General Linear Models

Regression

– O.Örsan Özener

The Linear Regression Model

The ***population regression line*****:**

$Test\ Score = \beta_0 + \beta_1 STR$

β_1 = slope of population regression line

$$= \frac{\Delta \text{Test score}}{\Delta STR}$$

= change in test score for a unit change in *STR*

- *Why are β_0 and β_1 "population" parameters*?
- We would like to know the population value of β_1.
- We don't know β_1, so must estimate it using data.

The Population Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + u_i,\ i = 1,\ldots,\ n$$

- We have *n* observations, (X_i, Y_i), $i = 1,.., n$.
- *X* is the ***independent variable*** or ***regressor***
- *Y* is the ***dependent variable***
- β_0 = ***intercept***
- β_1 = ***slope***
- u_i = the regression ***error***
- The regression error consists of omitted factors. In general, these omitted factors are other factors that influence *Y*, other than the variable *X*. The regression error also includes error in the measurement of *Y*.

The Ordinary Least Squares Estimator

How can we estimate β_0 and β_1 from data?

Recall that was the least squares estimator of μ_Y: $\bar{Y}$ solves,

$$\min_m \sum_{i=1}^{n} (Y_i - m)^2$$

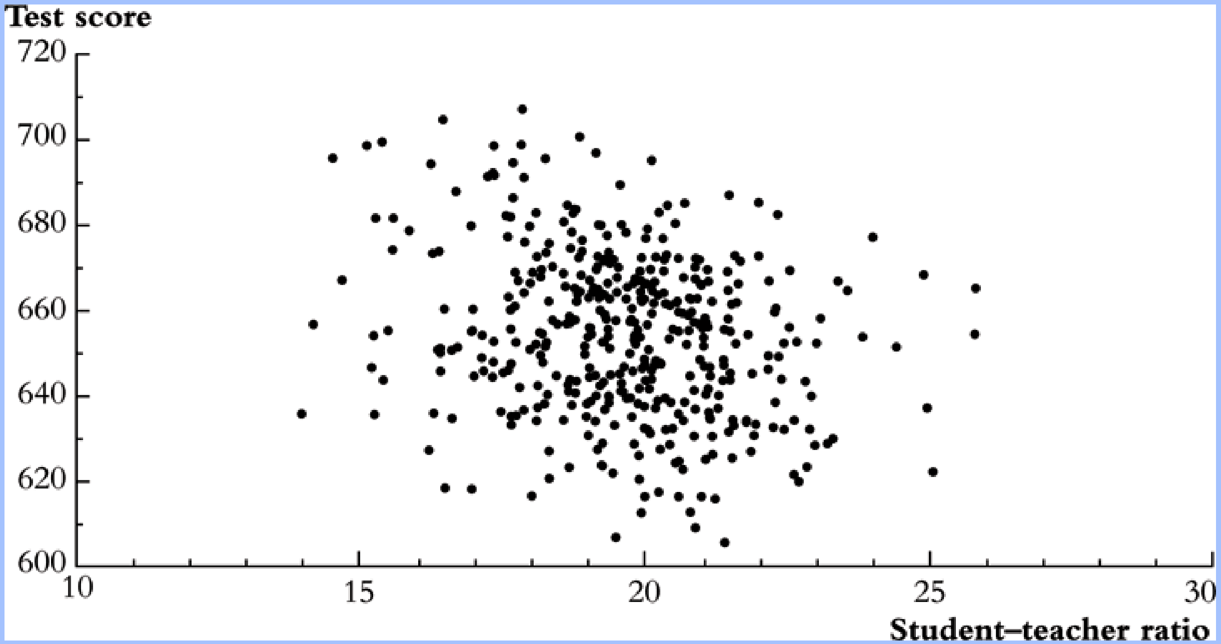
By analogy, **we will focus on the least squares (“*ordinary least squares*” or “*OLS*”) estimator of the unknown parameters β_0 and β_1.** The OLS estimator solves,

$$\min_{b_0, b_1} \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$$

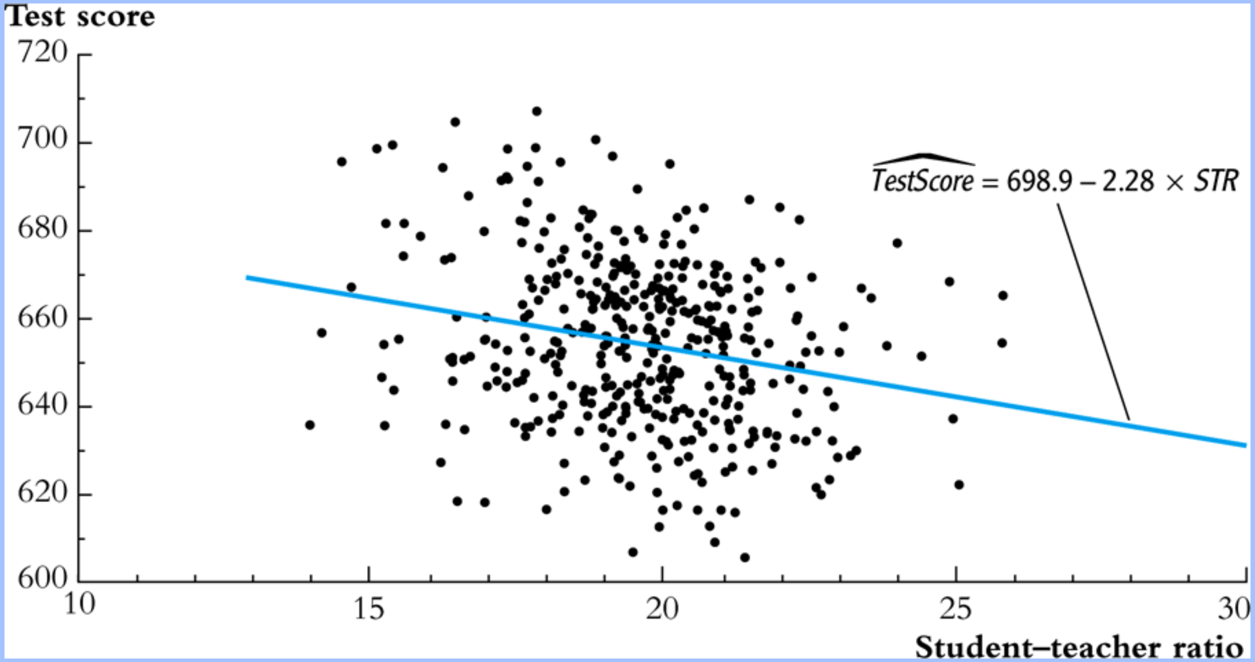
Mechanics of OLS

The population regression line: *Test Score* = $\beta_0 + \beta_1 STR$

$$\beta_1 = \frac{\Delta\text{Test score}}{\Delta STR} = ??$$



Application to the California *Test Score – Class Size* data

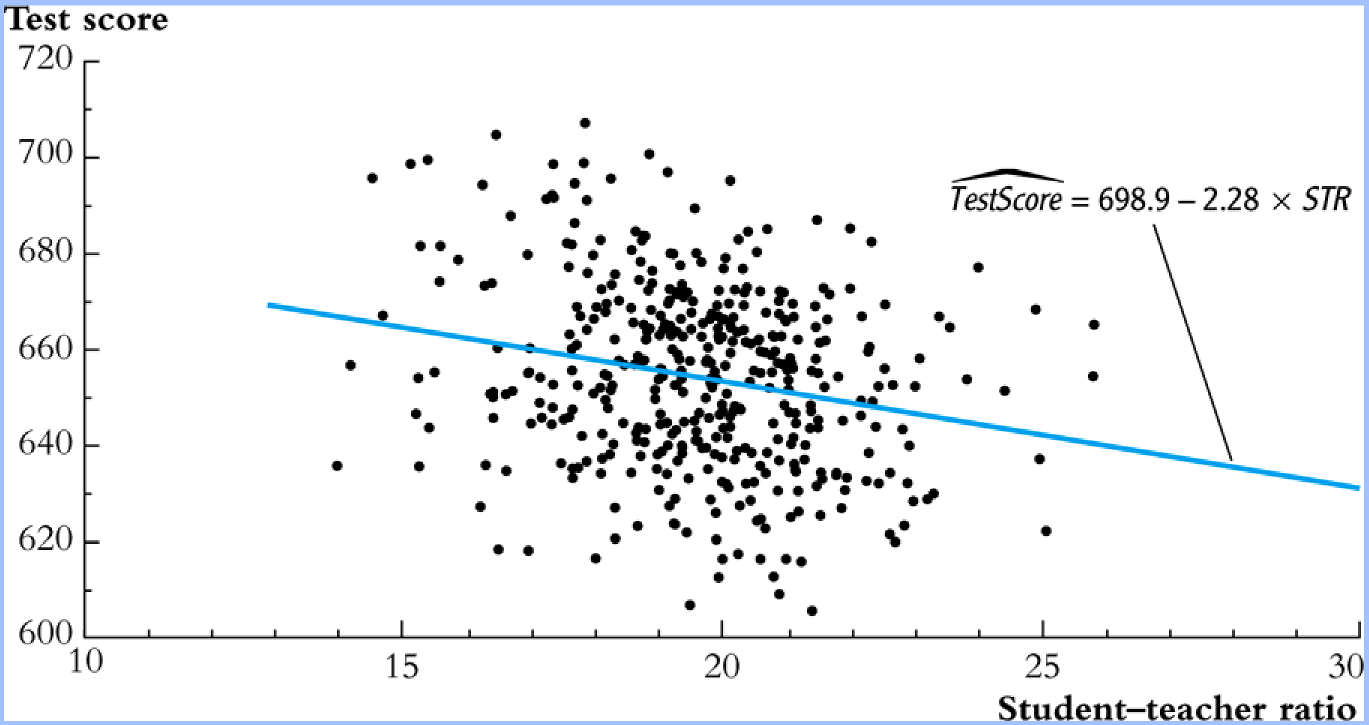


- Estimated slope = $\hat{\beta}_1$ = – 2.28
- Estimated intercept = $\hat{\beta}_0$= 698.9
- Estimated regression line: $\widehat{TestScore}$ = 698.9 – 2.28 × *STR*

