Non-Cross-Sectional Data

Econ 2560, Fall 2023

Prof. Josh Abel

(Chapters 10, 15.(1-2,7-8))

Introduction

- Have focused so far on "cross-sectional data"
  - Draw observations $i = 1, 2, ..., n$ from some population
  - Index i does not really mean anything: observations are independent, so ordering of the observations is irrelevant
- When data has a temporal dimension (e.g. $t = 2001, 2002, ..., 2023$) order does matter
- Economists frequently use two types of datasets with temporal dimensions
  1. "Pure" Time Series data
     - E.g. annual per capita income in MA for 2001-2023
  2. Panel data (containig both cross-sectional and temporal dimensions)
     - E.g. annual per capita income in for 50 states for 2001-2023
- These data structures introduce additional nuances to regression that affect both the point estimates (coefficients) and standard errors

Panel Data

- Consider "entities" $i = 1, 2, ..., n$ for time periods $t = 1, 2, ..., T$
- Assume this causal model:

$$Y_{it} = \beta_0 + \beta_1 \cdot X_{1,it} + \beta_2 \cdot X_{2i} + v_{it}$$

- X_{2i} represents **time-invariant heterogeneity** between the entities
  - I.e. differences between the entities that stay fixed over
  - May be difficult to measure, as it captures all such differences between entities
- If you cannot include X_{2i} in regression, $\hat{\beta}_1$ will be a biased estimator of β_1 if $\beta_2 \neq 0$ and $\text{corr}(X_{1,it}, X_{2i}) > 0$.

Hypothetical example

- Data may cover n firms across T years.
- Want to estimate "economies of scale"
  - I.e. how average total cost changes with quantity
- Model as industry-wide economies of scale (i.e. β_1 constant) with firm-specific fixed costs:

$$ATC_{it} = \beta_1 \cdot Q_{it} + \underbrace{\alpha_i}_{\beta_0+\beta_2 \cdot X_{2i}=AFC_i} + v_{it}$$

- Suppose larger firms have lower average fixed costs ($\text{corr}(Q_{it}, \alpha_i) < 0$), and we estimate the following "pooled" regression:

$$ATC_{it} = \hat{\beta}_0 + \hat{\beta}_1 \cdot Q_{it} + \hat{u}_{it}$$

- Is $\hat{\beta}_1$ an upward-, downward-, or un-biased estimator of β_1?

Hypothetical example

- Data may cover n firms across T years.
- Want to estimate "economies of scale"
  - I.e. how average total cost changes with quantity
- Model as industry-wide economies of scale (i.e. β_1 constant) with firm-specific fixed costs:

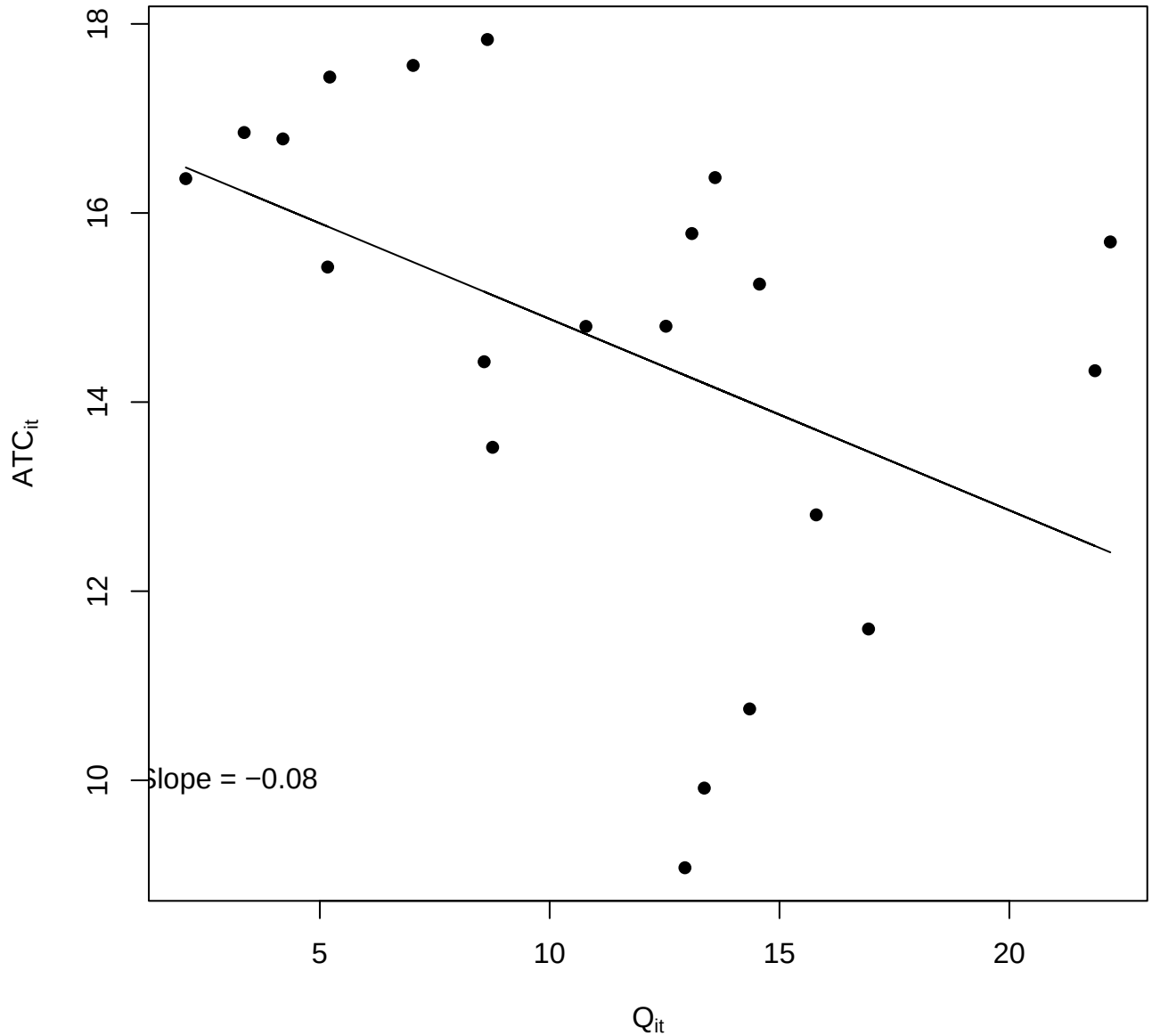
$$ATC_{it} = \beta_1 \cdot Q_{it} + \underbrace{\alpha_i}_{\beta_0+\beta_2 \cdot X_{2i}=AFC_i} + v_{it}$$

