonne "Carré Moyen" on retrouvent les deux quantités suivantes : $\frac{SCE}{k-1} = \frac{14.21}{2} = 7.105$ et $\frac{SCR}{N-k} = \frac{0.85}{24} = 0.035$ remarquez que $\frac{SCR}{N-k} = \hat{(\sigma)}^2$
Avec R, et avec la commande summary(mod1) on obtient la statistique de Fisher observée,200.2, avec les deux degrés de liberté, $k - 1 = 2$ et $N - k = 24$ la racine carrée de $\hat{(\sigma)}^2$ sur la ligne "Residual standard error", . On peut obtenir la SCR avec deviance(mod1) . En fait SCR est la seule "quantité" dont nous aurons besoin dans la suite de ce chapitre et donc il n'est pas necessaire d'obtenir les autres "quantités" fournies par SAS, comme la SCE par exemple.

2. UTILISATIONS DU TABLEAU : 2 indicateurs de la qualité du modèle peuvent être calculés à partir des données de ce tableau (en fait ces deux indicateurs sont fournis par SAS et R) :
- Le coefficient de détermination multiple, noté R^2 qui se calcule de la manière suivante : $R^2 = \frac{SCE}{SCT} = \frac{14.21}{15.06} = 0.9435$
SAS affiche ce R^2 en dessous du tableau d'analyse de la variance :"R-Square 0.9435"
R^2 mesure le pourcentage de la dispersion (de la variance) de la variable endogène expliqué par le modèle c'est à dire expliqué par les variables exogènes. Sur l'exemple : 94.35 % de la dispersion de Ln(Y) est expliqué par Ln(L) et Ln(K).
ATTENTION : en pratique il faut faire très attention au commentaire que l'on pourrait faire sur la valeur du R^2. Il faut d'abord savoir que la notion de "R^2 élevé" est relative : en général sur des données individuelles (les observations sont des individus comme des ménages, des pays, des firmes) le R^2 est plutôt plus faible que sur des données temporelles (les variables sont observées dans le temps). De plus il ne faut jamais écrire "le R^2 de ce modèle est élevé donc le modèle est bon" car le "modèle" en économétrie est issu de la théorie économique [17]. C'est la théorie économique qui nous indique quelles sont les variables explicatives de la variable endogène. La question est seulement " le modèle économique est-il invalidé ?" [18] et ce n'est pas avec le R^2 que nous allons répondre à cette question mais en faisant des tests (voir le paragraphe Tests). Eventuellement on peut détecter un problème de variable omise si R^2 est faible mais à nouveau la liste des variables explicatives est dictée par la théorie économique.
Ajoutons aussi que le R^2 a un énorme défaut : il augmente (ou reste constant) quand on ajoute une variable explicative même si cette dernière a un faible pouvoir explicatif. Pour corriger ce défaut, vous allez trouver dans les manuels

17. de plus nous avons vu que cela peut être le résultat d'un problème de multicolinéarité.
18. et non pas le modèle est - il validé ? car on ne peut jamais répondre avec certitude.

d'économétrie, un autre R^2 que l'on appelle R^2 ajusté dont la définition est la suivante :
$AdjR^2 = 1 - (1 - R^2)\frac{N-1}{N-k}$
SAS affiche ce R^2 ajusté dans "Adj R-Sq 0.9388 ". Notons que le R^2 ajusté est toujours inférieur au R^2. A nouveau on peut faire les même remarques sur le R^2 ajusté que sur le R^2 : un modèle est issu de la théorie économique et la liste des variables explicatives ne doit pas être issue d'indicateurs empiriques [19]. Nous reviendrons sur ces notions dans les sujets de TP et dans la suite du cours.
- Le second indicateur de la qualité du modèle est une statistique de Fisher
On calcule cette statistique pour tester H_0 : Tous les paramètres sont nuls sauf la constante c'est à dire ici $H_0 : \beta_1 = 0, \beta_2 = 0$,
On montre que si H_0 est vraie alors :

$$F = \frac{SCE/(k-1)}{SCR/(N-k)} \sim F(k-1, N-k) \tag{5}$$

On se fixe un risque de première espèce, en général 5 %, qui est la probabilité de rejeter H_0 alors qu'elle est vraie. SAS nous donne la valeur observée de cette statistique 200.25, ainsi que la probabiilté de dépasser la valeur observée. Quand cette proba est inferieure à 5 % (par défaut dans ce cours), on rejette H_0.
c'est le cas ici : SAS affiche une probabilité "<0.0001". Nous rejetons donc H_0.
Remarque : si le modèle est bien spécifié, c'est à dire en particulier s'il contient la liste des variables explicatives de la théorie économique, il n' y a aucune raison de penser que nous pourrions ne pas rejeter H_0.

V Tests sur les paramètres

On note p le nombre de contraintes (ou d'équations) dans H_0.

V.1 Test sur une seule équation : $p = 1$

Exemple 1 :

Sur la fonction de production Cobb-Dougals, $Ln(Y_i) = \beta_0 + \beta_1 Ln(L_i) + \beta_2 Ln(K_i) + u_i$, on s'intéresse à l'hypothèse $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$;on dit dans ce cas que le test est bilatéral car il contient deux régions de rejet [20]. Par défaut dan ce cours, les tests seront toujours bilatéraux sauf mention contraire. Il s'agit en fait de savoir si la variable LNL a un effet non nul significatif sur la variable endogène LNY ou encore si la variable LNL est significative. Quand le nombre de contraintes de H_0 est égal

19. sauf pour choisir parmi la liste des définitions possibles d'une même variable comme pour la variable choisie pour mesurer un facteur de production
20. voir http ://www.er.uqam.ca/nobel/r30574/PSY1300/C8P5.html

à 1, c'est à dire quand $p = 1$, il existe deux méthodes et deux statistiques de test équivalentes. Commençons par utiliser une statistique de Student qui est fournie par défaut par tous les logiciels pour $H_0 : \beta_1 = 0$.
Méthode 1 : Statistique de Student
On sait que

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{var}(\hat{\beta}_1)}} \text{suit une } St(N-k)$$

Reproduisons une partie du listing de SAS sur la fonction de production Cobb-Douglas.

Résultats estimés des paramètres

Variable	DF	Résultat estimé des paramètres	Erreur std	Valeur du test t	Pr > |t|
Intercept	1	1.17064	0.32678	3.58	0.0015
LNL	1	0.60300	0.12595	4.79	<.0001
LNK	1	0.37571	0.08535	4.40	0.0002

Dans ce listing de SAS et pour la variable LNL (dont β_1 est le paramètre), la "valeur du test t" est égale à 4.79. C'est la valeur observée de la statistique de Student. Nous avons déjà vu que 0.60300 est la valeur estimée de β_1. Dans la colonne "Erreur std", "erreur standard", traduit de l'anglais , on retrouve l'écart-type estimé de $\hat{\beta}_1$. Il s'agit en fait de $\sqrt{\hat{Var}\hat{\beta}_1}$. Remarquons que dans la matrice de variance -covariance des estimateurs nous avions identifié la valeur de cette variance ; elle était égale à 0.015864. Le calcul de la statistique observée est donc le suivant :
$\frac{\hat{\beta}_1}{\sqrt{\hat{var}(\hat{\beta}_1)}} = \frac{0.603}{0.12595} = 4.79$
Pour conclure sur H_0, soit on compare cette valeur observée (en toute rigueur il s'agit de la valeur absolue de cette valeur observée ce qui ne change rien ici car la valeur observée est positive) à la valeur lue dans une table, la valeur critique, de Student à (N-k) degré de liberté [21] soit on utilise la probabilité affichée par SAS à côté de la valeur observée 4.79. C'est la seconde solution que je privilégie dans ce cours. On lit, dans la colonne "$Pr > |t|$", une probabilité qui est "< 0.0001" ; elle est $<$ à 5% donc on rejette à 5% $H_0 : \beta_1 = 0$ et on conclut que LNL est une variable significative à 5%. Remarquons que cette variable est significative même à 1%.
Nous pouvons aussi utiliser un Intervalle de Confiance (IC) pour tester H_0 de la manière suivante :
- on calcule l'IC avec $IC = [\hat{\beta}_1 \pm table\sqrt{\hat{var}(\hat{\beta}_1)}]$
- si $O \notin IC$ alors on rejette H_0 .
Pour obtenir les IC :

21. pour une explication graphique voir le site internet http ://www.er.uqam.ca/nobel/r30574/PSY1300/C8P5.html

Avec SAS : proc reg data=tpfoad.cobb1 ;model lny=lnl lnk /clb ;run ;
Avec R : confint(mod1,level=0.95)
On obtient la même conclusion pour LNL : c'est une variable significative car $O \notin IC$; il en est de même pour LNK.
On montre que quand $p = 1$, on peut aussi utiliser une statistique de Fisher. C'est en fait la méthode utilisée par SAS dans l'instruction "test" que nous allons utiliser maintenant. Je voudrais ajouter que les manuels d'économétrie théoriques avec des applications ne présentent pas cette statistique de Fisher quand il y a une seule contrainte dans H_0. C'est le cas par exemple dans le manuel de Wooldridge, mentionné en bibliographie. Je présente cette statistique de Fisher car la commande "test" de SAS est très simple et elle utilise cette statistique de Fisher.
La syntaxe SAS est la suivante :

```
proc reg data=tpfoad.cobb1 ;
model LNY=LNL LNK ;test LNK=0 ;run ;
```

On teste ici $\beta_2 = 0$ c'est à dire la significativité de la variable capital en log. Remarquer que la contrainte s'écrit $LNK = 0$ avec le langage de SAS ;elle signifie "le paramètre de LNK est nul".
SAS présente le résultat du test comme un tableau d'analyse de la variance :

```
Test 1 Results for Dependent Variable LNY

                                 Carré     Valeur
Source              DF           moyen          F    Pr > F

Numerator            1         0.68767      19.38    0.0002
Denominator         24         0.03548
```

Si nous notons H_0 sour la forme $R\beta = c$ où R est une matrice connue de dimension (p,k) et c un vecteur connu de dimension (p,1) la statistique calculée par SAS est donnée par

$$F = \frac{(R\hat{\beta} - c)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - c)/p}{SCR/(N-k)} \text{ suit une } F(p, N-k) \qquad (6)$$

Cette formule sera appélée "formule 1 de Fisher" pour l'instant. Il y aura une autre formule de Fisher strictement équivalente dans le paragraphe "tests sur plusieurs équations : p> 1 " dans la suite de ce chapitre.
Pour le test qui nous intéresse, c'est à dire pour $H_0 : \beta_2 = 0$, les vecteurs et matrices utilisées dans la formule (**??**) sont

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \; R = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \text{ et } c = 0.$$

sur cet échantillon la statistique de test pour $H_0 : \beta_2 = 0$ est égale à 19.38 et le tableau du test décompose cette valeur de la manière suivante :

$$F = \frac{(R\hat{\beta} - c)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - c)/p}{SCR/(N-k)} = \frac{0.68767}{0.03548} = 19.38$$

A nouveau SAS affiche la probabilité de dépasser la valeur observée 19.38 ; elle est égale à 0.0002 comme pour la statistique de Student car ces deux statistiques sont équivalentes. En effet on sait qu'une Fisher à (1,N-K) degrés de liberté est le carré d'une Student à N-k degrés de liberté pour un test bilatéral. Ainsi le t-ratio au carré est égal à la statistique de Fisher (aux erreurs d'arrondis près) : $4.40^2 = 19.36$
Le test de Student est plus flexible car il permet de faire un test unilatéral. Nous étudierons ce cas en TD.
Avec le logiciel R :
Certains d'entre vous connaissent peut être la commande "linear.hypothesis(mod1,"LNK=0")" de R : quand vous l'exécutez vous obtenez la même valeur observée de la statistique de Fisher obtenue avec SAS mais R n'utilise pas la formule, "formule 1", que je présente dans ce paragraphe. En fait R utilise une formule 2 que je présenterai dans le paragraphe "tests sur plusieurs équations : p>1 " qui compare un modèle contraint et un modèle non contraint. Vous trouverez sur Moodle un programme R où je crée une fonction pour calculer la statistique de Fisher avec la formule 1 de ce paragraphe. Vous remarquerez que ce programme contient aussi tous les appels à cette fonction pour effectuer les tests où j'utilise cette formule 1 dans ce premier chapitre. Vous noterez que vous devez spécifier la matrice R et le vecteur c pour appeler cette fonction. Les résultats sont identiques à ceux obtenus par SAS. Donc dans la suite de ce paragraphe je vous renvoie au programme R que j'ai mis sur Moodle pour vérifier que vous obtenez les mêmes valeurs observées présentées dans ce cours avec SAS. Si vous avez des difficultés, nous en discuterons sur le forum.