Interpretation of the estimated slope and intercept

- $\widehat{Test\ Score}$ = 698.9 – 2.28 × *STR*
- Districts with one more student per teacher on average have test scores that are 2.28 points lower.
- That is, $\frac{\Delta \text{Test score}}{\Delta STR} = -2.28$
- The intercept (taken literally) means that, according to this estimated line, districts with zero students per teacher would have a (predicted) test score of 698.9. But this interpretation of the intercept makes no sense – it extrapolates the line outside the range of the data – here, the intercept is not economically meaningful.

Predicted values & residuals:



One of the districts in the data set is Antelope, CA, for which *STR* = 19.33 and *Test Score* = 657.8

predicted value: $\hat{Y}_{Antelope}$ = 698.9 – 2.28×19.33 = 654.8

residual: $\hat{u}_{Antelope}$ = 657.8 – 654.8 = 3.0

OLS regression: STATA output

```
regress testscr str, robust
Regression with robust standard errors                 Number of obs =     420
                                                       F(  1,   418) =   19.26
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0512
                                                       Root MSE      =  18.581
-------------------------------------------------------------------------
        |               Robust
testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------+----------------------------------------------------------------
    str |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
  _cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
-------------------------------------------------------------------------
```

$$\textit{Test Score} = 698.9 - 2.28 \times \textit{STR}$$

Measures of Fit

Two regression statistics provide complementary measures of how well the regression line “fits” or explains the data:

- The ***regression R^2*** measures the fraction of the variance of *Y* that is explained by *X*; it is unitless and ranges between zero (no fit) and one (perfect fit)
- The ***standard error of the regression* (*SER*)** measures the magnitude of a typical regression residual in the units of *Y*.

The *regression R*² is the fraction of the sample variance of Y_i "explained" by the regression.

$Y_i = \hat{Y}_i + \hat{u}_i$ = OLS prediction + OLS residual

→ sample var (Y) = sample var($\hat{Y}_i$) + sample var($\hat{u}_i$) (*why?*)

→ total sum of squares = "explained" SS + "residual" SS

*Definition of R*²: $R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$

- $R^2 = 0$ means $ESS = 0$
- $R^2 = 1$ means $ESS = TSS$
- $0 \le R^2 \le 1$
- For regression with a single *X*, R^2 = the square of the correlation coefficient between *X* and *Y*

The Standard Error of the Regression (SER)

The *SER* measures the spread of the distribution of *u*. The *SER* is (almost) the sample standard deviation of the OLS residuals:

$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(\hat{u}_i - \bar{\hat{u}})^2}$$

$$= \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}$$

SER = $\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}$

The *SER*:

has the units of *u*, which are the units of *Y*

measures the average “size” of the OLS residual (the average “mistake” made by the OLS regression line)

The ***root mean squared error*** (*RMSE*) is closely related to the *SER*:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2}$$

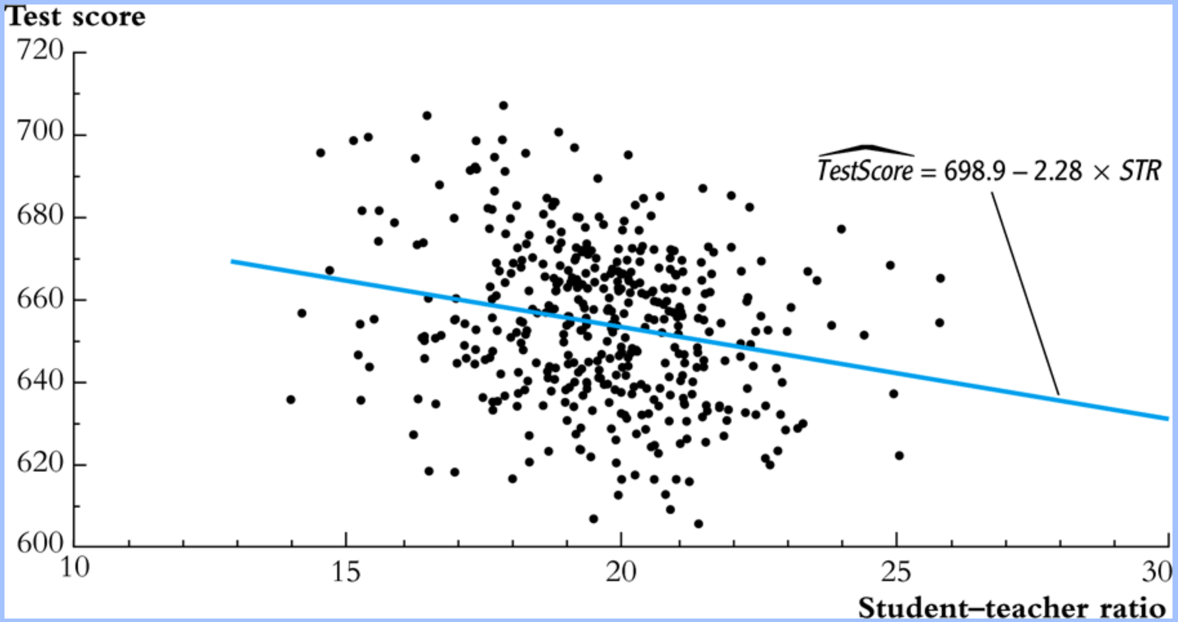
This measures the same thing as the *SER* – the minor difference is division by 1/*n* instead of 1/(*n*–2).

***Technical note*: why divide by *n*–2 instead of *n*–1?**

$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}$$

- Division by *n*–2 is a “degrees of freedom” correction – just like division by *n*–1 in , except that for the *SER*, two parameters have been estimated (β_0 and β_1, by $\hat{\beta}_0$ and $\hat{\beta}_1$), whereas in s_Y^2 only one has been estimated (μ_Y, by $\bar{Y}$).
- When *n* is large, it doesn’t matter whether *n*, *n*–1, or *n*–2 are used – although the conventional formula uses *n*–2 when there is a single regressor.

Example of the *R*² and the *SER*



Test Score = 698.9 – 2.28×*STR*, ***R*² = .05, *SER* = 18.6**

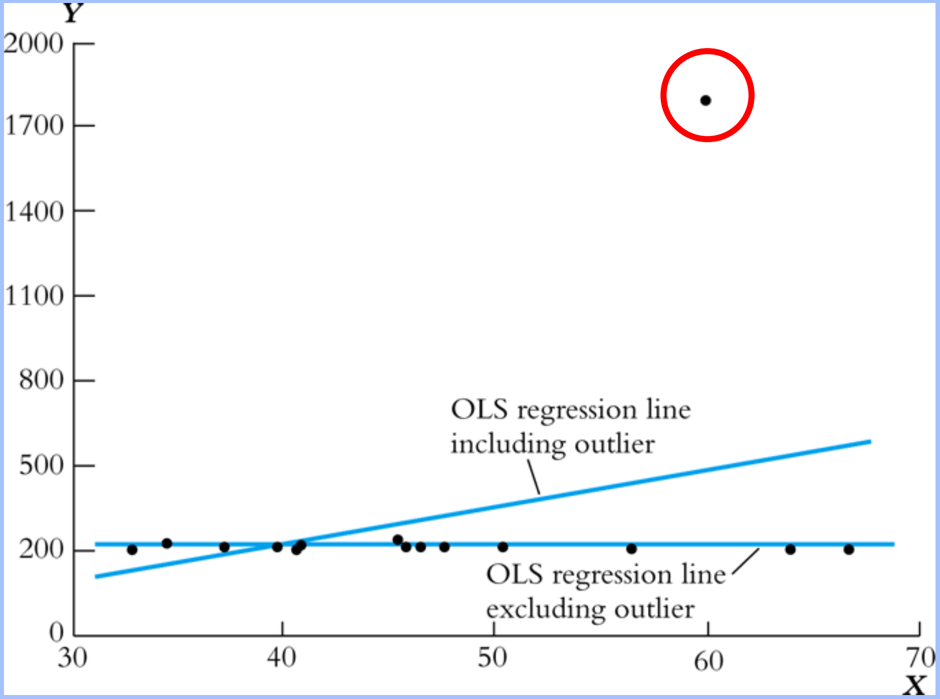
STR explains only a small fraction of the variation in test scores. Does this make sense? Does this mean the STR is unimportant in a policy sense?

