Regression with panel data: an Introduction

Professor Bernard Fingleton

What does panel (or longitudinal) data look like?

- Each of N individual’s data is measured on T occasions
- Individuals may be people, firms, countries etc
- Some variables change over time for t = 1,…,T
- Some variables may be fixed over the time period, such as gender, the geographic location of a firm or a person’s ethnic group
- When there are no missing data, so that there are NT observations, then we have a balanced panel (less than NT is called an unbalanced panel)
- Typically N is large relative to T, but not always

- Example of a simple panel

GDP pc   Log % no school   Log av. Yrs school

year	countriesx4	lnGDP_per_	lnno_sch_%	lnav_yrs_sch	fed[2]	fed[3]	fed[4]
1970	Argentina	2.226235129	2.174751721	1.771556762	0	0	0
1970	Australia	2.696003468	0.336472237	2.311544834	1	0	0
1970	Austria	2.413729352	1.458615023	1.947337701	0	1	0
1970	Bangladesh	0.099447534	4.453183829	-0.162518929	0	0	1
2000	Argentina	2.398482048	1.840549633	2.138889	0	0	0
2000	Australia	3.240990085	0.993251773	2.3580198	1	0	0
2000	Austria	3.164481031	0.587786665	2.174751721	0	1	0
2000	Bangladesh	0.521101364	4.028916757	0.896088025	0	0	1

- T = 2, t = 1…T time periods

Fixed effect dummies

- N = 4, n = 1,…,N individuals
- K = 5, k = 1,…,K independent variables

Notation

$Y_{it} =$ dependent variable value for individual i at time t

$X_{1it} =$ independent variable 1 value for individual i at time t

$X_{2it} =$ independent variable 2 value for individual i at time t

etc

$X_{Kit} =$ independent variable K value for individual i at time t

Why are panel data useful?

- With observations that span **both** time and individuals in a cross-section, more information is available, giving more efficient estimates.
- The use of panel data allows empirical tests of a wide range of hypotheses.
- With panel data we can control for :
  - Unobserved or unmeasurable sources of individual heterogeneity that vary across individuals but do not vary over time
  - omitted variable bias

Key Reading

- **Stock and Watson (2007), Chapter 10: Regression with panel data**
- **Baltagi(2002) Econometrics 3rd Edition**
- **Baltagi(2005) Econometric Analysis of Panel Data**

$$Y_{it} = \text{log GDP per capita}$$

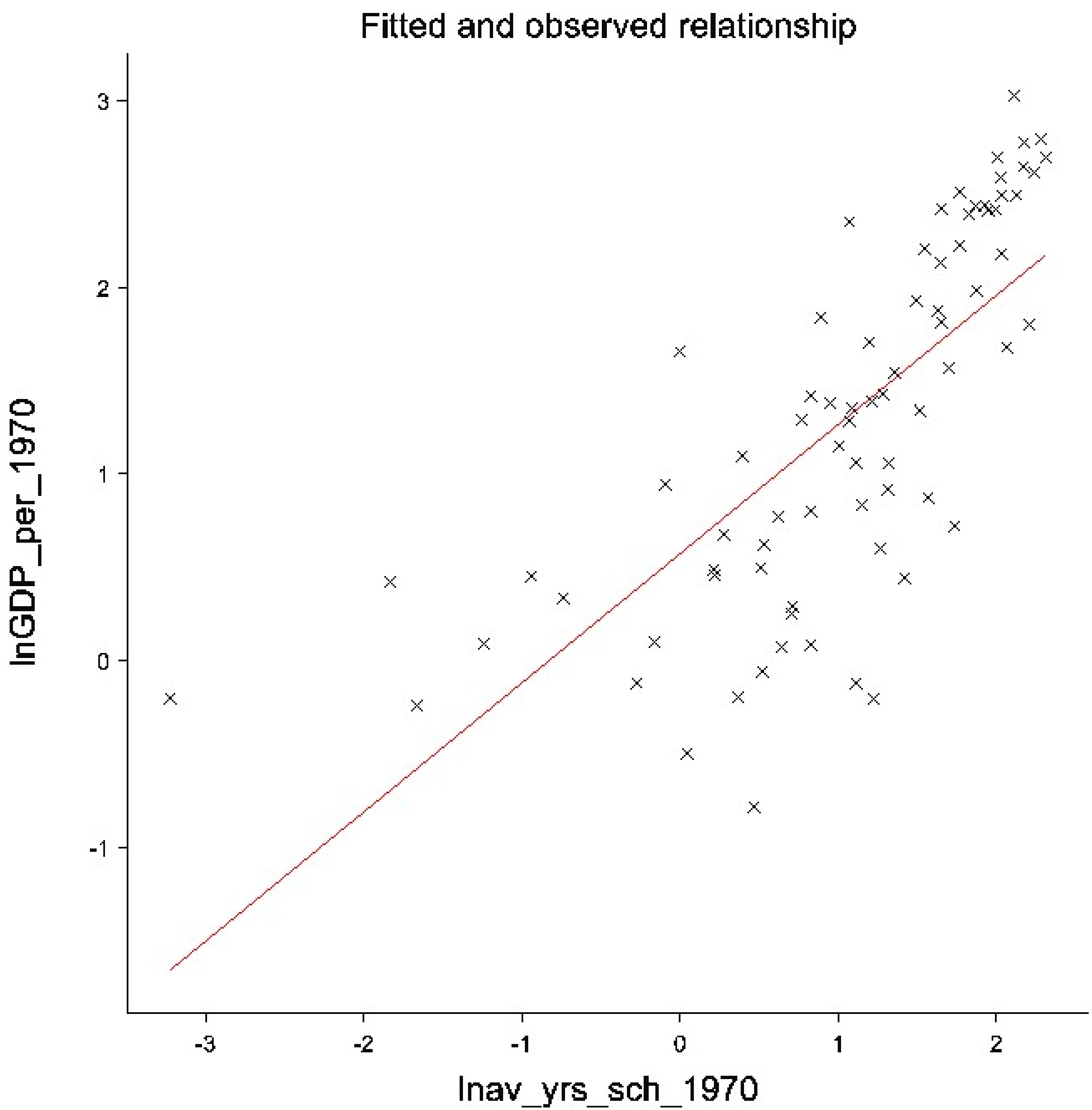
$$X_{1it} = \text{log average number of years with schooling}$$

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + u_i$$

$$i = 1, ..., N, \quad t = 1\ (1970)$$

```
Estimates of parameters
-----------------------

Parameter                 estimate          s.e.     t(75)
Constant                     0.571         0.109      5.24
lnav_yrs_sch_1970
                            0.6925        0.0746      9.28
```


Fitted and observed relationship
lnGDP_per_1970
lnav_yrs_sch_1970
3
2
1
0
-1
-3
-2
-1
0
1
2

$$Y_{it} = \text{ log GDP per capita}$$

$$X_{1it} = \text{ log average number of years with schooling}$$

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + u_i$$

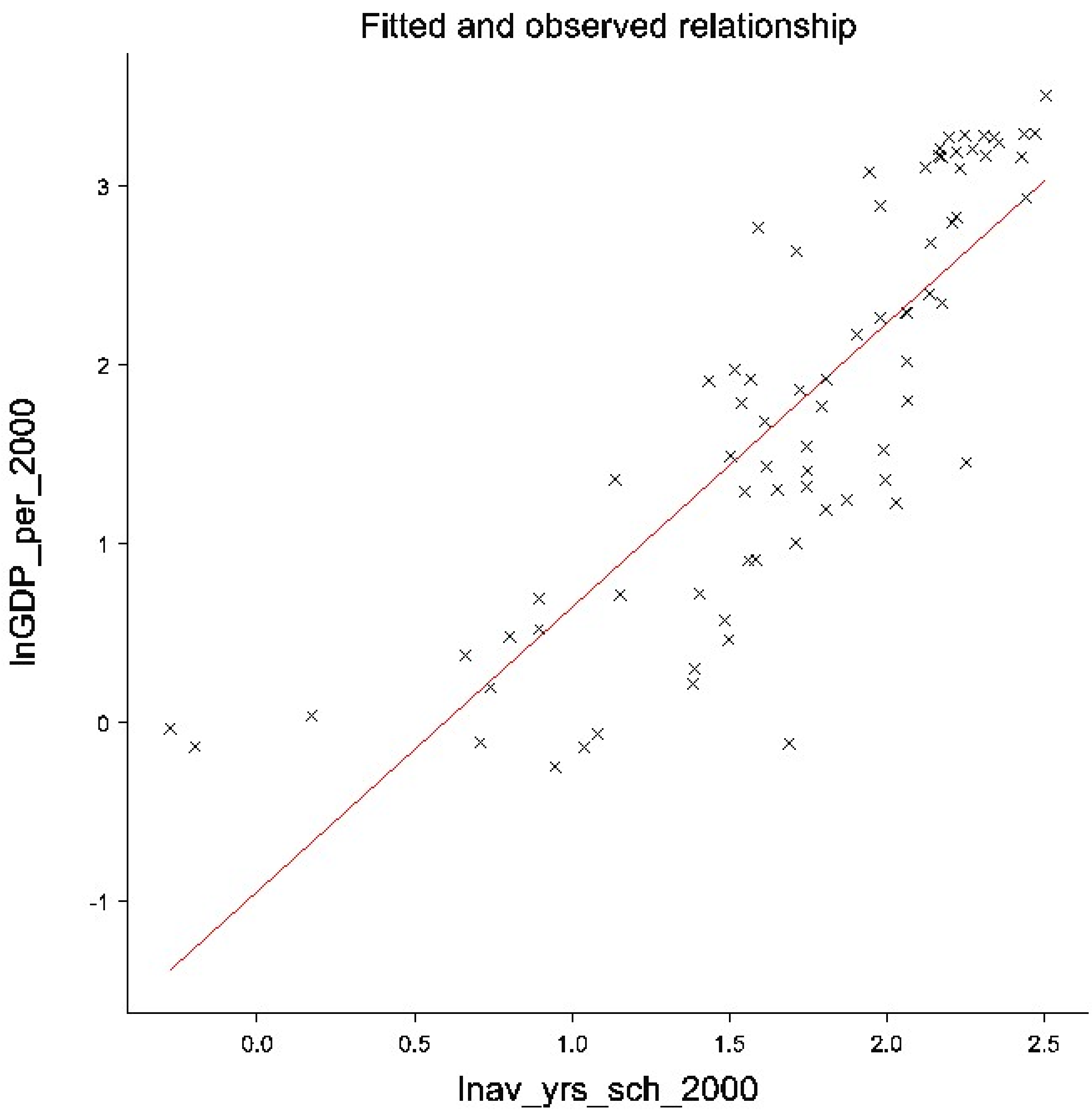
$$i = 1, ..., N, \quad t = 1, 2 \ (1970{,}2000)$$

