Tests d'hypothèses et intervalles de confiance dans la régression multiple
(SW Chap 7)

Plan de route

1. Test d'hypothèse et intervalle de confiance sur un coefficient
2. Test d'hypothèses conjointes sur plusieurs coefficients
3. Autres types d'hypothèses sur plusieurs coefficients
4. Variables d'intérêt, variables de contrôle et sélection des régresseurs

Test d'hypothèse et intervalle de confiance sur un coefficient (SW Section 7.1)

- Mêmes principes que dans une régression simple.
- On sait que le ratio $t = \dfrac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\text{var}(\hat{\beta}_1)}}$ suit approximativement la loi $N(0,1)$ lorsque n est assez grand (TCL).
- On peut donc procéder à des tests sur β_1 en utilisant la statistique t et les valeurs critiques de la loi N(0,1).
- Les intervalles de confiance se construisent de la même manière. Par exemple, à 95% : $\{\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)\}$.
- Idem pour les autres coefficients $\beta_2, \ldots, \beta_k$.

Exemple: Les données californiennes

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.650 PctEL$$
$$(8.7) \quad (0.43) \qquad (0.031)$$

- Dans ce modèle, le coefficient de STR est son effet marginal sur *TestScores*, si on maintient *PctEL* constant.
- Un intervalle de confiance à 95% pour ce coefficient est :

$$\{-1.10 \pm 1.96 \times 0.43\} = (-1.95, -0.26)$$

- Soit à tester l'hypothèse nulle $\beta_{STR} = 0$. Le *t* de Student est donné par: $t = -1.10/0.43 = -2.54$. On rejette donc la nulle (la nulle = l'hypothèse nulle) dans un test bilatéral au seuil de 5% de significativité.
- Idem pour un test unilatéral à gauche *(pourquoi ?)*

Écarts-types des coefficients dans la régression multiple
Exemple dans STATA

```
reg testscr str pctel, robust;

Regression with robust standard errors                Number of obs =     420
                                                      F(  2,   417) =  223.82
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.4264
                                                      Root MSE      =  14.464
------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -1.101296   .4328472    -2.54   0.011     -1.95213   -.2504616
       pctel |  -.6497768   .0310318   -20.94   0.000     -.710775   -.5887786
       _cons |   686.0322   8.728224    78.60   0.000     668.8754     703.189
------------------------------------------------------------------------------
```

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.650 PctEL$$
$$(8.7)\ \ (0.43) \qquad\qquad (0.031)$$

Remarque : les écart-types ci-dessus sont robustes à l'hétéroscédasticité.

Test d'hypothèses conjointes sur plusieurs coefficients (SW Section 7.2)

- Soit la variable *Expn* = dépenses gouvernementales par élève et considérons la régression multiple suivante:

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

- L'hypothèse nulle selon laquelle les ressources allouées à l'éducation n'ont pas d'impact sur les performances des élèves se formule comme suit :

$$H_0: \beta_1 = 0 \textbf{ et } \beta_2 = 0$$

- L'hypothèse alternative est :

$$H_1: \beta_1 \neq 0 \textbf{ ou } \beta_2 \neq 0$$

Hypothèses conjointes - Tests juxtaposés

- Une hypothèse nulle formulée comme ci-dessus est dite “conjointe” car elle impose plusieurs restrictions sur plusieurs coefficients.
- On dit aussi hypothèses "jointes" ou "composite".
- D’une manière générale, une hypothèse jointe peut imposer *q* restrictions ou contraintes sur les coefficients.
  - *Dans l’exemple précédent, q=2 :* $\boldsymbol{\beta_1 = 0}$ ***et*** $\boldsymbol{\beta_2 = 0}$.
- Dans un test au seuil de 5%, le bon sens suggère de rejeter la nulle dès lors que la statistique *t* associé à l’un des deux coefficients est plus grand que 1,96. Cette procédure intuitive s’appelle le « ***test juxtaposé*** ».

Hypothèses conjointes- test juxtaposé

- Cependant, le test juxtaposé conduit à rejeter la nulle avec plus de 5% des chances lorsqu'elle est vraie.
  - *Si les ressources allouées à l'éducation n'ont vraiment pas d'effet sur l'apprentissage (H_0 vraie), la procédure conclura le contraire avec plus de 5% de probabilité.*
- La taille d'un test est son taux réel de rejet de l'hypothèse nulle (H_0) lorsque celle-ci est vraie.
- Le niveau nominal d'un test est le taux de rejet souhaité de la nulle (celui utilisé pour choisir les valeurs critiques).
  - Un test dont la taille dépasse le niveau nominal est dit « invalide ».
  - Un test dont la taille est plus faible que le niveau nominal est dit « conservateur ».

Hypothèses conjointes- test juxtaposé

- Le test juxtaposé est invalide car il rejette la nulle avec plus de 5% de chance. Pour le voir, calculons la probabilité de rejeter l'hypothèse nulle.
- Pour simplifier, supposons que $\hat{\beta}_1$ et $\hat{\beta}_2$ sont indépendants (ceci rarement vrai dans la vraie vie). Soit t_1 et t_2 les t de Student associés aux deux coefficients:

$$t_1 = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \text{ et } t_2 = \frac{\hat{\beta}_2 - 0}{SE(\hat{\beta}_2)}$$

- Pour test au seuil de 5%, la règle de décision serait:

  Rejeter H_0: $\beta_1 = \beta_2 = 0$ lorsque $|t_1| > 1.96$ ou $|t_2| > 1.96$

Hypothèses conjointes- test juxtaposé

- En effet, si l'hypothèse nulle est vraie, la probabilité de la rejeter quand même est :

$$\Pr_{H_0}[|t_1| > 1.96 \text{ ou } |t_2| > 1.96]$$

$$= 1 - \Pr_{H_0}[|t_1| \leq 1.96 \text{ et } |t_2| \leq 1.96]$$

$$= 1 - \Pr_{H_0}[|t_1| \leq 1.96] \times \Pr_{H_0}[|t_2| \leq 1.96]$$

(Par indépendance supposée entre t_1 et t_2)

$$= 1 - (.95)^2$$

$$= .0975 = 9.75\% - \text{qui est plus grand que 5\%!!}$$

- La taille du test juxtaposé est de l'ordre de 9.75%. Ce test est donc invalide. Ceci reste vrai, quoique moins sévère, si t_1 et t_2 sont corrélés.