The Least Squares Assumptions

$$Y_i = \beta_0 + \beta_1 X_i + u_i,\ i = 1,\ldots,\ n$$

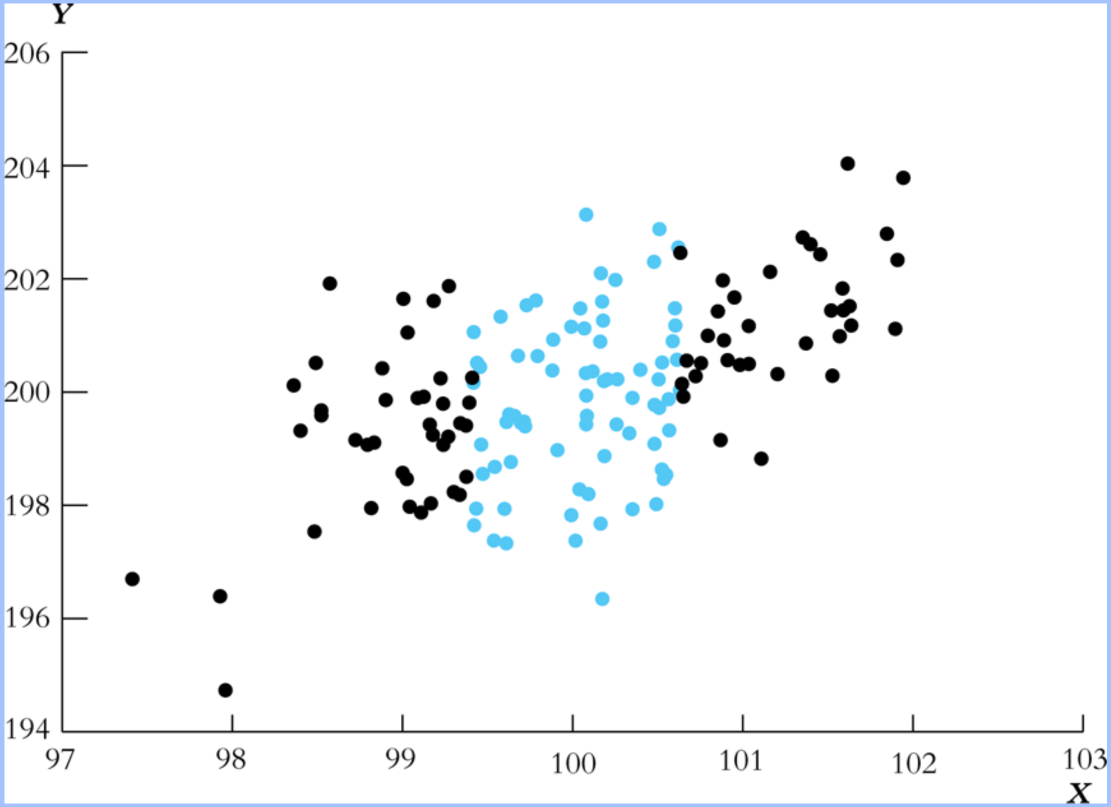
1. The conditional distribution of *u* given *X* has mean zero, that is, $E(u|X = x) = 0$.
   - *This implies that* $\hat{\beta}_1$ *is unbiased*
2. (X_i, Y_i), $i = 1,\ldots,n$, are i.i.d.
   - *This is true if* (*X*, *Y*) *are collected by simple random sampling*
   - *This delivers the sampling distribution of* $\hat{\beta}_0$ *and* $\hat{\beta}_1$
3. Large outliers in *X* and/or *Y* are rare.
   - *Technically, X and Y have finite fourth moments*
   - *Outliers can result in meaningless values of* $\hat{\beta}_1$

OLS can be sensitive to an outlier:



- *Is the lone point an outlier in X or Y?*
- In practice, outliers are often data glitches (coding or recording problems). Sometimes they are observations that really shouldn’t be in your data set. Plot your data!

The larger the variance of X, the smaller the variance of $\hat{\beta}_1$



The number of black and blue dots is the same. Using which would you get a more accurate regression line?

Regression when *X* is Binary

Sometimes a regressor is binary:

- X = 1 if small class size, = 0 if not
- X = 1 if female, = 0 if male
- X = 1 if treated (experimental drug), = 0 if not

Binary regressors are sometimes called "dummy" variables.

So far, β_1 has been called a "slope," but that doesn't make sense if X is binary.

How do we interpret regression with a binary regressor?

Interpreting regressions with a binary regressor

$Y_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X_i + u_i$, where X is binary ($X_i = 0$ or 1):

When $X_i = 0$, $Y_i = \beta_0 + u_i$

- the mean of Y_i is β_0
- that is, $E(Y_i|X_i=0) = \beta_0$

When $X_i = 1$, $Y_i = \beta_0 + \beta_1 + u_i$

- the mean of Y_i is $\beta_0 + \beta_1$
- that is, $E(Y_i|X_i=1) = \beta_0 + \beta_1$

so:

$$\begin{aligned}\beta_1 &= E(Y_i|X_i=1) - E(Y_i|X_i=0) \\ &= \text{population difference in group means}\end{aligned}$$

Example: Let $D_i = \begin{cases} 1 \text{ if } STR_i \le 20 \\ 0 \text{ if } STR_i > 20 \end{cases}$

OLS regression: $\widehat{TestScore} = \mathbf{650.0} + \mathbf{7.4} \times D$

$(1.3) \quad (\mathbf{1.8})$

Tabulation of group means:

Class Size	Average score ($\bar{Y}$)	Std. dev. (s_Y)	N
Small ($STR > 20$)	657.4	19.4	238
Large ($STR \ge 20$)	**650.0**	17.9	182

Difference in means: $\bar{Y}_{\text{small}} - \bar{Y}_{\text{large}}$ = 657.4 – 650.0 = **7.4**

Standard error $SE = \sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}$ **=** $\sqrt{\frac{19.4^2}{238} + \frac{17.9^2}{182}}$ = **1.8**

Omitted Variable Bias

The error *u* arises because of factors, or variables, that influence *Y* but are not included in the regression function. There are always omitted variables.

Sometimes, the omission of those variables can lead to bias in the OLS estimator.

Omitted variable bias, ctd.

The bias in the OLS estimator that occurs as a result of an omitted factor, or variable, is called ***omitted variable*** bias. For omitted variable bias to occur, the omitted variable "*Z*" must satisfy two conditions:

The two conditions for omitted variable bias

1. *Z* is a determinant of *Y* (i.e. *Z* is part of *u*); **and**

2. *Z* is correlated with the regressor *X* (*i.e.* corr(*Z*,*X*) ≠ 0)

Both *conditions must hold for the omission of Z to result in omitted variable bias*.

Omitted variable bias, ctd.

In the test score example:

1. English language ability (whether the student has English as a second language) plausibly affects standardized test scores: *Z* is a determinant of *Y*.
2. Immigrant communities tend to be less affluent and thus have smaller school budgets and higher *STR*: *Z* is correlated with *X*.

Accordingly, $\hat{\beta}_1$ is biased. What is the direction of this bias?

- *What does common sense suggest?*
- If common sense fails you, there is a formula…

Omitted variable bias, ctd.

A formula for omitted variable bias: recall the equation,

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\left(\frac{n-1}{n}\right)s_X^2}$$

where $v_i = (X_i - \bar{X})u_i \approx (X_i - \mu_X)u_i$. Under Least Squares Assumption #1,

$$E[(X_i - \mu_X)u_i] = \mathrm{cov}(X_i,u_i) = 0.$$

But what if $E[(X_i - \mu_X)u_i] = \mathrm{cov}(X_i,u_i) = \sigma_{Xu} \neq 0$?

Omitted variable bias, ctd.

Under LSA #2 and #3 (that is, even if LSA #1 is not true),

$$\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\xrightarrow{p} \frac{\sigma_{Xu}}{\sigma_X^2}$$

$$= \left(\frac{\sigma_u}{\sigma_X}\right) \times \left(\frac{\sigma_{Xu}}{\sigma_X \sigma_u}\right) = \left(\frac{\sigma_u}{\sigma_X}\right)\rho_{Xu}\ ,$$

where ρ_{Xu} = corr(X,u). If assumption #1 is correct, then ρ_{Xu} = 0, but if not we have….

The omitted variable bias formula:

$$\boxed{\hat{\beta}_1 \xrightarrow{p} \beta_1 + \left(\frac{\sigma_u}{\sigma_X}\right)\rho_{Xu}}$$

- If an omitted variable *Z* is ***both***:

1. a determinant of *Y* (that is, it is contained in *u*); ***and***
2. correlated with *X*,

   then $\rho_{Xu} \neq 0$ and the OLS estimator $\hat{\beta}_1$ is biased and is not consistent.

- For example, districts with few ESL students (1) do better on standardized tests and (2) have smaller classes (bigger budgets), so ignoring the effect of having many ESL students factor would result in overstating the class size effect. *Is this is actually going on in the CA data*?

