Séance 9.1: Regression linéaire multiple

Interprétation

Visseho Adjiwanou, PhD.

07 November 2022

Plan de présentation

- Rappel
- Interprétation des résultats
- Test d'hypothèses

Rappel

Spécification

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i$$

Où ϵ_i suit une loi normale de moyenne 0 et de variance σ^2. - On a k indépdnantes variables pour n observations $\{(Y_1, X_{11}, X_{12}, \ldots, X_{1k}), \ldots, (Y_n, X_{n1}, X_{n2}, \ldots, X_{nk})\}$.

Exemple: - Y peut être le poids à la naissance - X_1 l'age de la mère à la naissance de l'enfant - X_2 le sexe de l'enfant

Spécification

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} ; \boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix} ; \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} ; \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Estimation des paramètres

Hypothèses

	Hypothèses	Expression mathématique Linéaire simple	 Linéaire multiple	Violations
H1	Variable dépendante fonction linéaire d'un ensemble spécifique de variables indépendantes, plus une perturbation	$Y_i = ß_0 + ß_1X_i + \varepsilon_i$, $i = 1$ à N	$Y = Xß + \varepsilon$	Mauvaises variables indépendantes, Non linéarité, Paramètres changeant
H2	L'espérance mathématique de l'erreur est nulle	$E(\varepsilon_i) = 0$, pour tout i	$E\varepsilon = 0$	Intercept biaisé
H3	La variance de l'erreur est constante quel que soit i (homoscédasticité)	$Var(\varepsilon_i) = \sigma^2$	$E\varepsilon\varepsilon' = \sigma^2 I$	Hétéroscédasticité
	Les erreurs sont non corrélées (ou encore indépendantes)	$E(\varepsilon_i, \varepsilon_j) = 0$ pour $i \neq j$		Erreurs autocorrélées
H4	Les variables indépendantes sont observées sans erreur	X_i fixé	X fixé	Erreurs dans les variables X, Auto-régression, Équations simultanées
H5	Absence de colinéarité entre les variables indépendantes	$\sum(X_i - X)^2 \neq 0$	Rang de X = K<= N	Colinéarité parfaites
	Le nombre d'observations est supérieur au nombre des séries explicatives	N>= K		

Estimation des paramètres

- Les paramètres inconnus:
  - k (beta) + 1 (alpha) paramètres
  - σ^2
- On démontre que :

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

Variance-covariance de $Varcov(\beta^*) = \sigma^2(X'X)^{-1}$

Mais encore une fois, σ^2 n'est pas connu. Il est remplacé par:

$$s^2 = e'e/(T-k)$$

avec (e = Y-Y')

Interprétation

Exemple : déterminant du salaire aux USA en 1978

$Log_e(wage/hour) = \beta_0 + \beta_1 \text{education} + \beta_2 \text{experience} + \beta_3 \text{femme}$

- femme est une variable dichotomique (ou dummy)
- scolarité et expérience sont continues et reflètent les années

Résultat de l'estimation

Log(salaire/heure) = 0.5779 + 0.0780 * scolarite + 0.0134 * experience – 0.3356 * femme

- $\beta_0 = \text{Cste} = 0.5779$
- $\beta_1 = 0.0780$
- $\beta_2 = 0.0134$
- $\beta_3 = 0.3356$

Exemple : déterminant du salaire aux USA en 1978

- β_0 = Cste = 0,5779, donc $e^{0.5779}$ = 1,78 est le salaire moyen pour un homme (fem == 0) avec 0 année d'expérience (exp == 0) et 0 année d'études (école == 0)
- Un homme avec 6 ans d'éducation et 2 ans d'expérience gagnera combien?

- Un homme avec 6 ans d'éducation et 2 ans d'expérience gagnera en moyenne: Log(salaire / heure) = 0,5779 + 0,0780 * 6 + 0,0134 * 2 - 0,3356 * 0 = 1,0727 Le salaire horaire moyen pour cette personne est de 2,92 $

Exemple : déterminant du salaire aux USA en 1978

- β_0 = Cste = 0,5779, donc $e^{0.5779}$ = 1,78 est le salaire moyen pour un homme (fem == 0) avec 0 année d'expérience (exp == 0) et 0 année d'études (école == 0)
- Un homme avec 6 ans d'éducation et 2 ans d'expérience gagnera combien?

- Un homme avec 6 ans d'éducation et 2 ans d'expérience gagnera en moyenne: Log(salaire / heure) = 0,5779 + 0,0780 * 6 + 0,0134 * 2 - 0,3356 * 0 = 1,0727 Le salaire horaire moyen pour cette personne est de 2,92 $

- Pour une femme ayant les mêmes caractéristiques, salaire moyen/heure = 2,09 $

Exemple : déterminant du salaire aux USA en 1978

- $\beta_0 =$ Cste $=$ 0,5779, donc $e^{0.5779} =$ 1,78 est le salaire moyen pour un homme (fem $==$ 0) avec 0 année d'expérience (exp $==$ 0) et 0 année d'études (école $==$ 0)
- Un homme avec 6 ans d'éducation et 2 ans d'expérience gagnera combien?
- Un homme avec 6 ans d'éducation et 2 ans d'expérience gagnera en moyenne: Log(salaire / heure) $=$ 0,5779 $+$ 0,0780 * 6 $+$ 0,0134 * 2 - 0,3356 * 0 $=$ 1,0727 Le salaire horaire moyen pour cette personne est de 2,92 $
- Pour une femme ayant les mêmes caractéristiques, salaire moyen/heure $=$ 2,09 $
- La différence vaut 2.92$ - 2.09$ $=$ 0,0780$

Exemple : déterminant du salaire aux USA en 1978

Log(salaire / heure) = 0.5779 + 0.0780 * scolarite + 0.0134 * exp - 0.3356 * fem

- La question générale est la suivante: quel est l'effet d'une année d'études supplémentaire sur le log (salaire) en tenant compte du sexe et de l'expérience?
- En dérivant log(salaire/heure) par la scolarité, on trouve 0,0780.