CONCLUSION SUR EXEMPLE 1 : la méthodologie pratique est différente de la présentation pédagogique du cours. En cours, j'ai commenté les paramètres estimés avant de tester si les variables concernées étaient significatives ;en pratique on teste d'abord la significativité des variables , puis on commente les paramètres des variables significatives (pour les variables non significatives le seul commentaire à faire est qu'elles n'ont aucun effet sur la variable endogène ce qui est aussi une information intéressante).
Exemple 2 : Il y a toujours une seule équation ($p = 1$) dans H_0 mais elle contient une combinaison linéaire de plusieurs paramètres. Nous proposons de tester l'hypothèse $H_0 : \beta_1 + \beta_2 = 1$ qui est une hypothèse habituelle sur une fonction de production Cobb-Douglas. En effet cette hypothèse s'interprète comme un test sur les rendements d'échelle.
RAPPEL de microéconomie : Notons r les rendements d'échelle. Par définition, quand on multiplie par λ la quantité de chaque facteur utilisé alors la production est multipliée par λ^r.
On dit que les rendements sont constants si $r = 1$ c'est à dire si , quand on double par exemple la quantité de chaque facteur la production est aussi doublée. Les ren-

dements sont croissants si $r > 1$ c'est à dire si la production est plus que doublée quand la quantité de chaque facteur est doublée . Enfin les rendements sont décroissants si la quantité produite est moins que doublée quand la quantité de facteur est doublée . Avec une fonction de production Cobb-Douglas, le rendement d'échelle est égal à $\beta_1 + \beta_2$. Nous proposons de tester si les rendements d'échelle sont constants dans la branche des 27 entreprises de notre échantillon.
L'hypothèse alternative, notée H_1, sera $r \neq 1$, c'est à dire, les rendements ne sont pas constants ;ils peuvent être croissants ou décroissants. Nous procèderons donc à un test bilatéral.
Comme pour l'exemple 1, nous disposons de 2 méthodes équivalentes puisque le nombre de contrainte de H_0 est toujours égal à 1 : une statistique de Student ou une statistique de Fisher calculée directement avec SAS.
Méthode 1 : Statistique de Student
On pose $r = \beta_1 + \beta_2$, H_0 devient $r = 1$
On montre que $\dfrac{\hat{r} - r}{\sqrt{v\hat{a}r(\hat{r})}} \sim St(N-k)$

Si H_0 est vraie alors $\dfrac{\hat{r} - 1}{\sqrt{v\hat{a}r(\hat{r})}} \sim St(N-k)$

Pour calculer la valeur observée, $\dfrac{\hat{r} - 1}{\sqrt{v\hat{a}r(\hat{r})}}$ il faut d'abord calculer $\hat{r}$ et $\hat{Var}(\hat{r})$. On calcule $\hat{r}$ simplement avec $\hat{r} = \hat{\beta}_1 + \hat{\beta}_2 = 0.60300 + 0.37571 = 0.97871$
Pour calculer $\hat{Var}(\hat{r})$ il faut utiliser la formule $Var(X+Y) = VarX + VarY + 2Cov(X,Y)$. On obtient ainsi $\hat{Var}(\hat{r}) = \hat{Var}(\hat{\beta}_1) + \hat{Var}(\hat{\beta}_2) + 2\hat{Cov}(\hat{\beta}_1, \hat{\beta}_2)$
Ces trois valeurs se trouvent dans la matrice de Variance-Covariance que nous avons déjà présentée dans le paragraphe III.2 Propriétés des estimateurs. On peut donc remplacer ces 3 valeurs dans le calcul de $\hat{Var}(\hat{r})$:
$\hat{Var}(\hat{r}) = 0.0158644 + 0.0072839309 + 2(-0.009616201) = 0.003915929$
$\Rightarrow$ la valeur observée $= \dfrac{\hat{r} - 1}{\sqrt{v\hat{a}r(\hat{r})}} = \dfrac{0.97 - 1}{\sqrt{0.005}} = -0.34022$
En fait je me sers souvent de SAS comme d'une machine à calculer pour appliquer la formule de la valeur observée ci -dessus ; cela me permet d'avoir une trace de tous les calculs au même endroit , le programme SAS ;de plus, il faut utiliser SAS pour calculer la probabilité de dépasser la valeur observée si on n'a pas de table de Student sous les yeux. Donc je vous propose d'exécuter le programme SAS suivant

```
data calcul ;
r= 0.60300+ 0.37571 ;
var=0.0158644+0.0072839309+(2*-0.009616201 ) ;
vobs=(r-1)/sqrt(var) ;
p=(1-probt(abs(vobs),24))*2 ;
run ;
proc print data=calcul ;run ;
```

Je vous conseille d'aller lire la documentation SAS si vous ne connaissez pas la fonction "probt"
Le tableau calcul est le suivant :

```
Obs        r              var       vobs          p
1      0.97871       .003915929     -0.34022     0.73665
```

Remarque : il vaut mieux prendre toujours la valeur absolue de la valeur observée pour ne pas se tromper au cas où elle soit négative. pour calculer correctement la probabilité de dépasser la valeur observée avec un test bilaréral il faut multiplier cette probabilité$[1 - probt(abs(vobs), 24)]$ par 2.
On obtient une probabilité égale à 73.665% qui est beaucoup plus grande que 5% (ou même 10% : les seuils habituels sont égaux à 1, 5 ou 10%) le seuil que nous nous sommes fixés. On ne peut donc pas rejeter l'hypothèse selon laquelle les rendements sont constants dans cette branche. Ainsi nous pouvons écrire la fonction de production Cobb-Douglas de la manière suivante :
$Y_i = AL_i^{\beta_1} K_i^{1-\beta_1}$ car on n'a pas rejeté $H_0 : \beta_1 + \beta_2 = 1$ et donc $\beta_2 = 1 - \beta_1$.
Avec R :

```
mod1 <- lm(LNY   LNL+LNK, data =cobb1)
covb <- vcov(mod1)
print(covb)
coeff.mod1 <- coef(mod1)
print(coeff.mod1)
t <- (coeff.mod1[2]+coeff.mod1[3]-1)/sqrt(covb[2,2]
+covb[3,3]+2*covb[2,3])
print(t)
pvalue <- 2*(1-pt(abs(t),mod1$df))
print(pvalue)
```

Méthode 2 : Statistique de Fisher
On peut utiliser la commande test de la proc reg pour tester la même hypothèse de la manière suivante :

```
proc reg data=tpfoad.cobb1;
model LNY=LNL LNK;
test LNL+LNK=1;
run;
```

La sortie SAS est la suivante :

```
Test 1 Results for Dependent Variable LNY

                                 Carré     Valeur
Source                  DF       moyen         F    Pr > F

Numerator                1     0.00411      0.12    0.7366
Denominator             24     0.03548
```

La probabilité est égale à 73.66% comme dans la première méthode. On retrouve la relation entre Fisher et Student quand $p = 1$. On a : $\sqrt{0.12} = 0.34$ aux erreurs d'arrondis près.
On peut mettre un "label" dans l'instruction SAS "Test" pour distinguer les différents tests que nous avons effectués en ajoutant , par exemple, "rdt : " avant le mot "test" :

```
rdt : test LNL+LNK=1 ;
```

Le listing ne commence plus par "Test 1" mais par " rdt" ; je vous laisse exécuter cette ligne avec SAS et regarder ce que cela donne.

Exemple 3 Dans le paragraphe "Estimation" nous avons commenté les élasticités de la manière suivante :"Sur cet échantillon la quantité produite est plus sensible à une augmentation du facteur travail qu'à une augmentation du facteur capital car l'élasticité du travail est supérieure à celle du capital pour cette branche à condition que l'on rejette l'hypothèse d'égalité des deux élasticités et que donc la différence entre ces deux élasticités soit significative. Nous effectuerons ce test dans le paragraphe "Tests" (voir plus loin)".
Nous allons maintenant procéder à ce test. Nous souhaitons tester si la sensibilité de la quantité produite au facteur travail est la même que cette sensibilité au facteur capital. Nous nous proposons de tester l'égalité des deux élasticités soit $H_0 : \beta_1 = \beta_2$ ou encore $H_0 : \beta_1 - \beta_2 = 0$. Cet exemple ressemble beaucoup à l'exemple 2. Il s'agit seulement d'une autre combinaison linéaire des paramètres. Effectuons ce test avec les deux méthodes que nous avons déjà vues :
Méthode 1 : dans un premier temps, nous calculons la statisque observée de Student que nous notons toujours r. Le programme SAS est sensiblement le même que dans l'exemple 2. Je vous laisse le soin de noter les différences :

```
data calcul ;
r= 0.60300- 0.37571 ;
var=0.0158644+0.0072839309-(2*-0.009616201 ) ;
vobs=r/sqrt(var) ;
p=(1-probt(abs(vobs),24))*2 ;
run ;
proc print data=calcul ;run ;
```

Avec R :

```
mod1 <- lm(LNY   LNL+LNK, data =cobb1)
covb <- vcov(mod1)
coeff.mod1 <- coef(mod1)
print(coeff.mod1)
t <- (coeff.mod1[2]-coeff.mod1[3])/sqrt(covb[2,2]
+covb[3,3]-2*covb[2,3])
print(t)
pvalue <- 2*(1-pt(abs(t),mod1$df))
print(pvalue)
```

La sortie SAS de la procédure print est :

Obs	r	var	vobs	p
1	0.22729	0.042381	1.10407	0.28051

Ainsi la probabilité est supérieure à 5% , on ne rejette pas l'hypothèse selon laquelle les deux élasticités sont identiques.
Méthode 2 :
Le programme SAS est simple :

```
proc reg data=tpfoad.cobb1 ;
model LNY=LNL LNK ;
test LNL=LNK ;
run ;
```

La sortie SAS est la suivante :

Test 1 Results for Dependent Variable LNY

Source	DF	Carré moyen	Valeur F	Pr > F
Numerator	1	0.04325	1.22	0.2805
Denominator	24	0.03548		

Nous obtenons les mêmes résultats et les mêmes commentaires que pour la méthode 1.

V.2 Test sur plusieurs équations : $p > 1$

Exemple 4 Jusqu'à présent, nous avons supposé que la fonction de production était de type Cobb-Douglas mais il en existe au moins une autre appelée fonction de production "Translog" qui s'écrit de la manière suivante :
$Ln(Y_i) = \beta_0 + \beta_1 Ln(L_i) + \beta_2 Ln(K_i) + \beta_3(\frac{1}{2}[Ln(L_i)]^2) + \beta_4(\frac{1}{2}[Ln(K_i)]^2) + \beta_5(Ln(L_i) \times Ln(K_i)) + v_i$
La question que nous allons nous poser est : Quelle est la "meilleure" fonction de production pour ces 27 entreprises ?
Comparer ces deux fonctions de production revient à tester $H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$ qui contient 3 contraintes ($p = 3$). Dès que le nombre de contraintes de H_0 est strictement plus grand que 1 on ne peut plus utiliser la statistique de Student. Il faut utiliser une loi de Fisher. Il y a plusieurs formules équivalentes pour calculer la statistique observée. Nous nous proposons d'utiliser d'abord la statistique de l'instruction "Test" de SAS que nous avons déjà utilisée (voir la formule (**??**)) que nous

rappelons ci-après :

$$F = \frac{(R\hat{\beta} - c)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - c)/p}{SCR/(N-k)} \text{ suit une } F(p, N-k)$$

Sur le test qui nous intéresse ici, c'est à dire pour $H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$, on a

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_5 \end{pmatrix}$$

$$, R = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \text{ et } c = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Remarquons que cette statistique n'utilise que les estimations du modèle Translog ;il faut donc estimer ce modèle pour tester ces 3 contraintes .
Nous créons les 3 variables supplémentaires dans une étape data puis nous utilisons la proc reg de SAS :

```
data tpfoad.translog ;set tpfoad.cobb1 ;
LNL2=(1/2)*lnl*lnl ;
LNK2=(1/2)*lnk*lnk ;
LNLLNK=lnl*lnK ;
run ;
proc reg data=tpfoad.translog ;
model lnY=lnL lnK lnl2 lnk2 lnllnk ;
run ;
```

Les résultats d'estimation sont les suivants :

```
 Dependent Variable: LNY

Number of Observations Read          27
Number of Observations Used          27


                                  Analyse de variance

                                Somme des          Carré     Valeur
Source                     DF     carrés          moyen         F    Pr > F

Model                       5   14.38327        2.87665      88.85    <.0001
Error                      21    0.67993        0.03238
Corrected Total            26   15.06320


Root MSE              0.17994    R-Square     0.9549
```

Dependent Mean	7.44363	Adj R-Sq	0.9441
Coeff Var	2.41733		

Résultats estimés des paramètres

Variable	DF	Résultat estimé des paramètres	Erreur std	Valeur du test t	Pr > |t|
Intercept	1	0.94420	2.91075	0.32	0.7489
LNL	1	3.61364	1.54807	2.33	0.0296
LNK	1	-1.89311	1.01626	-1.86	0.0765
LNL2	1	-0.96405	0.70738	-1.36	0.1874
LNK2	1	0.08529	0.29261	0.29	0.7735
LNLLNK	1	0.31239	0.43893	0.71	0.4845

Commentaires :
Pour la fonction de production Translog la variable LNL est significative à 5% et la variable LNK est significative à 10%. Les 3 variables supplémentaires ne sont pas significatives [22].
Pour tester H_0 il faut ajouter l'instruction "test" :

```
proc reg data=tpfoad.translog ;
model lnY=lnL lnK lnl2 lnk2 lnllnk ;
test lnl2,lnk2,lnllnk ;
run ;
```

Remarquer que par défaut pour chaque paramètre la valeur est nulle dans l'instruction "test". Les résultats du test sont :

Test 1 Results for Dependent Variable LNY

Source	DF	Carré moyen	Valeur F	Pr > F
Numerator	3	0.05724	1.77	0.1841
Denominator	21	0.03238		

La probabilité est égale à 18.41% et donc on ne rejette pas H_0.
Ainsi , entre une fonction de production Cobb-Douglas très simple, et une fonction de production Translog, un peu plus générale on choisit de travailler sur la fonction Cobb-Douglas qui n'est pas une trop "mauvaise" simplification de la réalité.
Si nous avions rejetté H_0, cela aurait signifié que le modèle Cobb-Douglas était trop simple pour représenter la réalité économique et qu'il ne pouvait donc pas être