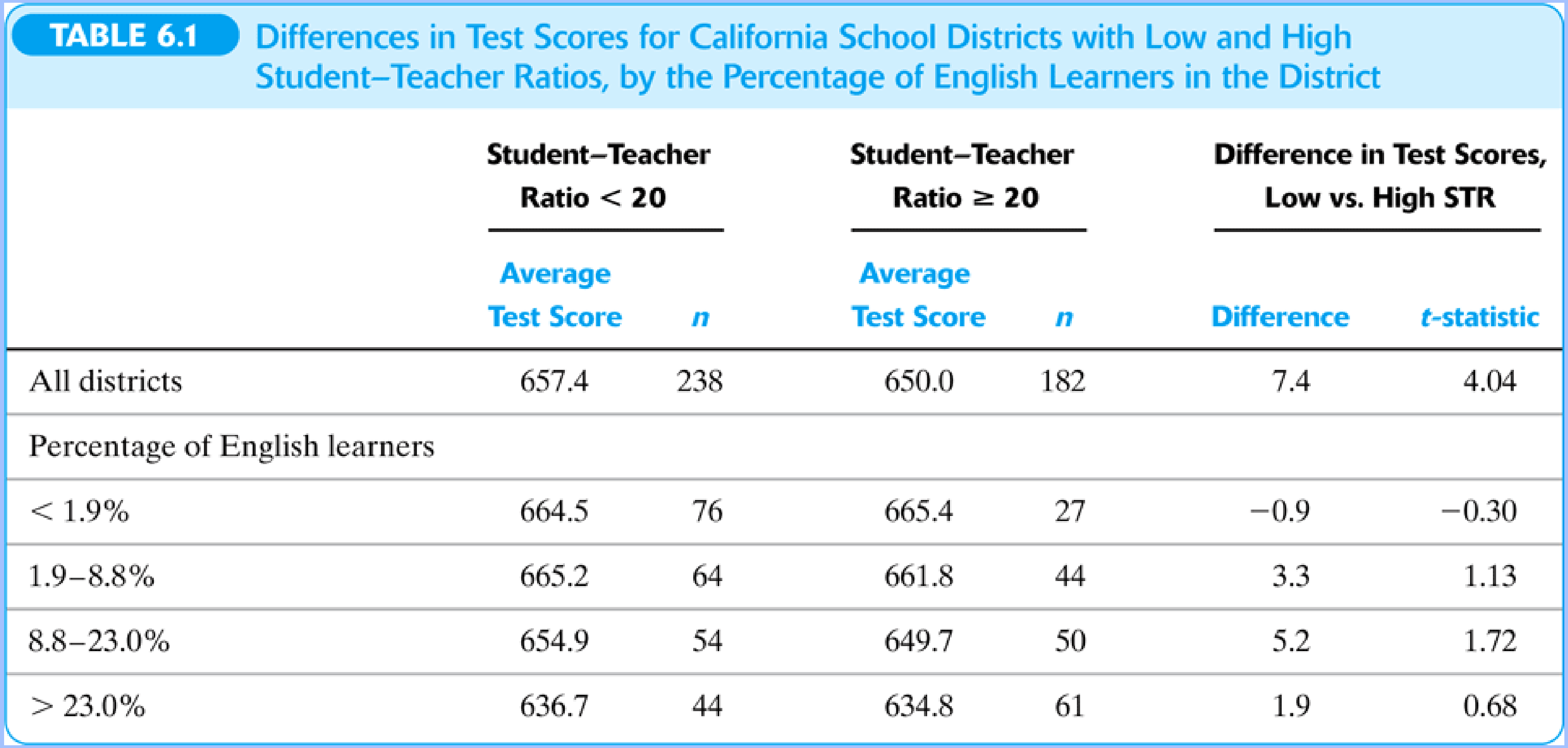


TABLE 6.1 Differences in Test Scores for California School Districts with Low and High Student–Teacher Ratios, by the Percentage of English Learners in the District

	Student–Teacher Ratio < 20		Student–Teacher Ratio ≥ 20		Difference in Test Scores, Low vs. High STR	
	Average Test Score	*n*	Average Test Score	*n*	Difference	*t*-statistic
All districts	657.4	238	650.0	182	7.4	4.04
Percentage of English learners						
< 1.9%	664.5	76	665.4	27	−0.9	−0.30
1.9–8.8%	665.2	64	661.8	44	3.3	1.13
8.8–23.0%	654.9	54	649.7	50	5.2	1.72
> 23.0%	636.7	44	634.8	61	1.9	0.68

- Districts with fewer English Learners have higher test scores
- Districts with lower percent *EL* (*PctEL*) have smaller classes
- Among districts with comparable *PctEL*, the effect of class size is small (recall overall “test score gap” = 7.4)

Causality and regression analysis

- The test score/*STR*/fraction English Learners example shows that, if an omitted variable satisfies the two conditions for omitted variable bias, then the OLS estimator in the regression omitting that variable is biased and inconsistent. So, even if *n* is large, $\hat{\beta}_1$ will not be close to β_1.

- This raises a deeper question: how do we define β_1? That is, what precisely do we want to estimate when we run a regression?

What precisely do we want to estimate when we run a regression?

We want to estimate the causal effect on *Y* of a change in *X*.

This is why we are interested in the class size effect. Suppose the school board decided to cut class size by 2 students per class. What would be the effect on test scores? This is a causal question (what is the causal effect on test scores of STR?) so we need to estimate this causal effect. Our aim is the estimation of causal effects using regression methods.

What, precisely, is a causal effect?

- “Causality” is a complex concept!
- We take a practical approach to defining causality:

  A causal effect is defined to be the effect measured in an ideal randomized controlled experiment.

Ideal Randomized Controlled Experiment

- *Ideal*: subjects all follow the treatment protocol – perfect compliance, no errors in reporting, etc.!
- *Randomized*: subjects from the population of interest are randomly assigned to a treatment or control group (so there are no confounding factors)
- *Controlled*: having a control group permits measuring the differential effect of the treatment
- *Experiment*: the treatment is assigned as part of the experiment: the subjects have no choice, so there is no "reverse causality" in which subjects choose the treatment they think will work best.

Back to class size:

Imagine an ideal randomized controlled experiment for measuring the effect on *Test Score* of reducing *STR*…

- In that experiment, students would be randomly assigned to classes, which would have different sizes.

- Because they are randomly assigned, all student characteristics (and thus u_i) would be distributed independently of STR_i.

- Thus, $E(u_i|STR_i) = 0$ – that is, LSA #1 holds in a randomized controlled experiment.

How does our observational data differ from this ideal?

- The treatment is not randomly assigned
- Consider *PctEL* – percent English learners – in the district. It plausibly satisfies the two criteria for omitted variable bias: $Z = PctEL$ is:
  1. a determinant of Y; ***and***
  2. correlated with the regressor X.
- Thus, the "control" and "treatment" groups differ in a systematic way, so corr(STR,$PctEL$) $\neq 0$

Return to omitted variable bias

Three ways to overcome omitted variable bias

1. Run a randomized controlled experiment in which treatment (*STR*) is randomly assigned: then *PctEL* is still a determinant of *TestScore*, but *PctEL* is uncorrelated with *STR*. (*This solution to OV bias is rarely feasible.*)

2. Adopt the "cross tabulation" approach, with finer gradations of *STR* and *PctEL* – within each group, all classes have the same *PctEL*, so we control for *PctEL* (*But soon you will run out of data, and what about other determinants like family income and parental education*?)

3. Use a regression in which the omitted variable (*PctEL*) is no longer omitted: include *PctEL* as an additional regressor in a multiple regression.

The Population Multiple Regression Model

- Consider the case of two regressors:

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

- Y is the *dependent variable*
- X_1, X_2 are the two *independent variables* (*regressors*)
- (Y_i, X_{1i}, X_{2i}) denote the i^{th} observation on Y, X_1, and X_2.
- β_0 = unknown population intercept
- β_1 = effect on Y of a change in X_1, holding X_2 constant
- β_2 = effect on Y of a change in X_2, holding X_1 constant
- u_i = the regression error (omitted factors)

Interpretation of coefficients in multiple regression

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

Consider changing X_1 by ΔX_1 while holding X_2 constant:

Population regression line ***before*** the change:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Population regression line, ***after*** the change:

$$Y + \Delta Y = \beta_0 + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2$$

Before: $Y = \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2 X_2$

After: $Y + \Delta Y = \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2 X_2$

Difference: $\Delta Y = \beta_1 \Delta X_1$

So:

$\beta_1 = \dfrac{\Delta Y}{\Delta X_1}$, **holding *$X_2$* constant**

$\beta_2 = \dfrac{\Delta Y}{\Delta X_2}$, **holding *$X_1$* constant**

β_0 = predicted value of Y when $X_1 = X_2 = 0$.

The OLS Estimator in Multiple Regression

- With two regressors, the OLS estimator solves:

$$\min_{b_0,b_1,b_2} \sum_{i=1}^{n}[Y_i - (b_0 + b_1 X_{1i} + b_2 X_{2i})]^2$$

- The OLS estimator minimizes the average squared difference between the actual values of *Y*$_i$ and the prediction (predicted value) based on the estimated line.
- This minimization problem is solved using calculus
- **This yields the OLS estimators of β_0 and β_1 .**

Example: the California test score data

Regression of *TestScore* against *STR*:

$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$

Now include percent English Learners in the district (*PctEL*):

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65 PctEL$$

- What happens to the coefficient on *STR*?