Exemple : déterminant du salaire aux USA en 1978

Log(salaire / heure) = 0.5779 + 0.0780 * scolarite + 0.0134 * exp - 0.3356 * fem

- La question générale est la suivante: quel est l'effet d'une année d'études supplémentaire sur le log (salaire) en tenant compte du sexe et de l'expérience?
- En dérivant log(salaire/heure) par la scolarité, on trouve 0,0780.
- Parce que nous utilisons log (salaire), nous dirons qu'une année supplémentaire d'études, **les autres facteurs maintenus constants** entraîne une augmentation du salaire de 7,8%

Une régression multiple est différente de plusieurs régressions simple

$Log_e(wage/hour) = \beta_0 + \beta_1$ education $+ \beta_2$ experience $+ \beta_3$ femme est différent de:

- $Log_e(wage/hour) = \beta_0 + \beta_1$*education
- $Log_e(wage/hour) = \beta_0 + \beta_2$*experience
- $Log_e(wage/hour) = \beta_0 + \beta_3$*femme

Résultats importants et interprétation

- Même si on ne reporte pas l'ensemble des résultats d'une régression, voici les principaux résultats et leur interprétation.

Source	SS	df	MS
Model	6.8604e+09	2	3.4302e+09
Residual	5.4747e+10	5355	10223566.4
Total	6.1608e+10	5357	11500396.2

Number of obs	=	5358
F(2, 5355)	=	335.52
Prob > F	=	0.0000
R-squared	=	0.1114
Adj R-squared	=	0.1110
Root MSE	=	3197.4

m19	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
v190	-803.7117	31.8308	-25.25	0.000	-866.113	-741.3104
v013	131.7199	29.0377	4.54	0.000	74.79422	188.6456
_cons	8138.445	149.8926	54.30	0.000	7844.595	8432.296

Résultats importants et interprétation

1. Qualité du modèle

- Model SS (SSM): Somme des carrés du modèle $= \sum(\hat{Y}_i - \bar{Y})^2$
- Residual SS (SSR) : Somme des carrés résiduelles $= \sum(Y_i - \hat{Y}_i)^2 = \sum[Y_i - (\alpha + \beta * X_i]^2$

Résultats importants et interprétation

1. Qualité du modèle

- Model SS (SSM): Somme des carrés du modèle $= \sum(\hat{Y}_i - \bar{Y})^2$
- Residual SS (SSR) : Somme des carrés résiduelles $=$ $\sum(Y_i - \hat{Y}_i)^2 = \sum[Y_i - (\alpha + \beta * X_i]^2$
- Total SS (SST) : Somme des carrées totale $= \sum(Y_i - \bar{Y})^2$

Résultats importants et interprétation

1. Qualité du modèle

- Model SS (SSM): Somme des carrés du modèle $= \sum(\hat{Y}_i - \bar{Y})^2$
- Residual SS (SSR) : Somme des carrés résiduelles $= \sum(Y_i - \hat{Y}_i)^2 = \sum[Y_i - (\alpha + \beta * X_i]^2$
- Total SS (SST) : Somme des carrées totale $= \sum(Y_i - \bar{Y})^2$
- Un modèle sera bon si la somme des carrés du modèle se rapproche de la somme des carrés total

Résultats importants et interprétation

- Les df sont les degrés de liberté. Ils permettent de corriger les modèles selon le nombre de variables dépendantes
- model df = nombre de variables indépendantes (sauf la constante)

Résultats importants et interprétation

- Les df sont les degrés de liberté. Ils permettent de corriger les modèles selon le nombre de variables dépendantes
- model df = nombre de variables indépendantes (sauf la constante)
- residual df = taille de l'échantillon - nombre de variables dépendantes y compris le terme constant

Résultats importants et interprétation

- Les df sont les degrés de liberté. Ils permettent de corriger les modèles selon le nombre de variables dépendantes
- model df = nombre de variables indépendantes (sauf la constante)
- residual df = taille de l'échantillon - nombre de variables dépendantes y compris le terme constant
- total df = model df + residual df

Résultats importants et interprétation

- Variances expliquées: en divisant les sommes des carrés par les df, on obtient les variances:
  - Model MS (mean square): Variance expliquée par le modèle
  - Residual Ms : variance résiduelle
  - Total MS: Variance totale

R carré: coéficient de détermination

A partir de ces variances, on calcule le R carré qui vaut:

$$R^2 = SSM/SST = 1 - SSR/SST$$

- Évalue la qualité de la relation linéaire entre Y et les variables indépendantes

R carré: coéficient de détermination

A partir de ces variances, on calcule le R carré qui vaut:

$$R^2 = SSM/SST = 1 - SSR/SST$$

- Évalue la qualité de la relation linéaire entre Y et les variables indépendantes
- Exprime la partie de la variance totale de Y expliquée par le modèle

R carré: coéficient de détermination

- Varie entre 0 et 1 mais est interprété en termes de pourcentage
- Si R^2 est proche de 1, nous avons un bon modèle