- Suppose larger firms have lower average fixed costs ($\text{corr}(Q_{it}, \alpha_i) < 0$), and we estimate the following "pooled" regression:

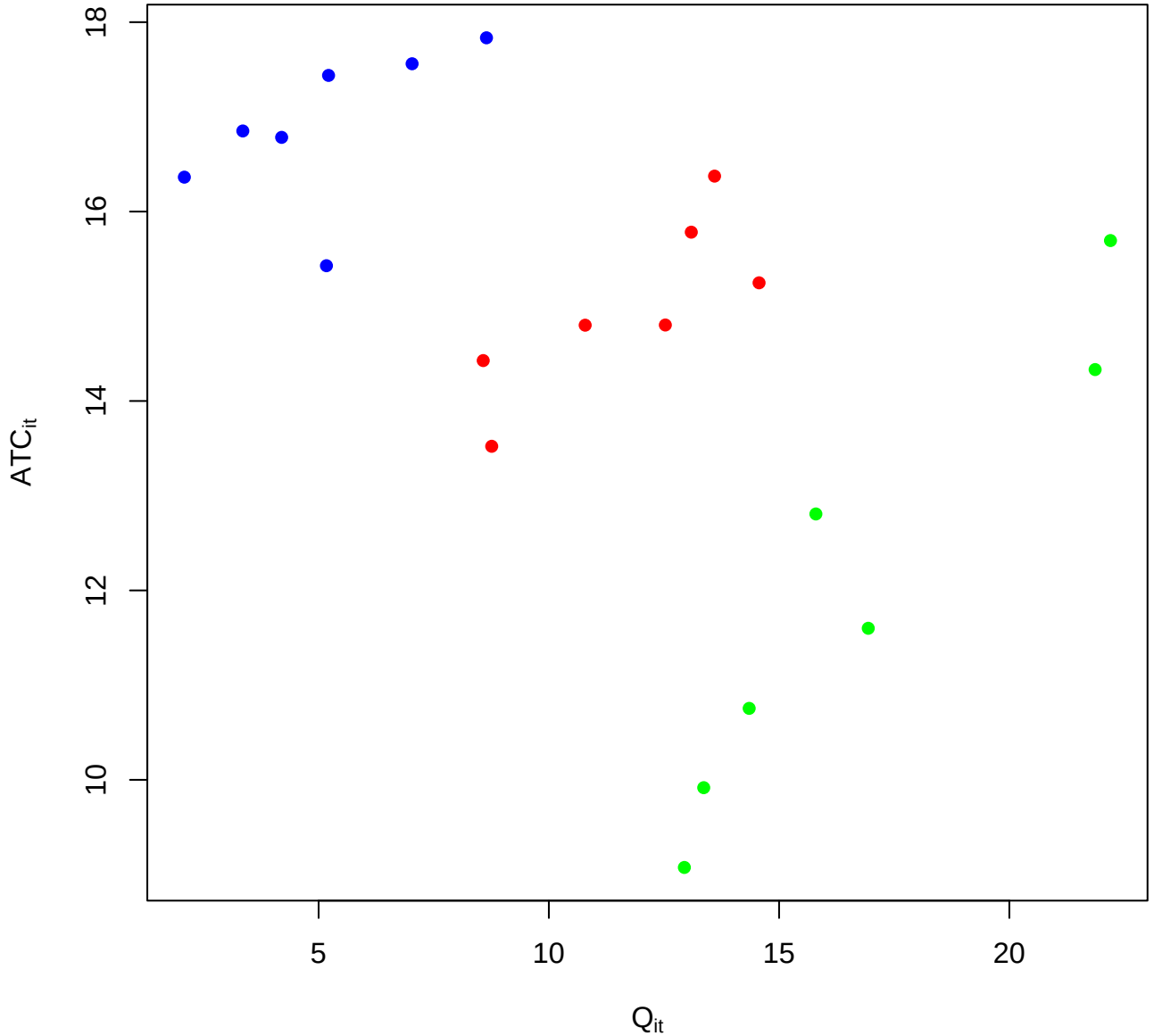
$$ATC_{it} = \hat{\beta}_0 + \hat{\beta}_1 \cdot Q_{it} + \hat{u}_{it}$$

- Is $\hat{\beta}_1$ an upward-, downward-, or un-biased estimator of β_1?
  - $\hat{\beta}_1 < \beta_1$

