TESTS de STUDENT ET FISHER

Rappelons les hyphothèses des MCO:

$\mathrm{H}_0^1 : n \geq k$

H_0^2 La matrice X des variables explicatives est de plein rang, alors ${}^tXX$ est de rang k et donc inversible.

Hypothèses sur les erreurs

H_1^1 : les erreurs sont des variables aléatoires, les variables explicatives sont non aléatoires

H_1^2 : les erreurs ont une espérance nulle $E(\epsilon_t) = 0$ pour tout t

H_1^3 : les erreurs ont une variance constante inconnue notée σ^2, $\forall t \rightarrow V(\epsilon_t) = \sigma^2$ on parlera d'homoscédasticité

H_1^4 : les covariances des erreurs sont nulles $Cov(\epsilon_t, \epsilon_{t+h}) = 0$ Non Autocorrélation

H_1^5: si n$\rightarrow \infty$ alors $\lim\limits_{n\rightarrow\infty} \frac{{}^tXX}{n} = V_X$ matrice finie définie positive.

H_1^6: les erreurs sont indépendantes et suivent la même loi inconnue (i..i.d.)

H_1^7 :les erreurs suivent la même loi Normale $N(0, \sigma^2)$

1 Les tests de student

Ils portent sur les coefficients a_i constants mais inconnus.

1.1 Le test de base

C'est le test d'un coefficient du modèle $Y=a_0+a_1X_1+ a_2X_2 + ...+a_kX_{k-1} +\epsilon$ modèle à k variables explicatives dont l'unité (correspondant au terme constant)
On veut tester si un coefficient est égal ou non à une valeur donnée le test

$$\begin{aligned} H_0 &: a_i = b \ , \quad \text{b étant une constante donnée} \\ H_1 &: a_i \neq b \end{aligned}$$

Le cas particulier important en économétrie correspond à b=0. Si on trouve que dans un modèle le test de Student indique un coefficient significatif (décision H1) alors on peut admettre la variable correspondante comme intéressante dans le modèle, elle apporte de l'explication.
Si on trouve que dans un modèle le test de Student indique un coefficient non significatif (décision H0) alors on peut admettre ce coefficient comme nul et donc la variable comme non intéressante dans le modèle. Cette seconde conclusion n'est pas toujours vrai quand il existe une forte colinéarité entre les variables. Pour être sur que la variable est négligeable dans le modèle, il faut construire un **nouveau modèle** sans cette variable et si le s du nouveau modèle est inférieur ou presque égal au modèle avec cette variable c'est qu'en effet elle n'avait pas de rôle explicatif. Si par contre le s a augmenté alors la variable était intéressante et la décision H0 était due seulement au problèmes de colinéarité. (Voir chapitre sur la colinéarité).

1.2 Construction du test

1.2.1 Choix de la variable aléatoire

$\overrightarrow{\widehat{a}} = (\, ^tXX)^{-1} \, ^tX\overrightarrow{Y}$ est l'estimateur des MCO de $\overrightarrow{a}$ c'est le meilleur si toutes les hypothèses de base sont vérifiées.

1.2.2 loi de la variable sous l'hypothèse $H_{\grave{a}}$

Nous avons vu (dans le capitre propriétés des MCO) qu'alors $\overrightarrow{\widehat{a}}$ suit une loi $N_k(\overrightarrow{a}, \sigma^2(\, ^tXX)^{-1}$ non dégénérée. En conséquence, chaque $\widehat{a_i}$ suit une loi $N(a_i, \sigma^2 h_{ii})$. Si on nomme h_{ij} l'élément sur la ligne i et la colonne j de la matrice $(\, ^tXX)^{-1}$, la variance de $\widehat{a_i}$ est sur la ligne i et colonne i de cette matrice.

$$\frac{\widehat{a_i} - a_i}{\sigma\sqrt{h_{ii}}} \text{ suit une loi N(0,1)}$$

σ étant inconnu nous avons vu que le meilleur estimateur de σ^2 était $s^2 = SCR/(n-k) = \sum_1^n e_t^2/(n-k)$, nous savons aussi que

$$\frac{(n-k)s^2}{\sigma^2} = \frac{\sum_1^n e_t^2}{\sigma^2} \text{ suit la loi du } \chi^2_{n-k}$$