R carré: coéficient de détermination

- Varie entre 0 et 1 mais est interprété en termes de pourcentage
- Si R^2 est proche de 1, nous avons un bon modèle
- Si R^2 est proche de 0, le modèle explique mal la variable dépendante

R carré: coéficient de détermination

- Varie entre 0 et 1 mais est interprété en termes de pourcentage
- Si R^2 est proche de 1, nous avons un bon modèle
- Si R^2 est proche de 0, le modèle explique mal la variable dépendante
- Utile pour la régression linéaire simple et la régression linéaire multiple

R carré: coéficient de détermination

- R^2 est affecté par l'ajout d'une nouvelle variable indépendante dans le modèle même si l'effet de cette variable n'est pas significatif
- R^2 ajusté corrige pour cela:

$$\text{Adjusted } R^2 = \frac{[(n-1)R^2 - k]}{[n-(k+1)]}$$

- Adjusted $R^2 <= R^2$

Les autres éléments du tableau

Les autres parties du tableau permettent de tester quelques hypothèses dont:

- Est-ce que le modèle dans sa globalité est significatif?
- Est-ce qu'une variable est significative
- Et bien plus...

Test d'hypothèses en régression linéaire classique

Introduction

- Passons maintenant au problème de l'utilisation du modèle de régression pour tester des hypothèses.
- Le type d'hypothèse le plus couramment testé avec l'aide du modèle de régression est qu'il n'y a pas de relation entre la variable explicative X et la variable dépendante Y.

Test pour une variable explicative: test de Student

En supposant que toutes les hypothèses des modèles de régression linéaire sont valides:

- L'équation de regression $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i$
- Ou: $E(Y_i|X) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki}$
- Tester que $\beta_i = 0$ signifie que
  - Il n'y a pas de relation entre X_i et Y
  - La valeur moyenne de Y_i ne dépend pas linéairement de X_i
  - La droite de régression de population est horizontale (en régression linéaire simple)

Test pour une variable explicative: test de Student

- Ce test est basé sur la variance de β_i
- En cas de régression multiple, le calcul de cette variance est compliqué et n'est pas présenté ici.
- En cas de MRLS, nous avons vu que:
- Variance de la pente : $Var(\beta^*) = \frac{\sigma_\epsilon^2}{\sum(X_i - \bar{X})^2}$

Test pour une variable explicative: test de Student

- Ce test est basé sur la variance de β_i
- En cas de régression multiple, le calcul de cette variance est compliqué et n'est pas présenté ici.
- En cas de MRLS, nous avons vu que:
- Variance de la pente : $Var(\beta^*) = \frac{\sigma_\epsilon^2}{\sum(X_i - \bar{X})^2}$
- Variance de l'intercept: $Var(\alpha^*) = \sigma_\epsilon^2[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}]$

Test pour une variable explicative: test de Student

- Mais comme σ^2 est inconnu, il est remplacé par son estimateur (s).
- Finalement, on a:
- Estimateur de la variance de la pente: $s_{\beta*}^2 = Var\hat{(}\beta^*) = \frac{s^2}{\sum(X_i - \bar{X})^2}$

Test pour une variable explicative: test de Student

- Mais comme σ^2 est inconnu, il est remplacé par son estimateur (s).
- Finalement, on a:
- Estimateur de la variance de la pente:
  $s^2_{\beta*} = Var\hat{(}\beta*) = \frac{s^2}{\sum(X_i - \bar{X})^2}$
- Estimateur de la variance de l'intercept:
  $s^2_{\alpha*} = s^2[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}]$

Test pour une variable explicative: test de Student

Comment faire le test?

1. Énoncez l'hypothèse nulle ($\beta_1 = \mu_1$)
2. Choisissez le niveau de signification (α, 0.1%, 1%, 5%)
3. *Comparer* β_1^* *à* μ_1
    - *Si* $\beta_1^* = \mu_1$, *pas besoin de tester, sinon testez*
4. *Type de test: test bilatéral ou test unilatéral*
    - *Toujours privilégier le test bilatéral*
    - *Hypothèse forte derrière le test unilatéral*
5. *Calculer les statistiques sous l'hypothèse nulle*
    - *t of Student (z normal dans l'exemple précédent)*

Test pour une variable explicative: test de Student

Comment faire le test?

1. Énoncez l'hypothèse nulle ($\beta_1 = \mu_1$)
2. Choisissez le niveau de signification (α, 0.1%, 1%, 5%)
    3. *Comparer β_1^* à μ_1*
        - *Si $\beta_1^* = \mu_1$, pas besoin de tester, sinon testez*
    4. *Type de test: test bilatéral ou test unilatéral*
        - *Toujours privilégier le test bilatéral*
        - *Hypothèse forte derrière le test unilatéral*
    5. *Calculer les statistiques sous l'hypothèse nulle*
        - *t of Student (z normal dans l'exemple précédent)*
6. Décision: Comparez le t calculé au t qui vous est donné par la

Test pour une variable explicative: test de Student

- Sous l'hypothèse nulle:

$$t^* = \frac{\beta_1^* - \mu_1}{s_{\beta_1^*}} \sim \text{une loi de Student avec n - (k+1) degrés de liberté}$$

- n est la taille de l'échantillon

- k est le nombre de variables indépendantes (excluant l'intercept)
- Décision:
  - Si |t*| > t(lu), rejeter l'hypothèse nulle
  - Si |t*| < (t(lu), ne rejeter pas l'hypothèse nulle

Exemple: Déterminant du nombre d'enfants

On estime le modèle:

$$Parite_i = \alpha + \beta_1 Urbain_i + \beta_2 Sans-education_i + \beta_3 College_i + \beta_4 Catholique_i +$$

$$\beta_5 Protestant_i + \beta_6 Animiste_i + \beta_7 Sans - religion_i + \beta_8 Age_i + \epsilon_i$$

- Où $Parite_i$ est le nombre d'enfants vivants de la femme i.