Hypothetical Example, Visualized



Hypothetical Example, Visualized



Panel data solutions

- In the presence of time-invariant heterogeneity, panel data offers 3 solutions:
  1. First Differences regression
  2. Demeaned regression
     - Equivalent to the above if $T = 2$
  3. Fixed Effect regression
     - Equivalent to Demeaned regression

First Differences panel regression

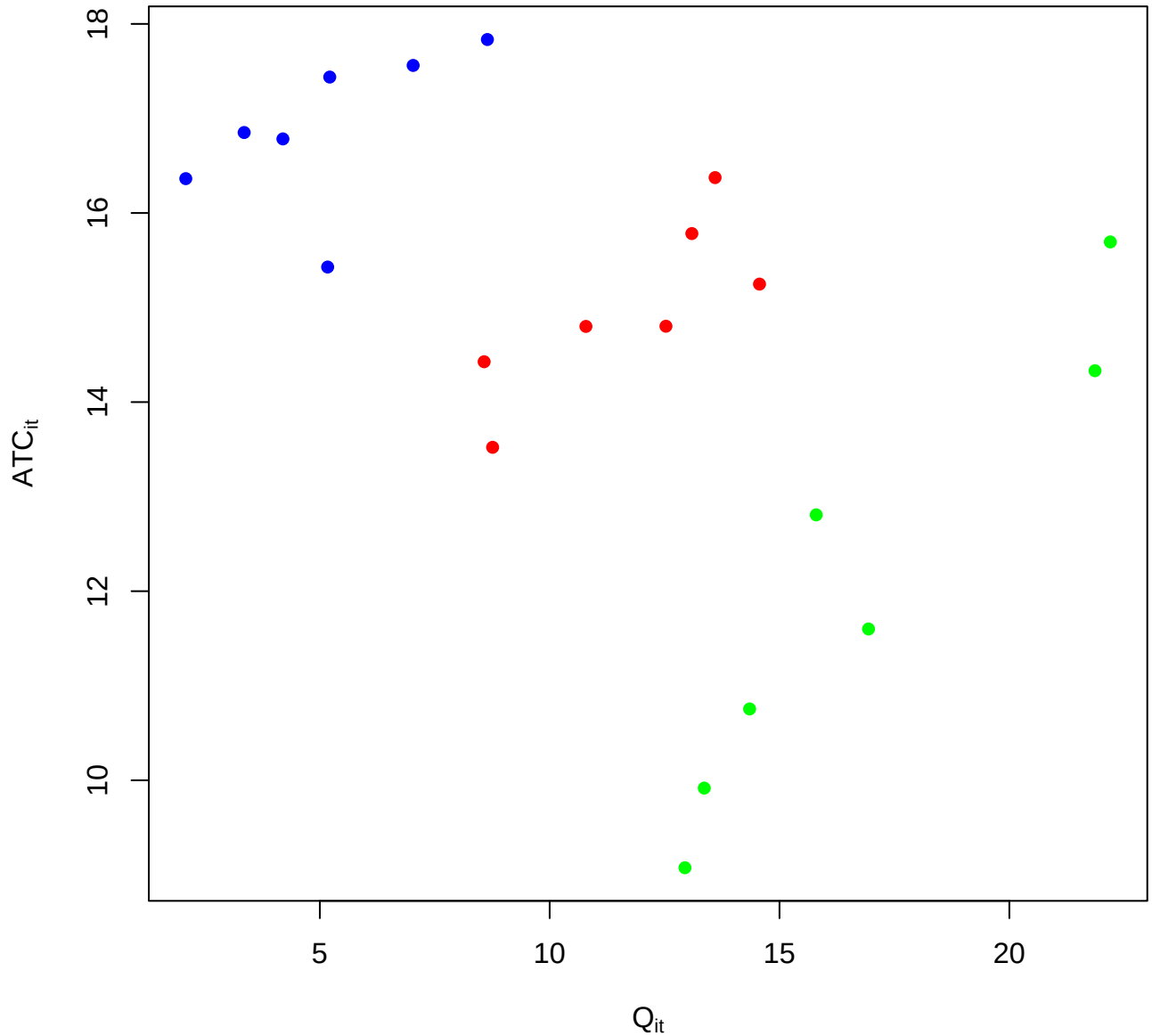
$$ATC_{it} = \beta_1 \cdot Q_{it} + \alpha_i + v_{it}$$

- Taking "First Differences" (1-period change) of both sides:

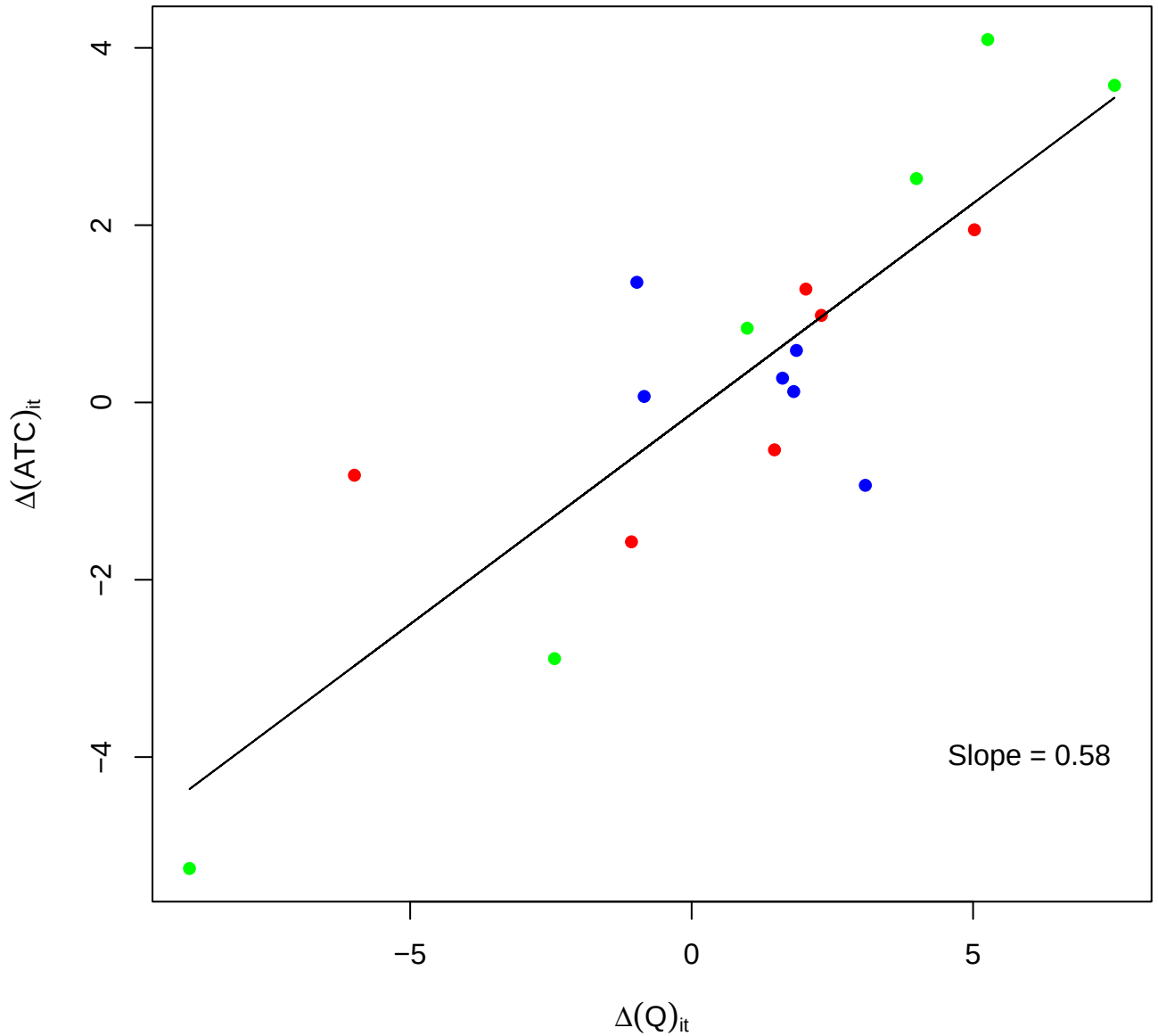
$$\begin{aligned}\Delta ATC_{it} &= \beta_1 \cdot \Delta X_{1,it} + \Delta \alpha_i + \Delta v_{it} \\ &= \beta_1 \cdot \Delta X_{1,it} + \Delta v_{it}\end{aligned}$$

- We have "differenced out" the omitted variable!
  - Because it has no impact on the *change in* Y_{it}, its omission from the FD regression causes no bias
- We can now estimate β_1 in an unbiased, consistent way with OLS

Hypothetical Example, First Differences Visualized



Hypothetical Example, First Differences Visualized



Demeaned panel regression

$$ATC_{it} = \beta_1 \cdot Q_{it} + \alpha_i + v_{it}$$

- Note the following:

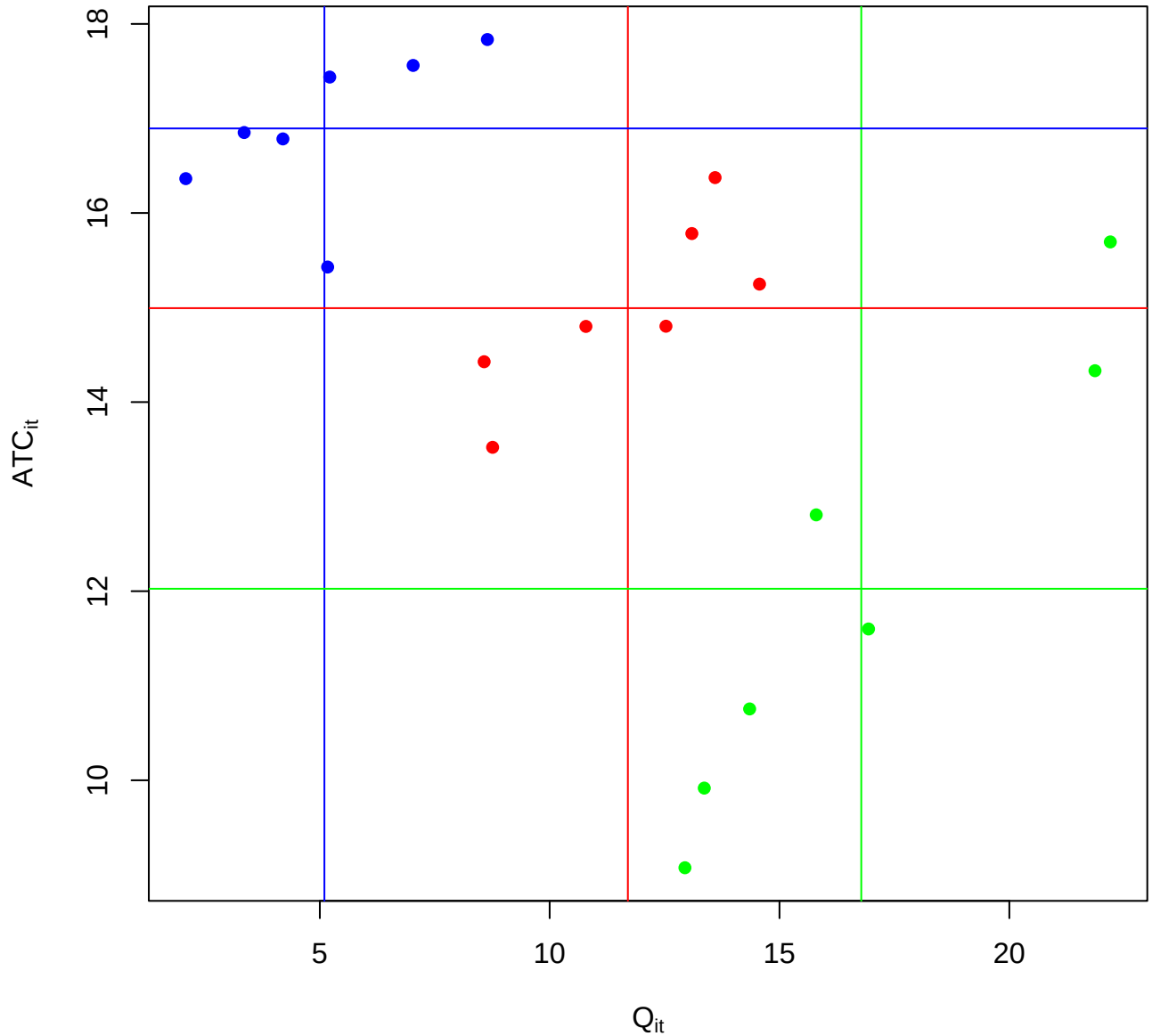
$$\overline{ATC}_i = \beta_1 \cdot \bar{Q}_i + \alpha_i + \bar{v}_i$$

