Cours d'économétrie II
Données de panel

Cours du 2 février 2006

Michel Juillard

Données de panel

Des données de panel contiennent des données sur plusieurs individus observés à différentes dates.
Exemples:

- Des données sur les 420 districts scolaires de Californie en 1999 et 2000, soit 840 observations en tout.
- Des données sur les 24 pays de l'Union européenne pendant trois ans, soit 72 observations.
- Des données sur 1000 individus pendant quatre mois, soit 4000 observations.

Notation

Les données sont indicées par un double indice:

i : l'individu, $i = 1, \ldots, n$, soit n individus

t : la période, $t = 1, \ldots, T$, soit T périodes

Pour un modèle avec une variable explicatives, les données sont

$$(X_{it}, Y_{it})$$

Avec k variables explicatives

$$(X_{1it}, X_{2it}, \ldots, X_{kit}, Y_{it})$$

Terminologie

- données longitudinales est synonyme de données de panel
- un panel équilibré (balanced panel en anglais) a le même nombre d'observations pour tous les individus
- un panel déséquilibré est un panel où il manque des observations pour certains individus

Utilité des panels

Les données en panel permettent de contrôler pour des facteurs qui

- varient entre les individus, mais ne varient pas au cours du temps,
- pourraient causer un biais d'omission si l'on en tenait pas compte,
- sont inobservables ou non disponibles et ne peuvent être inclus dans la régression.

Si ces facteurs ne varient pas au cours du temps, ils ne peuvent pas influencer la variation de Y *au cours du temps*.

Exemple

Accidents mortels de la circulation et impôts sur l'alcool aux Etats–Unis
Unité d'observation: une année dans un Etat des Etats–Unis:

- 48 Etats: $n = 48$,
- 7 ans (1982, . . . , 1988): $T = 7$,
- panel équilibré: $7 \times 48 = 336$ observations

Exemple (suite)

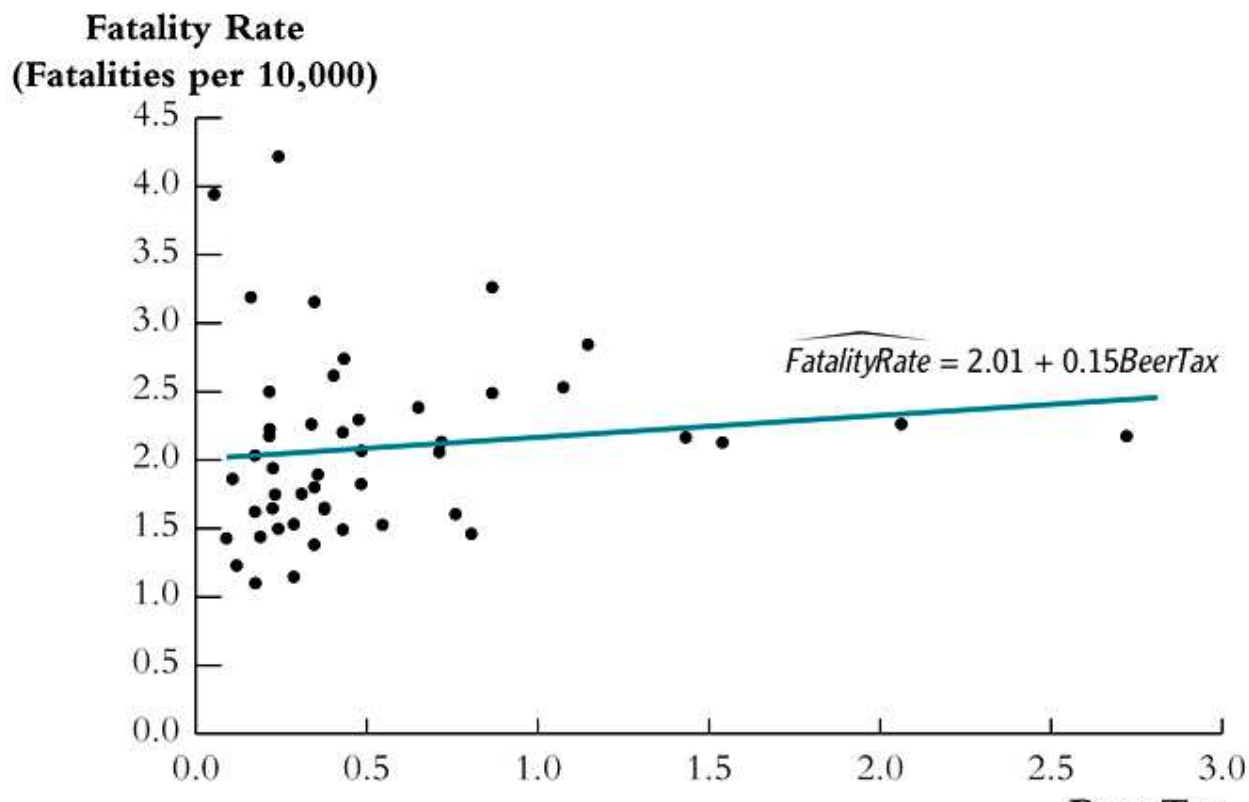
Variables:

