

Significativité statistique et qualité d'ajustement

Gilbert Ritschard

Département d'économétrie, Université de Genève

<http://mephisto.unige.ch>

Vidéo-conf avec Bamako, 19 novembre 2003

1 Objectifs

1. Comprendre
 - l'aléa d'échantillonnage
 - la variabilité des estimations d'un échantillon à l'autre (erreur standard)
2. Se familiariser avec la notion de degré de signification
3. Initiation à la modélisation
 - régression linéaire
 - régression logistique
 - ...
4. Savoir juger de la qualité d'ajustement
 - Distance par rapport à la cible
 - Gain par rapport à un modèle de référence (indépendance)

Significativité statistique et qualité d'ajustement

Gilbert Ritschard

Département d'économétrie, Université de Genève

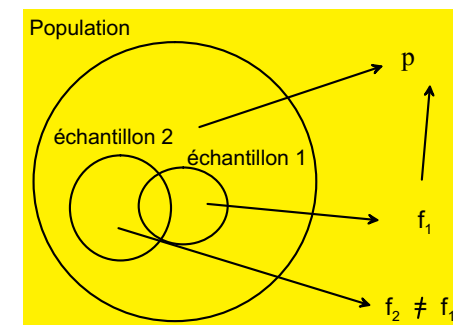
gilbert.ritschard@themes.unige.ch

Plan

- 1 Objectifs
- 2 Aléa d'échantillonnage
- 3 Erreur standard
- 4 Degré de signification
- 5 Modélisation
- 6 Qualité d'ajustement
- 7 Conclusion

<http://mephisto.unige.ch>

2 Aléa d'échantillonnage



Exemple : estimation de la proportion de filles

Soit une population avec $p = 56\%$ de filles.

On a tiré au hasard 5 échantillons de taille 5

i	échantillon				
	1	2	3	4	5
1	H	F	F	H	F
2	H	H	F	F	F
3	F	F	F	H	H
4	F	F	F	H	H
5	F	F	H	H	H
\hat{p}	60%	80%	80%	20%	40%

Ici, les estimations varient beaucoup d'un échantillon à l'autre

Plus l'échantillon est grand, moins les estimations varient.

Exemple de régression REGRESSION

/STATISTICS COEFF R ANOVA

/DEPENDENT nbgross

/METHOD=ENTER age agerapse amour rage2 .

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.208	.241		-.864	.388
	AGE âge (années)	.079	.009	.296	8.331	.000
	AGERAPSE Age au premier rapport sexuel	-.093	.013	-.256	-7.251	.000
	AMOUR Etes-vs déjà tombé amoureux (se)?	.617	.076	.248	8.090	.000
	RAGE2 nombre de jeunes de 20-30 ans dans la concession	-.006	.012	-.014	-.452	.652

a. Dependent Variable: NBGGROSS Nombre de grossesse

$$\text{nbgross} = -.208 + .079 \text{ age} - .093 \text{ agerapse} + .617 \text{ amour} - .006 \text{ rage2}$$

3 Erreur standard

La variabilité des estimations est évaluée par l'erreur standard

Voici un extrait d'output obtenu avec

DESCRIPTIVES

VARIABLES=agegros1

/STATISTICS=MEAN SEMEAN .

Descriptive Statistics

	N	Mean	
	Statistic	Statistic	Std. Error
AGEGROS1 âge à la première grossesse	458	18.48	.18
Valid N (listwise)	458		

Ici, l'erreur standard de .18 indique une faible variabilité de la moyenne

⇒ bonne confiance dans la moyenne trouvée de 18.48

4 Degré de signification

La statistique calculée (estimation, distance, ...) est-elle suffisamment éloignée de 0 ?

Degré de signification : probabilité d'obtenir une valeur plus éloignée de 0 pour un échantillon (en admettant qu'elle soit nulle dans la population.)

Exemple (cf régression) :

$$p(|\hat{\beta}_{\text{age}}| > .079) = 0\%$$

$$p(|\hat{\beta}_{\text{agerapse}}| > .093) = 0\%$$

$$p(|\hat{\beta}_{\text{amour}}| > .617) = 0\%$$

$$p(|\hat{\beta}_{\text{rage2}}| > .006) = 65\%$$

Le coefficient de `rage2` n'est pas statistiquement significatif : il y a 65% de chances d'obtenir .006 ou plus pour l'échantillon quand l'effet est nul dans la population.

Degré de signification (*p*-valeur, *p*-value, sig., prob., ...)