En utilisant la définition de la loi de student comme rapport d'une variable normale centrée réduite et de la racine d'un χ^2 divisé par son nombre de degrés de liberté et indépendant de la variable normale on trouve :

$$\frac{\frac{\widehat{a_i}-a_i}{\sigma\sqrt{h_{ii}}}}{\sqrt{\frac{\sum_1^n e_t^2}{\sigma^2(n-k)}}} = \frac{\widehat{a_i} - a_i}{\sqrt{s^2 h_{ii}}} = \frac{\widehat{a_i} - a_i}{s\sqrt{h_{ii}}} \text{ suit la loi de Student à (n-k) degrés de liberté}$$

Les variables $\widehat{a_i}$ et s^2 sont indépendantes car elles sont issues dans le cadre de la loi Normale de deux sous-ensembles orthogonaux H_k et H_{n-k}

$$\text{Si } H_0 \text{ vraie alors } \frac{\widehat{a_i} - b}{s\sqrt{h_{ii}}} = T_{n-k}$$
$$\text{Si } H_1 \text{ vraie alors } \frac{\widehat{a_i} - a_i}{s\sqrt{h_{ii}}} = T_{n-k}$$

1.2.3 Région critique

L'estimation des MCO donne la valeur de $\widehat{a_i}$ estimation dans l'échantitllon de taille n de a_i,on calcule $\frac{\widehat{a_i}-b}{s\sqrt{h_{ii}}} = \frac{\widehat{a_i}-b}{\text{écart-type estimé de } \widehat{a_i}} = T$
Si T est proche de 0, alors $\widehat{a_i}$ est proche de b et on décide H_0, sinon on décide H_1. la borne T_α est fournie par la table de Student au risque α.

$$\underline{\quad\quad} -T\alpha \underline{\quad\quad\quad\quad\quad\quad\quad\quad} 0 \underline{\quad\quad\quad\quad\quad\quad\quad\quad} T\alpha \underline{\quad\quad\quad\quad}$$

H_1 H$_0$ H$_1$

1.3 Cas le plus utilisé : b=0

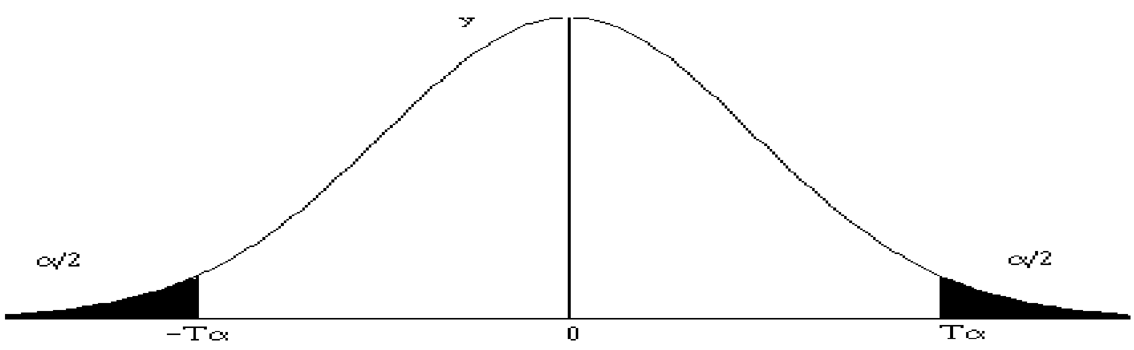
tester la nullité des coefficients de ce modèle. Ce test particulier de nullité des coefficients est donné directement dans RATS

$$Y = a_0 + a_1 X1 + \epsilon$$

```
Linear Regression - Estimation by Least Squares
Dependent Variable Y
Usable Observations    140      Degrees of Freedom   138
Centered R**2     0.915883      R Bar **2   0.915274
Uncentered R**2   0.993195      T x R**2     139.047
Mean of Dependent Variable      692268.63700
Std Error of Dependent Variable 206128.93876
Standard Error of Estimate       59999.61073
Sum of Squared Residuals         4.96794e+11
Regression F(1,138)                1502.5742
Significance Level of F           0.00000000
Log Likelihood                   -1737.93725
Durbin-Watson Statistic             0.120471

   Variable                     Coeff       Std Error      T-Stat     Signif
*******************************************************************************
1.  Constant                 -68443.58769  20269.23173     -3.37672  0.00095323
2.  X1                            7.40009      0.19091     38.76305  0.00000000
```