Estimates of parameters

Parameter	**estimate**	**s.e.**	**t(75)**
Constant	**0.571**	**0.109**	**5.24**
lnav_yrs_sch_1970	**0.6925**	**0.0746**	**9.28**

Estimates of parameters

Parameter	**estimate**	**s.e.**	**t(75)**
Constant	**-0.946**	**0.223**	**-4.23**
lnav_yrs_sch_2000	**1.589**	**0.123**	**12.87**



Fitted and observed relationship
lnGDP_per_2000
lnav_yrs_sch_2000
3
2
1
0
-1
0.0
0.5
1.0
1.5
2.0
2.5

$$Y_t = \beta_1 X_t + \beta_2 W_t + e_t \qquad (True)$$

$$Y_t = \beta_1 X_t + (\beta_2 W_t + e_t)$$

$$Y_t = \beta_1 X_t + v_t \qquad \text{(We estimate)}$$

If $Corr(X, W) \neq 0$ then $Cov(X, v) \neq 0$

$Y_{it} =$ log GDP per capita

$X_{1it} =$ log average number of years with schooling

W_i is omitted, so the estimate of β_1 is not consistent

Consider the model for time 1 and time 2, giving 2 equations

$$Y_{i2} = \beta_0 + \beta_1 X_{1i2} + (\beta_2 W_i + e_{i2})$$

$$Y_{i1} = \beta_0 + \beta_1 X_{1i1} + (\beta_2 W_i + e_{i1})$$

$$Y_{i2} - Y_{i1} = \beta_1 (X_{1i2} - X_{1i1}) + (e_{i2} - e_{i1})$$

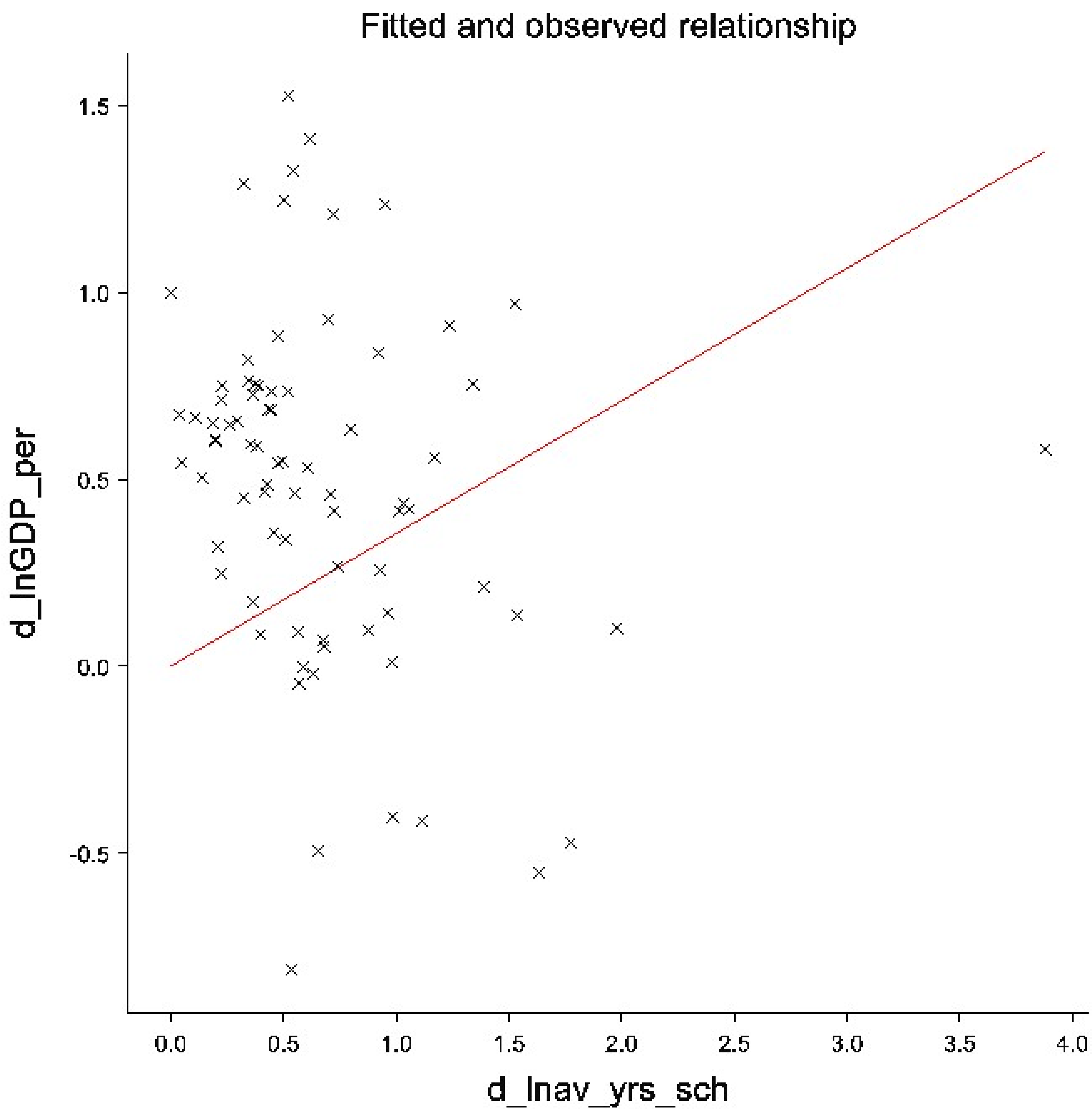
W_i is constant across time, but varies across countries

$e_{i2} - e_{i1}$ is independent of $X_{1i2} - X_{1i1}$ so the estimate β_1 is consistent

```
Estimates of parameters
-----------------------

Parameter          estimate         s.e.     t(76)
d_lnav_yrs_sch
                     0.3548       0.0772      4.59
```

Look what we are assuming here, that the slope of the line is constant
And does not vary over time
We also assume that differencing eliminates any correlation between
The explanatory variable and the residuals.
But for this to be the case the omitted variables have to be constant
Over time…..are there omitted variables that are not constant over time?



Fitted and observed relationship
1.5
1.0
0.5
0.0
-0.5
d_lnGDP_per
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
d_lnav_yrs_sch

Equivalent estimation methods

- Differencing is only applicable to the case where T = 2. More generally we have two options
- Dummy variables
  - One dummy variable for each individual, thus controlling for inter-individual heterogeneity
- The 'within' estimator
  - Each individual's value is a <u>deviation</u> from its own time-mean
  - This takes out the effect of differing individual levels as a result of inter-individual heterogeneity

- Both give the <u>same estimate</u> of β_1

Fixed Effects Regression: Estimation

- “dummy variables” is only practical when N isn’t too big, because one runs into computational problems. With N very large, we use of lots of degrees of freedom
- Note that with “dummy variables”, not all N can be included because of the dummy variable trap. Alternatively, we have to omit the constant.