- nombre de morts par accident de la circulation pour 10000 habitants
- impôts sur un paquet de bières
- autres (âge minimum pour la conduite, lois contre l'alcool au volant, . . .)

Données pour 1982

FIGURE 8.1 The Traffic Fatality Rate and the Tax on Beer

Panel a is a scatterplot of traffic fatality rates and the real tax on a case of beer (in 1988 dollars) for 48 states in 1982. Panel b shows the data for 1988. Both plots show a positive relationship between the fatality rate and the real beer tax.



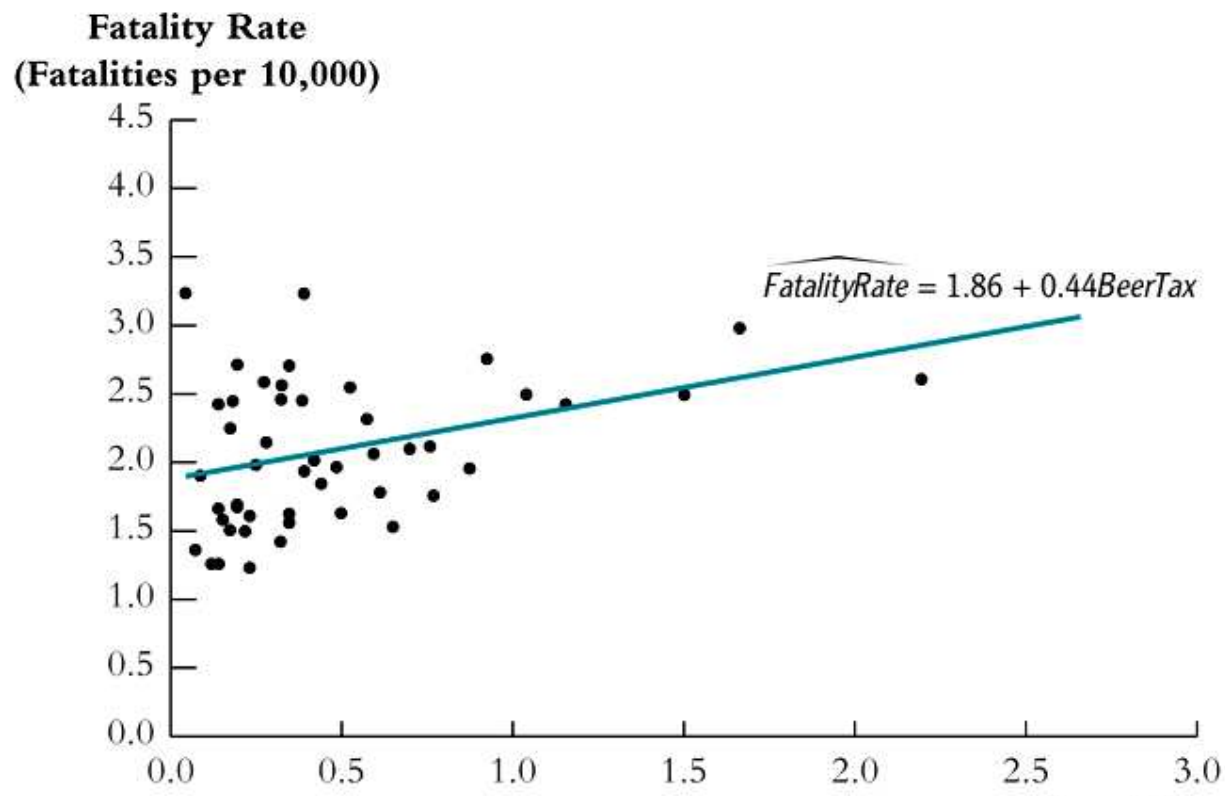
(a) 1982 data

(Source: Stock et Watson, 2003)

Données pour 1988

FIGURE 8.1 The Traffic Fatality Rate and the Tax on Beer

Panel a is a scatterplot of traffic fatality rates and the real tax on a case of beer (in 1988 dollars) for 48 states in 1982. Panel b shows the data for 1988. Both plots show a positive relationship between the fatality rate and the real beer tax.