Deux solutions possibles:

- La méthode de *Bonferroni*, qui consiste à utiliser une valeur critique différente de 1,96 (voir SW Appendice 7.1). Cette méthode est rarement utilisée en pratique.
- Le F-test, qui consiste à utiliser une statistique plus adaptée aux hypothèses jointes. Cette méthode est présentée ci-après.

Le *F*-test

- Le F-test permet de tester conjointement plusieurs hypothèses, ce qui évite de devoir juxtaposer des décisions prescrites par des tests séparées.
- Pour l'hypothèse conjointe "$\beta_1 = \beta_{1,0}$ et $\beta_2 = \beta_{2,0}$" à tester dans le cadre d'une régression multiple à deux variables explicatives, la formule de la statistique F est:

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1,t_2}^2}\right)$$

où $\hat{\rho}_{t_1,t_2}$ est un estimateur de la corrélation entre t_1 et t_2.

- On rejette la nulle lorsque F est plus grand qu'un seuil donnée (*mais comment choisir ce seuil?*).

Le *F*-test

- La statistique F (encore appelé F-stat) tient compte de la corrélation entre t_1 et t_2.
- Dans un modèle linéaire qui contient plus de variables explicatives, la statistique *F* peut aussi être calculée pour tester une ou plusieurs combinaisons linéaires de paramètres.
- Sa formule devient alors plus compliquée, mais demeure relativement sympathique en notation matricielle.
- Quelle est la distribution de la statistique *F*?
  - Cette distribution est nécessaire pour déterminer le seuil de rejet de l'hypothèse nulle.

Distribution de la *F*-stat en grand échantillon

- Considérons le cas spécial où t_1 et t_2 sont indépendant, de sorte que $\hat{\rho}_{t_1,t_2} \xrightarrow{p} 0$.
- En grand échantillon, on a :

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1,t_2}^2}\right) \cong \frac{1}{2}(t_1^2 + t_2^2)$$

- Sous la nulle, t_1 et t_2 suivent des lois normales centrées et réduite, et on a supposé qu'ils sont indépendants.
- Donc *F* est la moyenne de deux variables aléatoires indépendantes suivant la du khi-deux à un ddl.

- La loi du ***khi-deux*** à q degrés de liberté (χ^2_q) est la loi suivie par la somme des carrés de q variables aléatoires indépendantes de lois N(0,1).

En grand échantillon, F suit la loi χ^2_q***/q.***

La distribution χ^2_q**/*q***

q	valeur critique à 5%	
1	3.84	
2	3.00	(notre cas: q=2)
3	2.60	
4	2.37	
5	2.21	

La p-valeur du F-test

p-valeur = surface de la queue supérieure de χ^2_q/q, à partir de la valeur de F calculée

Mise en œuvre dans STATA

On utilise la commande "test" après une régression

Exemple : données californiennes

Dans la régression suivante, soit à tester si l'hypothèse jointe selon laquelle les coefficients de *STR* et de *expn_stu* sont tous les deux égaux à zéro. L'hypothèse alternative consiste à dire qu'au moins un de ces deux coefficients est non nul.

La régression multiple suivie du test d'hypothèse

```
reg testscr str expn_stu pctel, r;

Regression with robust standard errors                Number of obs =     420
                                                      F(  3,   416) =  147.20
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.4366
                                                      Root MSE      =  14.353

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -.2863992   .4820728    -0.59   0.553    -1.234001     .661203
    expn_stu |   .0038679   .0015807     2.45   0.015     .0007607    .0069751
       pctel |  -.6560227   .0317844   -20.64   0.000    -.7185008   -.5935446
       _cons |   649.5779   15.45834    42.02   0.000     619.1917    679.9641
------------------------------------------------------------------------------
```

NOTE
Commande du test

```
test str expn_stu;

 ( 1)  str = 0.0
 ( 2)  expn_stu = 0.0

       F(  2,   416) =    5.43
            Prob > F =    0.0047
```

q=2 restrictions

La valeur critique à 5% pour q=2 est 3.00
Stata calcule la p-valeur pour nous!

Le *F*-test sous l'hypothèse d'homoscédasticité

- La F-stat admet une expression simple lorsque les erreurs sont homoscédastiques
- On commence par estimer deux régressions : une régression sous la nulle (modèle contraint) et une autre sous l'alternative (modèle non contraint).
- Si le R^2 varie peu suivant les deux régressions, cela signifie que les contraintes imposées par la nulle ne peuvent être rejetées.
- Si le R^2 sous la nulle est significativement plus faible que sous l'alternative, alors on rejette la nulle.

Le *F*-test sous l'hypothèse d'homoscédasticité

Exemple: Les coefficients de *STR* et *Expn* sont-ils tous deux égaux à zéro?

Modèle non contraint (sous H_1):

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

Modèle contraint (sous H_0):

$$TestScore_i = \beta_0 + \beta_3 PctEL_i + u_i$$

- La qualité d'ajustement (R^2) du modèle non contraint sera meilleure que celle du modèle contraint.
- En fait, ceci revient à tester si l'augmentation du R^2 apportée par *STR* et *Expn* justifie leur présence dans le modèle non contraint.

Le *F*-test sous l'hypothèse d'homoscédasticité

$$F = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})/(n - k_{unrestricted} - 1)}$$

avec:

$R^2_{restricted}$ = R^2 du modèle contraint (sous H_0)

$R^2_{unrestricted}$ = R^2 du modèle non contraint (sous H_1)

q = nombre de restrictions ou contraintes sous H_0

$k_{unrestricted}$ = nombre de régresseurs sous H_1.