22. Attention :il faut tout de même procéder au test de Fisher

utilisé ; il aurait fallu alors et par exemple, calculer les élasticités pour le modèle Translog.
Attention : il ne faut surtout pas conclure à partir de la sortie du modèle translog sans faire de test. En particulier il peut être faux de dire : "Etant donné que les 3 variables supplémentaires du modèle translog ne sont pas significatives (leur probabilité est bien au dessus de 10%) nous ne rejettons pas que les 3 paramètres associés à ces variables soient nuls et donc nous allons travailler avec le modèle Cobb-Douglas". Le raisonnement n'est pas correct ; il faut procéder à un test de Fisher pour comparer ces deux modèles car on ne peut rien déduire des tests individuels de significativité des variables. Le test avec une Student et le test avec une Fisher sont équivalents seulement si le nombre de contraintes de H_0 est égal à 1 ce qui n'est pas le cas ici. Quand $p > 1$ alors tout peut se produire : on peut rejeter un paramètre est nul avec une Student et ne pas rejeter cette hypothèse quand ce paramètre est testé avec un ensemble de paramètre.
Revenons au test de comparaison des deux modèles et présentons la seconde méthode pour tester la même hypothèse. Pour utiliser la seconde formule de calcul de la statistique de Fisher, il faut introduire un peu de vocabulaire économétrique. Soient le Modèle Contraint (MC) le modèle qui vérifie toutes les contraintes de H_0 et le Modèle Non Contraint (MNC) qui ne vérifie pas H_0 (au moins une contrainte de H_0 n'est pas vérifiée).
On montre que si H_0 est vraie, alors

$$F = \frac{(SCR_c - SCR_{nc})/p}{SCR_{nc}/(N-k)} \sim F(p, N-k) \tag{7}$$

où SCR_c = SCR du modèle contraint et SCR_{nc} = SCR du modèle non contraint
Le modèle Cobb-Douglas est clairement le modèle contraint car il ne contient pas les 3 variables supplémentaires de la fonction "Translog". Le modèle non contraint est donc le modèle "Translog". Pour calculer la valeur observée $F = \frac{(SCR_c - SCR_{nc})/p}{SCR_{nc}/(N-k)}$, il faut "récupérer" la SCR de chaque modèle. Nous disposions déjà des résultats d'estimation du modèle Cobb-Douglas pour lequel $SCR_c = 0.85163$
Pour le modèle translog, il suffit de lire dans les sorties SAS ci dessus, $SCR_{nc} = 0.67993$
On obtient pour la valeur observée :
$F = \dfrac{(0.85163 - 0.67993)/3}{0.67993/(27-6)} = 1.7676$
Il faut ensuite calculer la probabilité de dépasser la valeur observée en ajoutant dans le programme SAS les lignes suivantes :

```
data calcul ;
vobs=1.7676 ;p=1-probf(vobs,3,21) ;
run ;
proc print ;run ;
```

On obtient la sortie suivante :

```
Obs     vobs        p
```

```
1     1.7676    0.18417
```

Donc on ne rejette pas H_0 et on choisit aussi d'utiliser le modèle Cobb-Douglas. Attention : la comparaison de la SCR de deux modèles n'est possible que si la variable endogène ou variable expliquée, est la même dans les deux modèles. Nous avons vu un test où cela n'était pas le cas : $H_o : \beta_1 + \beta_2 = 1$ dans le modèle Cobb-Douglas. Le modèle contraint n'a pas la même variable endogène. Dans ce cas nous ne pouvons pas utiliser la statistique de Fisher qui compare les deux SCR et nous sommes donc obligés d'utiliser la statistique de Fisher suivante :

$$F = \frac{(R\hat{\beta} - c)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - c)/p}{SCR/(N-k)} \text{ suit une } F(p, N-k)$$

Pour les utilisateurs du logiciel R :
- pour obtenir la SCR d'un modèle, utiliser la fonction deviance
- dans R il existe une fonction (anova) qui permet de comparer un modèle contraint et un modèle non contraint avec une statistique de Fisher . Le programme R pour comparer le modèle Cobb-Douglas et le modèle Translog est le suivant :

```
mod1 <- lm(LNY   LNL+LNK, data =cobb1)
LNL2 <- 0.5*cobb1$LNL*cobb1$LNL
LNK2 <- 0.5*cobb1$LNK*cobb1$LNK
LNLLNK <- cobb1$LNL*cobb1$LNK
mod2 <- lm(LNY   LNL+LNK+LNL2+LNK2+LNLLNK, data =cobb1)
anova(mod2,mod1)
```

Avec R pour calculer la statistique de Fisher précédente (formule 1) qui utilise seulement le modèle non contraint et qui donc ne nécessite pas l'estimation du modèle contraint , il faut créer une fonction (en tous cas je n'ai pas trouvé une telle fonction). Avec SAS c'est exactement le contraire : la commande test de la proc reg utilise la première formule de Fisher alors qu'il faut calculer la statistique qui compare les deux SCR (ou faire de l'anova avec SAS mais d'un point de vue pédagogique je prefère que vous calculiez la statistique observée. Nous en discuterons sur les forums du cours et/ou des TP si cela est nécessaire.

CONCLUSION :
ATTENTION à la méthodologie pratique :
Elle est différente de la présentation pédagogique que j'ai faite jusqu'à présent.
En cours jusqu'à présent nous avons :
- estimé une fonction Cobb-Douglas
- commenté les paramètres estimés avec la notion d'élasticité
- testé la significativité des variables
- testé les rendements d'échelle constants
- comparé une fonction Cobb-Douglas et une fonction Translog : nous n'avons pas rejeté la fonction Cobb-Douglas

En pratique il faut bien sûr commencer par le dernier point : comparer deux fonctions de production pour choisir de travailler sur une spécification du modèle. Ensuite et ensuite seulement , nous pouvons étudier le "meilleur" modèle du test de

comparaison de modèles. En clair, nous commençons par tester un modèle Cobb-Douglas contre un modèle Translog ; nous ne rejetons pas le modèle Cobb-Douglas. Nous commentons ensuite le modèle Cobb-Douglas : par exemple quelles sont les variables significatives ? nous commentons les élasticités de chaque facteur, nous testons "les rendements d'échelle sont - ils constants ?" ...
ATTENTION : pour comparer deux modèles il ne faut surtout pas comparer les R^2. Ici le modèle qui a le plus grand R^2 est le modèle Translog ($R^2 = 95.49\%$ contre 94.35% pour le modèle Cobb-Douglas). La différence entre les R^2 est plutôt faible sur notre échantillon.

REMARQUE SUR LA DEMARCHE ECONOMETRIQUE : En économétrie les données servent à invalider ou à ne pas invalider le modèle économique. On ne peut jamais valider un modèle, on ne peut jamais être sûr qu'il est correct. Sur cet exemple la fonction de production translog est invalidée cela ne veut pas dire que la fonction Cobb-Douglas est validée : il y a peut être une autre fonction de production qui serait préférée à la fonction de production Cobb-Douglas et que nous n'avons pas testée.

V.3 Tests sur la stabilité des paramètres ou test de Chow (1960)

Pour les utilisateurs de R : dans tous les paragraphes de ce chapitre 1 à partir de celui-ci, seuls les programmes SAS figurent dans le texte car il n'y aucune fonction nouvelle avec le logiciel R.
Le fichier de données de ce paragraphe concerne la consommation d'essence aux Etats-Unis entre 1960 et 1995.
FICHIER DE DONNEES : testchow
SOURCE : "Econométrie" W.H. Greene.
Liste des variables :

- annee : année d'observation
- G : conso d'essence (gasoline en anglais US)
- Pg : prix de l'essence
- Rt : revenu disponible par tête
- pop : population

Calculons la consommation d'essence par tête puis représentons graphiquement l'évolution de cette consommation sur la période d'observation avec le programme suivant :

```
data tpfoad.essence ;set tpfoad.testchow ;
consot=G/POP ;run ;
proc gplot data=tpfoad.essence ;
plot consot*annee ;run ;
```

Sur le graphique **??**, nous pouvons facilement remarquer que l'évolution de la

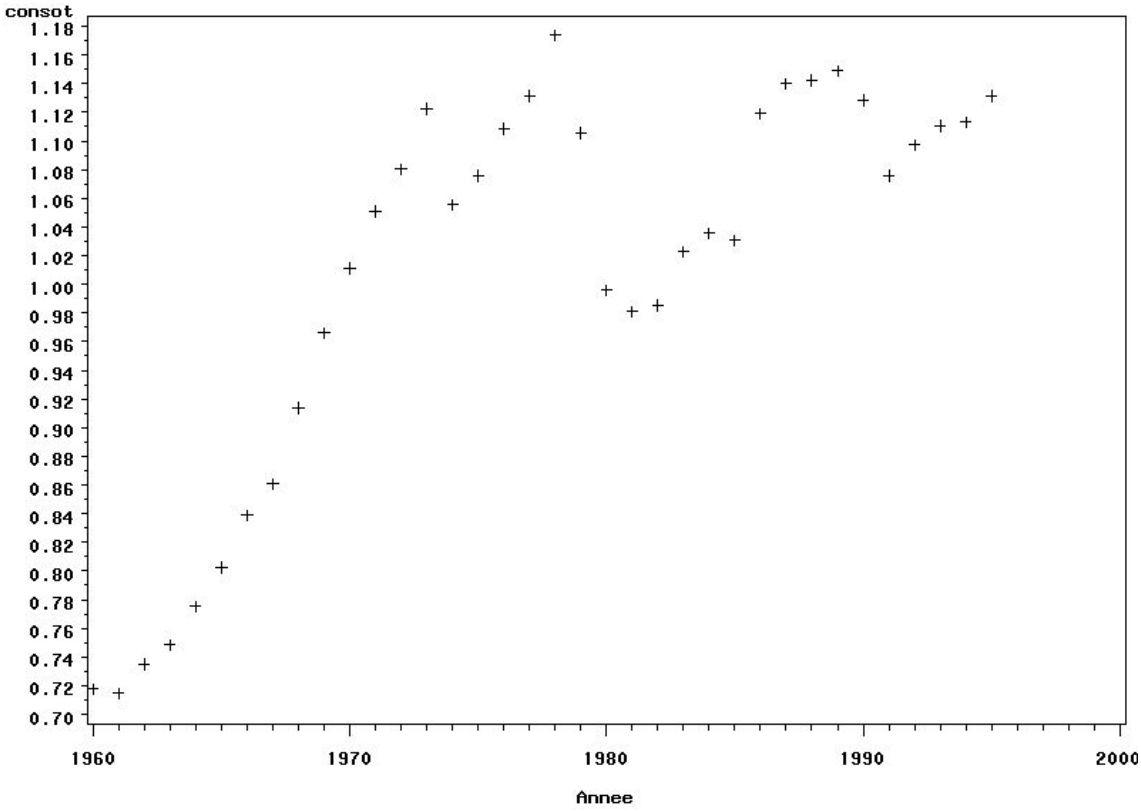

FIGURE 3 – Evolution de la consommation d'essence

consommation (par tête) n'est pas stable sur la période ; on distingue la cassure du premier choc pétrolier de 1973. Avant 1973, la consommation a un trend positif stable alors qu'après cette date son évolution devient plus chaotique.
Pour tester l'hypothèse selon laquelle les paramètres [23] ne sont pas restés stables sur la période 1960-1995, nous allons comparer un Modèle Contraint (MC) [24] et un Modèle Non Contraint (MNC) avec la statistique de Fisher que nous avons déjà étudiée.
Commençons par estimer un modèle de regression multiple sur toute la période qui est le MC et que nous écrirons :

$$Log(Consot)_t = \alpha + \beta Log(rt)_t + \gamma Log(pg)_t + u_t \text{ pour } t = 1960 - 1995 \qquad (8)$$

Nous avons transformé les variables avec la fonction Log afin que les paramètres estimés s'interprètent directement comme des élasticités (élasticité-prix et élasticité-revenu).
Le programme SAS est le suivant :

23. dans le modèle il y a 3 paramètres
24. la contrainte est : tous les paramètres sont stables ou identiques sur la période.

```
Data tpfoad.logessence ;set tpfoad.essence ;
lconsot=log(consot) ;lrt=log(rt) ;lpg=log(pg) ;run ;
proc reg data=tpfoad.logessence ;
model lconsot=lrt lpg ;run ;
```

La sortie SAS est la suivante :

Dependent Variable: lconsot

Number of Observations Read	36
Number of Observations Used	36

Analyse de variance

Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	2	0.73580	0.36790	174.56	<.0001
Error	33	0.06955	0.00211		
Corrected Total	35	0.80535			

Root MSE	0.04591	R-Square	0.9136
Dependent Mean	-0.00371	Adj R-Sq	0.9084
Coeff Var	-1237.90091		

Résultats estimés des paramètres

Variable	DF	Résultat estimé des paramètres	Erreur std	Valeur du test t	Pr > |t|
Intercept	1	-10.67585	0.79005	-13.51	<.0001
lrt	1	1.18584	0.08872	13.37	<.0001
lpg	1	-0.19577	0.03007	-6.51	<.0001

Commentaires :
- Toutes les variables sont significatives.
- Etant donné que la spécification est en Log, les paramètres s'interprètent comme des élasticités.
- L'élasticité revenu est égale à 1.18 ce qui signifie qu'une augmentation de 1% du revenu engendre une augmentation de la consommation d'essence de 1.18% pour cette période.
- L'élasticité prix est égale à -0.19 [25] : une augmentation de 1% du prix de l'essence

