# Panel data

Readings

Panel data

Examples

First differences

Fixed effects

Example: beertax

LSDV + Matrices

Models + estimators

# Readings

# Prep

```
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr

## Conflicts with tidy packages ----------------------------

## filter(): dplyr, stats
## lag():    dplyr, stats
```

# Readings

- These notes supplement readings
- Chapter 13+14
  - self-study: through 13.3
  - skip: Chow test
- Read 13+14 before next Tuesday
- First week: until (not including) derivations for FD
- Second week: FD, FE, LSDV-FWL

# Panel data

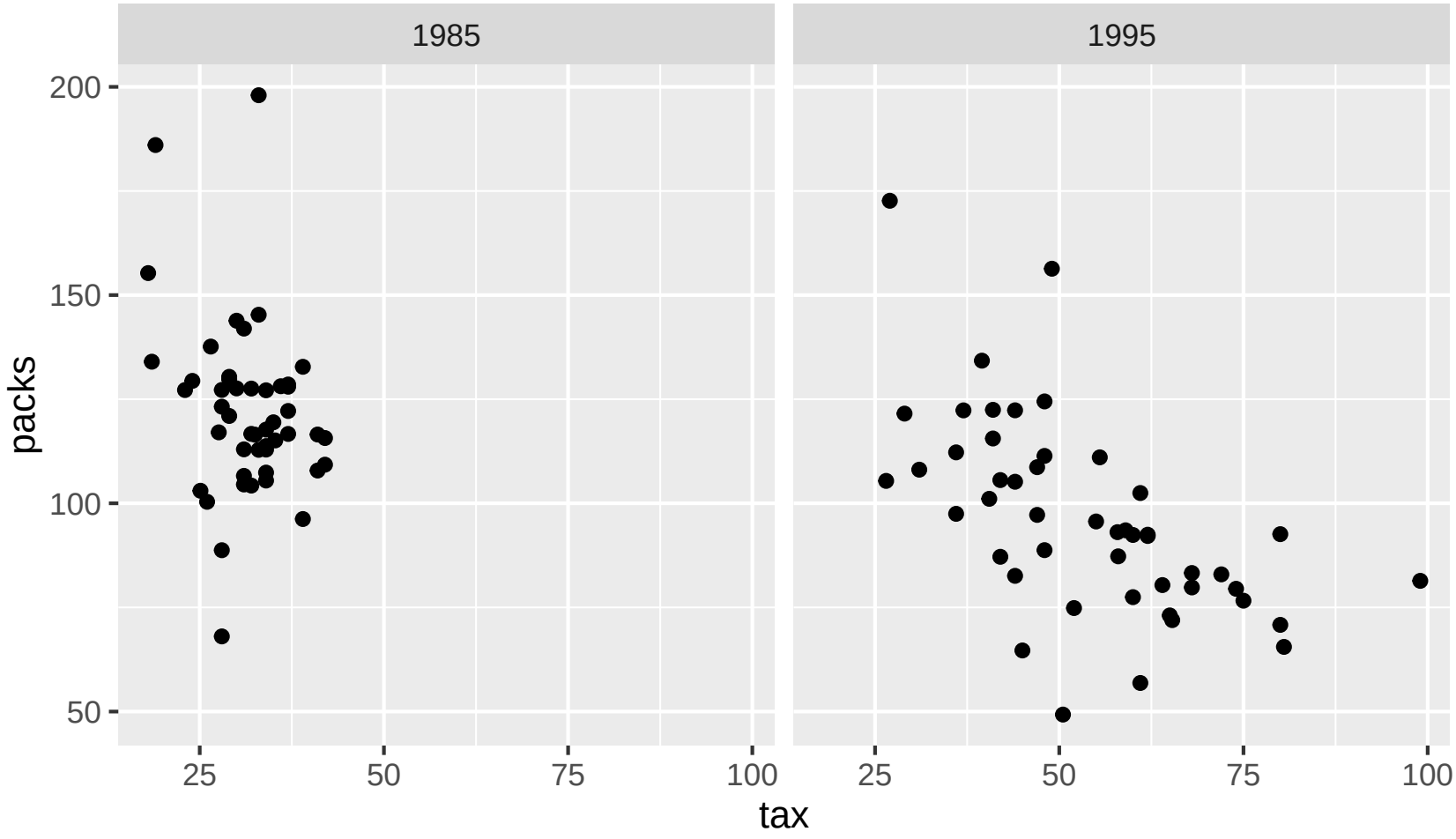
# Panel data

A **panel data** set is a data set that has two dimensions:

- cross-sectional
- time

# PD: Example

```
data("CigarettesSW",package="AER")
qplot(tax,packs,data=CigarettesSW) + facet_wrap(~year)
```



# PD: Example (2)

```
data("PSID7682",package="AER")
head(PSID7682 %>% select(id,year,wage,education),20)
```

```
##     id year wage education
## 1    1 1976  260         9
## 2    1 1977  305         9
## 3    1 1978  402         9
## 4    1 1979  402         9
## 5    1 1980  429         9
## 6    1 1981  480         9
## 7    1 1982  515         9
## 8    2 1976  475        11
## 9    2 1977  500        11
## 10   2 1978  525        11
## 11   2 1979  695        11
## 12   2 1980  810        11
## 13   2 1981  890        11
## 14   2 1982  912        11
```

# PD (3)

Panel data gives us **repeated measurements** on the same unit.

This week's PS:

- unit is family
- repeated measurements are the twins

# Sampling

- are rows independently distributed draws?
- are rows identically distributed draws?

# Sampling

- Data from different years may follow different distributions
- Data from the same unit is likely dependent across time

# Why use panels?

1. Panel data sets contain **more information**
   - smaller standard errors than comparable cross-sections
2. Estimate how (conditional) distributions change over time
3. Real innovation:
   - estimate models with heterogeneous individuals
   - Under E.1-E.4, all individuals are the same, conditional on $X_i$
     - $E[u_i|X_i] = 0$: people are identical, in expectation
   - Repeated measurements allows you to control for the fact that people are different:
     - **unobserved heterogeneity**

# Examples

# Intro

- To understand why panel data can be useful: examples.
- In each example:

$$y_{it} = \alpha_i + X_{it}\beta + u_{it}$$

  - Repeated measurements for each $i$ across $t$
  - $u_{it}$ will be an error terms as in the first part of this course
  - $\alpha_i$ represents a unit-specific error term:
    - unobserved (not in $X$) characteristics of the individual
    - time-invariant
- Questions:
  - What is in $\alpha_i$?
  - Is it correlated with $X_{it}$?

# Hint:

- Yes, it is correlated with $X_{it}$.
  - (But: Why?)
- Preview of panel data models:
  - $E(u_{it} + \alpha_i | X_{it}) \neq 0$

# Example 1: Birthweight and smoking

- Data on
  - birthweight of newborns ($y_{it}$)
  - smoking behavior of their mother $X_{1,it}$
  - $(i, t)$ refers to the $t-$th newborn of mother $i$

Linear panel data model:

$$y_{it} = \alpha_i + X_{it}\beta + u_{it}$$

where $X_{it}$ includes

- $X_{1it}$,
- age
- income,
- education

# Example 1: Questions

1. What are other factors that could influence a baby's birthweight?
2. Do you believe that those factors do not change over time?
3. Do you believe that those factors are correlated with smoking?

# Example 1: Answers

1: genetics, diet, exercise, healthy behavior, diligence in precautions for baby's outcomes

2: genetics: yes, diligence: probably not, healthy behavior: probably pretty persistent

3: genetics: tricky, should be different genes that make you tall and that make you smoke

The *unobserved heterogeneity* can be *specific*, unincluded variables, or *vaguer* terms such as "health awareness", "healthy behavior".

# Example 2: Skytrain

- Data on housing prices in two periods, $t = 1, 2$
- In between, a skytrain is built
- $D_i$ is a binary indicator: is skytrain $<$5min walk from house $i$?