- Plus la différence entre le R^2 du modèle non contraint et du modèle contraint est grande, plus les contraintes imposées par H_0 sont « contraignantes » et plus la valeur de F est grande.

Exemple: données californiennes

Modèle contraint

$$\widehat{TestScore} = 644.7 - 0.671 PctEL, \quad R^2_{restricted} = 0.4149$$
$$(1.0) \quad (0.032)$$

Modèle non contraint:

$$\widehat{TestScore} = 649.6 - 0.29 STR + 3.87 Expn - 0.656 PctEL$$
$$(15.5) \quad (0.48) \qquad (1.59) \qquad (0.032)$$

$$R^2_{unrestricted} = 0.4366, \; k_{unrestricted} = 3, \; q = 2$$

Donc
$$F = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})/(n - k_{unrestricted} - 1)}$$
$$= \frac{(.4366 - .4149)/2}{(1 - .4366)/(420 - 3 - 1)} = \mathbf{8.01}$$

Note: Sans l'hypothèse d'homoscédasticité: F = **5.43**…

Distribution suivie par la F-stat sous l'hypothèse d'homoscédasticité : la *loi F* ou *loi de Fisher*

- La distribution χ^2_q/q est en fait une approximation de la vraie loi suivie par la statistique *F*.
- En plus des hypothèses #1 à 4, supposons que :

  5. *u_i est homoscédastique, c'est-à-dire, var($u|X_1,\ldots,X_k$) ne dépend pas de X.*

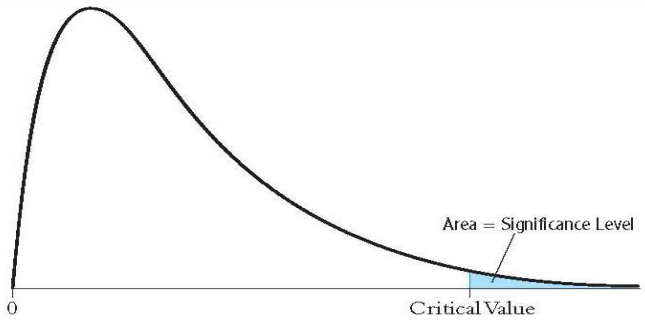
  6. *$u_1,\ldots,u_n$ suivent des lois normales.*

- Alors, la statistique *F* suit la loi de Fisher à (*q*,*n*-*k*-*1*) degrés de liberté (note $F_{q,n-k-1}$), où *q* est le nombre de contraintes et *k* le nombre de variables explicatives du modèle non contraint.

La distribution F (ou *loi de Fisher*)

- La loi de Fisher est encore appelée “la *loi F*”.
- La loi *F* est tabulée. Ses valeurs critiques peuvent être générées par la plupart des logiciels statistiques.
- Lorsque n $\rightarrow \infty$, la loi $F_{q,n-k-1}$ converge vers la loi χ^2_q/q. De ce fait, les distributions $F_{q,\infty}$ et χ^2_q/q sont identiques.
- Si *q* est assez petit par rapport à *n-k-1* (par exemple, *q=10* et *n>100*) les valeurs critiques de $F_{q,\infty}$ et χ^2_q/q sont peu différentes.
- Par défaut, les logiciels statistiques calculent la *p*-valeur du *F*-test en se basant sur la loi de Fisher.

TABLE 4 Critical Values for the $F_{m,\infty}$ Distribution



	Significance Level		
Degrees of Freedom	**10%**	**5%**	**1%**
1	2.71	3.84	6.63
2	2.30	3.00	4.61
3	2.08	2.60	3.78
4	1.94	2.37	3.32
5	1.85	2.21	3.02
6	1.77	2.10	2.80
7	1.72	2.01	2.64
8	1.67	1.94	2.51
9	1.63	1.88	2.41
10	1.60	1.83	2.32
11	1.57	1.79	2.25
12	1.55	1.75	2.18
13	1.52	1.72	2.13
14	1.50	1.69	2.08
15	1.49	1.67	2.04
16	1.47	1.64	2.00
17	1.46	1.62	1.97
18	1.44	1.60	1.93
19	1.43	1.59	1.90
20	1.42	1.57	1.88
21	1.41	1.56	1.85
22	1.40	1.54	1.83
23	1.39	1.53	1.81
24	1.38	1.52	1.79
25	1.38	1.51	1.77
26	1.37	1.50	1.76
27	1.36	1.49	1.74
28	1.35	1.48	1.72
29	1.35	1.47	1.71
30	1.34	1.46	1.70

This table contains the 90th, 95th, and 99th percentiles of the $F_{m,\infty}$ distribution. These serve as critical values for tests with significance levels of 10%, 5%, and 1%.

Digression: Un peu d'histoire des statistiques…

- La théorie du F-test dans le cas homoscédastique et la distribution $F_{q,n-k-1}$ reposent sur des hypothèses assez fortes, c'est-à-dire, difficile à satisfaire en pratique (*normalité et homoscédasticité des erreurs*).
- Cette théorie date du début du 20^{e} siècle, une époque où les ordinateurs n'étaient pas nés et les bases de données étaient limitées.
- La statistique F et la distribution $F_{q,n-k-1}$ constituent des percées majeure dans l'histoire de la statistique : une formule facile à calculer et une distribution élégante qu'on peut présenter sous–forme d'une table standard à usages multiples.

Digression: Un peu d'histoire des statistiques…

- Les hypothèses que nous jugeons fortes aujourd'hui représentaient un cout modeste par rapport à l'ampleur des percées scientifiques et leurs retombées pratiques.
- Avec la disponibilité d'ordinateurs modernes performants et des bases de données de grandes tailles, on peut désormais se passer des hypothèses de normalité des et d'homoscédasticité, en utilisant la formule du F qui est robuste à l'hétéroscédasticité et dont la distribution est $F_{q,\infty}$ en grand échantillon.
- Par défaut, les logiciels statistiques font l'hypothèse d'homoscédasticité et de normalité des erreurs.