Exemple: Déterminant du nombre d'enfants

Le tableau suivant présente les résultats de cette régression:

Variables	Coefficients	Erreur standard
Constant	-3.855	0.092
Urbain	-0.634	0.057
Sans-education	0.137	0.074
College	-0.738	0.097
Catholique	-0.095	0.051
Protestant	0.074	0.086
Animiste	0.040	0.067
Sans-religion	0.138	0.111
Age	0.261	0.002

- Référence?
- $N = 6444$

Test d'une variable continue

- Testez que $\beta_{age} = 0$
- Valeur de beta estimé: $\beta^*_{age} = 0,261 \neq 0$, test possible pour voir si l'âge n'est pas lié à la parité

Test d'une variable continue

- Testez que $\beta_{age} = 0$
- Valeur de beta estimé: $\beta^*_{age} = 0,261 \neq 0$, test possible pour voir si l'âge n'est pas lié à la parité
- t calculé: $t^* = \frac{(\beta^*_{age} - 0)}{s_{\beta^*_{age}}} = (0.261\text{-}0)/0.002 = 130.5$

Test d'une variable continue

- Testez que $\beta_{age} = 0$
- Valeur de beta estimé: $\beta^*_{age} = 0,261 \neq 0$, test possible pour voir si l'âge n'est pas lié à la parité
- t calculé: $t^* = \frac{(\beta^*_{age} - 0)}{s_{\beta^*_{age}}} = (0.261\text{-}0)/0.002 = 130.5$
- le t calculé: t^* suit une loi de Student avec (6444 - (8 + 1)) = 6435 degrés de liberté

Test d'une variable continue

- Testez que $\beta_{age} = 0$
- Valeur de beta estimé: $\beta^*_{age} = 0,261 \neq 0$, test possible pour voir si l'âge n'est pas lié à la parité
- t calculé: $t^* = \frac{(\beta^*_{age} - 0)}{s_{\beta^*_{age}}} = (0.261\text{-}0)/0.002 = 130.5$
- le t calculé: t^* suit une loi de Student avec (6444 - (8 + 1)) = 6435 degrés de liberté
- Considérons le seuil de significativité $\alpha = 1\%$

Test d'une variable continue

- Testez que $\beta_{age} = 0$
- Valeur de beta estimé: $\beta^*_{age} = 0,261 \neq 0$, test possible pour voir si l'âge n'est pas lié à la parité
- t calculé: $t^* = \frac{(\beta^*_{age} - 0)}{s_{\beta^*_{age}}}$ = (0.261-0)/0.002 = 130.5
- le t calculé: t^* suit une loi de Student avec (6444 - (8 + 1)) = 6435 degrés de liberté
- Considérons le seuil de significativité $\alpha = 1\%$
- t(lu) = 2,58

Test d'une variable continue

- Testez que $\beta_{age} = 0$
- Valeur de beta estimé: $\beta^*_{age} = 0,261 \neq 0$, test possible pour voir si l'âge n'est pas lié à la parité
- t calculé: $t^* = \frac{(\beta^*_{age} - 0)}{s_{\beta^*_{age}}} = (0.261\text{-}0)/0.002 = 130.5$
- le t calculé: t^* suit une loi de Student avec (6444 - (8 + 1)) = 6435 degrés de liberté
- Considérons le seuil de significativité $\alpha = 1\%$
- t(lu) = 2,58
- Décision : le t calculé $t^* > t(lu)$, on rejette l'hypothèse nulle

Test d'une variable continue

- Testez que $\beta_{age} = 0$
- Valeur de beta estimé: $\beta^*_{age} = 0,261 \neq 0$, test possible pour voir si l'âge n'est pas lié à la parité
- t calculé: $t^* = \frac{(\beta^*_{age} - 0)}{s_{\beta^*_{age}}} = (0.261\text{-}0)/0.002 = 130.5$
- le t calculé: t^* suit une loi de Student avec (6444 - (8 + 1)) = 6435 degrés de liberté
- Considérons le seuil de significativité $\alpha = 1\%$
- t(lu) = 2,58
- Décision : le t calculé $t^* > t(lu)$, on rejette l'hypothèse nulle
- Conclusion: l'âge a un effet significatif sur la parité

Test d'une variable dichotomique

- Comparer deux groupes distincts
- Test que $\beta_{catholique} = 0$

Test d'une variable dichotomique

- Comparer deux groupes distincts

- Test que $\beta_{catholique} = 0$
- Dans ce cas précis, vérifiez si les femmes catholiques ont une parité sensiblement différente de celle des femmes musulmanes.