(b) 1988 data

(Source: Stock et Watson, 2003)

Autres facteurs

Pourquoi y aurait–il davantage de morts par accident de la circulation dans les Etats qui ont des impôts plus élevés sur l'alcool?

- Qualité (âge) des automobiles
- Qualité des routes
- Tolérance culturelle vis à vis de boire et conduire
- Densité des voitures sur la route

Ces facteurs peuvent causer un biais d'omission s'ils sont corrélé avec les impôts sur l'alcool.

Densité du traffic

- Une densité du traffic plus élevée est associée avec davantage d'accidents mortels sur la route
- Les Etats de l'Ouest, moins peuplés, ont des impôts plus faibles sur l'alcool.
- Facteur peu susceptible des changer au cours du temps (en 7 ans)

Tolérance culturelle

- Une tolérance culturelle envers le fait de boire et conduire peut être associée à des accidents mortels plus nombreux
- Il peut y avoir une corrélation entre les impôts sur l'alcool et l'attitude culturelle
- Les attitudes culturelles ne changent en général pas d'une année à l'autre

Panel sur deux périodes

Modèle

$$FR_{it} = \beta_0 + \beta_1 BT_{it} + \beta_2 Z_i + u_{it} \qquad i = 1, \ldots, 48; \quad t = 1, 2$$

avec FR_{it} le taux d'accidents mortels (fatality rate), dans l'Etat i, à la période t, et BT_{it}, les impôts sur la bière (beer tax), dans l'Etat i, à la période t.
Z_i est un facteur qui ne varie pas au cours du temps.
Si Z n'est pas observé et que $\mathrm{corr}(\mathrm{BT}, \mathrm{Z}) \neq 0$, son omission entraîne un biais de l'estimateur $\hat{\beta}_1$

Eliminer Z

L'équation pour 1982:

$$FR_{i82} = \beta_0 + \beta_1 BT_{i82} + \beta_2 Z_i + u_{i82}$$

L'équation pour 1988

$$FR_{i88} = \beta_0 + \beta_1 BT_{i88} + \beta_2 Z_i + u_{i88}$$

Supposon que $\mathcal{E}\left(u_{it} | BT_{it}, Z_i\right) = 0$ (Supposons qu'il n'y a pas d'autres facteurs importants).
Différence entre 1982 et 1988

$$\Delta FR_i = \beta_1 \Delta BT_i + v_i$$

avec $v_i = u_{i88} - u_{i82}$. Bien que Z a disparu, $\mathrm{corr}\left(\Delta \mathrm{BT}, \mathrm{v}\right) = 0$.

Résultats

Données pour 1982

$$\widehat{FR}_i = \underset{}{2.01} + \underset{(0.15)}{0.15}\ BT_i \qquad (n = 48)$$
$$\qquad\qquad\quad (0.13)$$

Données pour 1988

$$\widehat{FR}_i = \underset{}{1.86} + \underset{(0.11)}{0.44}\ BT_i \qquad (n = 48)$$
$$\qquad\qquad\quad (0.13)$$