Multiple regression in 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
------------------------------------------------------------------------------
```

$$Test\ Score = 686.0 - 1.10 \times STR - 0.65 PctEL$$

Measures of Fit for Multiple Regression

Actual = predicted + residual: $Y_i = \hat{Y}_i + \hat{u}_i$

SER = std. deviation of $\hat{u}_i$ (with d.f. correction)

RMSE = std. deviation of $\hat{u}_i$ (without d.f. correction)

R^2 = fraction of variance of *Y* explained by *X*

$\bar{R}^2$ = "adjusted R^2" = R^2 with a degrees-of-freedom correction that adjusts for estimation uncertainty; $\bar{R}^2 < R^2$

SER and RMSE

As in regression with a single regressor, the *SER* and the *RMSE* are measures of the spread of the *Y*s around the regression line:

$$SER = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2}$$

R^2 and $\bar{R}^2$ (adjusted R^2)

The R^2 is the fraction of the variance explained – same definition as in regression with a single regressor:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS},$$

where $ESS = \sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2$, $SSR = \sum_{i=1}^{n}\hat{u}_i^2$, $TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

- The R^2 always increases when you add another regressor (*why?)* – a bit of a problem for a measure of “fit”

R*² and $\bar{R}^2$ *ctd.

The $\bar{R}^2$ (the “adjusted *R*²”) corrects this problem by “penalizing” you for including another regressor – the $\bar{R}^2$ does not necessarily increase when you add another regressor.

$$\textit{Adjusted } R^2\text{: } \bar{R}^2 = 1-\left(\frac{n-1}{n-k-1}\right)\frac{SSR}{TSS}$$

Note that $\bar{R}^2$ **<** *R*², however if *n* is large the two will be very close.

R^2 and $\bar{R}^2$ *ctd.*

The R^2 and adjusted R^2 tell you whether the regressors are good at predicting the values of the dependent variable. If the R^2 is nearly 1, then the regressors produce good predictions. If the R^2 is nearly 0, the opposite is true.

The R^2 and adjusted R^2 do NOT tell you whether:

- 1. An included variable is statistically significant,
- 2. The regressors are a true cause of the movements in the dependent variable,
- 3. There is omitted variable bias, or
- 4. You have chosen the most appropriate set of regressors

Measures of fit, ctd.

Test score example:

(1) $\widehat{TestScore} = 698.9 - 2.28 \times STR,$

$R^2 = .05,\ SER = 18.6$

(2) $\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65 PctEL,$

$R^2 = .426,\ \bar{R}^2 = .424,\ SER = 14.5$

- *What – precisely – does this tell you about the fit of regression (2) compared with regression (1)?*
- *Why are the R^2 and the $\bar{R}^2$ so close in (2)?*

The Least Squares Assumptions for Multiple Regression

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

1. The conditional distribution of u given the X's has mean zero, that is, $E(u_i | X_{1i} = x_1,\ldots, X_{ki} = x_k) = 0$.

2. $(X_{1i},\ldots,X_{ki},Y_i)$, $i = 1,\ldots,n$, are i.i.d.

3. Large outliers are unlikely: $X_1,\ldots, X_k$, and Y have four moments: $E(X_{1i}^4) < \infty,\ldots, E(X_{ki}^4) < \infty$, $E(Y_i^4) < \infty$.

4. There is no perfect multicollinearity.

Assumption #4: There is no perfect multicollinearity

Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors.

Example: Suppose you accidentally include *STR* twice:

```
regress testscr str str, robust
Regression with robust standard errors                 Number of obs =     420
                                                       F(  1,   418) =   19.26
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0512
                                                       Root MSE      =  18.581
-------------------------------------------------------------------------
        |               Robust
testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------+----------------------------------------------------------------
    str |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
    str |  (dropped)
  _cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
-------------------------------------------------------------------------
```

Multicollinearity, Perfect and Imperfect

Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors.

<u>Some more examples of perfect multicollinearity</u>

1. The example from before: you include *STR* twice,
2. Regress *TestScore* on a constant, *D*, and *B*, where: D_i = 1 if *STR* ≤ 20, = 0 otherwise; B_i = 1 if *STR* >20, = 0 otherwise, so $B_i = 1 – D_i$ and there is perfect multicollinearity.
3. Would there be perfect multicollinearity if the intercept (constant) were excluded from this regression?

The dummy variable trap

Suppose you have a set of multiple binary (dummy) variables, which are mutually exclusive and exhaustive – that is, there are multiple categories and every observation falls in one and only one category (Freshmen, Sophomores, Juniors, Seniors, Other). If you include all these dummy variables *and* a constant, you will have perfect multicollinearity – this is sometimes called ***the dummy variable trap***.

- *Why is there perfect multicollinearity here*?
- *Solutions to the dummy variable trap*:
  1. Omit one of the groups (e.g. Senior), or
  2. Omit the intercept
- *What are the implications of (1) or (2) for the interpretation of the coefficients?*

Perfect multicollinearity, ctd.

- Perfect multicollinearity usually reflects a mistake in the definitions of the regressors, or an oddity in the data

- If you have perfect multicollinearity, your statistical software will let you know – either by crashing or giving an error message or by "dropping" one of the variables arbitrarily

- The solution to perfect multicollinearity is to modify your list of regressors so that you no longer have perfect multicollinearity.

Imperfect multicollinearity

Imperfect and perfect multicollinearity are quite different despite the similarity of the names.

Imperfect multicollinearity occurs when two or more regressors are very highly correlated.

- Why the term "multicollinearity"? If two regressors are very highly correlated, then their scatterplot will pretty much look like a straight line – they are "co-linear" – but unless the correlation is exactly ±1, that collinearity is imperfect.

Imperfect multicollinearity, ctd.

Imperfect multicollinearity implies that one or more of the regression coefficients will be imprecisely estimated.

- The idea: the coefficient on X_1 is the effect of X_1 holding X_2 constant; but if X_1 and X_2 are highly correlated, there is very little variation in X_1 once X_2 is held constant – so the data don't contain much information about what happens when X_1 changes but X_2 doesn't. If so, the variance of the OLS estimator of the coefficient on X_1 will be large.
- Imperfect multicollinearity (correctly) results in large standard errors for one or more of the OLS coefficients.

Panel Data: What and Why

A ***panel dataset*** contains observations on multiple entities (individuals, states, companies…), where each entity is observed at two or more points in time.

Hypothetical examples:

- Data on 420 California school districts in 1999 *and again* in 2000, for 840 observations total.
- Data on 50 U.S. states, each state is observed in 3 years, for a total of 150 observations.
- Data on 1000 individuals, in four different months, for 4000 observations total.

Notation for panel data

A double subscript distinguishes entities (states) and time periods (years)

i = entity (state), n = number of entities,

so $i = 1,\ldots,n$

t = time period (year), T = number of time periods

so $t = 1,\ldots,T$

Data: Suppose we have 1 regressor. The data are:

$$(X_{it}, Y_{it}),\ i = 1,\ldots,n,\ t = 1,\ldots,T$$

Panel data notation, ctd.

Panel data with *k* regressors:

$$(X_{1it}, X_{2it},\ldots,X_{kit}, Y_{it}),\ i = 1,\ldots,n,\ t = 1,\ldots,T$$

n = number of entities (states)
T = number of time periods (years)

Some jargon…

- Another term for panel data is ***longitudinal data***
- ***balanced panel***: no missing observations, that is, all variables are observed for all entities (states) and all time periods (years)

Why are panel data useful?

With panel data we can control for factors that:

- Vary across entities but do not vary over time
- Could cause omitted variable bias if they are omitted
- Are unobserved or unmeasured – and therefore cannot be included in the regression using multiple regression

Here's the key idea:

If an omitted variable does not change over time, then any *changes* in *Y* over time cannot be caused by the omitted variable.

Example of a panel data set: Traffic deaths and alcohol taxes

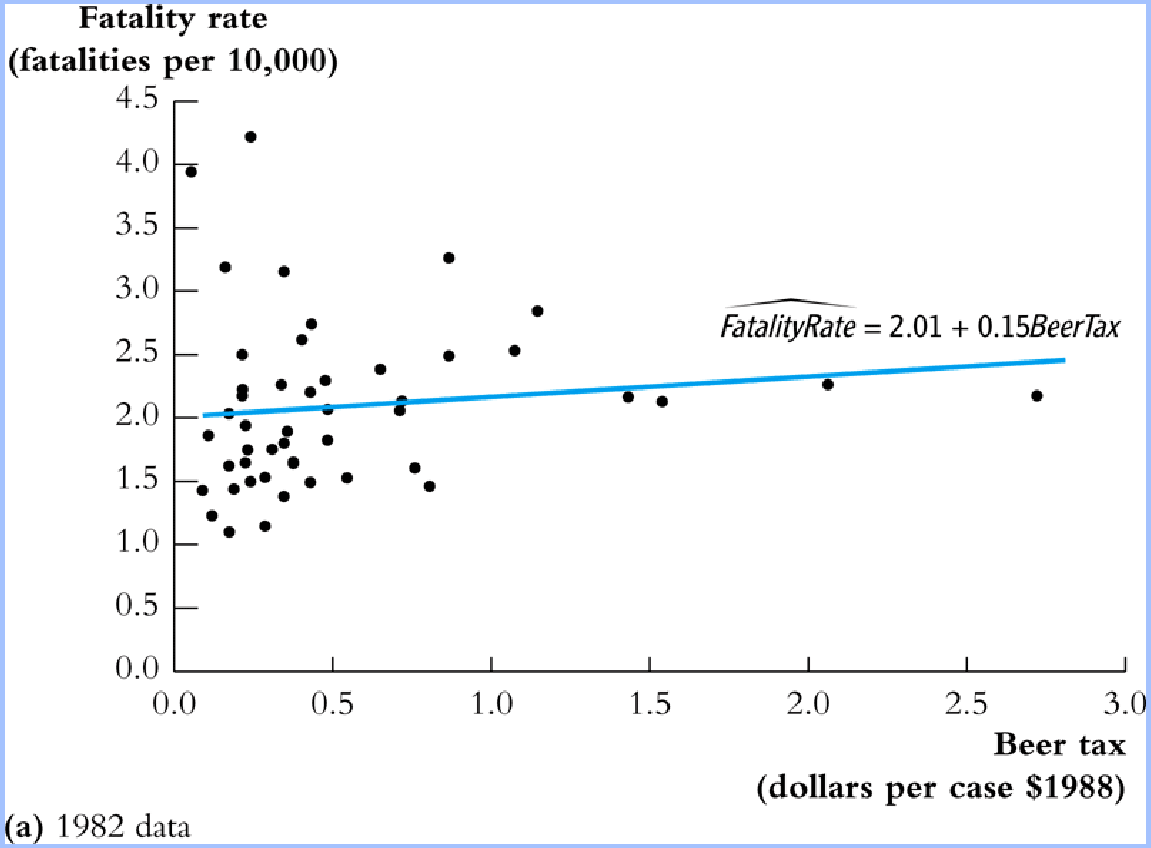
Observational unit: a year in a U.S. state

- 48 U.S. states, so n = # of entities = 48
- 7 years (1982,…, 1988), so T = # of time periods = 7
- Balanced panel, so total # observations = 7×48 = 336

Variables:

- Traffic fatality rate (# traffic deaths in that state in that year, per 10,000 state residents)
- Tax on a case of beer
- Other (legal driving age, drunk driving laws, etc.)

U.S. traffic death data for 1982:



(a) 1982 data

Higher alcohol taxes, more traffic deaths?

Why might there be higher *more* traffic deaths in states that have higher alcohol taxes?

Other factors that determine traffic fatality rate:

- Quality (age) of automobiles
- Quality of roads
- “Culture” around drinking and driving
- Density of cars on the road

These omitted factors could cause omitted variable bias.

Example #1: traffic density. Suppose:

I. High traffic density means more traffic deaths

II. (Western) states with lower traffic density have lower alcohol taxes

- Then the two conditions for omitted variable bias are satisfied. Specifically, “high taxes” could reflect “high traffic density” (so the OLS coefficient would be biased positively – high taxes, more deaths)
- Panel data lets us eliminate omitted variable bias when the omitted variables are constant over time within a given state.

Example #2:Cultural attitudes towards drinking and driving:

(i) arguably are a determinant of traffic deaths; and

(ii) potentially are correlated with the beer tax.

- Then the two conditions for omitted variable bias are satisfied. Specifically, “high taxes” could pick up the effect of “cultural attitudes towards drinking” so the OLS coefficient would be biased
- Panel data lets us eliminate omitted variable bias when the omitted variables are constant over time within a given state.

Panel Data with Two Time Periods

Consider the panel data model,