Test d'une variable dichotomique

- Comparer deux groupes distincts
- Test que $\beta_{catholique} = 0$
- Dans ce cas précis, vérifiez si les femmes catholiques ont une parité sensiblement différente de celle des femmes musulmanes.
- Le test est identique à celui effectué précédemment

Test d'une variable dichotomique

- Comparer deux groupes distincts
- Test que $\beta_{catholique} = 0$
- Dans ce cas précis, vérifiez si les femmes catholiques ont une parité sensiblement différente de celle des femmes musulmanes.
- Le test est identique à celui effectué précédemment
- Hypothèse nulle: $H_0 : \beta_{catholique} = 0$

Test d'une variable dichotomique

- Comparer deux groupes distincts
- Test que $\beta_{catholique} = 0$
- Dans ce cas précis, vérifiez si les femmes catholiques ont une parité sensiblement différente de celle des femmes musulmanes.
- Le test est identique à celui effectué précédemment
- Hypothèse nulle: $H_0 : \beta_{catholique} = 0$
- t calculé : $t^* = \frac{(\beta^*_{catholique} - 0)}{s_{\beta^*_{catholique}}}$ = (-0,095 - 0)/0,051 = -1.863

Test d'une variable dichotomique

- Comparer deux groupes distincts
- Test que $\beta_{catholique} = 0$
- Dans ce cas précis, vérifiez si les femmes catholiques ont une parité sensiblement différente de celle des femmes musulmanes.
- Le test est identique à celui effectué précédemment
- Hypothèse nulle: $H_0 : \beta_{catholique} = 0$
- t calculé : $t^* = \frac{(\beta^*_{catholique} - 0)}{s_{\beta^*_{catholique}}}$ = (-0,095 - 0)/0,051 = -1.863
- Décision: Seuil $\alpha = 1\%$ ==> t(lu) = 2,58, conclusion?

Test d'une variable dichotomique

- Comparer deux groupes distincts
- Test que $\beta_{catholique} = 0$
- Dans ce cas précis, vérifiez si les femmes catholiques ont une parité sensiblement différente de celle des femmes musulmanes.
- Le test est identique à celui effectué précédemment
- Hypothèse nulle: $H_0 : \beta_{catholique} = 0$
- t calculé : $t^* = \frac{(\beta^*_{catholique} - 0)}{s_{\beta^*_{catholique}}}$ = (-0,095 - 0)/0,051 = -1.863
- Décision: Seuil $\alpha = 1\%$ ==> t(lu) = 2,58, conclusion?
- Seuil $\alpha = 5\%$ ==> t(lu) = 1,96, conclusion?

Test d'une variable dichotomique

- Comparer deux groupes distincts
- Test que $\beta_{catholique} = 0$
- Dans ce cas précis, vérifiez si les femmes catholiques ont une parité sensiblement différente de celle des femmes musulmanes.
- Le test est identique à celui effectué précédemment
- Hypothèse nulle: $H_0 : \beta_{catholique} = 0$
- t calculé : $t^* = \frac{(\beta^*_{catholique} - 0)}{s_{\beta^*_{catholique}}}$ = (-0,095 - 0)/0,051 = -1.863
- Décision: Seuil $\alpha = 1\%$ ==> t(lu) = 2,58, conclusion?
- Seuil $\alpha = 5\%$ ==> t(lu) = 1,96, conclusion?
- Seuil $\alpha = 10\%$ ==> t(lu) = 1,64, conclusion?

Intervalle de confiance

- On parle d'intervalle de confiance en cas de test bilatéral
- Cet intervalle est donné par défaut à 95%, le complément à 1 de α
- L'intervalle de confiance peut également être utilisé pour le test d'hypothèse
  - Les valeurs en dehors de l'intervalle sont significativement différentes de β^*
  - Les valeurs à l'intérieur de l'intervalle ne sont pas significativement différentes de β^*

Conclusion: Une variable a un effet significatif si l'intervalle de confiance de ses estimations ne contient pas 0.

Intervalle de confiance

- La formule de l'intervalle de confiance est:

$$IC = \beta_s^* \pm t_\alpha * s_{\beta *}$$

$$[\beta_s^* - t_\alpha * s_{\beta *}, \beta_s^* + t_\alpha * s_{\beta *}]$$

- Exemple: intervalle de confiance de β_{age}^*

- [0.261 - 1.96(0.002), 0.261 + 1.96(0.002)]
- [0.257, 0.265]

Test pour plus d'une variable explicative

Introduction

- Utilisé pour tester plusieurs hypothèses à la fois, à la différence du test t, qui ne portait que sur une hypothèse.

Considérons:

- L'équation de regression
  $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i$
- L'un des tests les plus simples est l'égalité entre deux paramètres:
  - Hypothèse nulle: $H_0 : \beta_1 = \beta_2$;
  - Hypothèse alternative: $H_A : \beta_1 \neq \beta_2$

Introduction

- Ou pour tester:
  - Hypothèse nulle: $H_0 : \beta_1 = 1$ et $\beta_2 = 2$
  - Hypothèse alternative: $H_A : H_0$ n'est pas vrai).
  - Ce qui est différent de deux tests t -t test 1: $H_0 : \beta_1 = 1$ et $H_A : \beta_1 \neq 1$ -t test 2: $H_0 : \beta_2 = 2$ et $H_A : \beta_2 \neq 2$

F test: formulation

- Le test de ces hypothèses utilise le test F (Fischer) basé sur deux modèles:

1. Le modèle sans restriction ou sans contrainte

F test: formulation

- Le test de ces hypothèses utilise le test F (Fischer) basé sur deux modèles:

1. Le modèle sans restriction ou sans contrainte

- Contient tous les $(K + 1)$ paramètres à estimer

F test: formulation

- Le test de ces hypothèses utilise le test F (Fischer) basé sur deux modèles:

1. Le modèle sans restriction ou sans contrainte

- Contient tous les $(K + 1)$ paramètres à estimer

2. Le modèle restreint ou contraint

F test: formulation

- Le test de ces hypothèses utilise le test F (Fischer) basé sur deux modèles:

1. Le modèle sans restriction ou sans contrainte

- Contient tous les $(K + 1)$ paramètres à estimer

2. Le modèle restreint ou contraint

- Prendre en compte les contraintes imposées au modèle:

F test: formulation

- Le test de ces hypothèses utilise le test F (Fischer) basé sur deux modèles:

1. Le modèle sans restriction ou sans contrainte

- Contient tous les $(K + 1)$ paramètres à estimer

2. Le modèle restreint ou contraint

- Prendre en compte les contraintes imposées au modèle:
  - Contraintes: $(\beta_1 = \beta_2)$ ou $(\beta_1 = 1, \beta_2 = 2)$

F test: formulation

- Le test de ces hypothèses utilise le test F (Fischer) basé sur deux modèles:

1. Le modèle sans restriction ou sans contrainte

- Contient tous les $(K + 1)$ paramètres à estimer

2. Le modèle restreint ou contraint

- Prendre en compte les contraintes imposées au modèle:
  - Contraintes: $(\beta_1 = \beta_2)$ ou $(\beta_1 = 1, \beta_2 = 2)$
- Si c est le nombre de contraintes le modèle restreint aura $K +$ 1 - c paramètres

Exemples

Cas 1: $H_0 : \beta_1 = \beta_2$

- c $= 1$
- Modèle non contraint
  - UM: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i$
- Modèle contraint
  - RM: $Y_i = \alpha + \beta_1 X_{1i} + \beta_1 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i$
  - RM: $Y_i = \alpha + \beta_1 (X_{1i} + X_{2i}) + ... + \beta_k X_{ki} + \epsilon_i$
  - RM: $Y_i = \alpha + \beta_1 Z_i + ... + \beta_k X_{ki} + \epsilon_i$
    - Avec $Z_i = (X_{1i} + X_{2i})$

Exemples

Cas 2: $H_0 : \beta_1 = 1$ et $\beta_2 = 2$

- c = 2
- Modèle non contraint
  - UM: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i$
- Modèle contraint
  - RM: $Y_i = \alpha + 1X_{1i} + 2X_{2i} + ... + \beta_k X_{ki} + \epsilon_i$
  - RM: $Y_i - 1X_{1i} - 2X_{2i} = \alpha + ... + \beta_k X_{ki} + \epsilon_i$
  - RM: $T_i = \alpha + \beta_3 X_{3i} + ... + \beta_k X_{ki} + \epsilon_i$
    - Avec $T_i = Y_i - 1X_{1i} - 2X_{2i}$

F test

- Calculer la somme des carrés des résidus (SSR) de chaque modèle
- Calculer la statistique F

$$F^* = \frac{[SSR(RM) - SSR(UM)]/c}{SSR(UM)/(n-(k+1))} \sim Fischer_{c,n-(k+1)}$$

- Accéder à la valeur critique de F à partir de la table Fischer
- Règle de décision
  - Si $F* > F(lu)$, rejeter l'hypothèse nulle
  - Si $F* < F(lu)$, ne pas rejeter l'hypothèse nulle

Exemple

Source	SS	df	MS
Model	6.8604e+09	2	3.4302e+09
Residual	5.4747e+10	5355	10223566.4
Total	6.1608e+10	5357	11500396.2

Number of obs	=	5358
F(2, 5355)	=	335.52
Prob > F	=	0.0000
R-squared	=	0.1114
Adj R-squared	=	0.1110
Root MSE	=	3197.4

m19	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
v190	-803.7117	31.8308	-25.25	0.000	-866.113	-741.3104
v013	131.7199	29.0377	4.54	0.000	74.79422	188.6456
_cons	8138.445	149.8926	54.30	0.000	7844.595	8432.296

Exemple 1

$$Parite_i = \alpha + \beta_1 Urbain_i + \beta_2 Sans{-}education_i + \beta_3 College_i + \beta_4 Catholique_i +$$

$$\beta_5 Protestant_i + \beta_6 Animiste_i + \beta_7 Sans - religion_i + \beta_8 Age_i + \epsilon_i$$

- Hypothèse nulle:
  $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$

Exemple 1

$$Parite_i = \alpha + \beta_1 Urbain_i + \beta_2 Sans-education_i + \beta_3 College_i + \beta_4 Catholique_i +$$

$$\beta_5 Protestant_i + \beta_6 Animiste_i + \beta_7 Sans-religion_i + \beta_8 Age_i + \epsilon_i$$

- Hypothèse nulle:
  $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$
- Hypothèse alternative: H_A?

Exemple 1

$$Parite_i = \alpha + \beta_1 Urbain_i + \beta_2 Sans-education_i + \beta_3 College_i + \beta_4 Catholique_i +$$

$$\beta_5 Protestant_i + \beta_6 Animiste_i + \beta_7 Sans - religion_i + \beta_8 Age_i + \epsilon_i$$

- Hypothèse nulle:
  $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$
- Hypothèse alternative: H_A?
- Résultat:

Exemple 1

$$Parite_i = \alpha + \beta_1 Urbain_i + \beta_2 Sans-education_i + \beta_3 College_i + \beta_4 Catholique_i +$$

$$\beta_5 Protestant_i + \beta_6 Animiste_i + \beta_7 Sans - religion_i + \beta_8 Age_i + \epsilon_i$$

- Hypothèse nulle:
  $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$
- Hypothèse alternative: H_A?
- Résultat:
- SSR(UM) = 59968.90