Test d'hypothèses conjointes: résumé

- L'approche qui consiste à juxtaposer des tests séparés conduit à rejeter l'hypothèse nulle plus souvent que ce qu'indique le niveau nominal du test (e.g : niveau nominal de 5% => rejet de H_0 lorsque $t>1.96$). Le test juxtaposé est donc invalide.
- Lorsque n est grand, la distribution de la statistique F robuste à l'hétéroscédasticité est approximativement donné par χ^2_q/q (qui est aussi $F_{q,\infty}$).
- Si on fait l'hypothèse d'homoscédasticité et de normalité des erreurs, la distribution de F est donnée par la loi de Fisher $F_{q,n-k-1}$.

Test sur une combinaison linéaire de coefficients (SW Section 7.3)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i,\ i = 1,\ldots,n$$

Considérons le problème de test suivant :

$$H_0: \beta_1 = \beta_2 \quad \text{contre} \quad H_1: \beta_1 \neq \beta_2$$

L'hypothèse nulle impose une restriction unique sur deux coefficients. On a donc une seule contrainte (q=1) sous forme de combinaison linéaire (β_1-β_2=0).

Test sur une combinaison linéaire de coefficients

Ce genre d'hypothèse peut être testé selon deux méthodes:

1. ***Transformer l'équation de régression*** afin de simplifier la formulation du test.

2. ***Tester directement la contrainte*** en utilisant les résultats d'estimation du modèle non contraint.

Méthode 1: Transformer l'équation de régression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

Ajoutons et soustrayons $\beta_2 X_{1i}$:

$$Y_i = \beta_0 + (\beta_1 - \beta_2) X_{1i} + \beta_2 (X_{1i} + X_{2i}) + u_i$$

Ceci donne:

$$Y_i = \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + u_i$$

avec

$$\gamma_1 = \beta_1 - \beta_2 \quad \text{et} \quad W_i = X_{1i} + X_{2i}$$

Transformer l'équation de régression

(a) Équation de départ

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

(b) Équation transformée

$$Y_i = \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + u_i$$

où $\gamma_1 = \beta_1 - \beta_2$ and $W_i = X_{1i} + X_{2i}$

Ainsi, le test:

$$H_0: \beta_1 = \beta_2 \quad \text{contre} \quad H_1: \beta_1 \neq \beta_2$$

devient:

$$H_0: \gamma_1 = 0 \quad \text{contre} \quad H_1: \gamma_1 \neq 0$$

Les deux régressions ont les mêmes prédictions et le même R^2. Mais la formulation du test est simplifiée dans (b).

Méthode 2: Tester directement la contrainte

- On commence par estimer le modèle :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

- Ensuite, on se sert de la distribution suivie par $\hat{\beta}_1 - \hat{\beta}_2$ et des résultats d'estimation pour tester si $\beta_1 - \beta_2 = 0$. La démarche est similaire à ce qu'on a vu pour le test de différence entre deux moyennes.
- On calcule le *t* de Student : $t = (\hat{\beta}_1 - \hat{\beta}_2)/SE(\hat{\beta}_1 - \hat{\beta}_2)$. On rejette la nulle au seuil de 5% si $|t|>1{,}96$.
- ***Exemple dans STATA***

```
. regress testscore str expn pctel, robust
. test str=expn
```

Remarque

- On peut aussi utiliser le formalisme « modèle contraint » versus « modèle non contraint » vu plus tôt pour faire ce test. Ceci conduit à estimer deux modèles…
  - *Modèle non contraint :*

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

  - *Modèle contraint (c'est-à-dire, modèle si* $\beta_1 = \beta_2 = \beta_c$*) :*

$$Y_i = \beta_0 + \beta_c (X_{1i} + X_{2i}) + u_i$$

- Test basé sur la différence des R^2…

Intervalle de confiance joint pour plusieurs coefficients (SW Section 7.4)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, \quad i = 1,\ldots,n$$

- Un intervalle de confiance joint à 95% pour (β_1,β_2) est:
  - un ensemble dont les bornes sont calculées à partir des données et qui contient (β_1,β_2) avec 95% de chances.
  - de manière équivalente : un ensemble de paires (β_1,β_2) qui ne sont pas rejetées dans un test au seuil de 5%.
- Il s'agit d'un intervalle joint : ne pas le construire en croisant des intervalles obtenus séparément pour β_1 et β_2 (pour la même raison que dans le test conjoint).
- L'intervalle de confiance joint peut être construit à partir d'un *F*-test.

Intervalle de confiance joint

- Soit $F(\beta_{1,0},\beta_{2,0})$ la statistique F (version robuste à l'hétéroscédasticité) calculée pour tester l'hypothèse nulle

$$\beta_1 = \beta_{1,0} \text{ et } \beta_2 = \beta_{2,0}:$$

- Un intervalle de confiance joint à 95% pour (β_1,β_2) est donné par

$$\{(\beta_{1,0},\beta_{2,0}): F(\beta_{1,0},\beta_{2,0}) < 3.00\}$$

où 3.00 est la valeur critique de la loi $F_{2,\infty}$ au seuil de 5%.