1.3.1 Test de a_0

Rats donne le coefficient estimé = -68443.587 , son écart-type estimé $= s\sqrt{h_{ii}} =$ 20269.23173 et le rapport des deux $= \frac{\widehat{a_i} - 0}{s\sqrt{h_{ii}}} =$-3.37672
si on fait les tests au risque α =5%
pour le coefficient a_0 test H0 : $a_0 = 0$ contre H1: $a_0 \neq 0$ la valeur de la statistique T estimée est de -3.3767. ici n étant grand $T\alpha = 1.96$



On en déduit H1 le coefficient de la constante n'est pas nul donc la variable a un intéret dans le modèle

1.3.2 Test de a_1

Rats donne le rapport $\frac{\widehat{a1} - 0}{s\sqrt{h_{11}}} = 38.76305$ très supérieur à 1.96 on décide donc très nettement H_1, la variavle est très importante dans le modèle. Rappelons que ce n'est pas parce qu'un coefficient est grand qu'il est important dans le modèle, nous avons vu que sa grandeur dépend juste des unités et non de son intéret (voir chapitre sur RATS).

2 Test de Fisher

Il est utile pout tester des contraintes sur non plus un mais plusieurs coefficients.
On va présenter ce test à l'aide d'un exemple.

2.1 Exemples de tests de Fisher

Avec toujours le même exemple $Y = b_0 + b_1X_1 + b_2X_2 + \epsilon$ on va effectuer le test

$$\begin{aligned} H_0 &: \; b_1 = 1. \text{ et } b_2 = O.5 \\ H_1 &: \; \text{l'une des deux hypothèses au moin est fausse} \end{aligned}$$

MCO sur le modèle de base

```
Linear Regression - Estimation by Least Squares
Dependent Variable Y
Usable Observations    140      Degrees of Freedom   137
Centered R**2     0.999998      R Bar **2   0.999998
Uncentered R**2   1.000000      T x R**2     140.000
Mean of Dependent Variable      692268.63700
Std Error of Dependent Variable 206128.93876
Standard Error of Estimate         291.88445
Sum of Squared Residuals        11671924.919
Regression F(2,137)             34660910.1766
Significance Level of F           0.00000000
Log Likelihood                    -991.82521
Durbin-Watson Statistic             1.753973

   Variable                     Coeff       Std Error      T-Stat     Signif
*******************************************************************************
1.  Constant                 834.92234985 102.69414057      8.13018  0.00000000
2.  X1                         0.80325350   0.00288544    278.38191  0.00000000
3.  X2                         0.49984476   0.00020700   2414.74856  0.00000000
```

Si maintenant on utilise la contrainte $Y = b_0 + 1. * X_1 + 0.5 * X_2 + \epsilon$
$Y - X_1 - 0.5X_2 = b_0 + \epsilon$
On estime le modèle $Y - X_1 - 0.5X_2 = b_0 + \epsilon = Z = b_0 + \epsilon$ seule la constante est donc estimée dans cet exemple.