Exemple 1

$$Parite_i = \alpha + \beta_1 Urbain_i + \beta_2 Sans{-}education_i + \beta_3 College_i + \beta_4 Catholique_i +$$

$$\beta_5 Protestant_i + \beta_6 Animiste_i + \beta_7 Sans - religion_i + \beta_8 Age_i + \epsilon_i$$

- Hypothèse nulle:
  $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$
- Hypothèse alternative: H_A?
- Résultat:
- SSR(UM) = 59968.90
- SSR(RM) = 61414.84 (RM: $Parite_i = \alpha$

Exemple 1

$$Parite_i = \alpha + \beta_1 Urbain_i + \beta_2 Sans-education_i + \beta_3 College_i + \beta_4 Catholique_i +$$

$$\beta_5 Protestant_i + \beta_6 Animiste_i + \beta_7 Sans - religion_i + \beta_8 Age_i + \epsilon_i$$

- Hypothèse nulle:
  $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$
- Hypothèse alternative: H_A?
- Résultat:
- SSR(UM) $= 59968.90$
- SSR(RM) $= 61414.84$ (RM: $Parite_i = \alpha$
- C=?

Exemple 1

$$F^* = \frac{[61414.84 - 59968.90)]/8}{59968.90/(6444 - 9)} = 19.39$$

- F(lu) $= F_{8,6435} = 1.94$ pour $\alpha = 0.05$
- Décision:
  - Comme $F^* > F(lu)$, on rejette l'hypothèse nulle
- Conclusion: Les 8 variables indépendantes ont toutes un effet significatif sur la parité.

Exemple 2

$$Parite_i = \alpha + \beta_1 Urbain_i + \beta_2 Prim_i + \beta_3 Col_i + \beta_4 Catho_i + ... + \beta_8 Age_i + \epsilon_i$$

- Hypothèse nulle: $H_0 : \beta_2 = \beta_3$
- Il n'existe pas de différence entre l'effet de niveau d'enseignement primaire et l'effet de niveau d'enseignement secondaire

Exemple 2

$$Parite_i = \alpha + \beta_1 Urbain_i + \beta_2 Prim_i + \beta_3 Col_i + \beta_4 Catho_i + \ldots + \beta_8 Age_i + \epsilon_i$$

- Hypothèse nulle: $H_0 : \beta_2 = \beta_3$
- Il n'existe pas de différence entre l'effet de niveau d'enseignement primaire et l'effet de niveau d'enseignement secondaire
- Hypothèse alternative: $H_A : \beta_2 \neq \beta_3$

Exemple 2

$$Parite_i = \alpha + \beta_1 Urbain_i + \beta_2 Prim_i + \beta_3 Col_i + \beta_4 Catho_i + ... + \beta_8 Age_i + \epsilon_i$$

- Hypothèse nulle: $H_0 : \beta_2 = \beta_3$
- Il n'existe pas de différence entre l'effet de niveau d'enseignement primaire et l'effet de niveau d'enseignement secondaire
- Hypothèse alternative: $H_A : \beta_2 \neq \beta_3$
- RM: $Parite_i = \alpha + \beta_1 Urbain_i + \beta_2 (primaire_i + college_i) + \beta_4 * catholique_i + ... + \beta_8 * age_i + \epsilon_i$

Exemple 2

$$Parite_i = \alpha + \beta_1 Urbain_i + \beta_2 Prim_i + \beta_3 Col_i + \beta_4 Catho_i + ... + \beta_8 Age_i + \epsilon_i$$

- Hypothèse nulle: $H_0 : \beta_2 = \beta_3$
- Il n'existe pas de différence entre l'effet de niveau d'enseignement primaire et l'effet de niveau d'enseignement secondaire
- Hypothèse alternative: $H_A : \beta_2 \neq \beta_3$
- RM: $Parite_i = \alpha + \beta_1 Urbain_i + \beta_2 (primaire_i + college_i) + \beta_4 * catholique_i + ... + \beta_8 * age_i + \epsilon_i$
- Cacluler la nouvelle variable $educ = primaire_i + college_i$

Exemple 2

$$Parite_i = \alpha + \beta_1 Urbain_i + \beta_2 Prim_i + \beta_3 Col_i + \beta_4 Catho_i + ... + \beta_8 Age_i + \epsilon_i$$

- Hypothèse nulle: $H_0 : \beta_2 = \beta_3$
- Il n'existe pas de différence entre l'effet de niveau d'enseignement primaire et l'effet de niveau d'enseignement secondaire
- Hypothèse alternative: $H_A : \beta_2 \neq \beta_3$
- RM: $Parite_i = \alpha + \beta_1 Urbain_i + \beta_2 (primaire_i + college_i) + \beta_4 * catholique_i + ... + \beta_8 * age_i + \epsilon_i$
- Cacluler la nouvelle variable $educ = primaire_i + college_i$
- Estimer le modèle RM: $Parite_i = \alpha + \beta_1 Urbain_i + \beta_2 (educ_i) + \beta_4 Catho_i + ... + \beta_8 Age_i + \epsilon_i$

Exemple 2

- Résultat

$$F^* = \frac{[56164.15 - 55968.90)]/1}{55968.90/(6435)} = 22.45$$

- F(lu) : $F_{1,6435} = 3.84$ pour $\alpha = 0.05$
- F(lu) : $F_{1,6435} = 6.63$ pour $\alpha = 0.01$
- Conclusion:
  - Rejet de l'hypothèse nulle: la différence est significative dans les deux cas