- Therefore:

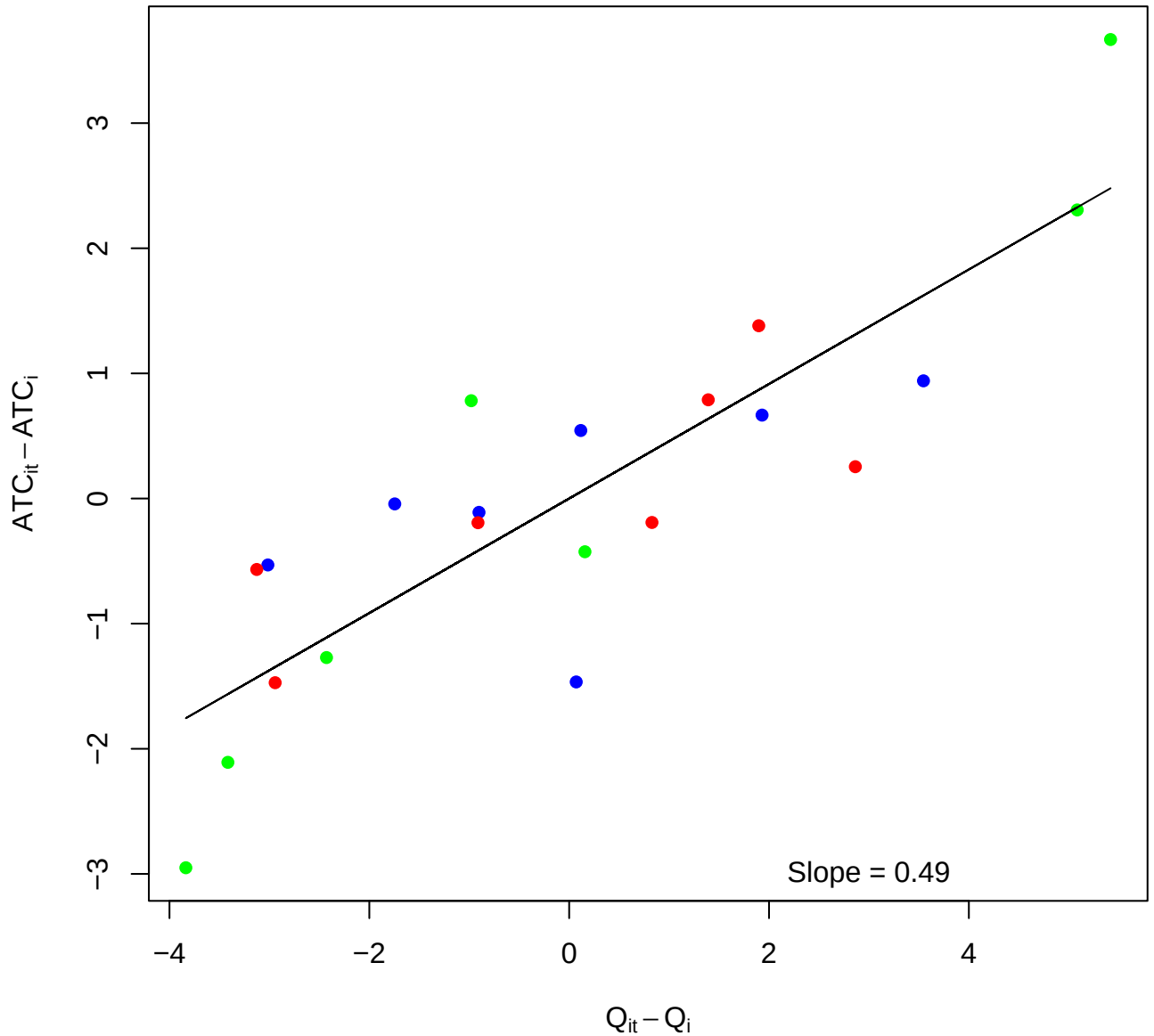
$$ATC_{it} - \overline{ATC}_i = \beta_1 \cdot (Q_{it} - \bar{Q}_i) + (\alpha_i - \alpha_i) + (v_{it} - \bar{v}_i)$$
$$ATC_{it} - \overline{ATC}_i = \beta_1 \cdot (Q_{it} - \bar{Q}_i) + (v_{it} - \bar{v}_i)$$