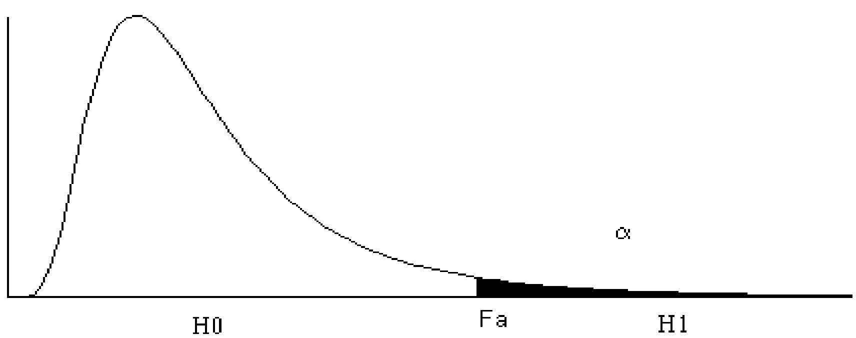
```
Linear Model - Estimation by Restricted Regression
Dependent Variable Y
Usable Observations    140      Degrees of Freedom   139
Mean of Dependent Variable      692268.63700
Std Error of Dependent Variable 206128.93876
Standard Error of Estimate        5307.36674
Sum of Squared Residuals        3915371695.7
Durbin-Watson Statistic             0.017286

   Variable                     Coeff       Std Error      T-Stat     Signif
*******************************************************************************
1.  Constant                 -19579.262998   448.554358     -43.64970  0.00000000
2.  X1                           1.000000     0.000000  1.60667e+09  0.00000000
3.  X2                           0.500000     0.000000  3.07521e+09  0.00000000
```

2.1.1 Construction du test de Fisher

Sous l'hypothèse H_1 on a la somme des carrés des résidus SCR_a (indice a car souvent on parle d'hypothèse alternative) et SCR_a/σ^2 suit une loi du χ^2_{n-k} ici χ^2_{140-3}
Sous l'hypothèse H_0 la somme des carrés des est SCR_0 et comme on a 1 seule variable explicative SCR_0/σ_1^2 suit une loi du χ^2_{n-1} ici χ^2_{140-1}. Dans le cas général si on a k variables explicatives et r contraintes, le modèle sous H_0 ne contient plus que k-r variables et donc le χ^2 aura n-(k-r) = n-k+r degrés de liberté.
La variable de Fisher est le rapport de deux χ^2 indépendants divisés par leur nombre de degrés de liberté. SI H_0 est vraie, sous les deux hypothèses la variance des erreurs est la même $\sigma^2 et \;\; (SCR_0 - SCR_a)/\sigma^2$ suit un χ^2 a $(n-k+r-(n-k)) = r$ degrés de liberté on a donc

$$\frac{SCR_0 - SCR_a}{\sigma^2 r} / \frac{SCR_a}{\sigma^2(n-k)} = F(r, n-k) = \frac{SCR_0 - SCR_a}{SCR_a} . \frac{n-k}{r}$$

Remarque : comme $SCR_0 > SCR_a$ le fisher est toujours positif.



Dans notre exemple les résultats de Rats donnent les estimations SCR_a =11671924.919, n-k=137 , r=2, SCR_0= 3915371695.7 et $F_\alpha = 19,5$ dans la table de Fisher (2,137)
$\frac{SCR_0 - SCR_a}{SCR_a} . \frac{n-k}{r}$ =22909.97 on decide donc très nettement H_1 l'une au moins des deux hypothèses n'est pas vérifiée.

2.2 Cas particulier: test de Fisher global

Il existe un test de Fisher qui dans un modèle avec terme constant permet de tester si tous les coefficients sont nuls sauf le terme constant.
Modèle à k-1 variables explicatives X_1, ...X_{k-1} autres que la constante

$$Y_t = a_0 + a_1 X_{1t} + a_2 X_{2t} + ... + a_{k-1} X_{k-1t} + \epsilon_t$$

H0: tous les coefficients nuls sauf la constante soit $a_0 \neq 0$ et $a_1 = a_2 = ... = a_{k-1} = 0$

H1: au moins un des $a_i \neq 0$ pour i=1 à k-1

- Sous l'hypothèse H1 on effectue les MCO sur l'échantillon global avec toutes les variables. La variable $SRC1/\sigma^2$ suit un χ^2_{n-k}

- Sous l'hypothèse H0 le modèle s'écrit $Y_t = a_0 + \epsilon_t$. on voit facilement que le résultat des MCO dans ce modèle est $\widehat{a_0} = \overline{Y}$ la moyenne des Y_t . On a donc $\widehat{Y_t} = \widehat{a_0} = \overline{Y}$. Les résidus de ce modèle sont $e_t = Y_t - \overline{Y}$ et $SCR0 = \sum(Y_t - \overline{Y})^2$. La variable SCR0/σ^2 suit un χ^2_{n-1} . Le Fisher est le rapport de deux χ^2 indépendants divisés par leurs degrés de liberté,soit

$$\frac{SCR0 - SCR1}{SCR1} * \frac{n-k}{k-1} = F_{(k-1),n-k} \text{ sous H0 vrai}$$

  et le test se fait comme un test de Fisher classique.