Relationship:

$$y_{it} = \alpha_i + D_i\beta + u_{it} + X_{it}\gamma$$

where $X_{it}$ is a vector of the characteristics of the house:

- number of bedrooms,
- bathrooms, and
- square footage, and the
- year that it was built
- ...

# Example 2: Skytrain: Questions

1. What does $\alpha_i$ capture?
2. Why is it correlated with D_i ?

# Example 2: Skytrain: Answers

1. Amenities (parks, schools, shopping); location (density, . . . )
2. Think about the decision to build the skytrain station. there are several ways in which this decision making process can induce correlation between $\alpha_i$ and $D_{it}$. If the social planner is trying to develop a new neighhbourhood, they may be looking for a spot with cheap (c.p.) plots. Alternatively, they may be targeting areas with high density, or with a lot of amenities, because it likely increases the usage rate. These are two opposing mechanisms. It is unlikely that they cancel out exactly.

Compare the **incinerator** example in Wooldridge.

# Example 3: Texting bans

- Existing literature:

$$P(death|driving + phone) = 4 \times P(death|driving + nophone)$$

- People continue to text. Why?

$$Y_{i,m} = \alpha_i + \delta_m + X_{im}\beta + \omega B_{im} + u_{im}$$

where:

- $i$ is state, $m$ is month
- $Y$ is (log of) traffic fatalities
- $X$ includes
  - population
  - proportion male
  - unemployment
  - gas tax
- $B$: is a texting ban in place?

# Example 3: Texting bans (2)

- What's in $\alpha_i$?
- Correlated with $X$?
- Correlated with $B$?

# Example 3: Texting bans

Finding: $\hat{\omega} = -0.0374$.

- Interpret this finding.

Details:

- No effect for "weak bans"
- No effect except for single-occupancy vehicles
- Effect starts when findings are announnced, disappears four months after ban in effect

# Example 4: Mafia and public spending

From PS8, AER(2014)

$$Y_{it} = \alpha_i + G_{it}\beta + \gamma_t + u_{it} + X_{it}\beta$$

- $i$ is an Italian province, $t$ is a year (1990-1999)
- $Y_{it}$ is the rate of growth
- $G_{it}$ is government spending on infrastructure in state $i$
- $X_{it}$: controls

# Example 4: Mafia

In terms of growth rates:

- What is captured by $\gamma_t$?
- What is captured by $\alpha_i$?

Why is $\alpha_i$ correlated with $G_{it}$?

Note: Paper uses panel data and IV.

# Example 5: A community-college teacher like me

Fairlie et al, AER(2014)

$$Y_{ic} = \alpha_i + \lambda_c + \beta 1 Z_{ic} + u_{ic} + X_{ic}\gamma$$

where

- $Y_{ic}$:
  - dropped course?
  - passed course | finishing
  - grade | finishing
  - good grade? | finishing
  - enrolled in a similar course subsequently?
- $Z_{ic}$ is an indicator for whether student $i$ and $j$ are part of the same minority
- $X_{ic}$ is a vector of controls
- What does $\alpha_i$ capture?
- What does $\lambda_c$ capture?

# Example 5: A

$\lambda_c$ and $\alpha_i$

> *control[s] for instructor fixed effects and minority-specific course fixed effects. The former controls for the possibility that minority students take courses from instructors who have systematically different grading policies from other instructors, while the latter controls for selection by comparative advantage where minority students are drawn to courses that are a particularly good match or in which minority students are drawn to courses that are a particularly good match.*

p. 2574, Fairlie, Hoffmann, Oreopolous (2014)

# Example 5: Findings

No findings without fixed effects. With fixed effects:

- dropped course?: $-0.02 * **$
- passed course | finishing: 0.012
- grade | finishing: $0.054 * **$
- good grade? | finishing $0.024 * **$
- enrolled in a similar course subsequently? $0.013*$