Exemple 3

$$Parite_i = \alpha + \beta_1 Urbain_i + \beta_2 Prim_i + \beta_3 Col_i + \beta_4 Catho_i + ... + \beta_8 Age_i + \epsilon_i$$

- Hypothèse nulle : $H_0 = \beta_2 = 2, \beta_3 = -5$
- RM: $Parite_i =$ $\alpha + \beta_1 Urbain_i + 2Prim_i - 5Col_i + \beta_4 Catho_i + ... + \beta_8 * age_i + \epsilon_i$
- Calculer une nouvelle variable dépendante: $Paritenew_i = Parite_i - 2 * primaire_i + 5 * college_i$

Exemple 3

$$Parite_i = \alpha + \beta_1 Urbain_i + \beta_2 Prim_i + \beta_3 Col_i + \beta_4 Catho_i + ... + \beta_8 Age_i + \epsilon_i$$

- Hypothèse nulle : $H_0 = \beta_2 = 2, \beta_3 = -5$
- RM: $Parite_i = \alpha + \beta_1 Urbain_i + 2Prim_i - 5Col_i + \beta_4 Catho_i + ... + \beta_8 * age_i + \epsilon_i$
- Calculer une nouvelle variable dépendante: $Paritenew_i = Parite_i - 2 * primaire_i + 5 * college_i$
- Estimer le nouvel modele:

Exemple 3

$$Parite_i = \alpha + \beta_1 Urbain_i + \beta_2 Prim_i + \beta_3 Col_i + \beta_4 Catho_i + ... + \beta_8 Age_i + \epsilon_i$$

- Hypothèse nulle : $H_0 = \beta_2 = 2, \beta_3 = -5$
- RM: $Parite_i =$ $\alpha + \beta_1 Urbain_i + 2 Prim_i - 5 Col_i + \beta_4 Catho_i + ... + \beta_8 * age_i + \epsilon_i$
- Calculer une nouvelle variable dépendante: $Paritenew_i = Parite_i - 2 * primaire_i + 5 * college_i$
- Estimer le nouvel modele:
- RM: $Paritenew_i = \alpha + \beta_1 * urbain_i + \beta_4 * catholique_i + ... + \beta_8 * age_i + \epsilon_i$

Exemple 3

$$Parite_i = \alpha + \beta_1 Urbain_i + \beta_2 Prim_i + \beta_3 Col_i + \beta_4 Catho_i + ... + \beta_8 Age_i + \epsilon_i$$

- Hypothèse nulle : $H_0 = \beta_2 = 2, \beta_3 = -5$
- RM: $Parite_i = \alpha + \beta_1 Urbain_i + 2 Prim_i - 5 Col_i + \beta_4 Catho_i + ... + \beta_8 * age_i + \epsilon_i$
- Calculer une nouvelle variable dépendante: $Paritenew_i = Parite_i - 2 * primaire_i + 5 * college_i$
- Estimer le nouvel modele:
- RM: $Paritenew_i = \alpha + \beta_1 * urbain_i + \beta_4 * catholique_i + ... + \beta_8 * age_i + \epsilon_i$
- Calculer la statistique de F et tester l'hypothèse nulle

Remarque sur F et t test

- Le test t et le test F sont similaires pour le test d'hypothèse sur 1 paramètre
- Deux tests t sont différents d'un test F car les hypothèses nulles sont différentes

Remarque sur F et t test

- Le test t et le test F sont similaires pour le test d'hypothèse sur 1 paramètre
- Deux tests t sont différents d'un test F car les hypothèses nulles sont différentes
- Par exemple: β_1 et β_2 peuvent ne pas être significatifs à partir du test t, mais être conjointement significatifs avec le F test

Remarque sur F et t test

- Le test t et le test F sont similaires pour le test d'hypothèse sur 1 paramètre
- Deux tests t sont différents d'un test F car les hypothèses nulles sont différentes
- Par exemple: β_1 et β_2 peuvent ne pas être significatifs à partir du test t, mais être conjointement significatifs avec le F test
- SSR(RM) $>$ SSR (UM)

Remarque sur F et t test

- Le test t et le test F sont similaires pour le test d'hypothèse sur 1 paramètre
- Deux tests t sont différents d'un test F car les hypothèses nulles sont différentes
- Par exemple: β_1 et β_2 peuvent ne pas être significatifs à partir du test t, mais être conjointement significatifs avec le F test
- SSR(RM) $>$ SSR (UM)
- Le modèle restreint et le modèle sans restriction doivent être basés sur le même échantillon avec le **même nombre d'observations**.

Degré de liberté

- Les degrés de liberté d'une statistique sont le nombre de grandeurs entrant dans le calcul de la statistique moins le nombre de contraintes reliant ces grandeurs.
- Par exemple, la formule utilisée pour calculer la variance de l'échantillon implique la statistique moyenne de l'échantillon.
- Cela impose une contrainte sur les données - étant donné la moyenne de l'échantillon, n'importe quel point de données peut être déterminé par les autres (N-1) points de données.
- Par conséquent, seules des observations non contraintes (N-1) sont disponibles pour estimer la variance de l'échantillon; le degré de liberté de la statistique de variance de l'échantillon est (N-1).

Références

- Wonnacott & Wonnacott. 1995. Statistique. Chapitre 9: Tests d'hypothèses.
- Fox (p190-197)
- Fox (207-224)
- Fox (232-235
- Fox (246-254
- Fox (p258-270)
- Fox (p385-417)
- Fox (p429-436)