25. le signe négatif est le signe attendu

provoque une diminution de la consommation d'essence de 0.19 %.
- On note que $SCR_C = 0.06955$
Pour tester l'hypothèse selon laquelle le modèle n'est pas stable sur la période, c'est à dire que les 3 paramètres du modèle ne sont pas les mêmes avant et après 1973, il faut estimer un modèle où les paramètres sont modifiés c'est à dire un modèle non contraint que nous pouvons écrire de la forme $Y = X\beta + u$ habituelle en découpant la période d'observation en deux sous-périodes : de 1960 à 1973 (période 1), date du premier choc pétrolier puis de 1974 à 1995 pour la période 2. Commençons par un bref rappel théorique avant d'estimer ce modèle non contraint. Sous forme matricielle le modèle s'écrit :

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \tag{9}$$

où l'indice 1 (resp. 2) signifie qu'il s'agit de données sur la première (resp. seconde) période.
L'estimateur non contraint est obtenu de la manière suivante :

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} X_1'X_1 & 0 \\ 0 & X_2'X_2 \end{bmatrix}^{-1} \begin{bmatrix} X_1'Y_1 \\ X_2'Y_2 \end{bmatrix}$$

Ce qui correspond aux paramètres estimés séparément sur les deux périodes. La somme totale des Carrés des Résidus de ce modèle non contraint est donc la somme des SCR de chaque sous période : $SCR_{nc} = SCR_1 + SCR_2$ où l'indice nc désigne le modèle non contraint comme précedemment et où l'indice 1 ou 2 correspond à la période.
Revenons à notre échantillon. Le modèle initial décomposé en deux sous périodes s'écrit :
$Log(Consot)_t = \alpha_1 + \beta_1 Log(rt)_t + \gamma_1 Log(pg)_t + u_t$ pour $t = 1960 - 1973$
$Log(Consot)_t = \alpha_2 + \beta_2 Log(rt)_t + \gamma_2 Log(pg)_t + u_t$ pour $t = 1974 - 1995$

Estimons le modèle séparement sur les deux sous-périodes.

```
Data P1 ;set tpfoad.logessence ;
if annee<=1973 ;run ;
proc reg data=p1 ;
model lconsot=lrt lpg ;run ;
```

La sortie SAS est la suivante :

```
Dependent Variable: lconsot

            Number of Observations Read          14
            Number of Observations Used          14
```

Analyse de variance

Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	2	0.32481	0.16241	262.51	<.0001
Error	11	0.00681	0.00061866		
Corrected Total	13	0.33162			

Root MSE	0.02487	R-Square	0.9795
Dependent Mean	-0.13830	Adj R-Sq	0.9757
Coeff Var	-17.98519		

ésultats estimés des paramètres

Variable	DF	Résultat estimé des paramètres	Erreur std	Valeur du test t	Pr > |t|
Intercept	1	-7.85866	1.55875	-5.04	0.0004
lrt	1	0.86832	0.17501	4.96	0.0004
lpg	1	0.56378	0.29181	1.93	0.0795

Nous notons $SCR_1 = 0.00681$.
De même pour la période 2 :

```
Data P2 ;set tpfoad.logessence ;
if annee > 1973 ;run ;
proc reg data=p2 ;
model lconsot=lrt lpg ;run ;
```

Number of Observations Used 22

Analyse de variance

Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	2	0.05077	0.02539	60.34	<.0001
Error	19	0.00799	0.00042070		
Corrected Total	21	0.05876			

Root MSE	0.02051	R-Square	0.8640

```
Dependent Mean        0.08194    Adj R-Sq     0.8497
Coeff Var            25.03217
```

Résultats estimés des paramètres

Variable	DF	Résultat estimé des paramètres	Erreur std	Valeur du test t	Pr > |t|
Intercept	1	-6.16733	0.61540	-10.02	<.0001
lrt	1	0.69968	0.06820	10.26	<.0001
lpg	1	-0.19985	0.01999	-10.00	<.0001

On note $SCR_2 = 0.00799$.
Avec l'écriture du modèle contraint et les deux listings de SAS correspondant aux deux périodes, l'hypothèse à tester s'écrit $H_0 : \alpha_1 = \alpha_2, \beta_1 = \beta_2, \gamma_1 = \gamma_2$ où l'indice 1 ou 2 représente la période. La statistique de test [26] est la suivante :

$$F = \frac{[SCR_c - (SCR_1 + SCR_2)]/p}{SCR_{nc}/(N-k)} \sim F(p, N-k) \tag{10}$$

où $p = 3$, le nombre de contraintes de H_0.
La valeur observée pour cette statistique de Fisher est ici égale à :

$$F = \frac{(0.06955 - (0.00681 + 0.00799))/3}{(0.00681 + 0.00799)/(36-6)} = 36.9932$$

Utilisons SAS pour calculer la probabilité de dépasser la valeur observée en ajoutant la ligne : p=1-probf(36.9932,3,30) ;
Cette probabilité est très petite [27] et donc on rejette H_0 qui suppose la stabilité des paramètres. Ainsi nous avons confirmation que la fonction de demande d'essence a été modifiée par le premier choc pétrolier. Nous ferons les commentaires des paramètres estimés à la fin du paragraphe suivant sur les variables indicatrices car à ce stade nous pouvons juste conclure que la demande d'essence a été modifiée après le premier choc pétrolier mais pour étudier les changements dans les paramètres il est plus facile d'utiliser des variables indicatrices. En effet nous avons obtenu des paramètres pour chaque sous-période mais nous ne savons pas encore si la différence entre les paramètres du revenu par exemple est significative. Nous verrons dans le paragraphe suivant un moyen très simple de répondre à ce type de questions.
Remarque pédagogique : dans la documentation de SAS et R, vous trouverez des procédures ou des fonctions qui effectuent ce type de test masi je préfère que vosu sachiez programmer les calculs et non pas que vosu utilisiez un logiciel comme une "boîte noire".

26. nous avons déjà utilisé cette formule pour comparer la fonction de production Cobb-Douglas et la fonstion Translog
27. elle est égale à 3.3286E-10

VI Les variables indicatrices

Les variables indicatrices ou variables muettes sont très utilisées en économétrie appliquée. Nous proposons ici quelques utilisations sur des fichiers de données économiques.
Définition :
Une variable indicatrice, notée D, ne peut prendre que deux valeurs 0 ou 1 par convention. Dans les paragraphes qui suivent nous allons détailler les différentes utilisations de ces variables indicatrices.

VI.1 Codage d'une var. quali à 2 modalités

Source : L'enquête Budget des Familles (BDF) de l'INSEE en 2001 comporte 10305 ménages. Nous nous intéressons ici aux 2284 ménages constitués d'une seule personne (célibataire, divorcé ou veuf). Nous souhaitons expliquer la dépense en "café et restaurants" de ces ménages.
FICHIER : depcafe
LISTE des variables :
- Depense : la dépense totale en cafés et restaurants réalisée par le ménage au cours de la période d'observation.
- sexe : 1 pour les hommes et 2 pour les femmes
- age codé en 5 modalités :
1. moins de 25 ans
2. de 25 à moins de 35 ans
3. de 35 à moins de 45 ans
4. de 45 à moins de 55 ans
5. de 55 à moins de 65 ans
- urban (mesure le degré d'urbanisation du lieu d'habitation) :
1. commune rurale
2. unité urbaine de moins de 20000 habitants
3. unité urbaine de 20000 à moins de 100000 habitants
4. unité de 100000 habitants et plus (hors région Parisienne)
5. Paris et sa région.
- revenu : revenu perçu par le ménage (pas d'indication sur les unités).

Comme toujours commençons par exécuter une "proc means" sur les variables.

```
proc means data=tpfoad.depcafe; run;
```

La sortie SAS est la suivante :

Variable	Nb	Moyenne	Écart-type	Minimum	Maximum
depense	2284	753.6094571	1264.71	0	13447.00

revenu	2284	8453.89	6183.75	0	68975.00
sexe	2284	1.6225919	0.4848444	1.0000000	2.0000000
age	2284	3.1773205	1.2660161	1.0000000	5.0000000
urban	2284	3.2784588	1.3171766	1.0000000	5.0000000

Je commence toujours par regarder les valeurs du minimum et du maximum pour chaque variable. Plusieurs commentaires sont à faire. Commençons d'abord par remarquer que le minimum du revenu est égal à 0. La première question à se poser est " Combien y a t-il de ménages avec un revenu nul ?" Pour y répondre il suffit d'exécuter :

```
data r ;set tpfoad.depcafe ;if revenu=0 ;run ;
```

Nous observons qu'il y a seulement 2 ménages (voir le journal de SAS) qui ont un revenu nul [28]. Etant donné que ce nombre est très faible comparé à la taille de l'échantillon, nous proposons de supprimer ces deux observations [29].
Remarquons ensuite que le minimum de la dépense est aussi nul ;certains ménages ne vont pas dans les cafés et restaurants (au moins sur cette période). Même programme SAS que ci-dessus :

```
data dep ;set tpfoad.depcafe ;if depense=0 ;run ;
```

Il y a cette fois-ci 690 ménages concernés. Dans ce cours de M1, nous allons nous contenter de supprimer ces 690 ménages mais la théorie et la pratique économétriques sur ce problème seront étudiées dans le cours "Econométrie 2" du Master 2. Nous pouvons signaler tout de même qu'en supprimant ces observations nous introduisons un biais de sélection.
Donc supprimons ces 690 ménages :

```
data depense ;set tpfoad.depcafe ;if depense ne 0 ;run ;
```

Le fichier contient 1594 ménages dont la dépense en "café et restaurant" est non nulle [30].
Remarquons enfin que la variable "sexe" est codée 1 ou 2 ;nous la transformons en variable indicatrice de la manière suivante :

```
data tpfoad.depense ;set depense ;s=sexe-1 ;run ;
```

Le codage de la nouvelle variable indicatrice notée s , est 0 pour les hommes et 1 pour les femmes. Le codage n'a en fait pas d'importance. Il faut toujours vérifier la création de la variable s avec une proc freq :

```
proc freq data= tpfoad.depense ;tables sexe s ;run ;
```

Etant donné que les effectifs sont identiques, les deux variables sont cohérentes [31].

28. Si, par exemple, on classe les ménages par revenu décroissant on observe des ménages avec un revenu faible mais nous n'avons pas d'indication sur la variable revenu donc nous nous contentons de supprimer les "revenus nuls".
29. il est inutile de les supprimer pour l'instant.
30. Il n'y a plus les 2 individus qui avaient un revenu nul ; ils avaient aussi une dépense nulle.
31. il aurait été plus judicieux d'appeler cette variable femme car elle vaut 1 pour les femmes.

Nous avons sauvé ce fichier dans le répertoire “TPFOAD" car il n'y a pas d'autres instructions à appliquer au fichier de données initial pour l'instant.
Nous souhaitons donc estimer une fonction de dépense et savoir en particulier s'il existe une différence significative entre les hommes et les femmes quant à la dépense en “café et restaurant". Le modèle est le suivant : pour $i = 1, \ldots, N$

$$Depense_i = \beta_0 + \beta_1 revenu_i + \beta_2 S_i + \beta_3 Age_i + \beta_4 Urban_i + u_i \qquad (11)$$

La sortie SAS est la suivante :

Number of Observations Used 1594

Analyse de variance

Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	4	751296604	187824151	127.61	<.0001
Error	1589	2338819400	1471881		
Corrected Total	1593	3090116004			

Root MSE	1213.21116	R-Square	0.2431
Dependent Mean	1079.82685	Adj R-Sq	0.2412
Coeff Var	112.35238		

Résultats estimés des paramètres

Variable	DF	Résultat estimé des paramètres	Erreur std	Valeur du test t	Pr > |t|
Intercept	1	273.06586	128.61148	2.12	0.0339
revenu	1	0.09468	0.00487	19.45	<.0001
s	1	-652.10031	61.90148	-10.53	<.0001
age	1	-30.48839	25.37573	-1.20	0.2297
urban	1	113.26455	24.27503	4.67	<.0001

Commentaires :
- Toutes les variables sont significatives sauf la variable “age".
- Remarquons que cette regression contient toutes les variables explicatives dont nous disposons dans le fichier. Nous commencerons toujours par le modèle complet

Cette variable indique donc les femmes.