- Ce procédé s'appelle "inversion d'un test" et produit un intervalle de confiance à 95%.
- Regardons de plus près l'ensemble:

$$\{(\beta_{1,0},\beta_{2,0}): F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1,t_2}^2}\right) \leq 3.00\}$$

Intervalle de confiance joint

- F est une forme quadratique en $\beta_{1,0}$ et $\beta_{2,0}$:

$$F = \frac{1}{2(1-\hat{\rho}^2_{t_1,t_2})} \times \left[t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2 \right]$$

$$= \frac{1}{2(1-\hat{\rho}^2_{t_1,t_2})} \times$$

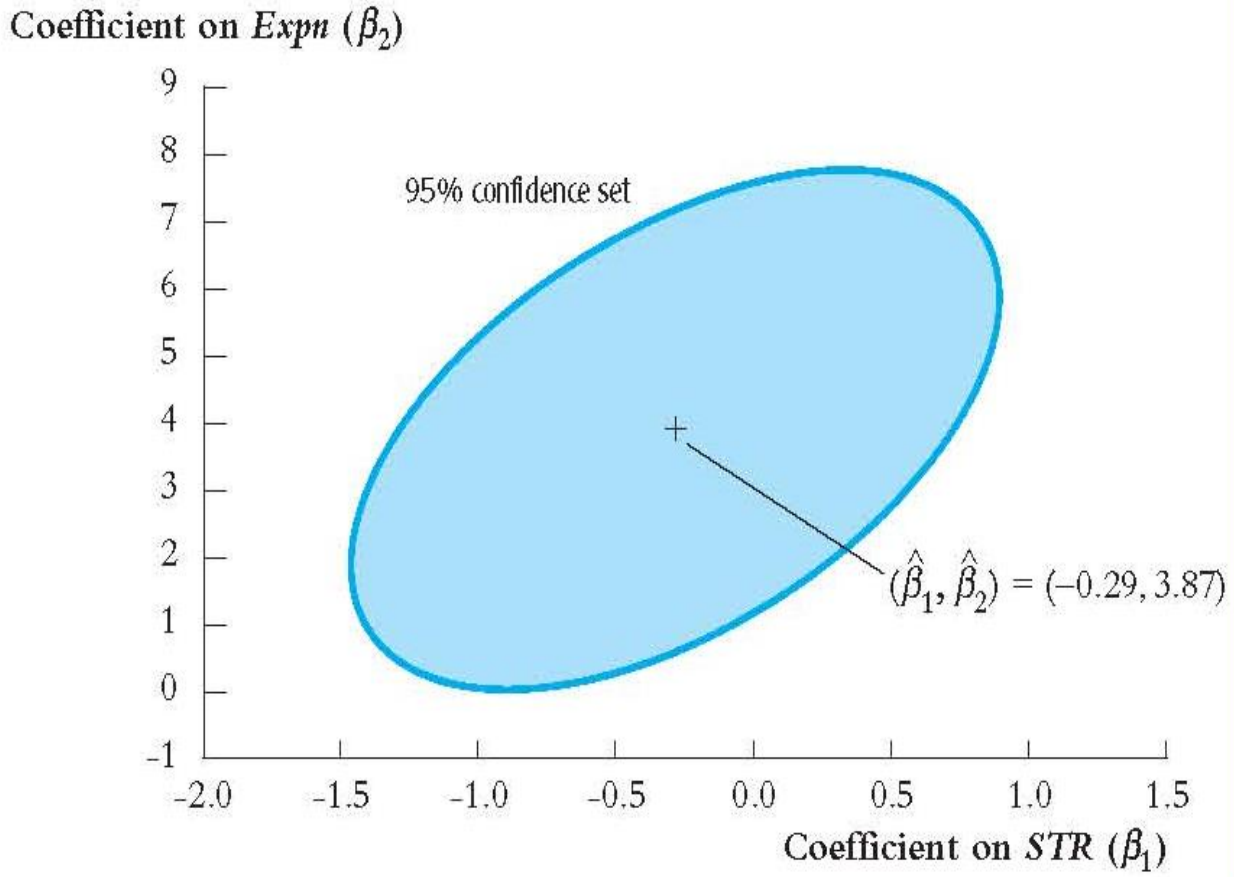
$$\left[\left(\frac{\hat{\beta}_2 - \beta_{2,0}}{SE(\hat{\beta}_2)} \right)^2 + \left(\frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} \right)^2 + 2\hat{\rho}_{t_1,t_2} \left(\frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} \right) \left(\frac{\hat{\beta}_2 - \beta_{2,0}}{SE(\hat{\beta}_2)} \right) \right]$$

- La frontière de l'ensemble $\{F = 3.00\}$ est une ellipse.
- Si la corrélation entre $\beta_{1,0}$ et $\beta_{2,0}$ est nulle, on obtient un cercle. Mais en aucun cas, un rectangle. (*Pourquoi j'évoque un rectangle ??)*

Intervalle de confiance obtenu par inversion d'un F-test

FIGURE 7.1 95% Confidence Set for Coefficients on *STR* and *Expn* from Equation (7.6)

The 95% confidence set for the coefficients on *STR* (β_1) and *Expn* (β_2) is an ellipse. The ellipse contains the pairs of values of β_1 and β_2 that cannot be rejected using the *F*-statistic at the 5% significance level.



Spécification du modèle linéaire
Variables d'internet, variables de contrôle et sélection des régresseurs (SW Section 7.5)

- On cherche un estimateur sans biais de l'effet de la taille de classe (*STR*) sur la qualité de l'apprentissage (*TestScore*), ***toutes choses égales par ailleurs***.
- Que veut dire « *toutes choses égales par ailleurs* » ?
  - *D'autres facteurs (comme l'apprentissage extra-scolaire, maîtrise de l'anglais, etc.) influencent également la qualité de l'apprentissage à l'école.*
  - *On veut mesurer la variation de TestScore qui dépend uniquement de STR, c'est-à-dire, si on « maintient les autres facteurs constants »*

Spécification du modèle (suite)

- Dans une expérience randomisée, on distribuerait les étudiants de manière aléatoire dans les écoles et les classes, de sorte que leurs caractéristiques (STR_i) seraient indépendantes des facteurs omis (u_i).
- L'hypothèse #1 serait alors satisfaite ($E(u_i|STR_i) = 0$) et l'estimateur MCO du coefficient de STR serait un estimateur sans biais de son effet causal sur *TestScore*.
- Avec des données observationnelles, il faut vérifier que les variables incluses dans le modèle ne sont pas corrélées aux variables omises.
- Idéalement, il faut identifier les variables omises (comme par exemple, *PctEL*) et les inclure dans le modèle.