$$FatalityRate_{it} = \beta_0 + \beta_1 BeerTax_{it} + \beta_2 Z_i + u_{it}$$

Z_i is a factor that does not change over time (density), at least during the years on which we have data.

- Suppose Z_i is not observed, so its omission could result in omitted variable bias.
- The effect of Z_i can be eliminated using $T = 2$ years.

The key idea:

Any *change* in the fatality rate from 1982 to 1988 cannot be caused by Z_i, because Z_i (by assumption) does not change between 1982 and 1988.

The math: consider fatality rates in 1988 and 1982:

$$FatalityRate_{i1988} = \beta_0 + \beta_1 BeerTax_{i1988} + \beta_2 Z_i + u_{i1988}$$
$$FatalityRate_{i1982} = \beta_0 + \beta_1 BeerTax_{i1982} + \beta_2 Z_i + u_{i1982}$$

Suppose $E(u_{it}|BeerTax_{it}, Z_i) = 0$.

Subtracting 1988 – 1982 (that is, calculating the change), eliminates the effect of Z_i…

$$FatalityRate_{i1988} = \beta_0 + \beta_1 BeerTax_{i1988} + \beta_2 Z_i + u_{i1988}$$

$$FatalityRate_{i1982} = \beta_0 + \beta_1 BeerTax_{i1982} + \beta_2 Z_i + u_{i1982}$$

so

$$FatalityRate_{i1988} - FatalityRate_{i1982} = \beta_1(BeerTax_{i1988} - BeerTax_{i1982}) + (u_{i1988} - u_{i1982})$$

- The new error term, $(u_{i1988} - u_{i1982})$, is uncorrelated with either $BeerTax_{i1988}$ or $BeerTax_{i1982}$.
- This “difference” equation can be estimated by OLS, even though Z_i isn’t observed.
- The omitted variable Z_i doesn’t change, so it cannot be a determinant of the *change* in Y
- This differences regression doesn’t have an intercept – it was eliminated by the subtraction step

Example: Traffic deaths and beer taxes

1982 data:

$$\underset{(.15)}{\textit{Fatality Rate}} = \underset{(.15)}{2.01} + \underset{(.13)}{0.15}\textit{BeerTax} \qquad (n = 48)$$

1988 data:

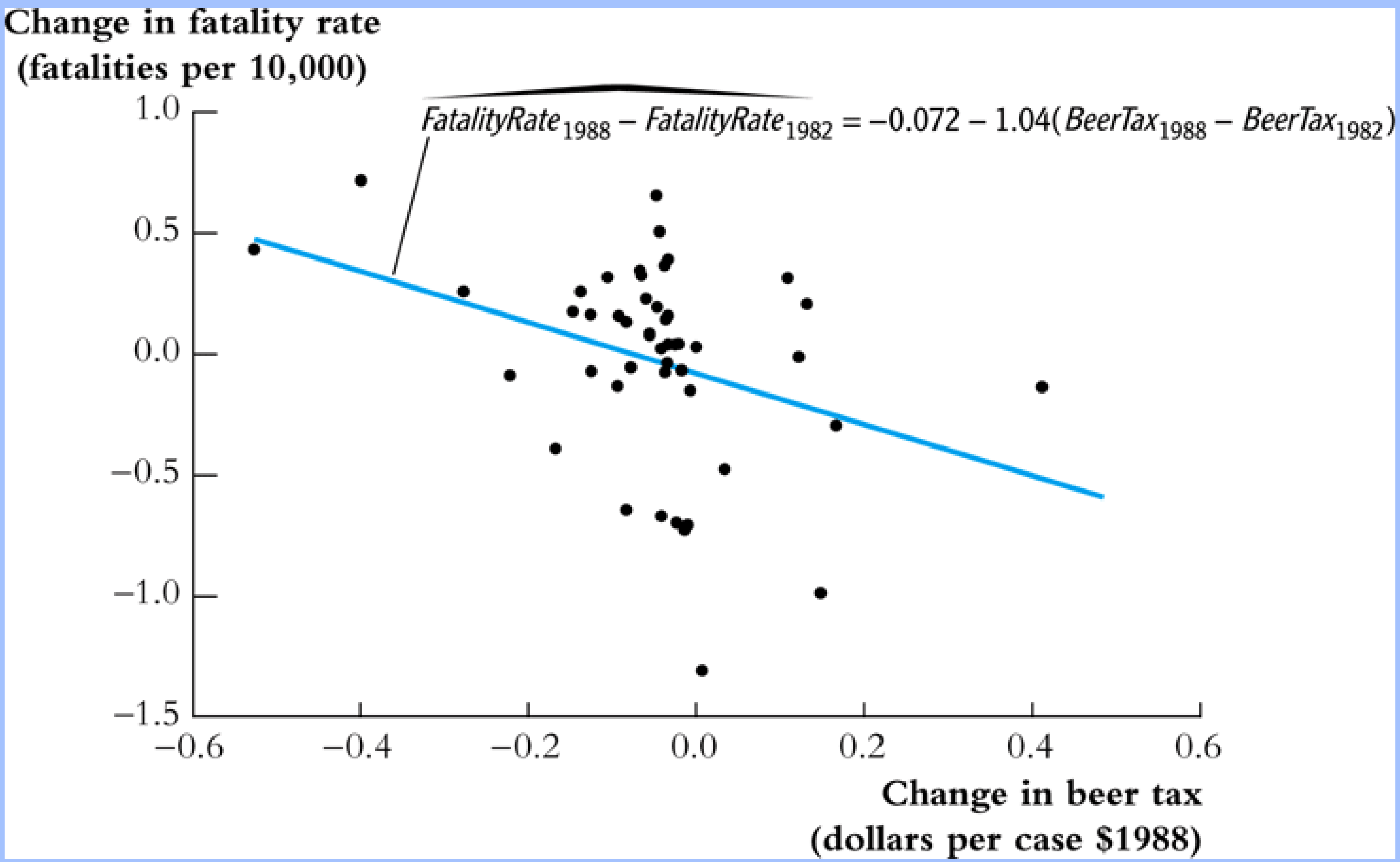
$$\textit{Fatality Rate} = \underset{(.11)}{1.86} + \underset{(.13)}{0.44}\textit{BeerTax} \qquad (n = 48)$$

Difference regression (n = 48)

$$FR_{1988} - FR_{1982} = \underset{(.065)}{-.072} - \underset{(.36)}{1.04}(\textit{BeerTax}_{1988} - \textit{BeerTax}_{1982})$$

An intercept is included in this differences regression allows for the mean change in FR to be nonzero – more on this later…

Δ*FatalityRate* v. Δ*BeerTax*:



Note that the intercept is nearly zero…

Fixed Effects Regression

What if you have more than 2 time periods (T > 2)?

$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it}$, i =1,…,n, T = 1,…,T

We can rewrite this in two useful ways:

1. "n-1 binary regressor" regression model
2. "Fixed Effects" regression model

We first rewrite this in "fixed effects" form. Suppose we have n = 3 states: California, Texas, and Massachusetts.

$$\boldsymbol{Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it}, i = 1,\ldots,n, T = 1,\ldots,T}$$

Population regression for California (that is, *i* = CA):

$$\begin{aligned} Y_{CA,t} &= \beta_0 + \beta_1 X_{CA,t} + \beta_2 Z_{CA} + u_{CA,t} \\ &= (\beta_0 + \beta_2 Z_{CA}) + \beta_1 X_{CA,t} + u_{CA,t} \end{aligned}$$

Or

$$Y_{CA,t} = \alpha_{CA} + \beta_1 X_{CA,t} + u_{CA,t}$$

- $\alpha_{CA} = \beta_0 + \beta_2 Z_{CA}$ doesn’t change over time
- α_{CA} is the intercept for CA, and β_1 is the slope
- The intercept is unique to CA, but the slope is the same in all the states: parallel lines.

For TX:

$$\begin{aligned} Y_{TX,t} &= \beta_0 + \beta_1 X_{TX,t} + \beta_2 Z_{TX} + u_{TX,t} \\ &= (\beta_0 + \beta_2 Z_{TX}) + \beta_1 X_{TX,t} + u_{TX,t} \end{aligned}$$

or

$$Y_{TX,t} = \alpha_{TX} + \beta_1 X_{TX,t} + u_{TX,t}, \text{ where } \alpha_{TX} = \beta_0 + \beta_2 Z_{TX}$$

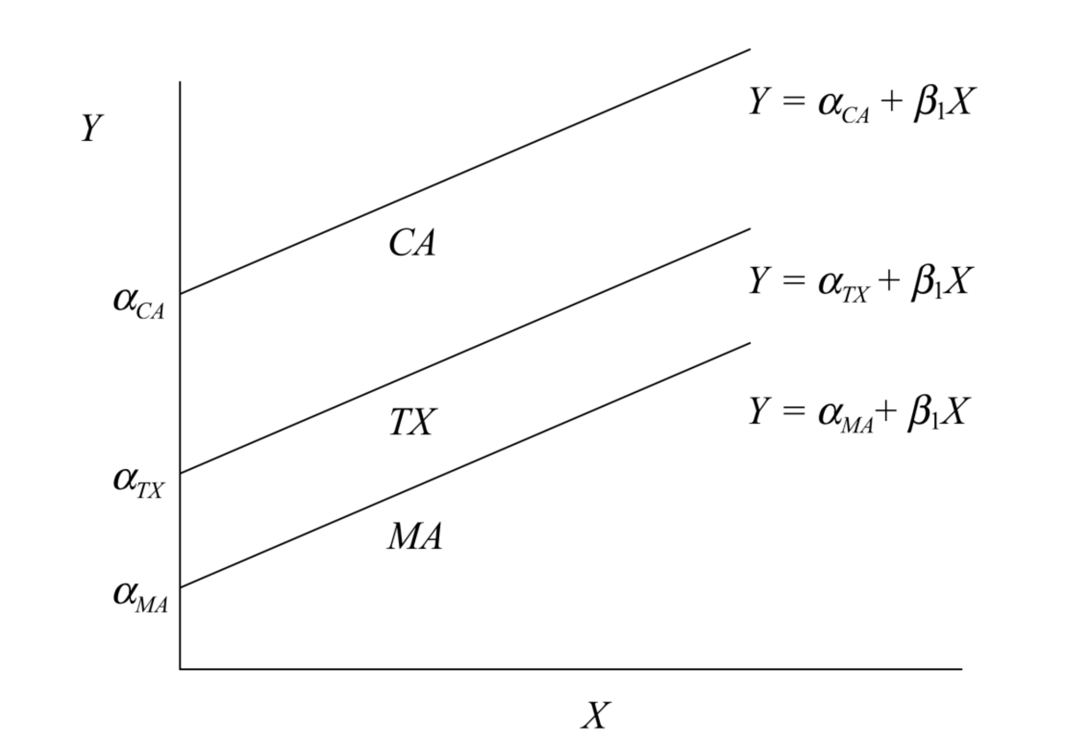
Collecting the lines for all three states:

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

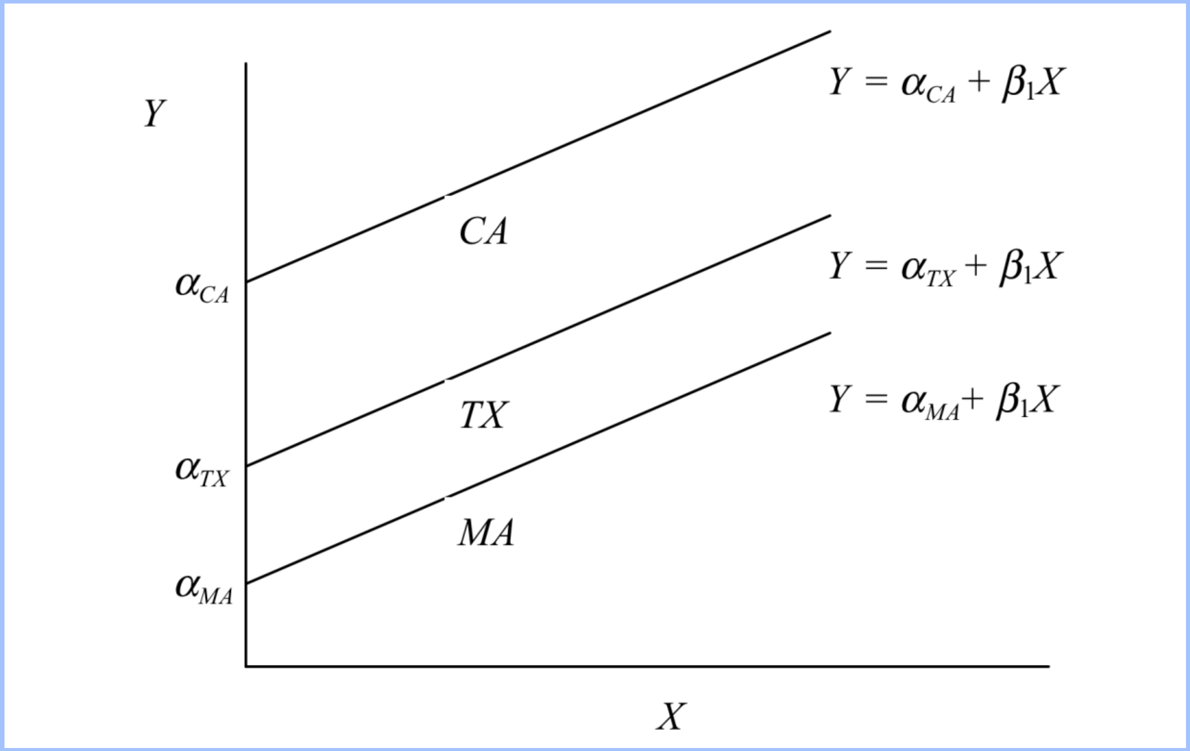
or

$$Y_{it} = \alpha_i + \beta_1 X_{it} + u_{it}, \; i = \text{CA, TX, MA}, \; T = 1,\ldots,T$$

The regression lines for each state in a picture


Y
α_{CA}
α_{TX}
α_{MA}
CA
TX
MA
$Y = \alpha_{CA} + \beta_1 X$
$Y = \alpha_{TX} + \beta_1 X$
$Y = \alpha_{MA} + \beta_1 X$
X

Recall that shifts in the intercept can be represented using binary regressors…



In binary regressor form:

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

- DCA_i = 1 if state is *CA*, = 0 otherwise
- DTX_t = 1 if state is *TX*, = 0 otherwise
- leave out DMA_i (*why*?)

Summary: Two ways to write the fixed effects model

1. "*n*-1 binary regressor" form

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

where $D2_i = \begin{cases} 1 \text{ for } i=2 \text{ (state \#2)} \\ 0 \text{ otherwise} \end{cases}$, etc.

2. "Fixed effects" form:

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

- α_i is called a "state fixed effect" or "state effect" – it is the constant (fixed) effect of being in state *i*

Fixed Effects Regression: Estimation

Three estimation methods:

1. "*n*-1 binary regressors" OLS regression
2. "Entity-demeaned" OLS regression
3. "Changes" specification, without an intercept (only works for *T* = 2)

- These three methods produce identical estimates of the regression coefficients, and identical standard errors.
- We already did the "changes" specification (1988 minus 1982) – but this only works for *T* = 2 years
- Methods #1 and #2 work for general *T*
- Method #1 is only practical when *n* isn't too big

1. “*n*-1 binary regressors” OLS regression

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \ldots + \gamma_n Dn_i + u_{it} \qquad (1)$$

where $D2_i = \begin{cases} 1 \text{ for } i\text{=2 (state \#2)} \\ 0 \text{ otherwise} \end{cases}$ etc.

- First create the binary variables $D2_i,\ldots,Dn_i$
- Then estimate (1) by OLS
- Inference (hypothesis tests, confidence intervals) is as usual (using heteroskedasticity-robust standard errors)
- This is impractical when *n* is very large (for example if *n* = 1000 workers)

2. "Entity-demeaned" OLS regression

The fixed effects regression model:

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

The entity averages satisfy:

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

Deviation from entity averages:

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

Entity-demeaned OLS regression, ctd.

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

or

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

where $\tilde{Y}_{it} = Y_{it} - \frac{1}{T}\sum_{t=1}^{T} Y_{it}$ and $\tilde{X}_{it} = X_{it} - \frac{1}{T}\sum_{t=1}^{T} X_{it}$

- $\tilde{X}_{it}$ and $\tilde{Y}_{it}$ are “entity-demeaned” data
- For *i*=1 and *t* = 1982, $\tilde{Y}_{it}$ is the difference between the fatality rate in Alabama in 1982, and its average value in Alabama averaged over all 7 years.

Entity-demeaned OLS regression, ctd.

$$\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{u}_{it} \qquad (2)$$

where $\tilde{Y}_{it} = Y_{it} - \frac{1}{T}\sum_{t=1}^{T} Y_{it}$, etc.

- First construct the entity-demeaned variables $\tilde{Y}_{it}$ and $\tilde{X}_{it}$
- Then estimate (2) by regressing $\tilde{Y}_{it}$ on $\tilde{X}_{it}$ using OLS
- This is like the "changes" approach, but instead Y_{it} is deviated from the state average instead of Y_{i1}.
- Standard errors need to be computed in a way that accounts for the panel nature of the data set (more later)
- This can be done in a single command in STATA