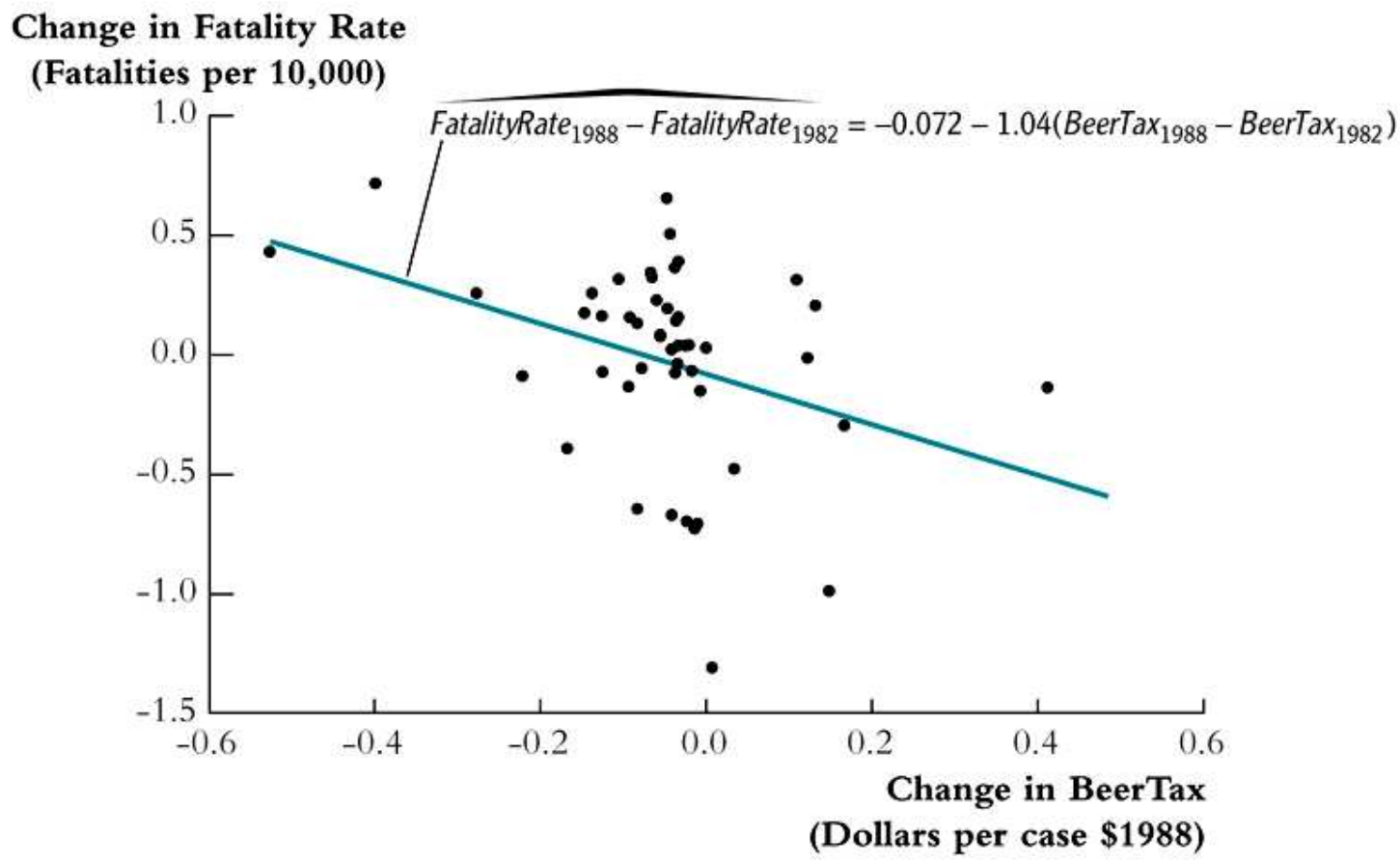
Différences 1982–1988

$$\widehat{\Delta FR}_i = - \underset{}{0.072} - \underset{(0.065)}{1.04}\ \Delta BT_i \qquad (n = 48)$$
$$\qquad\qquad\quad (0.36)$$

Représentation graphique

FIGURE 8.2 Changes in Fatality Rates and Beer Taxes, 1982–1988

This is a scatterplot of the *change* in the traffic fatality rate and the *change* in real beer taxes between 1982 and 1988 for 48 states. There is a negative relationship between changes in the fatality rate and changes in the beer tax.



(Source: Stock et Watson, 2003)

Régression à effet fixe

Comment faire lorsqu'on dispose d'observations sur plus de 2 périodes?
On peut écrire le modèle

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it}$$

comme

$$Y_{it} = \alpha_i + \beta_1 X_{it} + u_{it}$$

avec $\alpha_i = \beta_0 + \beta_2 Z_i$. L'effet fixe individuel inobservable devient un coefficient à estimer!

Estimation

Il existe trois méthodes pour estimer un modèle à effets fixes:

1. Représenter α_i par $n-1$ variables binaires (en pratique, que si n n'est pas trop grand)
2. Estimer en écart à la moyenne de chaque individu (au cours du temps)
3. Estimer sur les données en différence (seulement si $T=2$)

Les trois méthodes fournissent les mêmes valeurs estimées et les mêmes erreurs–type pour les coefficients.