Variables de contrôle

- Malheureusement, il n'est pas toujours possible d'observer les variables omises:
  - Par exemple, notre base de données n'indique pas le temps consacré par les parents à aider leurs enfants à faire les devoirs de maison.
- Une façon de contourner ce problème consiste à inclure des variables dites "de contrôle".
- Supposons que dans la régression de Y sur X, une variable importante Z a été omise car non disponible. Une variable de contrôle W est une variable disponible qu'on utilise pour remplacer Z parce qu'elle lui est corrélée, mais qui en elle-même n'a pas d'effet causal sur Y.

Variables de contrôle : Exemple

$$\widehat{TestScore} = 700.2 - 1.00STR - 0.122PctEL - 0.547LchPct,\ \bar{R}^2 = 0.773$$
$$(5.6)\ \ (0.27)\qquad (.033)\qquad\qquad (.024)$$

PctEL = Pourcentage d'élèves apprenant l'anglais
LchPct = Pourcentage d'élèves recevant de l'aide alimentaire
(Repas gratuits ou subventionnés accordés aux élèves provenant des familles à faibles revenus)

- Selon vous, quelles sont les variables d'intérêts? Les variables de contrôle ? Les variables omises qu'on cherche à contrôler?

Variables de contrôle : Exemple (suite)

$$\widehat{TestScore} = 700.2 - 1.00STR - 0.122PctEL - 0.547LchPct,\ \bar{R}^2 = 0.773$$
$$(5.6)\ \ (0.27) \qquad (.033) \qquad (.024)$$

- *STR* est la variable d'intérêt dans l'étude.
- *PctEL* peut avoir un effet causal: un enfant qui maîtrise mal l'anglais aura du mal à suivre ce que dit l'enseignant.
- Mais *PctEL* est aussi une variable de contrôle : les enfants d'immigrants sont plus concernés par les problèmes de langue et moins exposés à l'apprentissage extra-scolaire.
- *LchPct* pourrait aussi avoir un effet causal: ventre affamé n'a point d'oreille. Mais elle est surtout une variable qui contrôle l'effet du revenu des parents et de l'apprentissage extra-scolaire, non inclus dans le modèle.

Variables de contrôle (suite)

1. **Une bonne variable de contrôle peut être définit de manière équivalente comme:**
   i. Une variable sans effet causal dont l'inclusion dans la régression rend le terme d'erreur non corrélé avec la variable d'intérêt.
   ii. Si on maintient la variable de contrôle fixe, la variable d'intérêt et le terme d'erreur se comportent comme dans une expérience randomisée.
   iii. Parmi les observations ayant la même valeur pour la variable de contrôle, la variable d'intérêt est non corrélée aux variables omises.

Variables de contrôle (suite)

2. **Les coefficients des variables de contrôle n'ont généralement pas une interprétation causale**. *Par exemple*:

$$\widehat{TestScore} = 700.2 - 1.00STR - 0.122PctEL - 0.547LchPct,\ \bar{R}^2 0.773$$
$$(5.6)\ \ (0.27)\qquad (.033)\qquad\quad (.024)$$

- Le coefficient de *LchPct* a-t-il une interprétation causale?
- Si oui, on serait capable d'améliorer la qualité de l'apprentissage rien qu'en supprimant l'aide (*LchPct=0*) alimentaire (faites le calcul !)
- Est-ce logique ? Comment interpréter ce signe négatif ?

Indépendance de la moyenne conditionnelle (IMC)

- Soit Y_i la variable dépendante, X_i la variable explicative d'intérêt, Z_i la variable omise et W_i la variable incluse pour contrôler l'omission de Z_i.
- Pour que W_i soit une bonne variable de contrôle, il faut qu'elle soit corrélée à la variable omise Z_i.
- Ceci signifie que dans la régression de Y_i sur X_i et W_i, le terme d'erreur reste corrélé à W_i. Il y a donc violation de l'hypothèse #1 ($E(u_i|X_i,W_i) = 0$) et biais de variable omise.
- Mais si W_i est une bonne variable de contrôle, le biais de variable omise n'affectera que son coefficient et le coefficient de X_i sera épargné.

Indépendance de la moyenne conditionnelle (IMC)

- On vient de voir que la variable de contrôle est corrélée à la variable omise. C'est précisément ce qui fait d'elle une bonne variable de contrôle.
- On a vu aussi que l'hypothèse #1 est violée. Elle est remplacée par l'hypothèse d'***indépendance de la moyenne conditionnelle***, qui stipule que :

$$E(u_i|X_i,W_i) = E(u_i|W_i)$$

- Cette hypothèse requiert que l'espérance conditionnelle de u_i ne dépende pas de X_i. Lorsqu'elle est vérifiée, elle assure que l'estimateur MCO du coefficient de X est non biaisé et a une interprétation causale.
- ***On ne peut en dire de même du coefficient de W.***

Indépendance de la moyenne conditionnelle (IMC)

- Considérons le modèle linéaire suivant:

$$Y = \beta_0 + \beta_1 X + \beta_2 W + u$$

où X est la variable d'intérêt et W est une variable de contrôle effective, de sorte que l'hypothèse d'IMC est valide:

$$E(u_i|X_i, W_i) = E(u_i|W_i).$$

- En outre, supposons que les hypothèses #2, #3, et #4 sont valides. Alors:

1. β_1 admet une interprétation causale
2. $\hat{\beta}_1$ est sans biais
3. $\hat{\beta}_2$ est généralement biaisé (ne mesure pas l'effet "ceteris paribus" de W)

Indépendance de la moyenne conditionnelle (IMC)

1. β_1 a une interprétation causale

Si W est maintenu constant, l'effet attendu d'un changement de X sur Y est :

$$
\begin{aligned}
&E(Y|X = x+\Delta x, W=w) - E(Y|X = x, W=w) \\
&\quad = [\beta_0 + \beta_1(x+\Delta x) + \beta_2 w + E(u|X = x+\Delta x, W=w)] \\
&\qquad - [\beta_0 + \beta_1 x + \beta_2 w + E(u|X = x, W=w)] \\
&\quad = \beta_1 \Delta x + [E(u|X = x+\Delta x, W=w) - E(u|X = x, W=w)] \\
&\quad = \beta_1 \Delta x,
\end{aligned}
$$

puisque l'hypothèse d'IMC implique :

$E(u|X = x+\Delta x, W=w) = E(u|X = x, W=w) = E(u|W=w)$.