- Again, the omitted variable is gone!
  - Because it has the same effect on all observations within a firm, it is uncorrelated with deviations of Q from the firm's mean
- We can now estimate β_1 in an unbiased, consistent way with OLS

Hypothetical Example, Demeaned Regression Visualized



Hypothetical Example, Demeaned Regression Visualized



Fixed Effects panel regression

$$ATC_{it} = \beta_1 \cdot Q_{it} + \alpha_i + v_{it}$$

- Consider this regression:

$$ATC_{it} = \beta_0^{FE} + \beta_1^{FE} \cdot Q_{it} + \sum_{k=2}^{n} (\gamma_k \cdot D_{ik}) + u_{it},$$

where $D_{ik} = \begin{cases} 1 \text{ if } i = k \\ 0 \text{ if } i \neq k \end{cases}$.

- This regression gives each cluster (firm) its own constant. So:
  - $\beta_0^{FE} = \alpha_1$
  - $\beta_0^{FE} + \gamma_k = \alpha_k$ for $k > 2$
- α_i is no longer omitted – it is estimated by the FE regression!
- We can now estimate β_1 in an unbiased, consistent way with OLS

Fixed Effects panel regression (2)

$$ATC_{it} = \beta_0^{FE} + \beta_1^{FE} \cdot Q_{it} + \sum_{k=2}^{n} (\gamma_k \cdot D_{ik}) + u_{it}$$

- That equation contains "firm Fixed Effects"
  - "Fixed Effects" refers to taking a discrete/categorical variable and including an indicator variable for every value/category
- It is the same as the Demeaned regression
  - Note that $\overline{ATC}_i = \beta_0^{FE} + \gamma_i + \beta_1^{FE} \cdot \bar{Q}_i + \bar{u}_i$
- The mechanics of how the regression is estimated are essentially identical to any multivariate regression we've studied
  - Lone exception is SEs, discussed later
- Key is interpretation as "within estimator"
  - β_1^{FE} is estimated using only variation in Q_{it} *within* firms

Within and Between Estimators

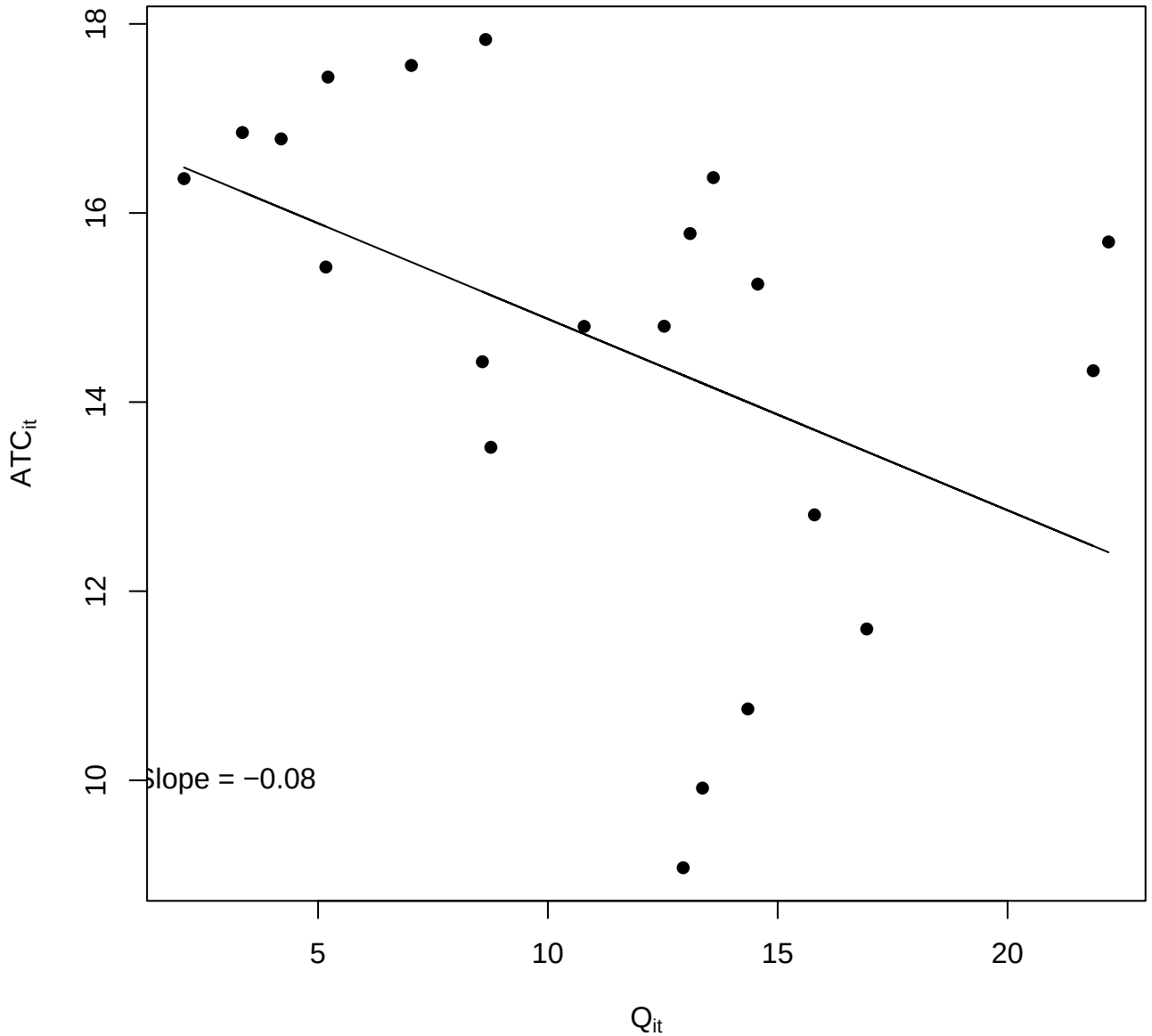
- It can be shown:

$$\beta_1^{Pooled} = \underbrace{\frac{var(X_{it} - \bar{X}_i)}{var(X_{it})}}_{\text{share of variation from within entity}} \cdot \beta_1^{Within} + \underbrace{\frac{var(\bar{X}_i)}{var(X_{it})}}_{\text{share of variation from between entities}} \cdot \beta_1^{Between},$$

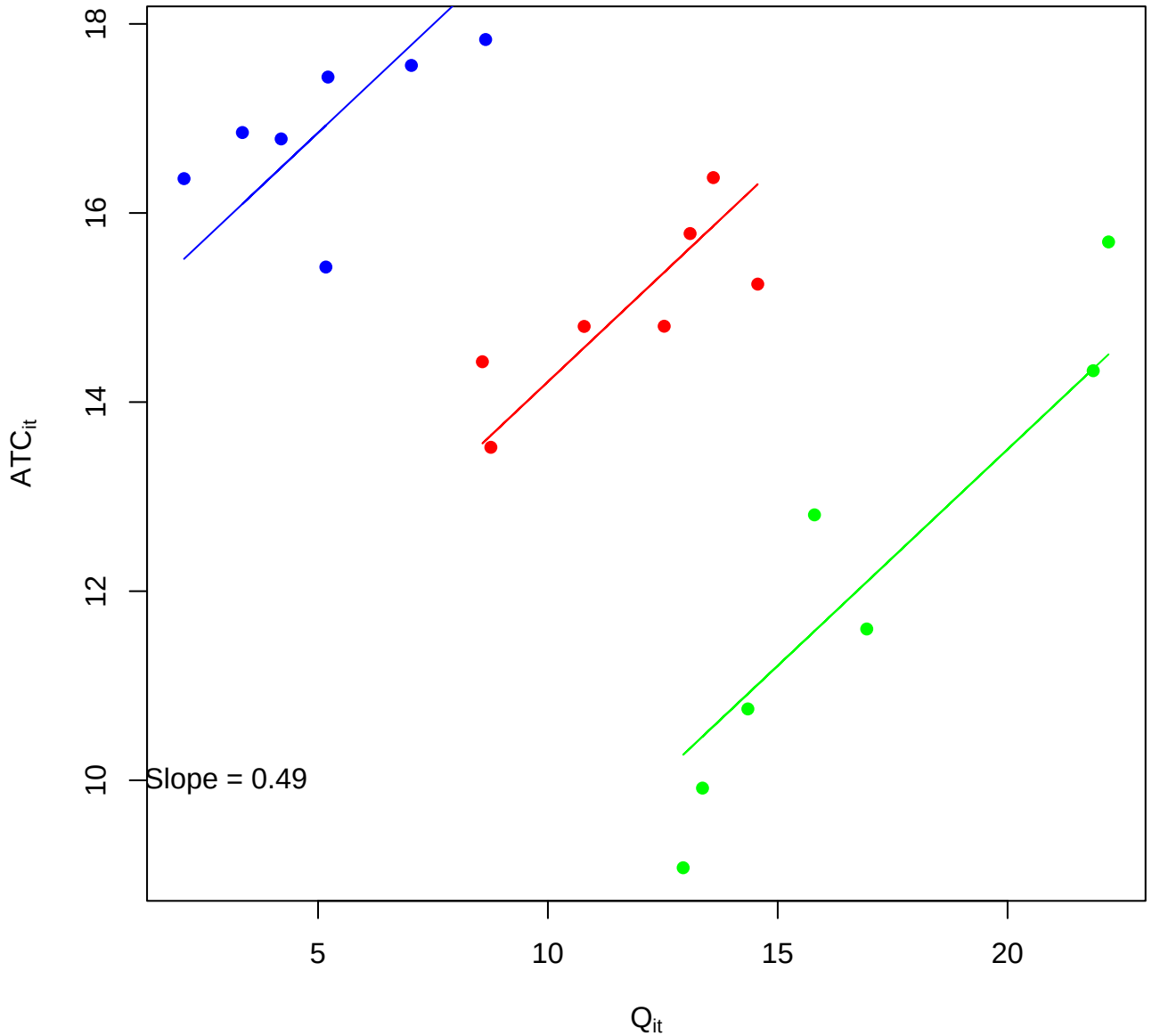
where

- β_1^{Pooled} is the coefficient from simply regressing Y_{it} on X_{it}
- β_1^{Within} is the coefficient from regressing Y_{it} on X_{it} with "i FEs"
- $\beta_1^{Between}$ is the coefficient from regressing $\bar{Y}_i$ (or Y_{it}) on $\bar{X}_i$