(celui qui contient toutes les variables explicatives) car un modèle réduit se heurterait à un problème de biais de variables omises. Il est facile de démontrer que l'oubli d'une variable significative dans la liste des variables explicatives conduit à des estimateurs biaisés (Voir Annexe du chapitre 1). En fait nous commencerons toujours par le modèle le plus complet possible ; il faut toujours se poser la question : " Y a t- il une variable ou des variables importante(s) qui pourraient influencer la dépense et que nous avons oubliées ?" Dans le modèle que nous estimons il y a la variable "revenu" ainsi que 3 variables "caractéristiques individuelles" qui tiennent compte du fait qu'à revenu identique les individus (les femmes et les hommes ou les habitants des grandes villes par rapport aux ruraux etc...) peuvent avoir des dépenses différentes.
Question : a t- on oublié une variable importante ? Si vous vous souvenez de notre exemple sur la consommation (ou dépense) d'essence , il y avait non seulement une variable revenu mais aussi une variable prix. Cette variable prix ne figure pas dans l'enquête BDF mais pour toute étude empirique sérieuse il faudrait trouver une autre source de données pour obtenir cette variable prix [32].
- remarquons que nous avons une illustration du fait que le R^2 peut sembler faible sur des données individuelles en particulier sur des fonctions de demande (de consommation) comme c'est le cas ici. Il y a sûrement d'autres facteurs sociologiques ou pschychologiques qui expliquent les comportements d'achat des ménages, variables qui ne figurent pas dans les bases de données "habituelles" en économie.
- le paramètre de la variable s s'interprète comme la différence de dépense entre les femmes (codées 1) et les hommes (codées 0). Ainsi les femmes ont une dépense moyenne en café et restaurants inférieure de 652 unités (de la variable dépense) par rapport à la dépense des hommes.
- On peut calculer une élasticité revenu. Reprenons la formule générale de l'élasticité :

$$\text{élasticité de Y par rapport à X } = \frac{\frac{dY}{Y}}{\frac{dX}{X}} = \frac{dLnY}{dLnX}$$

Nous avons utilisé le membre de droite quand nous avons estimé une fonction de production Cobb-Douglas car les variables étaient spécifiées en Log. Ici ce n'est pas le cas et nous allons donc utiliser le membre de gauche ce qui donne :

$$\text{élasticité de Y par rapport à X } = \frac{\frac{dY}{Y}}{\frac{dX}{X}} = \frac{\frac{dY}{dX}}{\frac{Y}{X}} = \frac{dY}{dX}\frac{X}{Y}$$

Le terme $\frac{dY}{dX}$ est ici la dérivée de la dépense par rapport au revenu c'est à dire simplement le paramètre du revenu ; pour quantifier l'élasticité précédente , nous

32. Pour pouvoir identifier un effet-prix il faut un vecteur de prix qui varie entre les différentes catégories de commune par exemple. Si cela n'est pas le cas et si les individus font face au même vecteur de prix alors nous ne pourrons pas identifier l'effet de cette variable qui se confond avec la constante du modèle.

allons remplacer $\frac{X}{Y}$ par $\frac{\bar{X}}{\bar{Y}}$ c'est à dire que nous allons évaluer l'élasticité au point moyen ;c'est la pratique habituelle. Sur cet échantillon, l'élasticité estimée de la dépense par rapport au revenu est donc égale à $0.09468 \times \frac{9440.35}{1079.83} = 0.82773$. Ce qui signifie que la dépense en café et restaurant augmente de 0.83% quand le revenu du ménage augmente de 1%.
- l'introduction de la variable "urban" pose problème car cette spécification repose sur une hypothèse implicite dont il faut être conscient [33]. On pourrait dans un premier temps recoder cette variable en une variable indicatrice qui vaudrait 1 si la commune est rurale et 0 sinon. On obtient :

```
data indic ;
set tpfoad.depense ;if urban=1 then rural=1 ;else rural=0 ;run ;
proc freq data=indic ;tables urban rural ;run ;
```

On remplace ensuite la variable "urban" par cette variable indicatrice rural.
La sortie SAS est la suivante :

Number of Observations Used 1594

Analyse de variance

Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	4	723034568	180758642	121.34	<.0001
Error	1589	2367081436	1489667		
Corrected Total	1593	3090116004			

Root MSE	1220.51930	R-Square	0.2340
Dependent Mean	1079.82685	Adj R-Sq	0.2321
Coeff Var	113.02917		

Résultats estimés des paramètres

Variable	DF	Résultat estimé des paramètres	Erreur std	Valeur du test t	Pr > |t|
Intercept	1	701.89262	91.48566	7.67	<.0001
revenu	1	0.09630	0.00488	19.72	<.0001
s	1	-645.63073	62.29374	-10.36	<.0001

33. ce sera l'objet du paragraphe suivant.

age	1	-44.32775	25.45834	-1.74	0.0818
rural	1	-152.56619	95.75643	-1.59	0.1113

Commentaires :
- le paramètre de la variable revenu est sensiblement le même donc l'élasticité est identique
- la variable Age est devenue significative à 10%
- la variable *rural* n'est pas significative même à 10%.
Ainsi l'utilisation d'une variable indicatrice "rural" n'est pas convaincante (a posteriori) puisque cette variable n'est pas significative. Pour expliquer la dépense il ne suffit pas de distinguer les deux catégories "rural" ou "pas". Il faut peut être introduire la taille des villes comme nous allons le faire dans le paragraphe suivant.

VI.2 Codage d'une variable qualitative à plusieurs modalités

Revenons sur le problème de l'introduction de la variable "urban" codée et 5 modalités et de l'hypothèse implicite que son introduction suppose.
La variable "urban" comporte les 5 modalités suivantes :
1. commune rurale
2. unité urbaine de moins de 20000 habitants
3. unité urbaine de 20000 à moins de 100000 habitants
4. unité de 100000 habitants et plus (hors région Parisienne)
5. Paris et sa région.
Revenons sur le modèle (**??**) qui contenait la variable "Urban". Il s'écrivait :

$$Depense_i = \beta_0 + \alpha Age_i + \beta_1 revenu_i + \beta_2 S_i + \beta_3 Urban_i + u_i$$

avec $i = 1, \ldots, N$
Pour comprendre l'hypothèse implicite de ce modèle nous allons détailler l'équation précédente de la dépense pour chaque modalité de la varialbe "urban". Nous obtenons :
Si urban=1, $Depense_i = \beta_0 + \alpha Age_i + \beta_1 revenu_i + \beta_2 S_i + \beta_3 + u_i$
Si urban=2, $Depense_i = \beta_0 + \alpha Age_i + \beta_1 revenu_i + \beta_2 S_i + \beta_3 + \beta_3 + u_i$
Si urban=3, $Depense_i = \beta_0 + \alpha Age_i + \beta_1 revenu_i + \beta_2 S_i + \beta_3 + \beta_3 + \beta_3 + u_i$
Si urban=4, $Depense_i = \beta_0 + \alpha Age_i + \beta_1 revenu_i + \beta_2 S_i + \beta_3 + \beta_3 + \beta_3 + \beta_3 + u_i$
Si urban=5, $Depense_i = \beta_0 + \alpha Age_i + \beta_1 revenu_i + \beta_2 S_i + \beta_3 + \beta_3 + \beta_3 + \beta_3 + \beta_3 + u_i$
En comparant ces 5 équations, nous pouvons dire que β_3 représente l'augmentation de la dépense consécutive à une "augmentation de 1" de la variable urban quelle que soit le degré d'urbanisation. Toutes choses égales par ailleurs, c'est à dire à revenu et sexe constants, passer d'une commune rurale à une commune urbaine de 20000 habitants augmente la dépense de β_3 ; cette augmentation de la dépense est identique quand on passe d'une commune de de 20000 habitants à une commune urbaine de 20000 à moins de 100000 habitants etc...Implicitement l'augmentation de la dépense est constante quand on introduit la variable urban dans le modèle. En fait il est fort possible que cette augmentation ne soit pas constante ; en tous cas il est absolument

nécessaire de tester cette hypothèse avant de l'imposer au modèle. Commençons par estimer un modèle où cette hypothèse n'est pas imposée, un modèle non contraint. Ce modèle va contenir 5 variables indicatrices , une pour chaque modalité de la variable urban. Le programme SAS ci dessous crée ces 5 variables et stocke le fichier dans la libriarie Tpfoad.

```
data tpfoad.idepense;set tpfoad.depense;
if urban= 1 then rur=1;else rur=0;
if urban=2 then ville1=1;else ville1=0;
if urban=3 then ville2=1;else ville2=0;
if urban=4 then ville3=1;else ville3=0;
if urban=5 then paris=1;else paris=0;
run;
```

Après avoir vérifié le codage avec la proc freq, nous estimons ensuite le modèle suivant :

$$Depense_i = \beta_0 + + \alpha Age_i + \beta_1 revenu_i + \beta_2 S_i + \alpha_1 rur_i + \alpha_2 ville1_i + \alpha_3 ville2_i + \alpha_4 ville3_i + \alpha_5 paris_i + u_i$$

avec $i = 1, \ldots, N$
Le programme SAS est le suivant :

```
proc reg data=tpfoad.idepense;
model depense=revenu s age rur ville1-ville3 paris;
run;
```

La sortie SAS est la suivante :

```
Number of Observations Read        1594
Number of Observations Used        1594


                              Analyse de variance

                                   Somme des          Carré     Valeur
Source                      DF        carrés          moyen          F    Pr > F

Model                        7     762162170      108880310      74.18    <.0001
Error                     1586    2327953835        1467815
Corrected Total           1593    3090116004


Root MSE           1211.53395    R-Square     0.2466
Dependent Mean     1079.82685    Adj R-Sq     0.2433
Coeff Var           112.19706

NOTE: Model is not full rank. Least-squares solutions for the parameters are not uniqu
```

NOTE: The following parameters have been set to 0, since the variables are a linear co

paris = Intercept - rur - ville1 - ville2 - ville3

Résultats estimés des paramètres

Variable	DF	Résultat estimé des paramètres	Erreur std	Valeur du test t	Pr > |t|
Intercept	B	964.33947	110.85429	8.70	<.0001
revenu	1	0.09373	0.00488	19.20	<.0001
s	1	-645.07691	61.88555	-10.42	<.0001
age	1	-35.36180	25.40937	-1.39	0.1642
rur	B	-422.55597	112.79889	-3.75	0.0002
ville1	B	-484.86934	108.37980	-4.47	<.0001
ville2	B	-436.81523	103.82810	-4.21	<.0001
ville3	B	-258.61848	83.87982	-3.08	0.0021
paris	0	0	.	.	.

SAS détecte un problème : “Model is not full rank" c'est à dire que la matrice X n'est pas de rang plein en colonne car il existe une combinaison linéaire entre les colonnes de cette matrice ;SAS nous donne la combinaison linéaire (CBL) sous la forme "paris = Intercept - rur - ville1 - ville2 - ville3" ce qui peut être écrit de la manière suivante :$rur + ville1 + ville2 + ville3 + paris = 1$ pour tou i. Nous nous heurtons à un problème de colinéarité exacte : il existe une CBL qui est vérifiée pour chaque observation ; le déterminant de $X'X$ est nul , on ne peut pas inverser la matrice$X'X$ et donc on ne peut pas calculer les paramètres estimés ; c'est pour cette raison que SAS élimine la variable “paris" en posant son paramètre à zéro. En effet une des solutions à ce problème de colinéarité exacte est de supprimer une des variables indicatrices. Ici SAS a choisi de supprimer la variable “ paris" mais c'est arbitraire. Une autre solution consiste à supprimer la constante (intercept) mais dans ce cas on ne peut pas interpréter le R^2 qui en particulier peut devenir négatif ; donc cette seconde solution est rarement utilisée en pratique. Revenons à la solution de supprimer une variable indicatrice ; la variable indicatrice supprimée est appelée modalité de référence car nous allons raisonner par rapport à cette modalité. Il n'est pas judicieux de choisir Paris comme modalité de référence et donc comme élément de comparaison ;nous préférons raisonner par rapport à une ville hors paris et supprimer ville2 ou ville3 par exemple ;en fait pour le premier essai la modalité supprimée n'a pas tellement d'importance. Nous y reviendrons. Estimons donc le modèle sans ville2 que nous écrivons
$Depense_i = \beta_0 + +\alpha Age_i + \beta_1 revenu_i + \beta_2 S_i + \alpha_1 rur_i + \alpha_2 ville1_i + \alpha_4 ville3_i + \alpha_5 paris_i + u_i$ avec $i = 1, \ldots, N$
Vérifions tout d'abord que l'hypothèse implicite que nous avons mentionnée plus haut n'est pas imposée dans ce modèle (qui est donc un modèle non contraint).

Nous allons pour cela “détailler" le modèle c'est à dire écrire une équation de dépense pour chaque modalité de la variable urban :
Si $rur_i = 1$, $Depense_i = \beta_0 + +\alpha Age_i + \beta_1 revenu_i + \beta_2 S_i + \alpha_1 + u_i$
Si $ville1 = 1$, $Depense_i = \beta_0 + +\alpha Age_i + \beta_1 revenu_i + \beta_2 S_i + \alpha_2 + u_i$
Si $ville3 = 1$,$Depense_i = \beta_0 + +\alpha Age_i + \beta_1 revenu_i + \beta_2 S_i + \alpha_4 + u_i$
Si $paris = 1$, $Depense_i = \beta_0 + +\alpha Age_i + \beta_1 revenu_i + \beta_2 S_i + \alpha_5 + u_i$
pour la modalité de référence ($ville2 = 1$), nous avons $Depense_i = \beta_0 + +\alpha Age_i + \beta_1 revenu_i + \beta_2 S_i + u_i$
L'augmentation de la dépense n'est donc pas supposée constante a priori dans ce modèle. En effet l'augmentation de la dépense est égale à $\alpha_2 - \alpha_1$ quand on passe d'une commune rurale à une commune de moins de 20000 habitants (passage de rur à ville1). Cette augmentation est égale à α_2 quand on passe d'une commune de moins de 20000 habitants à une commune de 20000 à moins de 100000 habitants etc... Les augmentations de la dépense sont donc a priori différentes dans ce modèle. De plus, en comparant ces équations entre elles il est facile de comprendre pourquoi la modalité supprimée est appelée “modalité de référence". En effet α_1 représente la différence entre la dépense d'une commune rurale et d'une commune de 20000 habitants à moins de 100000 habitants, α_2 la différence entre la dépense d'une commune de moins de 20000 habitants et d'une commune de 20000 habitants à moins de 100000 habitants etc...La comparaison se fait toujours par rapport à une commune de 20000 habitants à moins de 100000 habitants.
La sortie SAS est la suivante :

Number of Observations Read	1594
Number of Observations Used	1594

Analyse de variance

Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	7	762162170	108880310	74.18	<.0001
Error	1586	2327953835	1467815		
Corrected Total	1593	3090116004			

Root MSE	1211.53395	R-Square	0.2466
Dependent Mean	1079.82685	Adj R-Sq	0.2433
Coeff Var	112.19706		

Résultats estimés des paramètres

	Résultat estimé	Erreur

Variable	DF	des paramètres	std	Valeur du test t	Pr > |t|
Intercept	1	527.52424	115.32905	4.57	<.0001
revenu	1	0.09373	0.00488	19.20	<.0001
s	1	-645.07691	61.88555	-10.42	<.0001
age	1	-35.36180	25.40937	-1.39	0.1642
rur	1	14.25927	118.65606	0.12	0.9044
ville1	1	-48.05411	114.46383	-0.42	0.6747
ville3	1	178.19675	91.55871	1.95	0.0518
paris	1	436.81523	103.82810	4.21	<.0001