2.2.1 Remarque

Ce test est en fait le test du R^2 du modèle global sous H1 $Y_t = a_0 + a_1X_{1t} + a_2X_{2t} + ... + a_{k-1}X_{k-1t} + \epsilon_t$
Démonstration: le R^2 du modèle global est $R^2 = 1 - SCR1/\sum(Y_t - \overline{Y})^2 = 1 - SCR1/SCR0$ soit $SCR1/SCR0 = 1 - R^2$

$$F_{(k-1),n-k} = \frac{1 - SCR1/SCR0}{SCR1/SCR0} * \frac{n-k}{k-1} = \frac{1 - (1 - R^2)}{1 - R^2} * \frac{n-k}{k-1}$$

soit

$$F_{(k-1),n-k} = \frac{R^2}{1 - R^2} * \frac{n-k}{k-1}$$

On voit ainsi que le test peut s'écrire aussi
H0: $R^2 = 0$
H1: $R^2 \neq 0$
en effet sous l'hypothèse H0 le R^2 est 1-SCR0/$\sum(Y_t - \overline{Y})^2 = 1 - 1 = 0$ donc un modèle constitué seulement d'un terme constant et une erreur a un R^2 nul.
Voilà pourquoi ce test est peu regardé car si vous avez un R^2 grand il est évident qu'il est inutile de tester R^2 =0.

2.2.2 Exemple

Dans le modèle ci-dessous

```
Linear Regression - Estimation by Least Squares
Dependent Variable Y
Usable Observations    140      Degrees of Freedom   137
Centered R**2     0.999998      R Bar **2   0.999998
Uncentered R**2   1.000000      T x R**2     140.000
Mean of Dependent Variable      692268.63700
Std Error of Dependent Variable 206128.93876
Standard Error of Estimate         291.88445
Sum of Squared Residuals        11671924.919
Regression F(2,137)             34660910.1766
Significance Level of F           0.00000000
Log Likelihood                    -991.82521
Durbin-Watson Statistic             1.753973

   Variable                     Coeff       Std Error      T-Stat     Signif
*******************************************************************************
1.  Constant                 834.92234985 102.69414057      8.13018  0.00000000
2.  X1                         0.80325350   0.00288544    278.38191  0.00000000
3.  X2                         0.49984476   0.00020700   2414.74856  0.00000000
```

Rats donne le test de tous les coefficients nuls sauf la constante, c'est ici la ligne Regression F(2,137) , on retrouve bien avec k=3 et n=140 le résultat k-1=2 et n-k=137.Ici le résultat du test est F=34660910.1766 valeur très forte avec un niveau de significativité nul donc une décision H1 très nette, ce qui est normal car le $R^2 = 0.999998$ donc très loin de 0.
Nous voyons ici un niveau de significativité (significance level) = 0.000000 nul.
RAPPEL : le niveau de significativité est la probabilité pour que le Fisher soit superieur à la valeur trouvée dans le test. ici c'est la PRO$\{F > 34660910.1766\} = 0,0000$
D'autre part il est évident que chaque coefficient est très loin de 0 avec des 't de Student' de 278.38 et 2414.74 très nettement supérieurs à 1.96 , vous constatez que le niveau de significativité est aussi 0,00000

2.3 Si l'on n'est pas dans l'hypothèse de normalité

Si l'hypothèse H_1^7 n'est plus vérifiée et si l'on souhaite tester r contraintes comme ci-dessus, on montre que la même variable F est telle que sous l'hypothèse H_0 vraie alors

$$rF = \chi_r^2 \text{ si n est grand}$$