- Data layout using N-1 dummies
- N=77
- T=2

n154	lnGDP_per_70_00	lnav_yrs_sch_70_00	fed_70_00[2]	fed_70_00[3]
1.00	2.226	1.772	0.00000	0.00000
2.00	2.696	2.312	1.00000	0.00000
3.00	2.414	1.947	0.00000	1.00000
4.00	0.099	-0.163	0.00000	0.00000
78.00	2.398	2.139	0.00000	0.00000
79.00	3.241	2.358	1.00000	0.00000
80.00	3.164	2.175	0.00000	1.00000
81.00	0.521	0.896	0.00000	0.00000

- Output of a regression using N-1 dummies for fixed effects across 77 countries

```
Estimates of parameters
-----------------------

Parameter                        estimate        s.e.      t(76)
Constant                            1.619       0.333       4.85
lnav_yrs_sch_70_00                 0.3548      0.0772       4.59
fed_70_00[2]                        0.521       0.422       1.24
fed_70_00[3]                        0.439       0.421       1.04
fed_70_00[4]                       -1.439       0.438      -3.28
fed_70_00[5]                       -0.104       0.421      -0.25
fed_70_00[6]                        0.452       0.421       1.07
fed_70_00[7]                       -1.389       0.454      -3.06
fed_70_00[8]                       -1.197       0.422      -2.84
```

- **Etc, up to fed[77]**

- Output of a regression using N dummies for fixed effects across 77 countries

```
Estimates of parameters
-----------------------

Parameter                        estimate          s.e.      t(76)
lnav_yrs_sch_70_00                 0.3548        0.0772       4.59
fed_70_00[1]                        1.619         0.333       4.85
fed_70_00[2]                        2.140         0.348       6.15
fed_70_00[3]                        2.058         0.337       6.10
fed_70_00[4]                        0.180         0.299       0.60

and so on until fed[77]
```

- Interpretation, 77 regression lines,
- each with the same slope but
- different intercepts

- Consider the model for countries 1,2 and 3

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + (\beta_2 W_i + e_{it}) = (\beta_0 + \beta_2 W_i) + \beta_1 X_{1it} + e_{it}$$

for $i = 1, 2, 3$

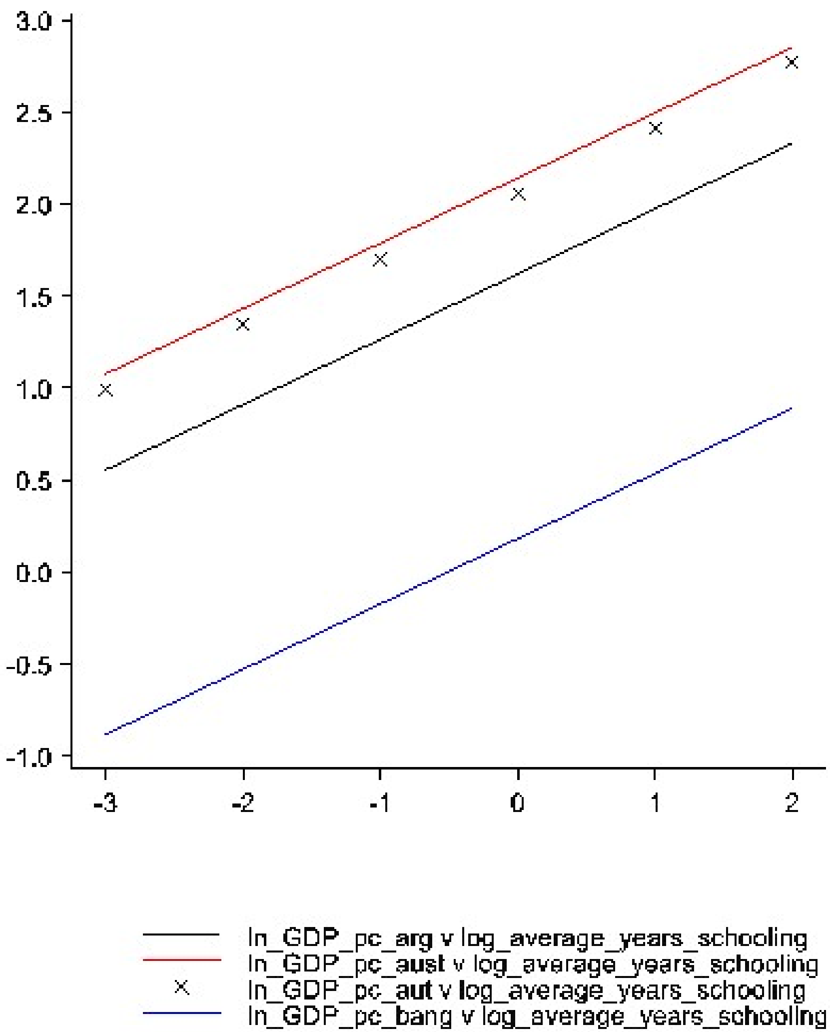
$$Y_{1t} = (\beta_0 + \beta_2 W_1) + \beta_1 X_{11t} + e_{1t} = \alpha_1 + \beta_1 X_{11t} + e_{1t}$$

$$Y_{2t} = (\beta_0 + \beta_2 W_2) + \beta_1 X_{12t} + e_{2t} = \alpha_2 + \beta_1 X_{12t} + e_{2t}$$

$$Y_{3t} = (\beta_0 + \beta_2 W_3) + \beta_1 X_{13t} + e_{3t} = \alpha_3 + \beta_1 X_{13t} + e_{3t}$$

$$Y_{it} = \alpha_i + \beta_1 X_{1it} + e_{it}$$

- Different intercepts   Same slope



3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-3
-2
-1
0
1
2
ln_GDP_pc_arg v log_average_years_schooling
ln_GDP_pc_aust v log_average_years_schooling
ln_GDP_pc_aut v log_average_years_schooling
ln_GDP_pc_bang v log_average_years_schooling

The within estimator

Calculate deviation from individual means, averaging over time

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

$$Y_{it} - \bar{Y}_{i.} = \beta_1 (X_{it} - \bar{X}_{i.}) + \nu_{it}$$

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

The within estimator (continued)

- Inference (hypothesis tests, confidence intervals) is as usual
- This is like the “differences” approach, but instead Y_{it} is subtracted from the average instead of from Y_{i1}.
- This can be done in a single command in PcGive and Gretl (and most other econometric packages)

Assumptions of fixed effects

1. The slopes of the regression lines are the same across states (countries)
2. The fixed effects capture entirely the time-constant omitted variables
   - This means we can soak up unmodelled heterogeneity across individuals/regions/countries and thus avoid misspecification error
   - But if there are <u>time-varying</u> omitted variables, their effects would not be captured by the fixed effects
   - Fixed <u>time</u> effects are also possible
     - But here we assume there are <u>no fixed effects</u> that cause GDP per capita to vary across time periods. These effects would have to be identical across all countries, a very strong assumption in this particular example

Disadvantage of fixed effects

- Fixed effects wipe out explanatory variables that do not vary within an individual (ie are time-invariant, such as gender, race)
- We are often interested in in the effects of these separate sources of individual heterogeneity

The error components model : random effects

- The alternative to the fixed effects model is the <u>random effects model</u>
  - In this the individual specific error components are chosen at random from a population of possible intercepts

Error components : random effects

We can write our model as an error components model, so that

$$Y_{it} = \alpha_i + \beta_1 X_{1it} + \ldots \beta_K X_{Kit} + e_{it}$$

becomes

$$Y_{it} = \beta_1 X_{1it} + \ldots \beta_K X_{Kit} + u_{it}$$

$$u_{it} = \alpha_i + e_{it}$$

$\alpha_i =$ individual specific components

e_{it} = remainder components, a 'traditional' error term

u_{it} = disturbance term

the disturbance term is a composite of the two error components

The random effects model

- In the fixed effects approach, we do not make any hypotheses about the individual specific effects
- beyond the fact that they exist — and that can be tested
- Once these effects are swept out by taking deviations from the group means, or by dummy variables, the remaining parameters can be estimated.

The random effects model

- the random effects approach attempts to model the individual effects as drawings from a probability distribution instead of removing them.
- In this the <u>individual effects</u> are part of the <u>disturbance term</u>, that is, zero-mean random variables, uncorrelated with the regressors.