# Example 6: Income and democracy

**PS6**

$$democracy_{it} = \alpha_i + GDP_{it}\beta + u_{it}$$

1. Reverse causality?
2. What is in $\alpha_i$?

# Takeaways

1. What is in $\alpha_i$?
   - "unobserved heterogeneity", or:
   - "fixed effect", or:
   - an intercept specific to the cross-section unit, or:
   - omitted variable that does not change over time
2. Key feature of panel data in economics: **Un**observed heterogeneity is generally correlated with the **observ**ables $X_{it}$

# Unobserved heterogeneity: problem

- 1: Geometrically: *sketch81* and *sketch82*
- 2: Omitted variables bias (333)

- Violation of E.???

# Problem (2)

- If estimation of $\beta$ using OLS,

$$y_{it} = X_{it}\beta + v_{it}$$

  then

$$v_{it} = \alpha_i + u_{it}.$$

- Even if $E[u_{it}|X_{it}] = 0$, still

$$E[v_{it}|X_{it}] = E[\alpha_i|X_{it}] = h(X_{it}) \neq 0$$

# Problem (3)

Why do we care about the **causal effect**, rather than the correlation?

# Problem (4)

- The OLS estimator will not converge to

$$\beta = E(y|X = x + 1) - E(y|X = x)$$

due unobserved heterogeneity correlated with $X_{it}$.

# First differences

# Solution: $T = 2$

- 1: Geometrically: *sketch83*
- 2: Algebraically: *whiteboard*

# Solution (2)

- Differenced expression is free of $\alpha_i$
- Apply OLS to time-differenced model

# New problem

Nothing is for free:

- We have solved a big problem: any time-invariant omitted variable is taken care of
- However, we need to impose strong conditions on error terms
- Exogeneity condition for OLS would be:

$$E[u_{it}|X_{it}] = 0$$

Right?

# New problem

That condition is not sufficient for the consistency of OLS after differencing.

- We need:

$$E(\Delta u_i | \Delta X_i) = 0$$

- Not implied by $E[u_{it}|X_{it}]$
- Impose **strict exogeneity**

$$E(u_{is}|X_{it}) = 0 \forall (s,t).$$

# New problem: texting bans

- Back to texting ban example
- Assumption that $E(u_{i1}|B_{i2}) = 0$ rules out ...

# New problem: texting bans

. . . political response.

- Imagine a sudden increase in traffic fatilties in state $i$ in period 1
- Politicians respond by imposing the texting ban.
- Ruled out by *strict exogeneity*: the assumption is restrictive

# $T > 2$

How to extend to $T > 2$?