Indépendance de la moyenne conditionnelle (IMC)

2. $\hat{\beta}_1$ ***est sans biais***

3. $\hat{\beta}_2$ ***est généralement biaisé***

Preuve:

Reprenons le modèle linéaire,

$$Y = \beta_0 + \beta_1 X + \beta_2 W + u$$

Sans perte de généralité, supposons que :

$$E(u|W) = \gamma_0 + \gamma_2 W.$$

Sous l'hypothèse d'IMC, on a:

$$E(u|X, W) = E(u|W) = \gamma_0 + \gamma_2 W. \qquad (*)$$

Définissons donc:

$$v = u - E(u|X, W) \qquad (**)$$

Indépendance de la moyenne conditionnelle (IMC)

Si on combine (*) and (**) on obtient:

$$u = E(u|X, W) + v$$
$$= \gamma_0 + \gamma_2 W + v, \qquad (***)$$

Notez que $E(v|X, W) = 0$, et donc, v apparaît comme l'erreur de la régression de u sur X et W. Substituons (***) dans le modèle initial. On obtient:

$$Y = \beta_0 + \beta_1 X + \beta_2 W + u \qquad (+)$$
$$= \beta_0 + \beta_1 X + \beta_2 W + (\gamma_0 + \gamma_2 W + v) \quad \text{selon } (***)$$
$$= (\beta_0 + \gamma_0) + \beta_1 X + (\beta_2 + \gamma_2) W + v$$
$$= \delta_0 + \beta_1 X + \delta_2 W + v \qquad (++)$$

où $\delta 0 = \beta_0 + \gamma_0$ et $\delta 2 = \beta_2 + \gamma_2$.

Indépendance de la moyenne conditionnelle (IMC)

- Comme $E(v|X, W) = 0$, l'hypothèse #1 est satisfaite dans le modèle (++). Ceci signifie que les estimateurs MCO de δ_0, β_1, and δ_2 sont sans biais dans la régression (++).
- Comme les régresseurs de (+) et (++) sont identiques, cela signifie que les estimateurs MCO de (+) satisfont :

$$E(\hat{\beta}_1) = \beta_1 \text{ et}$$

$$E(\hat{\beta}_2) = \delta_2 = \beta_2 + \gamma_2 \neq \beta_2.$$

Donc, $\hat{\beta}_1$ est sans biais et $\hat{\beta}_2$ est biaisé.

Variable de contrôle: résumé

Soit *Y* une variable dépendante, *X* une variable d'intérêt, *Z* une variable omise et *W* une une variable de contrôle qui satisfait l'hypothèse d'indépendance de la moyenne conditionnelle. Si on régresse Y sur X et W, alors:

- L'estimateur MCO du coefficient de la variable d'intérêt, $\hat{\beta}_1$, est sans biais.
- L'estimateur MCO du coefficient de la variable de contrôle, $\hat{\beta}_2$, est biaisé. Ceci est dû à la corrélation entre la variable de contrôle W et la variable omise Z. Mais cette corrélation est nécessaire pour que W soit une variable de contrôle effective.

Implications pour la sélection de variables et la spécification du modèle

1. Identifier les variables d'intérêt.

2. Identifier les variables omises qui peuvent potentiellement induire le biais d'omission de variable.

3. Si les variables omises sont disponibles, les inclure. Sinon, en inclure d'autres qui leur sont corrélées et qui serviront de variables de contrôle. Les variables de contrôle sont dites effectives si elles satisfont l'hypothèse d'IMC. On obtient ainsi un modèle de base ou "benchmark".

Spécification du modèle (suite)

4. Spécifier également une série de modèles alternatifs plausibles, en jouant sur la liste des régresseurs. Ces modèles alternatifs servent à *l'analyse de sensibilité*.

5. Estimer tous les modèles
   - L'ajout d'un certain régresseur modifie-t-il les estimés des coefficients des variables d'intérêt?
   - Quels coefficients sont statistiquement significatifs? Sont-ils des effets causaux ?
   - Privilégier le jugement et éviter des recettes mécaniques.
   - Ne pas avoir pour seul objectif de maximiser le R2.

Digression au sujet du R^2

On peut facilement tomber dans le piège de la maximisation du R^2 ou du $\bar{R}^2$. Mais ne perdons pas de vue l'objectif initial, qui est d'estimer sans biais l'effet causal de la taille de classe sur l'apprentissage.

- Un R^2 (ou $\bar{R}^2$) élevé signifie que les régresseurs expliquent une grande portion de la variance de Y.
- Un R^2 (ou $\bar{R}^2$) élevé ne signifie pas que le biais de variable omise a été éliminé.
- Un R^2 (ou $\bar{R}^2$) élevé ne signifie pas que l'estimation de l'effet causal d'une variables d'intérêt est non biaisé.
- Un R^2 (ou $\bar{R}^2$) élevé ne signifie pas que les coefficients des variables incluses sont statistiquement significatifs.

Exemple: Les données californiennes (SW Section 7.6)

1. Identification de la variable d'intérêt:

   STR

2. Identification des variables importantes dont l'omission entrainerait le « biais de variable omise »

   Maîtrise de l'anglais

   Apprentissage extra-scolaire

   Aide parental pour devoirs de maison

   Revenu des parents

   Compétences du maître

   etc… la liste est longue!

3. Si les variables omises sont disponibles, les inclure. Sinon, recourir à des variables de contrôle…

   La plupart des variables omises sont difficiles à mesurer. Nous aurons recours à des variables de contrôle (tel que PctEl, qui a aussi une interprétation causale) et des variables qui mesurent de la richesse.