The random effects model

- The composite disturbance term means that OLS is not appropriate
- We therefore use GLS (generalised least squares)
- There are various GLS estimators, but all are asymptotically efficient as T and N become large
    - Gretl uses the Swamy and Arora(1972) estimator of the random effects model, which is also the default in Stata

The random effects model

- the fixed-effects estimator “always works”, but at the cost of not being able to estimate the effect of time-invariant regressors.
  - This is because time-invariant regressors are perfectly correlated with the fixed effect dummies
- the random-effects estimator : time-invariant regressors can be estimated,
- but if individual effects (captured by the disturbance) are correlated with explanatory variables, then the random-effects estimator would be inconsistent, while fixed-effects estimates would still be valid.
- In contrast, the fixed effects are explicit (dummy) variables and can be correlated with the other X variables

The random effects model

- The random effects specification is appropriate if we assume the data are a representative and large sample of individuals N drawn at random from a large population
- Each individual effect is modelled as a random drawing from a probability distribution with mean 0 and with constant variance
- We are assuming that the composite disturbance term u has a value for a particular individual at a specific time which is made up of two components

The random effects model

- Two components
- A <u>random intercept term</u>, which measures the extent to which an individual's intercept differs from the overall intercept
  - This varies across individuals but is constant over time, reflecting the individual specific effect which is time-constant
- A '<u>traditional' random error</u>
  - this varies across individuals and across time and represents other unmodeled effects occurring at random

- The random effects model

$$Y_{it} = \alpha_i + \beta_1 X_{1it} + \ldots \beta_K X_{Kit} + e_{it}$$

$$Y_{it} = \beta_1 X_{1it} + \ldots \beta_K X_{Kit} + u_{it}$$

$$\alpha_i \sim iid(0, \sigma_\alpha^2)$$

$$e_{it} \sim iid(0, \sigma_e^2)$$

$$u_{it} = \alpha_i + e_{it}$$

$$\operatorname{cov}(\alpha_i ; e_{it}) = 0$$

$$\operatorname{cov}(X_1, \ldots, X_K ; u_{it}) = 0$$

• The random effects model

for OLS to be BLUE (the best linear unbiased estimator)

we require that

$E(u_{it}^2) =$ a constant σ_u^2 for all i and t

$E(u_{it}, u_{is}) = 0$ for s $\neq$ t

$E(u_{it}, u_{jt}) = 0$ for i $\neq$ j

If these assumptions are not met, and they are unlikely to be met

in the context of panel data, OLS is not the most efficient estimator.

Greater efficiency may be gained using generalized least

squares (GLS), taking into account the covariance structure of the error term.

• The random effects model

$$\alpha_i \sim iid(0, \sigma_\alpha^2)$$

$$e_{it} \sim iid(0, \sigma_e^2)$$

$$u_{it} = \alpha_i + e_{it}$$

$$\text{cov}(u_{it}, u_{js}) = \text{var}(u_{it}) = \sigma_\alpha^2 + \sigma_e^2 \text{ for } i = j \text{ and } t = s$$

$$\text{cov}(u_{it}, u_{js}) = \sigma_\alpha^2 \text{ for } i = j \text{ and } t \neq s$$

$$\text{cov}(u_{it}, u_{js}) = 0 \text{ for } i \neq j$$

thus there is serial correlation over time between disturbances

of the same individual

these variances and covariances form the elements

of an NT by NT variance-covariance matrix Ω

which is the basis of GLS estimation (ie weighted least squares)

Error Covariance structure

OLS

		i =1	i =1	i =2	i =2
		t =1	t =2	t =1	t =2
i =1	t =1	σ_u^2	0	0	0
i =1	t =2	0	σ_u^2	0	0
i =2	t =1	0	0	σ_u^2	0
i =2	t =2	0	0	0	σ_u^2

GLS

		i =1	i =1	i =2	i =2
		t =1	t =2	t =1	t =2
i =1	t =1	$\sigma_\alpha^2 + \sigma_e^2$	σ_α^2	0	0
i =1	t =2	σ_α^2	$\sigma_\alpha^2 + \sigma_e^2$	0	0
i =2	t =1	0	0	$\sigma_\alpha^2 + \sigma_e^2$	σ_α^2
i =2	t =2	0	0	σ_α^2	$\sigma_\alpha^2 + \sigma_e^2$

The random effects model

- We gain degrees of freedom
- We can introduce time invariant regressors (gender, race, religion etc) which are not wiped out by the presence of the fixed effect dummies
- Greater efficiency may be gained using generalized least squares (GLS), taking into account the covariance structure of the error term.

data set

- From Baltagi(2005) ‘the Econometric Analysis of Panel Data, 3rd Edition, page 25
- Consider the factors determining the gross output of US states
- Data comprises annual observations for 48 contiguous states over 1970-1986

Data layout

STATE	ST_ABB	st_number	YR	Public_CAP
ALABAMA	AL	1	1970	15032.67
ALABAMA	AL	1	1971	15501.94
ALABAMA	AL	1	1972	15972.41
ALABAMA	AL	1	1973	16406.26
ALABAMA	AL	1	1974	16762.67
ALABAMA	AL	1	1975	17316.26
ALABAMA	AL	1	1976	17732.86
ALABAMA	AL	1	1977	18111.93
ALABAMA	AL	1	1978	18479.74
ALABAMA	AL	1	1979	18881.49
ALABAMA	AL	1	1980	19012.34
ALABAMA	AL	1	1981	19118.52
ALABAMA	AL	1	1982	19118.25
ALABAMA	AL	1	1983	19122
ALABAMA	AL	1	1984	19257.47
ALABAMA	AL	1	1985	19433.36
ALABAMA	AL	1	1986	19723.37
ARIZONA	AZ	2	1970	10148.42
ARIZONA	AZ	2	1971	10560.54
ARIZONA	AZ	2	1972	10977.53
ARIZONA	AZ	2	1973	11598.26
ARIZONA	AZ	2	1974	12129.06
ARIZONA	AZ	2	1975	12929.06

$$\ln Y_{it} = \beta_0 + \beta_1 \ln K_{1it} + \beta_2 \ln K_{2it} + \beta_3 \ln L_{it} + \beta_4 Unemp_{it} + u_{it}$$

$i = 1,...,48$

$t = 1,...,17$

$Y =$ gross state output

$K_1 =$ public capital which includes highways and streets, water and sewage facilities, public buildings and structures

$K_2 =$ private capital stock

$L =$ labour input

$Unemp =$ state unemployment rate, to capture business cycle effects

- Fixed effects

```
Model 1: Fixed-effects estimates using 816 observations
Included 48 cross-sectional units
Time-series length = 17
Dependent variable: lnGrossStatePro

      VARIABLE       COEFFICIENT        STDERROR      T STAT    P-VALUE

  lnPublic_CAP         -0.0261497        0.0290016    -0.902    0.36752
  lnPrivateCapita       0.292007         0.0251197    11.625  <0.00001 ***
  lnEMP                 0.768159         0.0300917    25.527  <0.00001 ***
  UNEMP                -0.00529774       0.000988726  -5.358  <0.00001 ***


Test for differing group intercepts -
  Null hypothesis: The groups have a common intercept
  Test statistic: F(47, 764) = 75.8204
  with p-value = P(F(47, 764) > 75.8204) = 1.16445e-253
```

Fixed effects

- Hypothesis of individual specific heterogeneity given by F test
- This tests the null that all intercepts are the same
- Rejecting the null means that one needs to model individual heterogeneity
- One cannot simply pool the data and treat it as a single regression with just one intercept

• Random effects

```
Model 2: Random-effects (GLS) estimates using 816 observations
Included 48 cross-sectional units
Time-series length = 17
Dependent variable: lnGrossStatePro

      VARIABLE       COEFFICIENT        STDERROR      T STAT   P-VALUE

  const                 2.13541          0.133461     16.000  <0.00001 ***
  lnPublic_CAP          0.00443859       0.0234173     0.190   0.84971
  lnPrivateCapita       0.310548         0.0198047    15.681  <0.00001 ***
  lnEMP                 0.729671         0.0249202    29.280  <0.00001 ***
  UNEMP                -0.00617247       0.000907282  -6.803  <0.00001 ***


Breusch-Pagan test -
  Null hypothesis: Variance of the unit-specific error = 0
  Asymptotic test statistic: Chi-square(1) = 4134.96
  with p-value = 0

Hausman test -
  Null hypothesis: GLS estimates are consistent
  Asymptotic test statistic: Chi-square(4) = 9.52542
  with p-value = 0.0492276
```