Commentaires :
- Dans ce modèle les variables non significatives sont : age et les deux variables indicatrices *rur* et *ville*1. Quand *rur* n'est pas significative cela signifie que la différence entre la dépense d'une commune rurale et la dépense d'une commune de référence c'est à dire une commune de 20000 habitants à moins de 100000 habitants n'est pas significative. De même quand *ville*1 n'est pas significative cela signifie que la différence entre la dépense d'une commune de moins de 20000 habitans et celle d'une commune de 20000 habitants à moins de 100000 habitants n'est pas significative.
- Pour commenter les paramètres estimés des variables indicatrices de la variable "urban", nous allons " détailler" le modèle c'est à dire écrire une équation de dépense pour chaque modalité de la variable "urban" de la manière suivante :
Le modèle est le suivant :
$Depense = 527.52424 + 0.09373 Revenu - 645.07691 S + 178.19675 ville3 + 436.81523 Paris + \hat{u}$.
Por interpréter le modèle nous avons enlevé les deux variables non significatives c'est à dire *rur* et *ville*1 ;de plus nous avons omis l'indice i pour alléger les notations
Si $rur = 1$, $Depense = 527.52424 + 0.09373 Revenu - 645.07691 S + \hat{u}$
Si $ville1 = 1$, $Depense = 527.52424 + 0.09373 Revenu - 645.07691 S + \hat{u}$
Si $ville2 = 1$, $Depense = 527.52424 + 0.09373 Revenu - 645.07691 S + \hat{u}$
Si $ville3 = 1$, $Depense = 527.52424 + 0.09373 Revenu - 645.07691 S + 178.19675 + \hat{u}$
Si $paris = 1$, $Depense = 527.52424 + 0.09373 Revenu - 645.07691 S + 436.81523 + \hat{u}$
Etant donné que deux variables indicatrices *rur* et *ville*1 ne sont pas significatives, on voit bien dans les équations précédentes que les paramètres estimés s'interprètent non seulement par rapport à la modalité de référence mais aussi par rapport aux variables non significatives. Comment interpréter la valeur du paramètre estimé 178.19675 ? ce paramètre est la différence entre l'équation de ville3=1 et les 3 équations de rur=1 ou ville1=1 ou ville2=1. On peut donc dire que les habitants d'une unité urbaine de plus de 100000 habitants dépensent 178.19675 unités monétaires de plus que les habitants d'une unité rurale ou d'une unité urbaine de plus petite taille. De même pour la dernière équation, les habitants de Paris dépensent 436.81523 unités monétaires de plus que les habitants d'une commune rurale ou d'une unité urbaine de moins de 100000 habitants.
Pour terminer ce paragraphe remarquons que la variable "age" est aussi codée en plusieurs modalités et que nous pouvons faire le même raisonnement que pour la

variable “urban". A titre d'exercice creér les 5 variables indicatrices de l'age puis introduire 4 variables indicatrices en prenant comme modalité de référence les individus entre 35 et 45 ans. Commenter l'effet de l'age sur la dépense en café et restaurant.

REPONSE

Le programme SAS est le suivant :

```
data tpfoad.idepense ;set tpfoad.depense ;
if urban= 1 then rur=1 ;else rur=0 ;
if urban=2 then ville1=1 ;else ville1=0 ;
if urban=3 then ville2=1 ;else ville2=0 ;
if urban=4 then ville3=1 ;else ville3=0 ;
if urban=5 then paris=1 ;else paris=0 ;
if age=1 then m25=1 ;else m25=0 ;
if age=2 then m35=1 ;else m35=0 ;
if age=3 then m45=1 ;else m45=0 ;
if age=4 then m55=1 ;else m55=0 ;
if age=5 then m65=1 ;else m65=0 ;
proc reg data=tpfoad.idepense ;
model depense= revenu s m25 m35 m55 m65 rur ville1 ville3 paris ;
run ;
```

Nous avons juste ajouté des lignes au programme de création des variables indicatrices de la variable “urban" sans changer le nom du fichier.

Après avoir vérifié le codage avec la proc freq, nous avons estimé ensuite le modèle dont la sortie SAS est la suivante :

```
Number of Observations Read        1594
Number of Observations Used        1594
```

Analyse de variance

Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	10	780306915	78030692	53.48	<.0001
Error	1583	2309809089	1459134		
Corrected Total	1593	3090116004			

Root MSE	1207.94618	R-Square	0.2525
Dependent Mean	1079.82685	Adj R-Sq	0.2478
Coeff Var	111.86480		

Résultats estimés des paramètres

Variable	DF	Résultat estimé des paramètres	Erreur std	Valeur du test t	Pr > |t|
Intercept	1	290.84840	109.63477	2.65	0.0081
revenu	1	0.09369	0.00493	19.02	<.0001
s	1	-647.65967	61.72993	-10.49	<.0001
m25	1	141.68035	105.42891	1.34	0.1792
m35	1	292.20985	86.03032	3.40	0.0007
m55	1	201.05378	86.88547	2.31	0.0208
m65	1	29.26077	100.15344	0.29	0.7702
rur	1	5.37726	118.43842	0.05	0.9638
ville1	1	-53.39530	114.15410	-0.47	0.6400
ville3	1	176.55120	91.44871	1.93	0.0537
paris	1	425.52436	103.65795	4.11	<.0001

Commentaires :
- Les deux modalités ville3 et paris sont toujours les seules variables significatives pour la variable “urban".
- Les paramètres estimés de ville3 et paris sont à peu près identiques.
- les variables indicatrices significatives pour l'âge sont m35 et m55. Etant donné que la modalité de référenec est m45 et que m25 et m65 ne sont pas ignificatives on peut dire que la dépense en café et restaurant des individus entre 25 et 35 ans est supérieure de 292 unités par rapport aux individus de moins de 25 ans, entre 35 et 45 ans et des individus de 55 à 65 ans. De plus, les individus entre 45 à 5(ans ont une dépense supérieure de 201 unités par rapprot aux individus de moins de 25 ans, entre 35 et 45 ans et des individus de 55 à 65 ans.
Pour l'instant nous avons seulement étudié le cas où au premier essai nous obtenons des variables indicatrices significatives. Que se passe-til si cela n'est pas le cas et si au premier essai aucune variable indicatrice de l'âge est significative ? Prenons une autre modalité de référence comme m25 par exemple ; on obtient :

Number of Observations Read	1594
Number of Observations Used	1594

Analyse de variance

Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	10	780306915	78030692	53.48	<.0001
Error	1583	2309809089	1459134		
Corrected Total	1593	3090116004			

Root MSE	1207.94618	R-Square	0.2525
Dependent Mean	1079.82685	Adj R-Sq	0.2478
Coeff Var	111.86480		

Résultats estimés des paramètres

Variable	DF	Résultat estimé des paramètres	Erreur std	Valeur du test t	Pr > |t|
Intercept	1	432.52875	119.56175	3.62	0.0003
revenu	1	0.09369	0.00493	19.02	<.0001
s	1	-647.65967	61.72993	-10.49	<.0001
m35	1	150.52950	105.95926	1.42	0.1556
m45	1	-141.68035	105.42891	-1.34	0.1792
m55	1	59.37343	109.90071	0.54	0.5891
m65	1	-112.41958	119.59098	-0.94	0.3473
rur	1	5.37726	118.43842	0.05	0.9638
ville1	1	-53.39530	114.15410	-0.47	0.6400
ville3	1	176.55120	91.44871	1.93	0.0537
paris	1	425.52436	103.65795	4.11	<.0001

Imaginons que la sortie SAS ci-dessus représente le premier essai d'estimation avec ces variables indicatrices de l'âge et de l'urbanisation. Ainsi c'est le premier listing qu'il nous faut commenter. Il ne faut surtout pas écrire "Les 4 variables indicatrices de l'âge ne sont pas significatives donc l'âge n'a aucun effet sur la dépense en café et restaurant". Effectivement aucune variable de l'âge n'est significative mais on ne peut pas conclure que l'âge n'a pas d'effet sur la dépense. Les seuls commentaires corrects sont :
- il n'y a pas de différence significative entre la dépense d'un individu de moins de 25 ans et un individu entre 25 et 35 ans car m35 n'est pas significative. [34]
- il n'y a pas de différence significative entre la dépense d'un individu de moins de 25 ans et un individu entre 35 et 45 ans car m45 n'est pas significative
- etc...

Ainsi si lors de notre premier essai nous observons qu'aucune variable indicatrice de l'âge n'est significative, nous devons changer de modalité de référence et d'éventuellement de toutes les essayer (sauf une) si nous n'avons pas de chance et que c'est seulement à la dernière modalité de référence que nous avons une variable indicatrice significative. Si , après avoir essayé toutes les modalités de référence nous n'avons toujours pas de variable indicatrice significative alors et alors seulement nous pouvons conclure que l'âge n'a pas d'effet significatif sur la dépense.
Revenons à la spécification du modèle où m45 est la modalité de référence ; nous retrouvons les résultats du modèle précédent que nous reproduisons en partie ci-dessous :

34. Souvenez vous que nous raisonnons par rapport à la modalité de référence

Variable	DF	Résultat estimé des paramètres	Erreur std	Valeur du test t	Pr > |t|
Intercept	1	290.84840	109.63477	2.65	0.0081
revenu	1	0.09369	0.00493	19.02	<.0001
s	1	-647.65967	61.72993	-10.49	<.0001
m25	1	141.68035	105.42891	1.34	0.1792
m35	1	292.20985	86.03032	3.40	0.0007
m55	1	201.05378	86.88547	2.31	0.0208
m65	1	29.26077	100.15344	0.29	0.7702
rur	1	5.37726	118.43842	0.05	0.9638
ville1	1	-53.39530	114.15410	-0.47	0.6400
ville3	1	176.55120	91.44871	1.93	0.0537
paris	1	425.52436	103.65795	4.11	<.0001

Commentaire de l'effet de l'âge :
- Les individus qui ont entre 35 et 45 ans dépensent 292.20985 unités monétaires de plus que les individus qui ont entre 35 et 45 ans (ainsi que ceux qui ont moins de 25 ans car m25 n'est pas significative et ceux qui ont entre 55 et 65 ans car m65 n'est pas significative).
- Les individus qui ont entre 45 et 55 ans dépensent 201.05378 unités monétaires de plus que les individus qui ont entre 35 et 45 ans (ainsi que ceux qui ont moins de 25 ans car m25 n'est pas significative et ceux qui ont entre 55 et 65 ans car m65 n'est pas significative).
A ce stade de notre étude ce modèle représente la meilleure spécification de la dépense en café et restaurants.

VI.3 Interactions entre variable quantitative et indicatrice

On utilise une variable "interaction" quand on veut tester l'existence d'un effet différent de la variable quantitative pour chaque modalité d'une variable indicatrice. Sur notre exemple, on souhaite tester si l'effet du revenu sur la dépense (et donc l'élasticité du revenu) est différent pour les hommes et les femmes. On créee une variable interaction qui est le produit de la variable *revenu* par la variable indicatrice S et on l'ajoute au modèle. Le programme SAS est donc :

```
data tpfoad.inter ;set tpfoad.idepense ;
rs=revenu*s ;run ;
proc reg data=tpfoad.inter ;
model depense= revenu s rs m25 m35 m55 m65 rur ville1 ville3 paris ;
run ;
```

La sortie SAS est donnée ci-après :

```
Number of Observations Read        1594
```

Number of Observations Used 1594

Analyse de variance

Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	11	835695809	75972346	53.31	<.0001
Error	1582	2254420196	1425044		
Corrected Total	1593	3090116004			

Root MSE	1193.75222	R-Square	0.2704
Dependent Mean	1079.82685	Adj R-Sq	0.2654
Coeff Var	110.55034		

Résultats estimés des paramètres

Variable	DF	Résultat estimé des paramètres	Erreur std	Valeur du test t	Pr > |t|
Intercept	1	8.00991	117.46124	0.07	0.9456
revenu	1	0.12709	0.00724	17.55	<.0001
s	1	-95.60557	107.52913	-0.89	0.3741
rs	1	-0.05915	0.00949	-6.23	<.0001
m25	1	104.92665	104.35672	1.01	0.3148
m35	1	250.87622	85.27753	2.94	0.0033
m55	1	185.44475	85.90102	2.16	0.0310
m65	1	-12.14802	99.19919	-0.12	0.9025
rur	1	22.34680	117.07836	0.19	0.8487
ville1	1	-41.12755	112.82990	-0.36	0.7155
ville3	1	177.69795	90.37433	1.97	0.0494
paris	1	430.10108	102.44255	4.20	<.0001

Commentaires :
- pour les variables indicatrices de l'âge et de l'urbanisation, la liste des variables significatives est identique au modèle sans la variable interaction "rs". Les paramètres estimés sont comparables
- la variable s n'est plus significative
- la variable d'interaction qui nous intéresse ici est significative ce qui signifie que l'on ne peut pas supposer que l'effet du revenu sur la dépense est le même pour les hommes et pour les femmes. Pour identifier cette différence, nous procédons de la même manière que d'habitude ; nous détaillons le modèle pour les hommes et pour les femmes de la manière suivante :
Pour les hommes, $depense = 8 + 0.12709 Revenu + \ldots$

Pour les femmes, $depense = 8 + 0.12709Revenu - 0.05915Revenu + \ldots$ ou encore $depense = 8 + 0.06794Revenu + \ldots$
où les points de suspension correspondent aux variables et aux paramètres estimés de l'âge et de l'urbanisation qui ne nous intéressent pas dans ces commentaires.
En comparant les deux équations, on constate que l'effet du revenu sur la dépense en café et restaurant est un peu plus faible pour les femmes que pour les hommes toutes choses égales par ailleurs c'est à dire à âge et degré d'urbanisation constants.
Nous pouvons utiliser les effets du revenu pour calculer une élasticité estimée du revenu pour les hommes et pour les femmes (au point moyen). Le programme SAS de calcul est le suivant :

```
*elasticité pour h et f;
proc sort data=tpfoad.inter out=tri;by s;run;
proc means data=tri;var depense revenu;by s;run;
data elast;
ehom=0.12709*(9032.85/1420.64);
efem=0.06794*(9726.07/840.8591);run;
proc print data=elast;
run;
```

Remarque : Avant de faire une proc means avec l'instruction "by s;", il faut que le fichier soit trié selon cette variable "S". De plus , les élasticités estimées sont calculées au point moyen des hommes puis au point moyen des femmes.
Dans le fichier SAS elast, nous trouvons que l'élasticité revenu pour les hommes est égale à 0.80808 et celle des femmes à 0.75585 ce qui est un peu plus faible.
Conclusion : à nouveau la meilleure spécification de la dépense est un modèle qui contient les variables indicatrices de l'âge, de l'urbanisation, une variable revenu, une variable indicatrice pour le sexe et une variable interaction "revenu*s". En général les économètres laissent les variables non significatives dans le modèle car le fait qu'elles ne soient pas significatives est une information à part entière. De plus, on peut montrer qu'il n'y a aucun biais à laisser une variable non significative dans unerégression.
Pour terminer il faudrait commenter les paramètres des variables indicatrices comme nous l'avons déjà fait. Il n'y a aucune difficulté.