# Fixed effects

# Geometrically

Picture of problem and solution: *sketch-83*

# Algebra

*whiteboard*

# Example: beertax

# Introduction

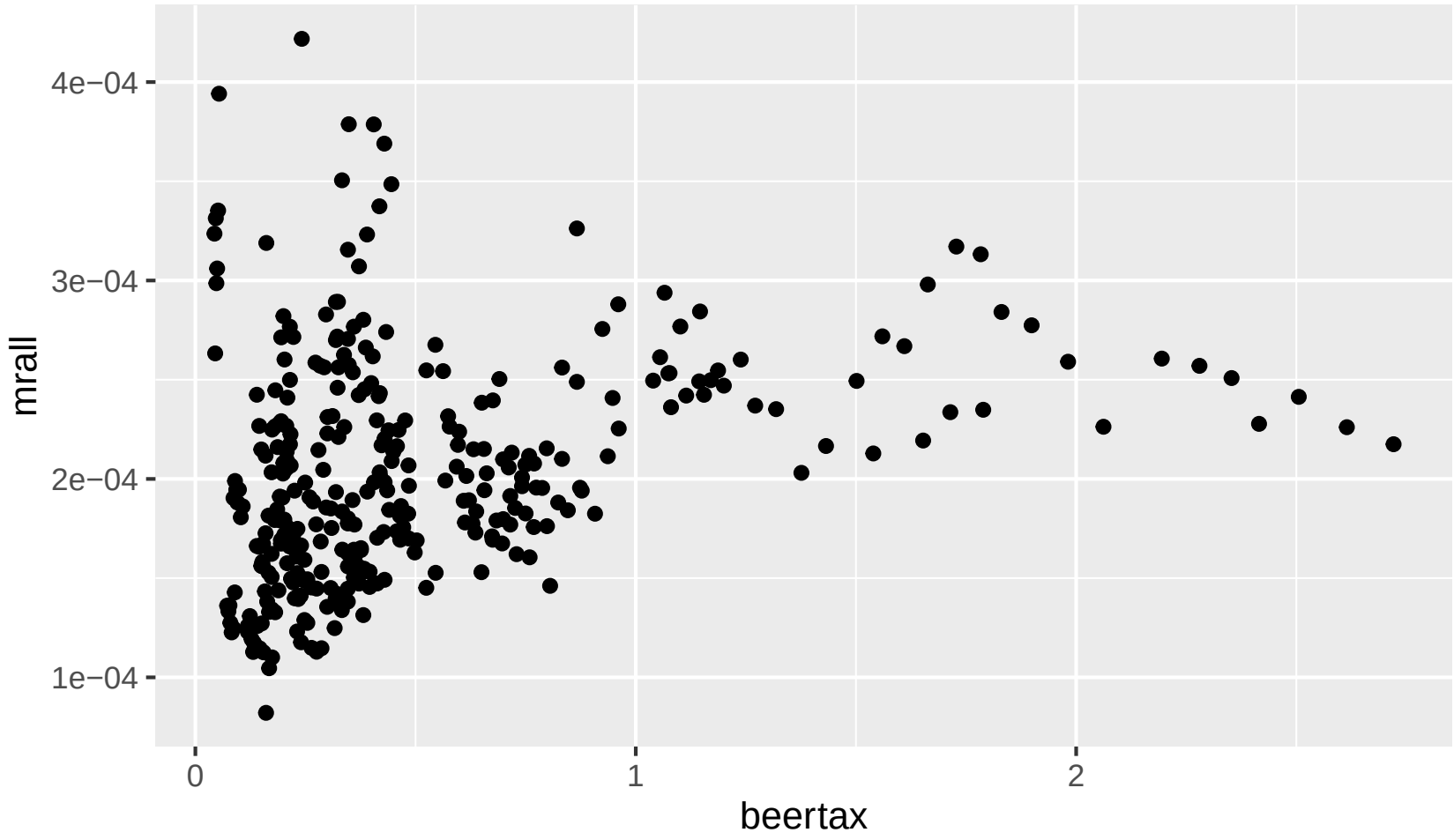
from Stock and Watson.

```
## Load the fatality data
library(haven)
beer_fatality <- read_dta("fatality.dta")
summary(beer_fatality)
```

```
##      state            year         spircons          unra
##  Min.   : 1.00   Min.   :1982   Min.   :0.790   Min.
##  1st Qu.:18.75   1st Qu.:1983   1st Qu.:1.300   1st Qu.
##  Median :30.50   Median :1985   Median :1.670   Median
##  Mean   :30.19   Mean   :1985   Mean   :1.754   Mean
##  3rd Qu.:42.50   3rd Qu.:1987   3rd Qu.:2.013   3rd Qu.
##  Max.   :56.00   Max.   :1988   Max.   :4.900   Max.
##
##      perinc          emppop          beertax              s
##  Min.   : 9514   Min.   :42.99   Min.   :0.04331   Min.
##  1st Qu.:12086   1st Qu.:57.69   1st Qu.:0.20885   1st Q
```