Random effects

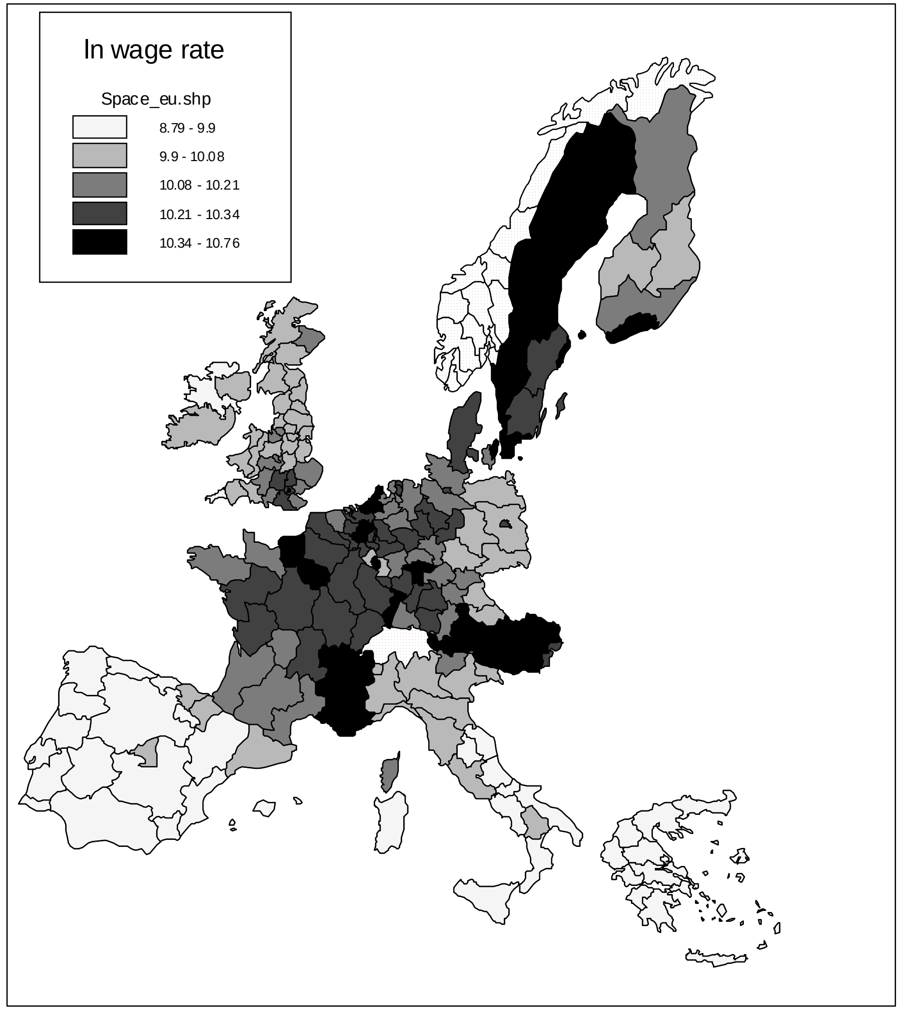
- The Breusch–Pagan test is the counterpart to the *F*-test for the fixed effects model.
- The null hypothesis is that the variance of the random intercept error component equals zero

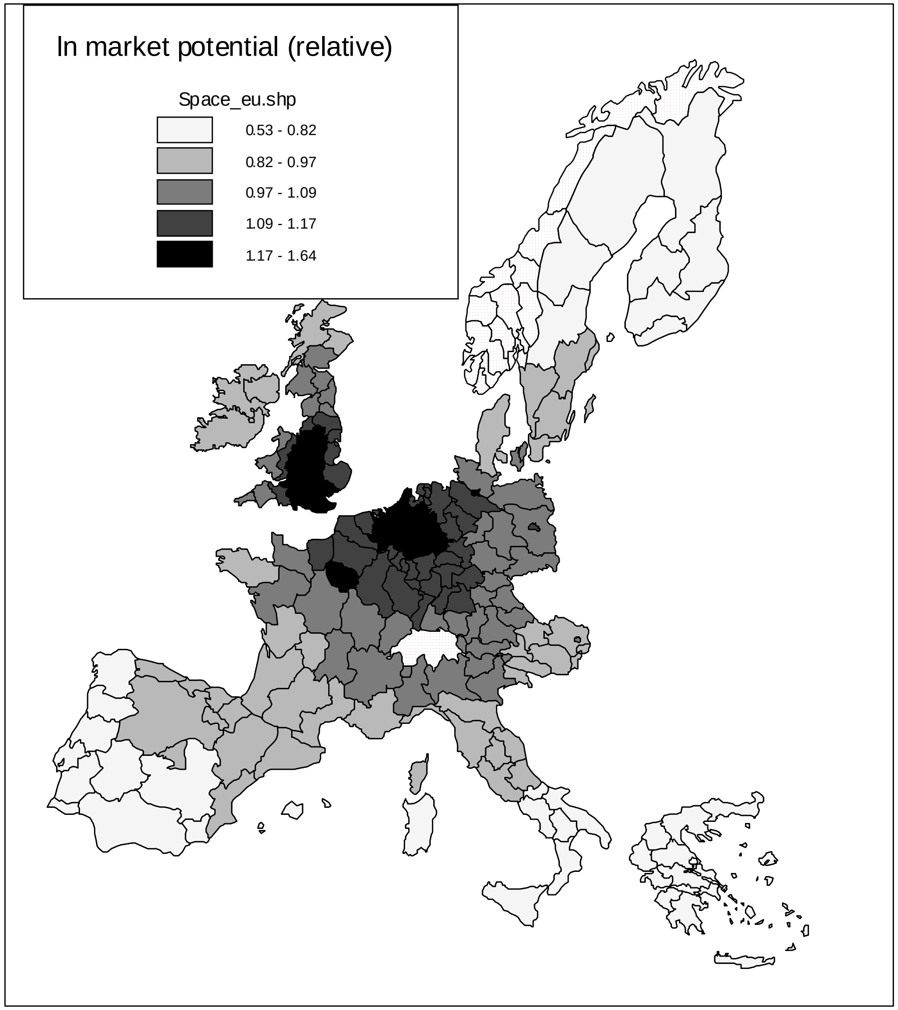
Random effects

- The Hausman test examines the consistency of the GLS (random effects) estimates.
- The null hypothesis is that the random effects estimates are consistent —

that is, that the disturbances and Xs are independent

- The test is based on a measure, *H*, of the "distance" between the fixed-effects and random-effects estimates
- *H* follows a Chi-squared distribution with degrees of freedom equal to the number of time-varying regressors in the matrix *X*.
- If the value of *H* is "large" this suggests that the random effects estimator is not consistent and the fixed-effects model is preferable.


ln wage rate
Space_eu.shp
8.79 - 9.9
9.9 - 10.08
10.08 - 10.21
10.21 - 10.34
10.34 - 10.76



ln market potential (relative)
Space_eu.shp
0.53 - 0.82
0.82 - 0.97
0.97 - 1.09
1.09 - 1.17
1.17 - 1.64

- Data layout 255 EU regions

CODE	NAME	lnGVApw	ln_adj_p_g	lns	lnMPa	lnHed	year_1995(1)-2003(9)	CZ	Eesti
AT11	Burgenland	10.57923	-2.89944	-1.34807	9.297275	1.811864	1	0	0
AT12	Niederösterreic	10.70796	-2.8929	-1.43019	9.25791	2.035079	1	0	0
AT13	Wien	10.93456	-3.11427	-1.71378	10.17165	2.625214	1	0	0
AT21	Kärnten	10.67353	-2.94139	-1.46771	9.291567	1.949171	1	0	0
AT22	Steiermark	10.62469	-3.00101	-1.43836	9.236481	2.175438	1	0	0
AT31	Oberösterreich	10.70373	-2.91851	-1.46739	9.284781	1.891296	1	0	0
AT32	Salzburg	10.76274	-2.86082	-1.47764	9.330464	2.21531	1	0	0
AT33	Tirol	10.70657	-2.83274	-1.32531	9.334951	1.827362	1	0	0
AT34	Vorarlberg	10.7661	-2.85989	-1.42534	9.581739	1.872076	1	0	0
BE10	Région de Brux	11.00334	-2.9784	-1.7951	10.7211	2.893847	1	0	0
BE21	Prov. Antwerpe	10.96942	-2.9369	-1.61928	9.706519	2.588043	1	0	0
BE22	Prov. Limburg (	10.80939	-2.83605	-1.50151	9.653046	2.285049	1	0	0
BE23	Prov. Oost-Vlaa	10.81587	-2.94067	-1.53618	9.660629	2.595616	1	0	0
BE24	Prov. Vlaams B	11.00496	-2.84727	-1.63594	9.754418	2.875688	1	0	0
BE25	Prov. West-Vla	10.76332	-2.94501	-1.51328	9.632094	2.397464	1	0	0
BE31	Prov. Brabant V	10.95707	-2.74243	-1.60496	9.803192	2.989779	1	0	0
BE32	Prov. Hainaut	10.74821	-3.01044	-1.80255	9.576635	2.400279	1	0	0
BE33	Prov. Liège	10.76181	-3.00145	-1.72772	9.612037	2.532512	1	0	0
BE34	Prov. Luxembo	10.64172	-2.79525	-1.51756	9.503033	2.491419	1	0	0
BE35	Prov. Namur	10.67695	-2.86625	-1.81862	9.519095	2.682265	1	0	0
CH01	Région lémanic	11.06256	-2.66348	-1.44457	9.482571	2.615551	1	0	0
CH02	Espace Mittella	10.94564	-2.99998	-1.37025	9.483919	2.432667	1	0	0
CH03	Nordwestschwe	11.10428	-2.62383	-1.49313	9.764912	2.457364	1	0	0
CH04	Zürich	11.12237	-2.66996	-1.50482	9.8488	2.652089	1	0	0
CH05	Ostschweiz	10.98804	-2.74371	-1.4115	9.464399	2.257391	1	0	0
CH06	Zentralschweiz	11.09763	-2.76917	-1.54572	9.566808	2.396922	1	0	0
CH07	Ticino	10.83974	-2.88874	-1.21977	9.524018	2.322914	1	0	0
CZ01	Praha	9.404299	-3.06956	-1.1449	9.492706	2.5983	1	1	0
CZ02	Strední Cechy	8.805276	-3.02197	-1.19799	9.240543	1.383584	1	1	0
CZ03	Jihozápad	8.896247	-2.99863	-0.87192	9.244868	1.713004	1	1	0
CZ04	Severozápad	8.919282	-2.98325	-1.27958	9.274141	1.291539	1	1	0