VI.4 Exercice sur les variables indicatrices et les interactions : retour sur le fichier de consommation d'essence

Nous avons estimé deux modèles sur cet échantillon :

- un modèle contraint pour lequel les paramètres sont stables sur la période :

$Log(Consot)_t = \alpha + \beta Log(rt)_t + \gamma Log(pg)_t + u_t$ pour $t = 1960 - 1995$

- un modèle non contraint en estimant les paramètres sur deux sous périodes :

$$\begin{cases} Log(Consot)_t = \alpha_1 + \beta_1 Log(rt)_t + \gamma_1 Log(pg)_t + u_t & \text{pour } t = 1960 - 1973 \\ Log(Consot)_t = \alpha_2 + \beta_2 Log(rt)_t + \gamma_2 Log(pg)_t + u_t & \text{pour } t = 1974 - 1995 \end{cases}$$

Questions et réponses :

1. Ecrire le modèle non contraint sous forme matricielle puis estimer ce modèle.
   Réponse : Pour le modèle non contraint sous forme matricielle on obtient la matrice X suivante (peu importe l'ordre des variables ou colonnes) :

$$X = \begin{pmatrix} 1 & 0 & lrt_{1960} & 0 & lpg_{1960} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & lrt_{1973} & 0 & lpg_{1973} & 0 \\ 0 & 1 & 0 & lrt_{1974} & 0 & lpg_{1974} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & lrt_{1995} & 0 & lpg_{1995} \end{pmatrix}$$

   Ainsi pour estimer ce modèle il faut créer les 6 variables de cette matrice X avec le programme SAS suivant :

```
data tpfoad.chow ;set tpfoad.logessence ;
if annee <=1973 then D1=1 ;else d1=0 ;
if annee>=1974 then D2=1 ;else D2=0 ;
lpg1=lpg*D1 ;lpg2=lpg*D2 ;lrt1=lrt*D1 ;lrt2=lrt*D2 ;
run ;
proc reg data=tpfoad.chow ;
model lconsot=D1 D2 lrt1 lrt2 lpg1 lpg2 /noint ;
run ;
```

   Commentaires :
   - parmi les 6 variables créées dans le fichier tpfoad.chow il y a deux variables indicatrices, D1 et D2 ;les 4 autres variables sont des interactions.
   - dans l'écriture matricielle il n'y a pas de constante ;pour enlever la constante avec SAS il faut utiliser l'option "noint" (no intercept) de l'instruction model. Dans ce cas , le R^2 ne peut plus être interprété de la manière habituelle.
   - cette écriture matricielle nous permet d'écrire le MNC sous la forme suivante : $Log(Consot)_t = \alpha_1 D1_t + \alpha_2 D2_t + \beta_1 Lrt1_t + \beta_2 Lrt2_t + \gamma_1 Lpg1_t + \gamma_2 Lpg2_t + u_t$ pour $t = 1960 - 1995$
   La sortie SAS est la suivant :

```
 Dependent Variable: lconsot

Number of Observations Read          36
Number of Observations Used          36


NOTE: No intercept in model. R-Square is redefined.

                                  Analyse de variance

                                  Somme des          Carré     Valeur
```

Source	DF	carrés	moyen	F	Pr > F
Model	6	0.79105	0.13184	267.27	<.0001
Error	30	0.01480	0.00049328		
Uncorrected Total	36	0.80585			

Root MSE	0.02221	R-Square	0.9816
Dependent Mean	-0.00371	Adj R-Sq	0.9780
Coeff Var	-598.87751		

Résultats estimés des paramètres

Variable	DF	Résultat estimé des paramètres	Erreur std	Valeur du test t	Pr>|t|
D1	1	-7.85866	1.39187	-5.65	<.0001
D2	1	-6.16733	0.66638	-9.25	<.0001
lrt1	1	0.86832	0.15628	5.56	<.0001
lrt2	1	0.69968	0.07385	9.47	<.0001
lpg1	1	0.56378	0.26057	2.16	0.0386
lpg2	1	-0.19985	0.02165	-9.23	<.0001

Nous remarquons que la SCR de ce modèle est exactement la même, 0.01480, que pour le MNC estimé dans le paragraphe sur le test de Chow. Ce qui est un moyen de vérifier que je ne me suis pas trompée. Nous avons donc obtenu une deuxième méthode pour estimer le MNC, la première étant d'estimer les paramètres sur deux sous périodes (la première méthode est celle présentée ci dessus en estimant le modèle sur deux sous périodes).

2. Avec SAS comparer le modèle précédent au modèle contraint c'est à dire celui où tous les paramètres sont restés stables sur la période.
Réponse : il suffit d'ajouter l'instruction "test D1=D2,lrt1=lrt2,lpg1=lpg2 ;" dans le programme précédent et nous obtenons :

The REG Procedure
Test 1 Results for Dependent Variable lconsot

Source	DF	Carré moyen	Valeur F	Pr > F
Numerator	3	0.01825	37.00	<.0001
Denominator	30	0.00049328		

On retrouve la même valeur pour la statistique de Fisher que dans le paragraphe sur le test de Chow aux erreurs d'arrondis près : cette statistique était

égale à 36.9932.

3. Dans la question précédente nous avons testé la stabilité des paramètres en posant l'égalité des paramètres des deux sous périodes. Trouver une spécification qui permette de tester les 3 contraintes séparement (avec une Student) sans faire de calcul.
   Réponse : L'intuition est la suivante : prenons la constante ; pour estimer une constante différente pour les deux périodes nous avons "séparé" la constante en deux parties à l'aide de deux indicatrices. On peut aussi garder la constante et ajouter un terme pour la première ou la seconde période. dans ce cas aussi nous aurons une constante différente pour les deux périodes. De même pour tester la satbilité de la variable lrt on peut introduire $lrt1$ et $lrt2$ ou ici introduire lrt et une partie supplémentaire pour la première période $lrt1$. Il suffit de faire le même raisonnement pour la dernière variable $lpgt$. On obtient le modèle suivant :
   $Log(Consot)_t = \delta_0 + \delta_1 D1_t + \delta_2 Lrt_t + \delta_3 Lrt1_t + \delta_4 Lpg_t + \delta_5 Lpg1_t + u_t$ pour $t = 1960 - 1995$

   La sortie SAS de ce modèle est donnée ci-après :

```
Number of Observations Read          36
Number of Observations Used          36
```

Analyse de variance

Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr>F
Model	5	0.79056	0.15811	320.53	<.0001
Error	30	0.01480	0.00049328		
Corrected Total	35	0.80535			

Root MSE	0.02221	R-Square	0.9816
Dependent Mean	-0.00371	Adj R-Sq	0.9786
Coeff Var	-598.87751		

Résultats estimés des paramètres

Variable	DF	Résultat estimé des paramètres	Erreur std	Valeur du test t	Pr>|t|

Intercept	1	-6.16733	0.66638	-9.25	<.0001
D1	1	-1.69133	1.54317	-1.10	0.2818
lrt	1	0.69968	0.07385	9.47	<.0001
lrt1	1	0.16865	0.17285	0.98	0.3370
lpg	1	-0.19985	0.02165	-9.23	<.0001
lpg1	1	0.76364	0.26147	2.92	0.0066

Sur cette spécification on peut tester la stabilité individuelle d'un paramètre : par exemple la constante peut être supposée stable individuellement car la variable $D1$ n'est pas significative ;de même le paramètre de lrt peut être supposé stable individuellement car la variable $lrt1$ n'est pas significative. Par contre le paramètre de la variable lpg ne peut pas être supposé stable.
Attention il s'agit de tests individuels. Pour effectuer un test de plusieurs contraintes il faut utiliser une statistique de Fisher comme dans la question suivante.

4. Avec SAS comparer ce modèle à un modèle contraint où tous les paramètres sont restés stables.
Le programme est le suivant :

```
proc reg data=m8 ;
model lconsot= d1 lrt lrt1 lpg lpg1 ;
test d1=0,lrt1=0,lpg1=0 ;
run ;
```

The REG Procedure Model: MODEL1

Test 1 Results for Dependent Variable lconsot

Source	DF	Carré moyen	Valeur F	Pr > F
Numerator	3	0.01825	37.00	<.0001
Denominator	30	0.00049328		

La probabilité associée à ce test est très faible, on rejette l'hypothèse de stabilité des paramètres.

5. Commenter les paramètres estimés .
Réponse : Dans le paragraphe où nous avons étudié la stabilité des paramètres sur ce fichier (test de Chow) j'ai indiqué à la fin du paragraphe que :“ Nous ferons les commentaires des paramètres estimés à la fin du paragraphe suivant sur les variables indicatrices car à ce stade nous pouvons juste conclure que la demande d'essence a été modifiée après le premier choc pétrolier mais

pour étudier les changements dans les paramètres il est préférable d'uiliser des variables indicatrices. En effet nous avons obtenu des paramètres pour chaque sous-période mais nous ne savons pas encore si la différence entre les paramètres du revenu par exemple est significative. Nous verrons dans le paragraphe suivant un moyen très simple de répondre à ce type de questions. ". C'est l'objet de cette question.
Avec la spécification que nous avons retenue nous avons vu qu' il est très simple de tester la stabilité de la constante par exemple [35]. Il suffit de tester $H_0 : D1 = 0$ en regardant la probabilité de dépasser la valeur observée dans le listing de SAS. Cette probabilité est égale à 0.2818 donc on ne rejette pas l'hypothèse de stabilité de la constante après le premier choc pétrolier.
De même pour tester la stabilité de l'élasticité-revenu il faut tester $H_0 : lrt1 = 0$;la probabilité est égale à 0.3370 donc on ne rejette pas l'hypothèse de stabilité de l'élasticité-revenu après le premier choc pétrolier. L'élasticité-revenu est donc égale à 0.69968 (environ 0.7) sr la période.
Enfin l'élasticité-prix n'est pas stable sur la période car la probabilité de la variable lpg1 est égale à 0.0066. Ainsi avant le premier choc pétrolier l'élasticité-prix était de +0.56379 (-0.19985 + 0.76364) alors qu'elle était égale à -0.19985 après le premier choc [36]. Une élasticité-prix positive peut se produire dans deux situations : un bien de première nécessité ou un bien de luxe (voir élasticité de la demande sur wikipédia) ;ici on constate donc que l'essence était un bien de première nécéssité c'est à dire un bien dont le prix élevé nous oblige à renoncer à d'autres biens (si le revenu reste constant). Le fait que l'élasticité soit devenue négative montre que les individus ont adpaté leur comportement de consommation après le premier choc pétrolier.

VII Conclusion du chapitre 1 : la spécification - retour sur le fichier de la demande de café et restaurant

Plusieurs questions distinctes se posent à l'économètre appliqué ; nous tentons de répondre à certaines d'entre elles dans cette conclusion

La première question qu'un économètre peut se pose est la suivante : Faut - il transformer les variables en log ?
Ce choix dépend de l'interprétation des paramètres que l'on souhaite comme indiqué dans le tableau suivant [37] :

Modèle	Var. endogène	Var. explicative	interprétation du paramètre
niveau-niveau	Y	X	$\Delta Y = \beta \Delta X$
Log-Log	Log(Y)	Log(X)	$\%\Delta Y = \beta\ \%\Delta X$
Log-niveau	Log(Y)	X	$\%\Delta Y = (100\beta)\ \Delta X$

35. dans le modèle qui contient lpg1 et lrt1
36. mêmes paramètres estimés que dans le modèle estimé séparément sur les deux sous périodes
37. Source Wooldridge

Remarque il n'y a pas de ligne pour un modèle "Niveau - Log" car cette spécification n'est pas souvent utilisée en économie (nous pouvons en discuter sur les forums).
De plus, et de manière standard dans les études appliquées [38], les variables exprimées en unités monétaires (en euros, en dollar ...) comme le salaire ou encore les ventes d'une entreprise, sont transformées en log ainsi que les variables telles que la population, le nombre de salariés ou encore le nombre total d'étudiants qui sont des nombres entiers "grands". Par contre les variables mesurées en années (l'âge l'experience ou l'ancienneté) sont laissées en niveau. Les variables exprimées en pourcentage comme le taux de chômage par exemple, sont en général laissées en niveau. Ceci n'est pas une règle absolue mais plutôt une observation de la littérature empirique.