# beertax:plot

```
library("ggplot2")
qplot(beertax,mrall,data=beer_fatality)
```



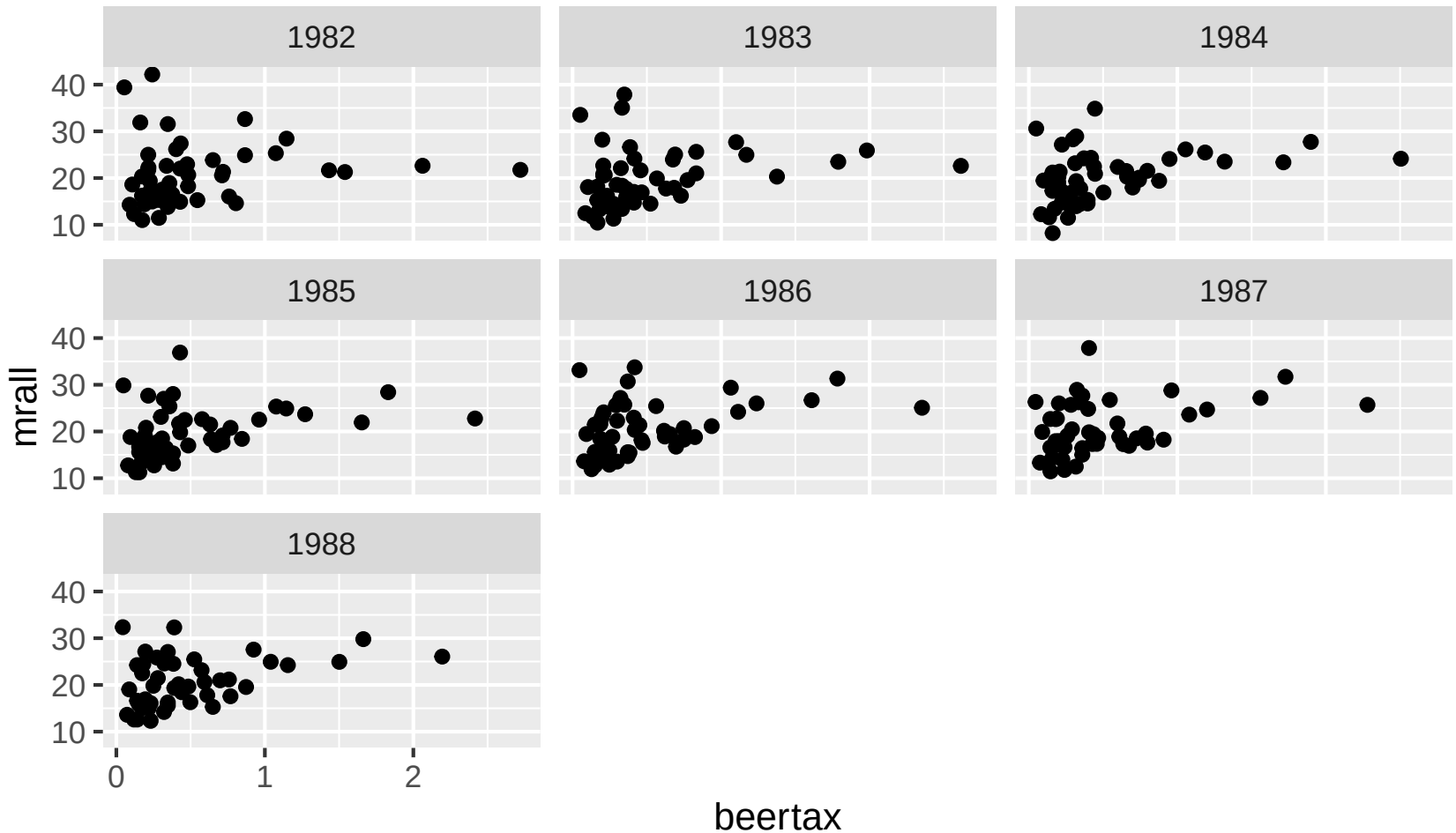
# beertax: OLS

```
## Scale up fatailities for readability
beer_fatality$mrall <- beer_fatality$mrall*100000
ols_reg <- lm(mrall~beertax,data=beer_fatality)
summary(ols_reg)
```

```
##
## Call:
## lm(formula = mrall ~ beertax, data = beer_fatality)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9060  -3.7768  -0.9436   2.8548  22.7643
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.5331     0.4357  42.539  < 2e-16 ***
## beertax       3.6461     0.6217   5.865 1.08e-08 ***
## ---
```