Probabilité que la «distance» entre

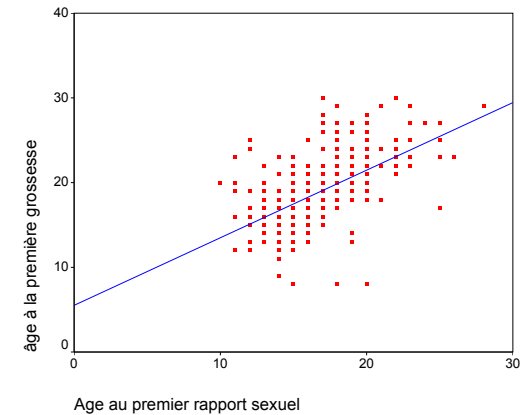
- ce que l'on observe (échantillon)
 - valeur du coefficient de *rage2*; nbre de grossesses; ...
- et ce que l'on attend sous l'hypothèse H_0 :
 - coefficient de *rage2* nul; nbre de grossesses indépendant de *age*, *agerapse*, *amour*, *rage2*; nbre de grossesses régi par le modèle postulé; ...

soit due au hasard de l'échantillonnage.

Utilisation

degré de signification < 5% (ou 10%) ⇒ rejet de l'hypothèse
 ≥ 5% (ou 10%) ⇒ hypothèse raisonnable

Régression : $agegros1 = 5.5 + 0.8 \cdot agerapse + \epsilon$



Pour *agerapse* = 14, la prédiction est : $agegros1 = 5.5 + .8 \cdot 14 = 16.7$ ans

5 Modélisation

Description mathématique permettant de prédire les valeurs prises par une grandeur :

Exemples (avec variable réponse quantitative) :

Moyenne indépendante de facteurs explicatifs l'âge à la première grossesses (*agegros1*) distribuée selon une loi normale $N(\mu = 18, \sigma = 3)$

$$agegros1 = 18 + \epsilon \quad \epsilon \sim N(0, \sigma_\epsilon = 3)$$

On prédit en moyenne 18 ans (l'écart type des observations autours de 18 étant de 3 ans)

Régression la valeur attendue de *agegros1* dépend linéairement de l'âge au premier rapport sexuel

$$agegros1 = 5.5 + 0.8 \cdot agerapse + \epsilon \quad \epsilon \sim N(0, \sigma_\epsilon)$$

Régression logistique

Variable réponse dichotomique :

«A eu rapport sexuel avant 16 ans» (oui = 1, non = 0)

p probabilité du oui

Régression logistique postule que $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$ dépend linéairement de facteurs explicatifs (genre, quartier, nbre de jeunes dans concession)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	SEXE(1)	-.345	.119	8.453	1	.004	.708
	QUARTIER			2.850	2	.241	
	QUARTIER(1)	.061	.195	.100	1	.752	1.063
	QUARTIER(2)	-.170	.159	1.150	1	.284	.844
	RAGE2	.036	.022	2.680	1	.102	1.037
	Constant	-1.225	.163	56.429	1	.000	.294

a. Variable(s) entered on step 1: SEXE, QUARTIER, RAGE2.

$$\text{logit}(p) = -1.225 - .345 \text{ homme} + .061 \text{ Niarela} - .17 \text{ Sicorini} + .36 \text{ rage2}$$

Régression logistique (suite)

$$\text{logit}(p) = -1.225 - .345 \text{ homme} + .061 \text{ Niarela} - .17 \text{ Sicorini} + .36 \text{ rage2}$$

Exemple : Homme, Badangiara Coura, 2 jeunes dans concession \Rightarrow

$$\text{logit}(p) = -1.225 - .345 \cdot 1 + .061 \cdot 0 - .17 \cdot 0 + .36 \cdot 2 = -.85$$

$$p = \frac{1}{1 + \exp(-\text{logit}(p))}$$

$$= \frac{1}{1 + \exp(.85)} = \frac{1}{1 + 2.34} = .29 = \boxed{29\%}$$

Pour une fille, $\text{logit}(p) = -1.225 + .72 = -0.505 \Rightarrow p = \frac{1}{1 + \exp(.505)} = 38\%$

Qualité d'ajustement en régression

$D(M, O)$ est mesuré par l'erreur standard de régression

$D(I, M)$ est évalué par le R^2 qui mesure en fait $\frac{D(I, O) - D(M, O)}{D(I, O)}$

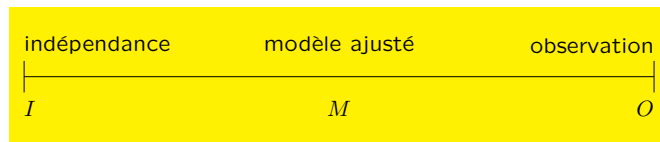
Régression pour le nombre de grossesses

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.363 ^a	.132	.128	.974

a. Predictors: (Constant), RAGE2 nombre de jeunes de 20-30 ans dans la concession, AMOUR Etes-vs déjà tombé amoureux (se)?, AGERAPSE Age au premier rapport sexuel, AGE âge (années)