```
. xtreg vfrall beertax, fe vce(cluster state)

Fixed-effects (within) regression               Number of obs      =       336
Group variable: state                           Number of groups   =        48
R-sq:  within  = 0.0407                         Obs per group: min =         7
       between = 0.1101                                        avg =       7.0
       overall = 0.0934                                        max =         7
                                                F(1,47)            =      5.05
corr(u_i, Xb)  = -0.6885                        Prob > F           =    0.0294

                                 (Std. Err. adjusted for 48 clusters in state)
------------------------------------------------------------------------------
             |               Robust
      vfrall |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     beertax |  -.6558736   .2918556    -2.25   0.029    -1.243011   -.0687358
       _cons |   2.377075   .1497966    15.87   0.000     2.075723    2.678427
------------------------------------------------------------------------------
```

- The panel data command xtreg with the option fe performs fixed effects regression. The reported intercept is arbitrary, and the estimated individual effects are not reported in the default output.
- The fe option means use fixed effects regression
- The vce(cluster state) option tells STATA to use clustered standard errors – more on this later

***Example*, ctd. For *n* = 48, *T* = 7:**

$$\text{Fatality Rate} = \underset{(.29)}{-.66}\textit{BeerTax} + \textit{State fixed effects}$$

- Should you report the intercept?
- How many binary regressors would you include to estimate this using the "binary regressor" method?
- Compare slope, standard error to the estimate for the 1988 v. 1982 "changes" specification ($T = 2$, $n = 48$) (*note that this includes an intercept – return to this below*):

$$FR_{1988} - FR_{1982} = \underset{(.065)}{-.072} - \underset{(.36)}{1.04}(BeerTax_{1988} - BeerTax_{1982})$$

By the way… how much do beer taxes vary?

Beer Taxes in 2005

Source: Federation of Tax Administrators

http://www.taxadmin.org/fta/rate/beer.html

	EXCISE TAX RATES ($ per gallon)	SALES TAXES APPLIED	OTHER TAXES
Alabama	$0.53	Yes	$0.52/gallon local tax
Alaska	1.07	n.a.	$0.35/gallon small breweries
Arizona	0.16	Yes	
Arkansas	0.23	Yes	under 3.2% - $0.16/gallon; $0.008/gallon and 3% off-10% on-premise tax
California	0.20	Yes	
Colorado	0.08	Yes	
Connecticut	0.19	Yes	
Delaware	0.16	n.a.	
Florida	0.48	Yes	2.67¢/12 ounces on-premise retail tax

Georgia	0.48	Yes	$0.53/gallon local tax
Hawaii	0.93	Yes	$0.54/gallon draft beer
Idaho	0.15	Yes	over 4% - $0.45/gallon
Illinois	0.185	Yes	$0.16/gallon in Chicago and $0.06/gallon in Cook County
Indiana	0.115	Yes	
Iowa	0.19	Yes	
Kansas	0.18	--	over 3.2% - {8% off- and 10% on-premise}, under 3.2% - 4.25% sales tax.
Kentucky	0.08	Yes*	9% wholesale tax
Louisiana	0.32	Yes	$0.048/gallon local tax
Maine	0.35	Yes	additional 5% on-premise tax

Maryland	0.09	Yes	$0.2333/gallon in Garrett County
Massachusetts	0.11	Yes*	0.57% on private club sales
Michigan	0.20	Yes	
Minnesota	0.15	--	under 3.2% - $0.077/gallon. 9% sales tax
Mississippi	0.43	Yes	
Missouri	0.06	Yes	
Montana	0.14	n.a.	
Nebraska	0.31	Yes	
Nevada	0.16	Yes	
New Hampshire	0.30	n.a.	
New Jersey	0.12	Yes	
New Mexico	0.41	Yes	

New York	0.11	Yes	$0.12/gallon in New York City
North Carolina	0.53	Yes	$0.48/gallon bulk beer
North Dakota	0.16	--	7% state sales tax, bulk beer $0.08/gal.
Ohio	0.18	Yes	
Oklahoma	0.40	Yes	under 3.2% - $0.36/gallon; 13.5% on-premise
Oregon	0.08	n.a.	
Pennsylvania	0.08	Yes	
Rhode Island	0.10	Yes	$0.04/case wholesale tax
South Carolina	0.77	Yes	
South Dakota	0.28	Yes	
Tennessee	0.14	Yes	17% wholesale tax
Texas	0.19	Yes	over 4% - $0.198/gallon, 14% on-premise and $0.05/drink on airline sales

Utah	0.41	Yes	over 3.2% - sold through state store
Vermont	0.265	no	6% to 8% alcohol - $0.55; 10% on-premise sales tax
Virginia	0.26	Yes	
Washington	0.261	Yes	
West Virginia	0.18	Yes	
Wisconsin	0.06	Yes	
Wyoming	0.02	Yes	
Dist. of Columbia	0.09	Yes	8% off- and 10% on-premise sales tax
U.S. Median	$0.188		

Regression with Time Fixed Effects

An omitted variable might vary over time but not across states:

- Safer cars (air bags, etc.); changes in national laws
- These produce intercepts that change over time
- Let S_t denote the combined effect of variables which changes over time but not states (“safer cars”).
- The resulting population regression model is:

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

Time fixed effects only

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

This model can be recast as having an intercept that varies from one year to the next:

$$\begin{aligned} Y_{i,1982} &= \beta_0 + \beta_1 X_{i,1982} + \beta_3 S_{1982} + u_{i,1982} \\ &= (\beta_0 + \beta_3 S_{1982}) + \beta_1 X_{i,1982} + u_{i,1982} \\ &= \lambda_{1982} + \beta_1 X_{i,1982} + u_{i,1982}, \end{aligned}$$

where $\lambda_{1982} = \beta_0 + \beta_3 S_{1982}$ Similarly,

$$Y_{i,1983} = \lambda_{1983} + \beta_1 X_{i,1983} + u_{i,1983},$$

where $\lambda_{1983} = \beta_0 + \beta 3 S_{1983}$, etc.

Two formulations of regression with time fixed effects

1. "*T*-1 binary regressor" formulation:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \delta_2 B2_t + \ldots \delta_T BT_t + u_{it}$$

where $B2_t = \begin{cases} 1 \text{ when } t\text{=2 (year \#2)} \\ 0 \text{ otherwise} \end{cases}$, etc.

2. "Time effects" formulation:

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

Time fixed effects: estimation methods

1. "*T*-1 binary regressor" OLS regression

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \delta_2 B2_{it} + \ldots \delta_T BT_{it} + u_{it}$$

- Create binary variables *B2*,…,*BT*
- *B2* = 1 if *t* = year #2, = 0 otherwise
- Regress *Y* on *X*, *B2*,…,*BT* using OLS
- Where's *B1*?

2. "Year-demeaned" OLS regression

- Deviate Y_{it}, X_{it} from *year* (not state) averages
- Estimate by OLS using "year-demeaned" data

Estimation with both entity and time fixed effects

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

- When $T = 2$, computing the first difference and including an intercept is equivalent to (gives exactly the same regression as) including entity and time fixed effects.
- When $T > 2$, there are various equivalent ways to incorporate both entity and time fixed effects:
  - entity demeaning & $T - 1$ time indicators (this is done in the following STATA example)
  - time demeaning & $n - 1$ entity indicators
  - $T - 1$ time indicators & $n - 1$ entity indicators
  - entity & time demeaning