- Quand les variables sont spécifiées en niveau faut -il ajouter X^2 dans la liste des variables explicatives ? En fait on ajoute X^2 si on veut tenir compte de la croissance ou décroissance des effets marginaux [39] de X sur Y (ou sur Log(Y)).
Revenons à la pratique : Dans tous les modèles que nous avons estimés jusqu'à présent, nous avons supposé implicitement que chaque variable explicative avait un effet constant sur la variable endogène. On peut se demander si cette hypothèse est correcte ; en particulier, sur le fichier de dépense en café et restaurants, l'effet de la variable revenu n'est peut être pas constant. On peut imaginer un effet de saturation : la croissance de la dépense serait de moins en moins grande avec l'augmentation du revenu ou encore , l'augmentation de la dépense est décroissante avec le revenu : pour les individus avec un revenu " faible" l'augmentation d'une unité du revenu produit un effet plus grand sur la dépense que pour les individus avec un revenu " plus élevé". Ainsi on peut représenter un tel modèle, modèle 2 en rouge, sur le graphique suivant.
Le modèle qui ne contient pas le revenu au carré est la droite en noir. Si le "vrai modèle" est le modèle 2 qui contient le revenu au carré nous savons déjà que les paramètres estimés d'un modèle qui ne contient pas le revenu au carré, modèle 1, seront biaisés.
Sur ce graphique, on peut facilement comprendre que nous disposons d'un moyen de détection [40] de l'omission de la variable X^2 (ici revenu au carré) en étudiant le signe des résidus. Si le "vrai" modèle est le modèle 2 alors le nuage des points sera "réparti autour" de la courbe du modèle 2 ;si nous estimons le modèle 1, les résidus seront d'abord négatifs puis positfs et à nouveau négatifs. Etudions le signe des résidus sur le modèle sans interaction du revenu et du sexe pour simplifier [41]. Pour sauver les résidus de ce modèle dans un fichier il faut utiliser l'instruction "output" puis éxecuter le programme suivant pour obtenir le graphique des résidus :

38. Source Wooldridge
39. ou dérivées
40. détection graphique
41. vous pouvez essayer sur le dernier modèle de dépense en café et restaurant que nous avons estimé dans le paragraphe sur les interactions c'est à dire le modèle qui contient le croisement "rs=revenu*s"

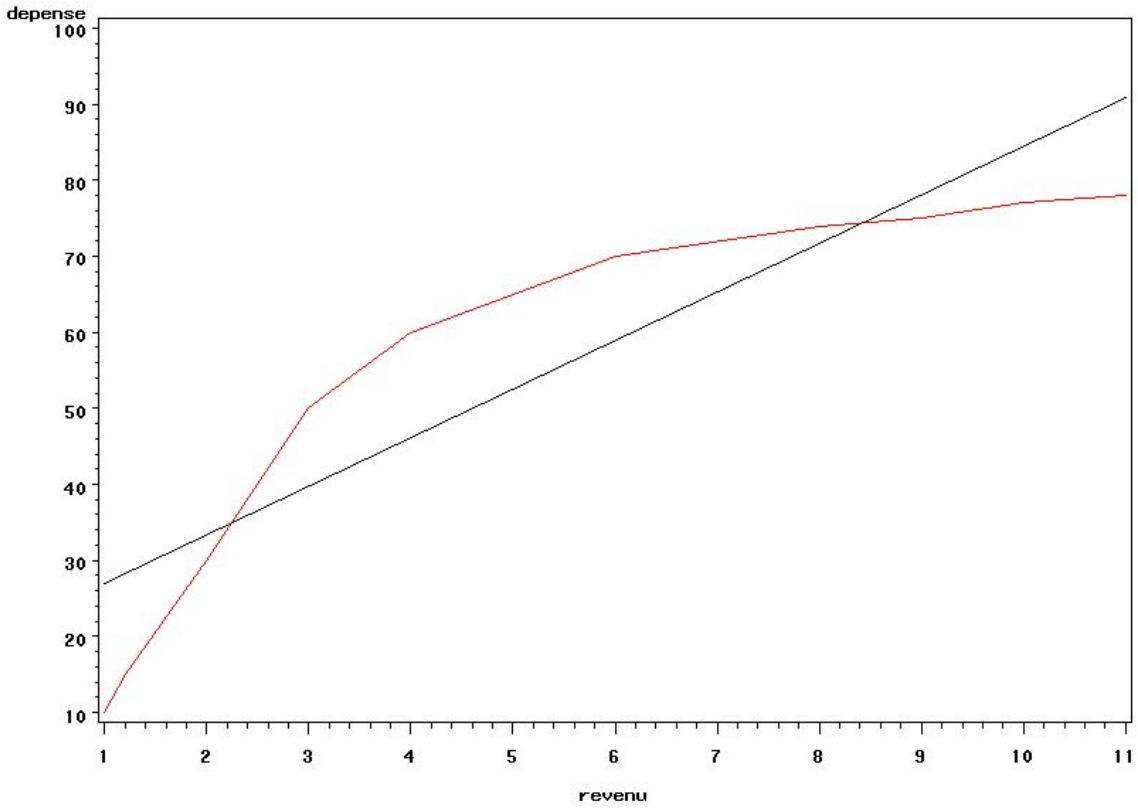

FIGURE 4 – Problème de spécification : M1 contre M2

```
proc reg data=tpfoad.inter ;*peu importe le fichier ;
model depense= revenu s m25 m35 m55 m65 rur ville1 ville3 paris ;
output out=outres r=res ;
quit ;
run ;
proc gplot data=outres ;
plot res*revenu ;run ;
```

Nous obtenons le graphique suivant :

Deux commentaires : tout d'abord ce graphique **??** illustre parfaitement la difficulté de confronter la théorie et la pratique car il n'est pas clair du tout que le signe des résidus soit négatif, positif puis à nouveau négatif [42].
Enfin et à nouveau il est préférable d'utiliser un argument économique pour introduire le revenu au carré : la dépense en café et restaurant augmente peut être (nous allons le tester) plus rapidement pour les individus avec un revenu "faible".
Pour modéliser un effet non constant du revenu nous pouvons ajouter une variable $revenu^2$ au modèle [43] et simplement tester si elle est significative.

42. par contre il est tout à fait clair que la dispersion des résidus est une fonction croissante du revenu. Ce problème d'hétéroscédasticité sera étudié dans le chapitre 2
43. qui reste linéaire dans les paramètres.

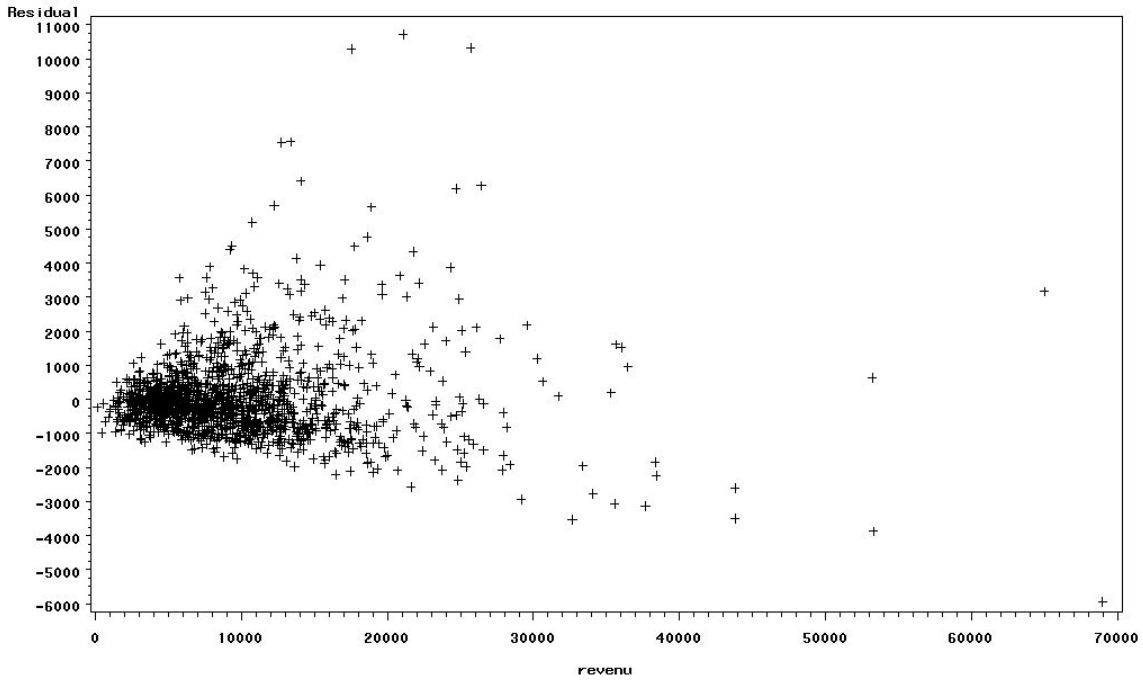

FIGURE 5 – Graphique des résidu du M1

Donc si on ajoute la variable revenu2=revenu*revenu dans le modèle on obtient :

```
Number of Observations Read        1594
Number of Observations Used        1594


                                Analyse de variance

                                Somme des          Carré     Valeur
Source                     DF      carrés          moyen          F    Pr > F

Model                      11   820593152       74599377      52.00    <.0001
Error                    1582  2269522852        1434591
Corrected Total          1593  3090116004

Root MSE           1197.74410    R-Square     0.2656
Dependent Mean     1079.82685    Adj R-Sq     0.2604
Coeff Var           110.92001

                  Résultats estimés des paramètres

                  Résultat estimé         Erreur
Variable     DF     des paramètres           std    Valeur du test t    Pr > |t|
```

Intercept	1	14.83384	120.54259	0.12	0.9021
revenu	1	0.14197	0.01034	13.73	<.0001
revenu2	1	-0.00000134	2.532843E-7	-5.30	<.0001
s	1	-665.48070	61.30088	-10.86	<.0001
m25	1	204.08905	105.19975	1.94	0.0526
m35	1	306.74364	85.34780	3.59	0.0003
m55	1	182.62251	86.22184	2.12	0.0343
m65	1	28.71725	99.30762	0.29	0.7725
rur	1	-5.97826	117.45767	-0.05	0.9594
ville1	1	-56.57769	113.19157	-0.50	0.6173
ville3	1	176.84598	90.67637	1.95	0.0513
paris	1	424.63590	102.78261	4.13	<.0001

Commentaires :
- les variables non significatives sont : s, m25, m65, rur et ville1.
- le programme SAS pour calculer l'élasticité estimée au point moyen est le suivant :

```
*elast ;
proc means data=tpfoad.inter ;var depense revenu ;run ;
data elastM2 ;
d= 0.14197-(2*0.00000134* 9440.35) ;
rap=9440.35/1079.83 ;
elast2=d*rap ;run ;
proc print data=elastM2 ;run ;
*elastM2= 1.01998 ;
```

L'élasticité estimée du revenu [44] pour le M2 est égale à 1.01998 ce qui est un peu plus élevé que l'estimation de l'élasticité revenu du Modèle 1 qui était de 0.82773 et que nous savons biaisée. Ainsi on peut arrondir l'élasticité revenu à 1.

Conclusion sur la spécification :
Pour introduire un effet non linéaire du revenu nous avons introduit le revenu au carré. Nous aurions pu essayer d'introduire des tranches de revenu avec des variables indicatrices. En fait nous aurions essayé la modélisation avec des variables indicatrices si la variable revenu n'était pas significative ce qui n'est le cas ici.

VIII Annexe du chapitre 1

VIII.1 Propriétés des MCO en échantillon fini

1. L'estimateur des MCO est sans biais :
   $\hat{\beta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + u) = \beta + X'X)^{-1}X'u$. Nous prenons

44. au point moyen

l'espérance et nous obtenons :
$E(\hat{\beta}/X) = \beta + E(X'X)^{-1}X'u/X)$. Le second terme est nul par l'hypothèse H_1 et donc $E(\hat{\beta}/X) = \beta$.

2. Calcul de la Variance de $\hat{\beta}$:
Nous avons $\hat{\beta} = \beta + X'X)^{-1}X'u$ et donc $Var(\hat{\beta}/X) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'/X] =$
$E[(X'X)^{-1}X'uu'X(X'X)^{-1}/X] = (X'X)^{-1}X'E(uu'/X)X(X'X)^{-1} = (X'X)^{-1}X'(\sigma^2 I)X(X'X)^{-1}$
$= \sigma^2(X'X)^{-1}$

VIII.2 Biais des variables omises

source W. H. Greene

Supposons que le "vrai" modèle ou le modèle "correct" soit :
$Y = X_1\beta_1 + X_2\beta_2 + u$ où X_1 et X_2 ont respectivement k_1 et k_2 colonnes. Si on regresse Y sur X_1 en omettant X_2 l'estimateur est :
$b_1 = (X_1'X_1)^{-1}X_1'Y = (X_1'X_1)^{-1}X_1'(X_1'\beta_1 + X_2'\beta_2 + u) = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 +$
$(X_1'X_1)^{-1}X_1'u$
En prenant l'espérance, on obtient :
$E(b_1/X) = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2$.
Ansi b_1 est biaisé à moins que $X_1'X_2 = 0$ ou $\beta_2 = 0$.