6 Qualité d'ajustement



En notant : modèle ajusté M , sans prédictors I , observations O
on distingue deux types d'indicateurs :

1. divergence par rapport aux données $D(M, O)$
2. gain par rapport au modèle sans prédictors $D(I, M) = D(I, O) - D(M, O)$

Mesure de la divergence

optique moindres carrés : $D(M, O) =$ somme de carrés d'écart

optique maximum de vraisemblance : $D(M, O) = -2\text{LogLik}$

Statistiques du khi-2

Pour autres types de modélisation (régression logistique, modèles de Cox, ...), distance aux observations mesurée par le -2LogLik ($-2LL = -2$ fois logarithme de la vraisemblance)

$$-2LL(I) - [-2LL(M)] \sim \chi_d^2 \text{ loi du khi-carré}$$

avec d degrés de liberté, $d =$ différence du nombre de paramètres

Omnibus Tests of Model Coefficients

Step		Chi-square	df	Sig.
1	Step	13.217	4	.010
	Block	13.217	4	.010
	Model	13.217	4	.010

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1816.085	.007	.011

Statistiques du khi-2 (suite)

Selon les contextes, on utilise parfois d'autres statistiques du khi-2 (Pearson, Score, ...)

Le principe est toujours le même :

Distance par rapport à une hypothèse

- indépendance (modèle sans facteurs explicatifs)
- égalité de distributions
- égalité de courbe de survie (cf. Kaplan-Meier)
- ...

Pour juger d'une statistique du khi-2, il faut savoir quelle est l'hypothèse testée!

Exemple : Tableau croisé : hypothèse d'indépendance

RAPOCANB fréquence des rapports sexuels occasionnels * SANTEGO auto-évaluation état santé
Crosstabulation

Count

		SANTEGO auto-évaluation état santé				Total
		1 Très bonne	2 bonne	3 moyenne	4 mauvaise	
RAPOCANB	1 une fois	18	3	6	2	29
fréquence	2 plusieurs fois	117	76	36	3	232
des rapports	3 souvent	58	36	17	3	114
sexuels	4 pas de rapport	33	1	5	0	39
occasionnels	5 non qu'avec	341	169	148	12	670
	6 non jamais de	352	148	92	10	602
	rapports sexuels					
Total		919	433	304	30	1686

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	44.844 ^a	15	.000
Likelihood Ratio	48.247	15	.000
Linear-by-Linear Association	.730	1	.393
N of Valid Cases	1686		

a. 4 cells (16.7%) have expected count less than 5. The minimum expected count is .52.

Exemple : Statistique de Hosmer-Lemeshow (mesure $D(M, O)$)

Hypothèse : le modèle de régression logistique est correct pour les données ($D(M, O)$ petit).

Hosmer and Lemeshow Test

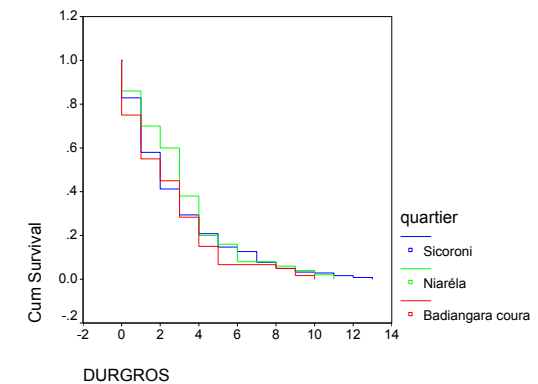
Step	Chi-square	df	Sig.
1	7.434	8	.491

Ici, comme le degré de signification est $> .05$, la distance est petite, ce qui indique un bon ajustement aux données.

Exemple : KM : hypothèse d'égalité de courbes de survie

Durée (durgros) entre premier rapport sexuel et 1ère grossesse selon quartier

Survival Functions



Et voici les statistiques de test fournies par la procédure KM :

Test Statistics for Equality of Survival Distributions for QUARTIER

	Statistic	df	Significance
Log Rank	1.73	2	.4208
Breslow	3.04	2	.2192
Tarone-Ware	2.49	2	.2873

7 Conclusion

Evaluation d'un modèle :

1. Ajustement global : khi-2, R^2 , etc.

Principe de parcimonie : retenir le modèle le plus simple qui donne un ajustement satisfaisant

Avec trop de paramètres, le modèle « colle » trop à l'échantillon, et manque de robustesse.

Retenir modèle avec plus petit AIC ou BIC.

2. Significativité individuelle des effets
3. L'examen des résidus (écarts prédiction-observation) peut fournir des informations utiles.