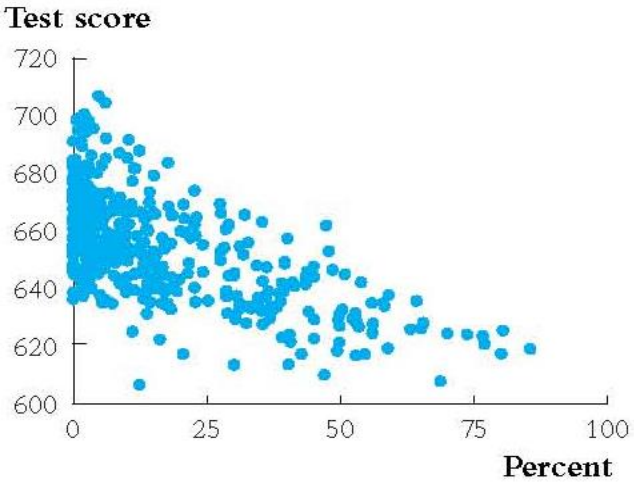
4. Modèles alternatifs pour l'analyse de sensibilité

   La richesse peut se mesurer de plusieurs façons. On essayera donc plusieurs mesures de richesses afin de d'identifier celle qui corrige le mieux le biais de variable omise.

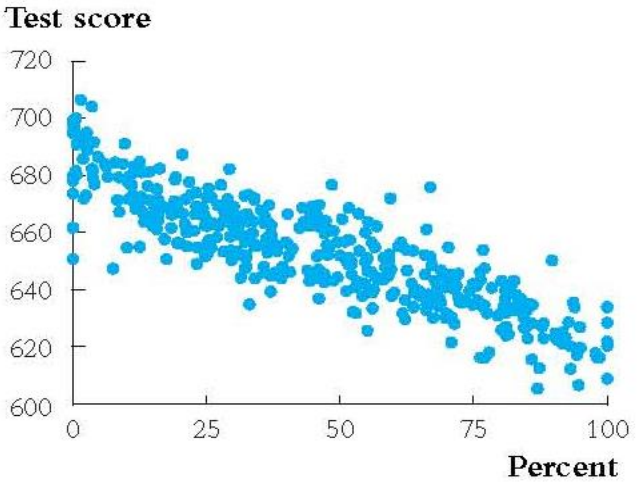
5. Estimer tous les modèles spécifiés.

Données californiennes: nuages de points

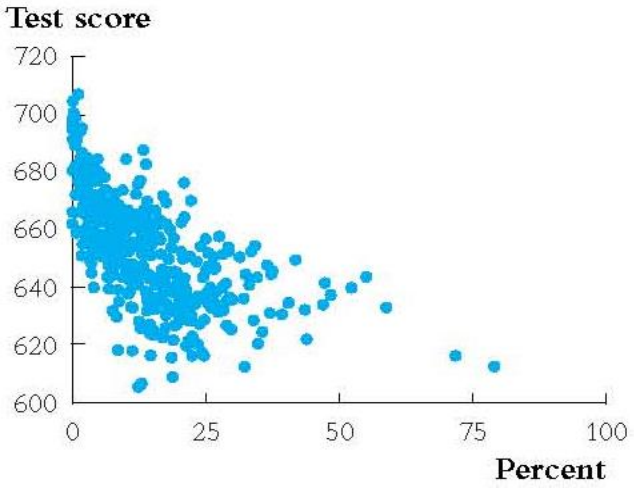
FIGURE 7.2 Scatterplots of Test Scores vs. Three Student Characteristics



(a) Percentage of English language learners



(b) Percentage qualifying for reduced price lunch



Présentation des résultats

- Lorsqu'on a plusieurs régressions, il faut rapporter les résultats sous forme de tableaux synthétiques. Pour des raisons de lisibilité, éviter de les rapporter sous-forme d'équations.
- Un tableau de régression devrait contenir:
  - Les coefficients estimés
  - Les écarts-types de ces coefficients
  - Les tests de significativité des coefficients
  - Les mesures de la qualité d'ajustement du modèle
  - Le nombre d'observations
  - Les *F-stats* et les résultats de *F-tests* au seuil de 5%
  - Autres informations pertinentes.

Exemple

Dependent variable: average test score in the district.

Regressor	(1)	(2)	(3)	(4)	(5)
Student–teacher ratio (X_1)	−2.28** (0.52)	−1.10* (0.43)	−1.00** (0.27)	−1.31** (0.34)	−1.01** (0.27)
Percent English learners (X_2)		−0.650** (0.031)	−0.122** (0.033)	−0.488** (0.030)	−0.130** (0.036)
Percent eligible for subsidized lunch (X_3)			−0.547** (0.024)		−0.529** (0.038)
Percent on public income assistance (X_4)				−0.790** (0.068)	0.048 (0.059)
Intercept	698.9** (10.4)	686.0** (8.7)	700.2** (5.6)	698.0** (6.9)	700.4** (5.5)
Summary Statistics					
SER	18.58	14.46	9.08	11.65	9.08
$\overline{R}^2$	0.049	0.424	0.773	0.626	0.773
n	420	420	420	420	420

These regressions were estimated using the data on K-8 school districts in California, described in Appendix 4.1. Heteroskedasticity-robust standard errors are given in parentheses under coefficients. The individual coefficient is statistically significant at the *5% level or **1% significance level using a two-sided test.

La régression multiple: résumé

- La régression multiple permet d'estimer l'effet d'une variable explicative *X* sur une variable dépendante Y, *toutes choses égales par ailleurs*.
- Le modèle doit inclure tous les régresseurs pertinents pour expliquer Y. Sinon, on fait face au biais de variable omise.
- Si une variable importante n'est pas disponible, on doit la remplacer par des variables de contrôle.
- Il n'y a pas de recette magique pour déterminer les régresseurs à inclure dans le modèle: il faut tester plusieurs spécifications et utiliser le jugement.
    - Une approche consiste à spécifier un modèle de base et ensuite, faire une analyse de sensibilité en testant les variantes de ce modèle de base.