Variables binaires

Imaginons que nous n'avons que trois Etats: Californie (CA), Texas (TX) et Massachusetts (MA). Nous obtenons les équations suivantes:

$$\begin{aligned} Y_{CA,t} &= \alpha_{CA} + \beta_1 X_{CA,t} + u_{CA,t} \\ Y_{TX,t} &= \alpha_{TX} + \beta_1 X_{TX,t} + u_{TX,t} \\ Y_{MA,t} &= \alpha_{MA} + \beta_1 X_{MA,t} + u_{MA,t} \end{aligned}$$

Trois droites parallèles avec la même pente (β_1).
On peut représenter les trois α_i à l'aide de deux variables binaires: DCA = 1, s'il s'agit de la Californie, 0 autrement
DTX = 1, s'il s'agit du Texas, 0 autrement.

Variables binaires (suite)

On obtient le modèle

$$Y_{it} = \beta_0 + \gamma_{CA} DCA_i + \gamma_{TX} DTX_i + \beta_1 X_{it} + u_{it}$$

et

$$\begin{aligned} \alpha_{CA} &= \beta_0 + \gamma_{CA} \\ \alpha_{TX} &= \beta_0 + \gamma_{TX} \\ \alpha_{MA} &= \beta_0 \end{aligned}$$

En général, il faut $n - 1$ variables binaires.

Variables binaires (suite)

Modèle général:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \ldots + \gamma_n Dn_i + u_i$$

- Peut être estimé par les MCO
- Les tests et les intervalles de confiance se calculent de la manière habituelle (en utilisant les erreurs-type robustes à l'hétéroscédasticité)
- Difficile à utiliser pour un très grand nombre d'individus

Données centrées

Modèle à effets fixes

$$Y_{it} = \alpha_i + \beta_1 X_{it} + u_{it}$$

Moyennes au cours du temps pour chaque individu:

$$\frac{1}{T}\sum_{t=1}^{T} Y_{it} = \alpha_i + \beta_1 \frac{1}{T}\sum_{t=1}^{T} X_{it} + \frac{1}{T}\sum_{t=1}^{T} u_{it}$$

Modèle en écarts à la moyenne:

$$\left(Y_{it} - \frac{1}{T}\sum_{t=1}^{T} Y_{it}\right) = \beta_1 \left(X_{it} - \frac{1}{T}\sum_{t=1}^{T} X_{it}\right) + \left(u_{it} - \frac{1}{T}\sum_{t=1}^{T} u_{it}\right)$$

Données centrées (suite)

$$\widetilde{Y}_{it} = \beta_1 \widetilde{X}_{it} + \tilde{u}_{it}$$

avec

$$\begin{aligned}
\widetilde{Y}_{it} &= Y_{it} - \frac{1}{T}\sum_{\tau=1}^{T} Y_{it} \\
\widetilde{X}_{it} &= X_{it} - \frac{1}{T}\sum_{\tau=1}^{T} X_{it} \\
\tilde{u}_{it} &= u_{it} - \frac{1}{T}\sum_{\tau=1}^{T} u_{it}
\end{aligned}$$

Par exemple, pour $i = 1$ et $t = 1$, Y_{it} représente la différence entre le taux de mortalité par accident en Alabama en 1982 et le taux moyen de mortalité dans cet Etat entre 1982 et 1988.

Estimation

$$\widetilde{Y}_{it} = \beta_1 \widetilde{X}_{it} + \tilde{u}_{it}$$

- Calculer $\widetilde{Y}_{it}$ et $\widetilde{X}_{it}$
- Estimer par les MCO
- Les tests et les intervalles de confiance se calculent de la manière habituelle (en utilisant les erreurs-type robustes à l'hétéroscédasticité)

Exemple

$$\widehat{FR}_{it} = \underset{(0.20)}{-\ 0.66}\ BT_{it} + \text{ effets fixes}$$

A comparer avec “différences 1982–1988”

$$\widehat{\Delta FR}_i = -\ \underset{(0.065)}{0.72}\ -\ \underset{(0.36)}{1.04}\ \Delta BT_i$$

Effets fixes temporels

Une variable omise peut être identique dans les différents Etats, mais varier au cours du temps. Par exemple la sécurité des véhicles ou la réglementation nationale.
On introduit des constantes qui changent avec le temps, mais pas les individus