- With fixed effects PL(Poland) is aliased, because
- It is perfectly collinear with the dummies for fixed effects

```
Model 1: Fixed-effects estimates using 2295 observations
Included 255 cross-sectional units
Time-series length = 9
Dependent variable: lnGVApw
Omitted due to exact collinearity: PL

             coefficient   std. error   t-ratio   p-value
  -------------------------------------------------------
  const       6.62246      0.0650836    101.8      0.000  ***
  lnMPa       0.387843     0.00660828    58.69     0.000  ***

Test for differing group intercepts -
  Null hypothesis: The groups have a common intercept
  Test statistic: F(254, 2039) = 202.104
  with p-value = P(F(254, 2039) > 202.104) = 0
```

```
Model 2: Random-effects (GLS) estimates using 2295 observations
Included 255 cross-sectional units
Time-series length = 9
Dependent variable: lnGVApw

             coefficient   std. error   t-ratio    p-value
  ------------------------------------------------------------
  const        6.68620     0.0713099     93.76    0.000     ***
  lnMPa        0.389746    0.00663348    58.75    0.000     ***
  PL          -1.31447     0.113139     -11.62    2.32E-030 ***


Breusch-Pagan test -
  Null hypothesis: Variance of the unit-specific error = 0
  Asymptotic test statistic: Chi-square(1) = 7978.71
  with p-value = 0
```

Other issues

- Dynamic panels
- Fixed time effects

Dynamic panel models

error components model with lagged dependent variable

$Y_{it} = \delta Y_{it-1} + \beta_1 X_{it} + u_{it}$

$u_{it} = \alpha_i + e_{it}$

$\alpha_i \sim iid(0, \sigma_\alpha^2)$

$e_{it} \sim iid(0, \sigma_e^2)$

problem

Y_{it} depends on $u_{it} = \alpha_i + e_{it}$, hence on α_i

Y_{it-1} also depends on α_i hence Y_{it}

because α_i at t is the same as α_i at t-1

in other words

since u_{it} includes α_i, then Y_{it-1} is bound to be correlated with u_{it},

because the value of α_i affects Y_{it} at all t

This makes OLS biased and inconsistent

even if e_{it} is not serially correlated,

see Baltagi(2005) Econometric Analysis of Panel Data, ch. 8

Dynamic panel models

- Solution :
- Use first differences to eliminate the individual effects (heterogeneity)
- use an instrumental variable for the endogenous first differenced lagged values of the dependent variable
- The instrument should be correlated with the first differenced lagged values of the dependent variable but uncorrelated with the first differenced error
- Proposed by Anderson and Hsiao(1981)
- Many alternatives, notably Arellano and Bond(1991)

• Dynamic panel models

error components model with lagged dependent variable

$$Y_{it} = \delta Y_{it-1} + \beta_1 X_{it} + u_{it} \quad i = 1,\ldots,N; t = 1,\ldots,T$$

$$u_{it} = \alpha_i + e_{it}$$

$$\alpha_i \sim iid(0, \sigma_\alpha^2)$$

$$e_{it} \sim iid(0, \sigma_e^2)$$

first difference to get rid of the α_i

$$Y_{it} - Y_{it-1} = \delta(Y_{it-1} - Y_{it-2}) + \beta_1(X_{it} - X_{it-1}) + (e_{it} - e_{it-1})$$

$$\Delta Y_{it} = \delta \Delta Y_{it-1} + \beta_1 \Delta X_{it} + \Delta e_{it}$$

Anderson and Hsaio(1981) suggest Y_{it-2} as an instrument for ΔY_{it-1}

Y_{it-2} will not correlate with Δe_{it} provided e_{it} is not serially correlated

• Dynamic panel models

- Does MP retain its significance in the presence of the
- Lagged dependent variable?
- Anderson-Hsiao estimator

```
Model 3: TSLS estimates using 1785 observations
Dependent variable: d_lnGVApw
Instruments: const d_lnMPa lnGVApw_2

               coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------------
  const        -0.00400154   0.00357919   -1.118    0.2636
  d_lnMPa       0.0365759    0.0203616     1.796    0.0724    *
  d_lnGVApw_1   0.877190     0.0624823    14.04     9.00E-045 ***

Hausman test -
  Null hypothesis: OLS estimates are consistent
  Asymptotic test statistic: Chi-square(1) = 199.626
  with p-value = 2.51984e-045

First-stage F-statistic (1, 1782) = 406.743
  A value < 10 may indicate weak instruments
```

• Dynamic panel models

- Does MP retain its significance in the presence of the
- Lagged dependent variable?
- Anderson-Hsiao estimator with two rhs endogenous variables

```
Model 7: TSLS, using 1785 observations
Dependent variable: d_lnGVApw
Instrumented: d_lnMPa d_lnGVApw_1
Instruments: const ne PL HU CZ lnGVApw_2

               coefficient   std. error   t-ratio    p-value
  ----------------------------------------------------------
  const        -0.0194414    0.00802026   -2.424    0.0153    **
  d_lnMPa       0.200562     0.0809965     2.476    0.0133    **
  d_lnGVApw_1   0.823811     0.0685286    12.02     2.74e-033 ***

Mean dependent var   0.042329   S.D. dependent var   0.056240
Sum squared resid    7.325627   S.E. of regression   0.064116
R-squared            0.085723   Adjusted R-squared   0.084697
F(2, 1782)           115.9455   P-value(F)           4.60e-48
```

• Dynamic panel models

- Does MP retain its significance in the presence of the
- Lagged dependent variable?
- Anderson-Hsiao estimator with two rhs endogenous variables
- The Hausman test shows that we need to use instruments
- The Sargan test indicates that the instruments are valid,
- i.e. independent of the errors

```
Hausman test -
  Null hypothesis: OLS estimates are consistent
  Asymptotic test statistic: Chi-square(2) = 215.869
  with p-value = 1.33216e-047

Sargan over-identification test -
  Null hypothesis: all instruments are valid
  Test statistic: LM = 6.29168
  with p-value = P(Chi-Square(3) > 6.29168) = 0.0982503
```

Introducing Time Fixed Effects

- An omitted variable might vary over time but not across regions/countries/individuals:
- E.G. legislation at EU level (employment, environment etc.)
- These produce intercepts that change over time
- The resulting regression model is:

$$Y_{it} = \phi_t + \beta_1 X_{1it} + \varepsilon_{it}$$

Fixed Time Effects

- The fixed time effects are introduced in exactly the same way as the individual fixed effects, with N-1 dummies (plus constant) or N (without constant) or demeaning
- In this case, the dummies are set to 1 for a specific time period, and zero otherwise
  - For example, the dummy variable for 1970 would have 1s for all the EU regions for 1970, and zeros for all other times
  - In contrast a region specific fixed effect has 1s for the region for all times, and zeros for all the other regions.
- Demeaning is with reference to time means not region means.

Fixed time effects : fixed effects model

Model 4: Fixed-effects estimates using 2295 observations
Included 255 cross-sectional units
Time-series length = 9
Dependent variable: lnGVApw
Omitted due to exact collinearity: PL CZ Eesti HU Lietuva Latvija Slovenija
SK

	coefficient	std. error	t-ratio	p-value	
const	6.57181	0.837659	7.845	6.92E-015	***
lnMPa	0.391029	0.0890567	4.391	1.19E-05	***
dt_2	0.0528402	0.00806929	6.548	7.35E-011	***
dt_3	0.0439357	0.0173945	2.526	0.0116	**
dt_4	-0.0177550	0.0371898	-0.4774	0.6331	
dt_5	-0.00321803	0.0419597	-0.07669	0.9389	
dt_6	0.00484411	0.0559099	0.08664	0.9310	
dt_7	0.00999319	0.0655615	0.1524	0.8789	
dt_8	0.0344413	0.0695263	0.4954	0.6204	
dt_9	0.0484496	0.0693092	0.6990	0.4846	

```
Test for differing group intercepts -
  Null hypothesis: The groups have a common intercept
  Test statistic: F(254, 2031) = 56.3848
  with p-value = P(F(254, 2031) > 56.3848) = 0

Wald test for joint significance of time dummies
  Asymptotic test statistic: Chi-square(8) = 164.676
  with p-value = 1.68164e-031
```