# beertax: Plot for each year:

```
p <- qplot(beertax,mrall,data=beer_fatality)
p <- p + facet_wrap(~year)
p
```



# beertax: Fixed effects results

```r
# Fixed effects estimator
fe_reg <- lm(mrall~beertax+as.factor(state),data=beer_fatal
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2015). stargazer: Well-Formatted Regress

##  R package version 5.2. http://CRAN.R-project.org/packag
```

```r
stargazer(ols_reg,fe_reg,
          type = "text",
          keep = "beertax",
          keep.stat = c("n"))
```

```
##
## ===============================================
```

# LSDV + Matrices

# Intro

By FWL, we know that $y$ on $X_1$ and $X_2$ is equivalent to

1. regressing $y$ on $X_1$, call residuals $y^*$
2. regressing $X_2$ on $X_1$, call residuals $X_2^*$
3. regressing $y^*$ on $X_2^*$

Equivalence is for $\hat{\beta}_2$:

# FWL for PD

Panel data regression model

$$y_{it} = \alpha_i + X_i t\beta + u_{it}$$

can be rewritten

$$y_{it} = \sum_j \alpha_j 1\{i == j\} + X_{it}\beta + u_{it}$$

# FWL for PD (2)

Running OLS of $y$ on $\{1\{i == j\}, j\}$ and X is equivalent to:

1. $y$ on $\{1\{i == j\}, j\}$, call residuals $y^*$
2. $X$ on $\{1\{i == j\}, j\}$, call residuals $X^*$
3. $y^*$ on $X^*$

# Dummies

How do you get the residuals of the regression of y on Z $= \{$all dummies$\}$?

# Dummies (2)

Well, the

- coefficient estimates are $(Z'Z)^{-1}Z'y$, so the
- predictions are $Z(Z'Z)^{-1}Z'y$, so the
- residuals are

$$y - Z(Z'Z)^{-1}Z'y = [I - Z(Z'Z)^{-1}Z']y$$

# Dummies (3)

Let's investigate what

$$Z(Z'Z)^{-1}Z'$$

looks like.

# Dummies (4)

What is the dimension of $Z$?

# Dummies (5)

What is the dimension of $Z'Z$?

# Dummies (6)

- What does $Z'Z$ look like?
- What does $(Z'Z)^{-1}$ look like?

# Dummies (7)

Therefore,

$$Z(Z'Z)^{-1}Z' = \frac{1}{T}ZZ'$$

# Dummies (8)

- What does $ZZ'$ look like?

# Dummies (9)

Conclusion:

$$I - Z(Z'Z)^{-1}Z' = \begin{pmatrix} 1-1/T & -1/T & -1/T & 0 & 0 \\ -1/T & 1-1/T & -1/T & 0 & 0 \\ -1/T & -1/T & 1-1/T & 0 & 0 \\ 0 & 0 & 0 & 1-1/T & -1/T \\ 0 & 0 & 0 & -1/T & 1-1/T \\ 0 & 0 & 0 & -1/T & -1/T \end{pmatrix}$$

Q: What does A_T do?

# LSDV: Conclusions

FE estimator is equivalent to running a regression on the explanatory variables $X$ and a set of dummies, one for each country.

- Gauss-Markov: under E.4, LSDV=FE is efficient
- LSDV=FE estimator is OLS applied to transformed data
  1. Subtract country-specific averages
  2. Apply OLS

# Models + estimators

# Intro

see *sketch85.png*

# FE v FD

# Overview

Table 1: Overview of panel data estimators.

| | OLS | RE | FE | FD | LD |
|---|---|---|---|---|---|
| Consistent under RE | x | x | x | x | x |
| Consistent under FE | | | x | x | x |
| Efficient under RE | | x | | | |
| Efficient under FE | | | x | | |