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + \beta_3 S_t + u_{it}$$

Estimation

Modèles uniquement à effets fixes temporels

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_3 S_t + u_{it}$$

Les modèles à effets fixes temporelles peuvent s'estimer

- en ajoutant T-1 variables binaires
- en centrant les données autour de la moyenne des individus par période

Effets fixes individuels et temporels

Le modèle

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + \beta_3 S_t + u_{it}$$

peut s'estimer

1. à l'aide de variables binaires

$$\begin{aligned} Y_{it} &= \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \ldots + \gamma_n Dn_i \\ &\quad \delta_2 B2_t + \ldots + \delta_T BT_t + u_{it} \end{aligned}$$

2. en centrant les variables autour de la moyenne globale (à travers le temps et les individus)

$$\widetilde{Y}_{it} = Y_{it} - \frac{1}{nT} \sum_{j=1}^{n} \sum_{\tau=1}^{T} Y_{j\tau}$$

Hypothèses

H_1 : le modèle

$$Y_{it} = \alpha_i + \beta_1 X_{it} + u_{it}$$

H_2 : $\mathcal{E}\left(u_{it} | X_{i1}, \ldots, X_{iT}\alpha_i\right) = 0$

H_3 : $(X_{i1}, \ldots, X_{iT}, Y_{i1}, \ldots, Y_{iT})$ est i.i.d.

H_4 : (X_{it}, u_{it}) a des quatrièmes moments finis

H_5 : il n'y a pas de multicolinéarité parfaite (lorsqu'il y a plusieurs variables explicatives)

H_6 : $\mathrm{corr}\left(\mathrm{u_{it}}, \mathrm{u_{is}} | \mathrm{X_{it}}, \mathrm{X_{is}}, \alpha_\mathrm{i}\right) = 0$

$$\mathcal{E}\left(u_{it}|X_{i1},\ldots,X_{iT}\alpha_i\right)=0$$

- $u_{it}=0$ étant donné l'effet fixe pour l'individu i et toute l'histoire de X pour cet individu.
- Il ne peut y avoir d'effets retardés omis. Les effets retardés doivent être modélisés explicitement.
- Il ne peut non plus y avoir d'effets vers le futur. Par exemple, un taux d'accidents particulièrement élevé une année ne doit pas entraîner une hausse des impôts sur la bière les années suivantes.

$(X_{i1}, \ldots, X_{iT}, Y_{i1}, \ldots, Y_{iT})$ est i.i.d.

- L'hypothèse est satisfaite si les individus sont tirés au hasard dans la population et qu'on enregistre l'histoire de ces individus
- Il n'est pas nécessaire que les observations successives au cours du temps pour le même individu soient i.i.d. Ce serait hautement irréaliste.

$$\mathrm{corr}\left(u_{it}, u_{is} | X_{it}, X_{is}, \alpha_i\right) = 0$$

- Nouvelle hypothèse propre aux panels.
- Les erreurs ne sont pas corrélées au travers du temps pour un même individu.
- Les facteurs omis de l'équation ne doivent pas être corrélés au cours du temps.
  Exemples:
  - Conditions météorologiques
  - Ouverture d'une nouvelle autoroute
  - Fluctuations du traffic dues aux conditions économiques locales

Si H_6 n'est pas satisfaite

- Autocorrélation des erreurs.
- $\hat{\beta}_1$ demeure non–biaisé et convergent
- l'erreur–type de l'estimateur est fausse.
- Intuitivement, lorsque les erreurs sont autocorrélées, on ne dispose pas d'autant d'information qu'on le croit.
- Solution: utiliser une formule qui corrige les erreurs–type poru l'hétéroscédasticité et l'autocorrélation (HAC).

Application

Quels sont les effets des différentes mesures contre l'alcool au volant?

- sanctions pénales
- âge minimum pour consommer de l'alcool
- impôts sur l'alcool

Données

Un panel équilibré $(n = 48, T = 7)$

Variables:

- taux de mortalité par accident de la circulation (pour 10000 habitants)
- impôts sur la bière
- âge minimum pour consommer de l'alcool
- sanctions pénales
  - prison
  - service d'intérêt général
  - amende
- miles parcourus par conducteur
- taux de chômage
- revenu réel par habitant

Utilité d'un panel

Effets manquants, variables d'un Etat à l'autre, mais constants au cours du temps

- attitudes culturelles vis–à–vis de boire et conduire
- qualité des routes
- âge des véhicules

Effets manquants, constants parmi les Etats, mais changeants au cours du temps

- amélioration de la sécurité des véhicules
- modification nationale de l'attitue vis–à–vis de boire et conduire

Résultats

TABLE 8.1 Regression Analysis of the Effect of Drunk Driving Laws on Traffic Deaths

Dependent Variable: Traffic Fatality Rate (Deaths Per 10,000).