Fixed time effects : random effects model

Model 5: Random-effects (GLS) estimates using 2295 observations
Included 255 cross-sectional units
Time-series length = 9
Dependent variable: lnGVApw

	coefficient	std. error	t-ratio	p-value	
const	7.21009	0.508853	14.17	9.69E-044	***
lnMPa	0.346095	0.0537823	6.435	1.50E-010	***
PL	-1.46110	0.0631224	-23.15	1.25E-106	***
CZ	-1.29706	0.0845054	-15.35	1.14E-050	***
Eesti	-1.56843	0.233672	-6.712	2.41E-011	***
HU	-1.35210	0.0909194	-14.87	8.23E-048	***
Lietuva	-1.88865	0.233829	-8.077	1.06E-015	***
Latvija	-1.86960	0.233820	-7.996	2.03E-015	***
Slovenija	-0.726502	0.232720	-3.122	0.0018	***
SK	-1.42925	0.118351	-12.08	1.37E-032	***
dt_2	0.0530195	0.00806326	6.575	6.00E-011	***
dt_3	0.0517129	0.0123133	4.200	2.78E-05	***
dt_4	0.000563398	0.0233600	0.02412	0.9808	
dt_5	0.0175588	0.0261415	0.6717	0.5019	
dt_6	0.0327592	0.0343703	0.9531	0.3406	
dt_7	0.0428219	0.0401111	1.068	0.2858	
dt_8	0.0692849	0.0424762	1.631	0.1030	
dt_9	0.0831829	0.0423467	1.964	0.0496	**

Fixed time effects : random effects model

```
Breusch-Pagan test -
  Null hypothesis: Variance of the unit-specific error = 0
  Asymptotic test statistic: Chi-square(1) = 6793.21
  with p-value = 0

Hausman test -
  Null hypothesis: GLS estimates are consistent
  Asymptotic test statistic: Chi-square(9) = 0.40073
  with p-value = 0.999988
```

Panel data Application: Drunk Driving Laws and Traffic Deaths

Some facts

- Approx. 40,000 traffic fatalities annually in the U.S.
- 1/3 of traffic fatalities involve a drinking driver
- 25% of drivers on the road between 1am and 3am have been drinking (estimate)
- A drunk driver is 13 times as likely to cause a fatal crash as a non-drinking driver (estimate)
- Drunk driving causes massive externalities (sober drivers are killed, etc.). There is ample justification for governmental intervention

The role of alcohol taxes

Public policy issues

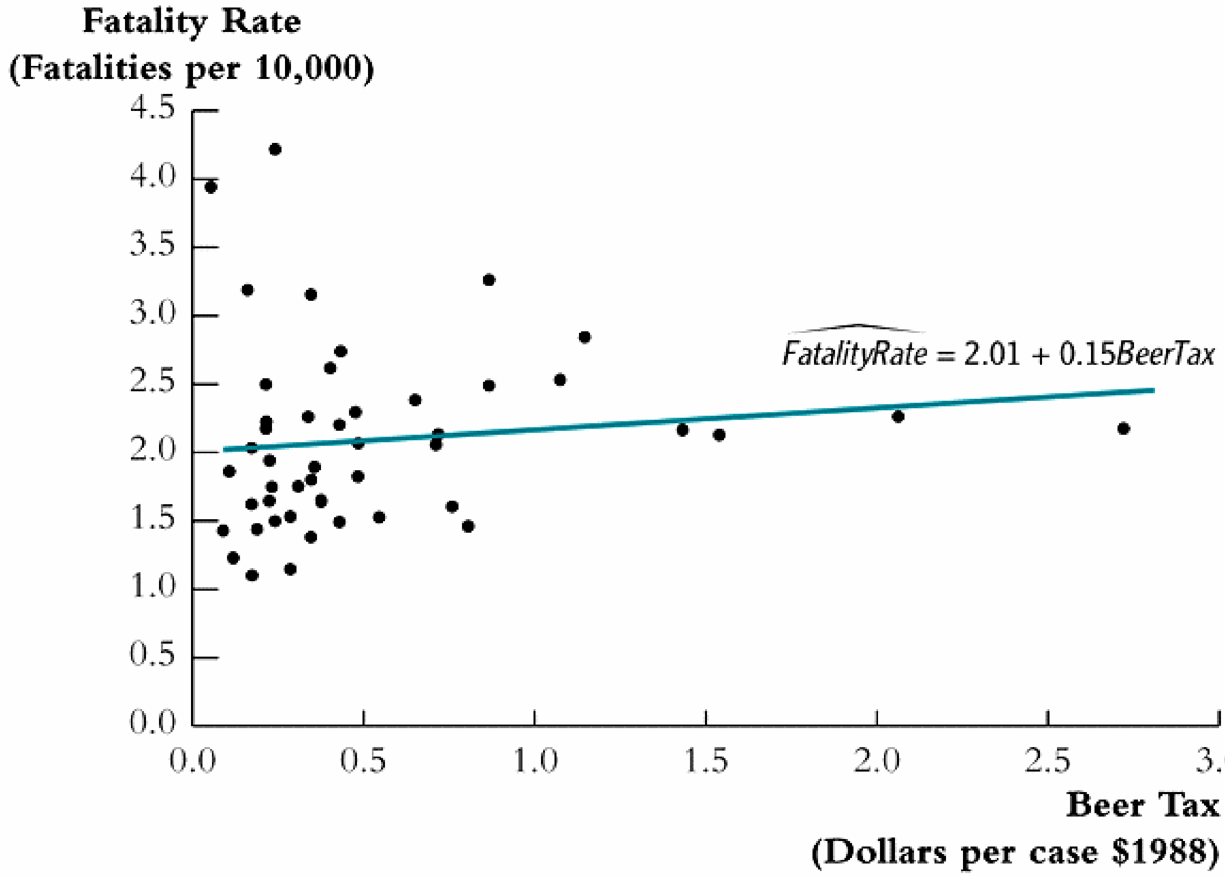
- Are there any effective ways to reduce drunk driving? If so, what?
- What are effects of specific laws:
  - mandatory punishment
  - minimum legal drinking age
  - <u>economic interventions (alcohol taxes)</u>

Data

- 48 U.S. states, so N = number of states = 48
- 7 years (1982, ... , 1988), so T = number of time periods = 7
- Balanced panel, so total number of observations = 7 × 48 = 336
- Variables:
- Traffic fatality rate *FR* (number of traffic deaths in that state in that year, per 10,000 state residents)
- Tax on beer
- Other variables (legal driving age, drunk driving laws, etc.)

FIGURE 8.1 **The Traffic Fatality Rate and the Tax on Beer**

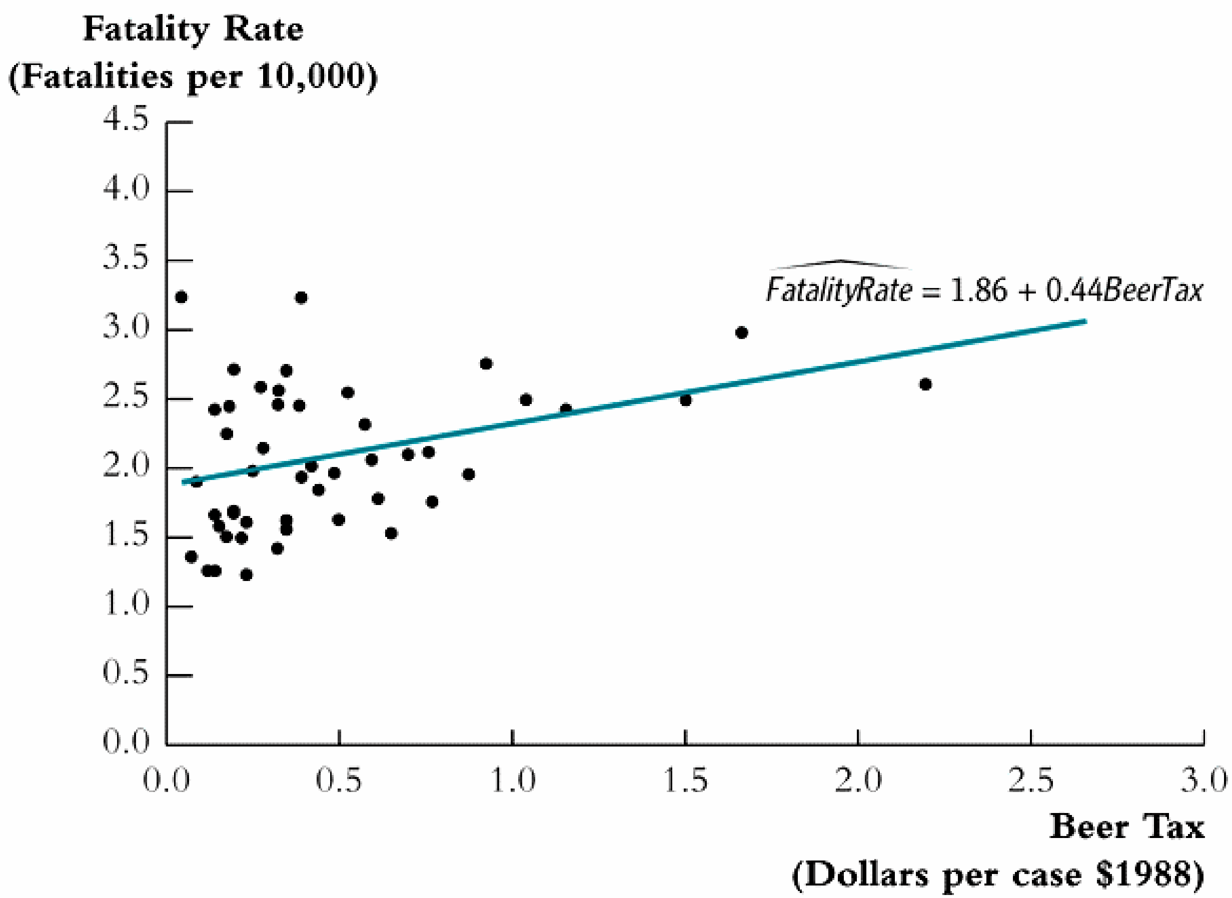
Panel a is a scatterplot of traffic fatality rates and the real tax on a case of beer (in 1988 dollars) for 48 states in 1982. Panel b shows the data for 1988. Both plots show a positive relationship between the fatality rate and the real beer tax.