Regressor	(1)	(2)	(3)	(4)	(5)	(6)
Beer tax	0.36** (0.05)	−0.66** (0.20)	−0.64* (0.25)	−0.45* (0.22)	−0.70** (0.25)	−0.46* (0.22)
Drinking age 18				0.028 (0.066)	−0.011 (0.064)	
Drinking age 19				−0.019 (0.040)	−0.078 (0.049)	
Drinking age 20				0.031 (0.046)	−0.102* (0.046)	
Drinking age						−0.002 (0.017)
Mandatory jail?				0.013 (0.032)	−0.026 (0.065)	
Mandatory community service?				0.033 (0.115)	0.147 (0.137)	
Mandatory jail or community service?						0.031 (0.076)
Average vehicle miles per driver				0.008 (0.008)	0.017 (0.010)	0.009 (0.008)
Unemployment rate				−0.063** (0.012)		−0.063** (0.012)
Real income per capita (logarithm)				1.81** (0.47)		1.79** (0.45)
State effects?	no	yes	yes	yes	yes	yes
Time effects?	no	no	yes	yes	yes	yes

These regressions were estimated using panel data for 48 U.S. states from 1982 to 1988 (336 observations total), described in Appendix 8.1. Standard errors are given in parentheses under the coefficients, and *p*-values are given in parentheses under the *F*-statistics. The individual coefficient is statistically significant at the *5% level or **1% significance level.

(Source: Stock et Watson, 2003)

Résultats (suite)

TABLE 8.1 Regression Analysis of the Effect of Drunk Driving Laws on Traffic Deaths

Dependent Variable: Traffic Fatality Rate (Deaths Per 10,000).

	(1)	(2)	(3)	(4)	(5)	(6)
***F*-statistics and *p*-values Testing Exclusion of Groups of Variables:**						
Time effects = 0			2.47 (0.024)	11.44 (<0.001)	2.28 (0.037)	11.59 (<0.001)
Drinking age coefficients = 0				0.48 (0.696)	2.09 (0.102)	
Jail, community service coefficients = 0				0.17 (0.845)	0.59 (0.557)	
Unemployment rate, income per capita = 0				38.29 (<0.001)		40.12 (<0.001)
$\overline{R}^2$	0.090	0.889	0.891	0.926	0.893	0.926

These regressions were estimated using panel data for 48 U.S. states from 1982 to 1988 (336 observations total), described in Appendix 8.1. Standard errors are given in parentheses under the coefficients, and *p*-values are given in parentheses under the *F*-statistics. The individual coefficient is statistically significant at the *5% level or **1% significance level.

(Source: Stock et Watson, 2003)

Discussion

- Le signe de l'effet des impôts sur la bière change lorsqu'on introduit des effets fixes individuels.
- Les effets fixes temporels sont significatifs, mais ne changent pas les résultats de l'estimation de manière importante.
- L'effet estimé de l'impôt sur la bière diminue lorsqu'on introduit les autres mesures dissuasives.
- La seule mesure qui apparaît statistiquement significative est l'impôt sur la bière.
- Les variables d'environnement économique apparaissent comme importante.

Utiliser des variables binaires

On peut utiliser des variables binaires dans d'autres contextes que les données en panel pour contrôler des effets de groupe. C'est utile si l'on soupçonne que des groupes d'observations sont affectés par des effets non–observés constants pour les observations d'un groupe.

Exemple: si la politique scolaire était décidée au niveau du comté et que chaque comté soit composé de plusieurs districts scolaires.

Données de panel: résumé

Avantages des effets fixes:

- permettent de contrôler pour des effets non–observés fixes au cours du temps ou à travers les individus
- Davantage d'observations fournit davantage d'information (valeurs estimées plus précises)
- Extension de la méthodologie de la régression
- Inférence se conduit de la manière habituelle

Limitations:

- Les observations doivent varier au cours du temps.
- Les effets retardés peuvent être importants
- Les erreurs–types peuvent être sous–estimées si les erreurs sont autocorrélées