(a) 1982 data

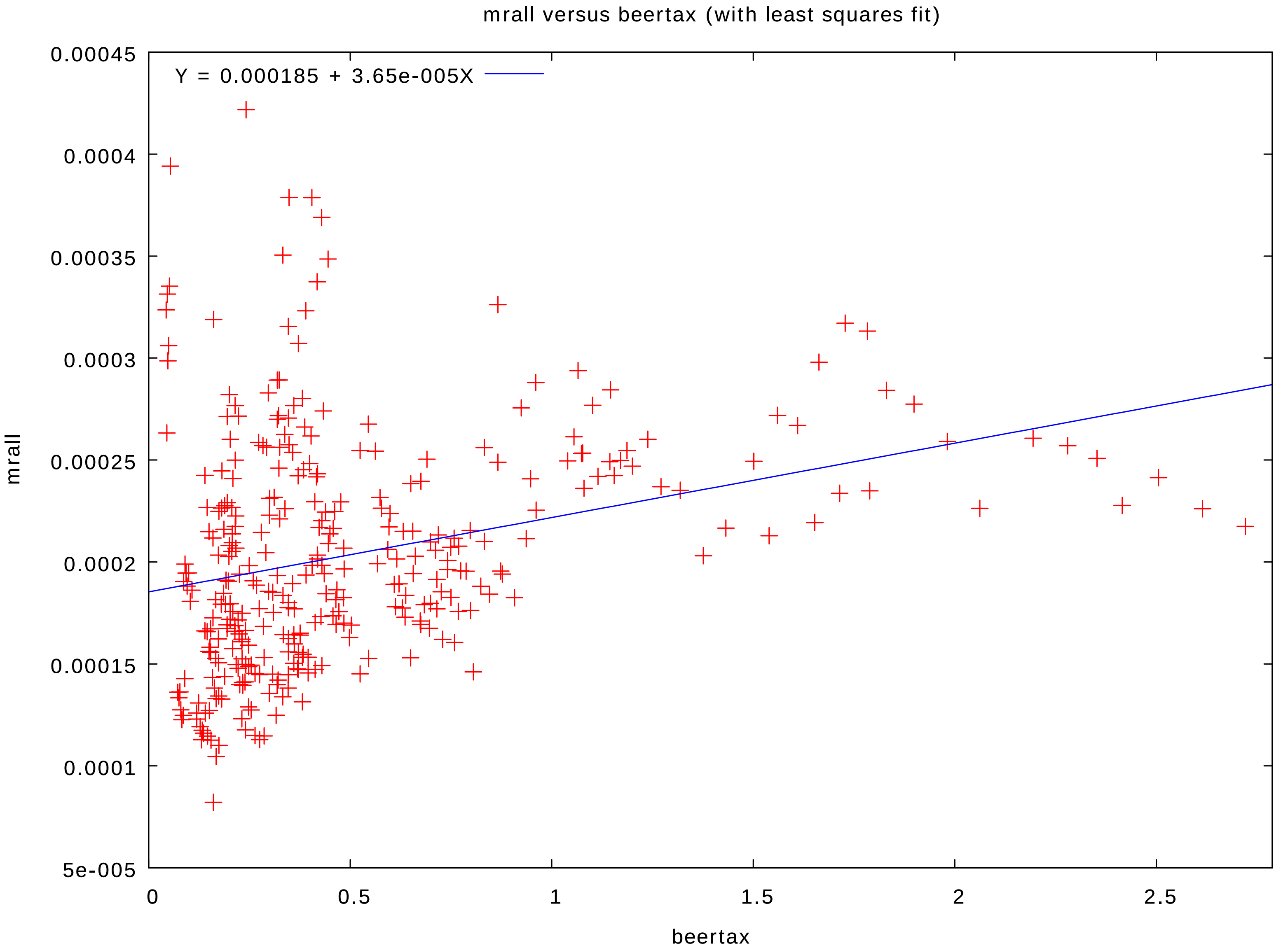
FIGURE 8.1 The Traffic Fatality Rate and the Tax on Beer

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

Fatalities increase with beer tax : all observations



Inference ?

- Higher alcohol taxes are associated with more traffic deaths.
- Higher alcohol taxes leading causally to more traffic deaths is implausible.
- Why might there be higher traffic death rates in states with higher alcohol taxes?

Inference ?

- Likely explanation is that other factors that also determine the traffic fatality rate in any state are not 'controlled for' in the simple regression of *FR* on beer tax.

- By omitting these factors, it is likely that the regression model that underlies these scatter plots is misspecified as a result of omitted variable bias.

Possible omitted variables

- Potential omitted variables (OV) bias from variables that vary across states but are constant over time:
  - culture of drinking and driving
  - quality of roads
  - Average age of cars
- Thus, use state fixed effects

- Potential OV bias from variables that vary over time but are constant across states:
  - improvements in auto safety over time
  - changing national attitudes towards drink driving
- Thus use time fixed effects

Regression with State and Time Fixed Effects

with both state effects α_i and time effects ϕ_t, the model is

$$Y_{it} = \alpha_i + \phi_t + \beta_1 X_{1it} + \varepsilon_{it}$$

example : Traffic deaths

```
Model 2: Fixed-effects, using 336 observations
Included 48 cross-sectional units
Time-series length = 7
Dependent variable: mrall

             coefficient   std. error   t-ratio    p-value
  --------------------------------------------------------
  const       2.42847      0.108120     22.46     1.12e-064 ***
  beertax    -0.639980     0.197377     -3.242    0.0013    ***
  dt_2       -0.0799029    0.0383537    -2.083    0.0381    **
  dt_3       -0.0724206    0.0383517    -1.888    0.0600    *
  dt_4       -0.123976     0.0384418    -3.225    0.0014    ***
  dt_5       -0.0378645    0.0385879    -0.9813   0.3273
  dt_6       -0.0509021    0.0389737    -1.306    0.1926
  dt_7       -0.0518038    0.0396235    -1.307    0.1921

Mean dependent var   2.040444   S.D. dependent var   0.570194
Sum squared resid    9.919301   S.E. of regression   0.187883
R-squared            0.908927   Adjusted R-squared   0.891425
F(54, 281)           51.93379   P-value(F)           9.6e-118
```

example : Traffic deaths

```
Test for differing group intercepts -
  Null hypothesis: The groups have a common intercept
  Test statistic: F(47, 281) = 53.1926
  with p-value = P(F(47, 281) > 53.1926) = 2.93879e-114

Wald test for joint significance of time dummies
  Asymptotic test statistic: Chi-square(6) = 12.0701
  with p-value = 0.0604241
```

example : Traffic deaths with time dummies eliminated

```
Model 1: Fixed-effects, using 336 observations
Included 48 cross-sectional units
Time-series length = 7
Dependent variable: mrall

             coefficient   std. error   t-ratio    p-value
  --------------------------------------------------------
  const        2.37707     0.0969699    24.51     2.35e-072 ***
  beertax     -0.655874    0.187850     -3.491    0.0006    ***

Mean dependent var   2.040444   S.D. dependent var   0.570194
Sum squared resid    10.34537   S.E. of regression   0.189859
R-squared            0.905015   Adjusted R-squared   0.889129
F(48, 287)           56.96916   P-value(F)           2.0e-120
```

Drunk Driving and Traffic Deaths Empirical Analysis: Main Results

- Sign of beer tax coefficient changes when fixed state effects are included
- Fixed time effects are marginally significant and do not have big impact on the estimated coefficients
- Is the effect of beer tax the same when other laws are included as regressor?
- Are there other policy variables that have an impact is the tax on beer – such as minimum drinking age, sentencing policy, etc?
- Which economic variables are also a cause of variation in fatality rates (e.g income) and why?