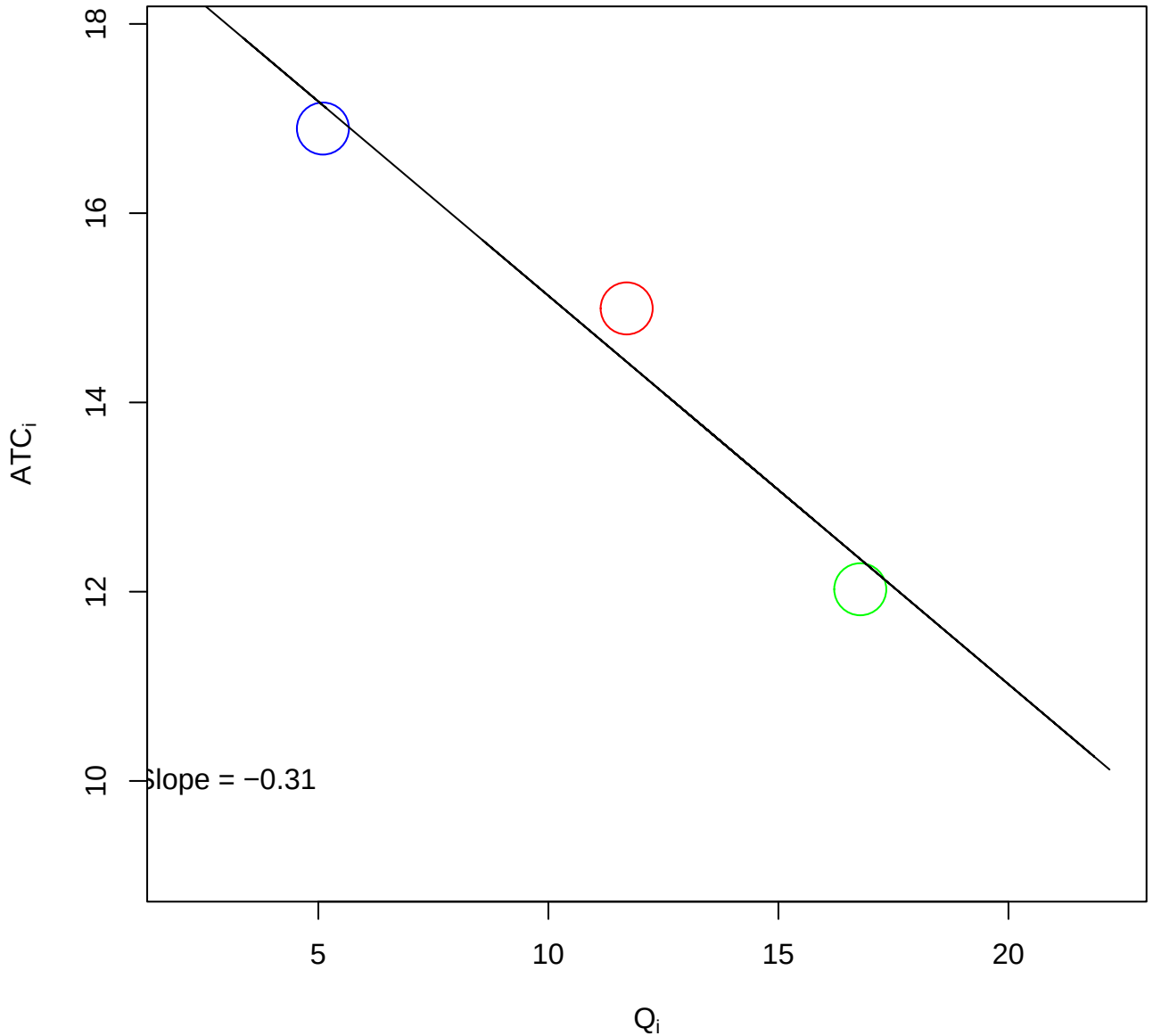
Pooled regression



Within (or "FE" or demeaned) regression



Between regression



Within and Between Estimators Demonstrated

$$\underbrace{\beta_1^{Pooled}}_{-0.08} = \underbrace{\frac{var(X_{it} - \bar{X}_i)}{var(X_{it})}}_{12.4/43.0=29\%} \cdot \underbrace{\beta_1^{Within}}_{0.49} + \underbrace{\frac{var(\bar{X}_i)}{var(X_{it})}}_{30.6/43.0=71\%} \cdot \underbrace{\beta_1^{Between}}_{-0.31}$$

Fixed Effects in non-panel settings

- We have introduced FEs in the context of panel data, but the ideas apply to any "hierarchical data"
- In a cross-sectional study of individuals, can include FEs for things like state or occupation (or state-by-occupation!)
- In time series data, can include FEs for year
  - Can only do this if your data's frequency is higher than annual. Why?
- In panel data, can include things like state FEs, year FEs, or state-by-year FEs

Traffic Fatalities on DUI Jail Time, Results

	Dependent variable: Fatality Rate				
Jail	35.27	2.13	2.13	-4.23	-4.26
State FEs?	N	Y	Y	Y	Y
"Change" States?	N	N	Y	N	Y
Unemp Ctrl?	N	N	N	Y	N

Goldin (2014)

TABLE 1—RESIDUAL GENDER DIFFERENCES IN EARNINGS AND THE ROLE OF OCCUPATION

Sample	Variables included	Coefficient on female	Standard error	R^2
Full-time	Basic	−0.248	0.00101	0.112
Full-time	Basic, time	−0.193	0.00100	0.163
Full-time	Basic, time, education	−0.247	0.000905	0.339
Full-time	Basic, time, education, occupation	−0.192	0.00104	0.453
All	Basic	−0.320	0.00105	0.102
All	Basic, time	−0.196	0.000925	0.353
All	Basic, time, education	−0.245	0.000847	0.475
All	Basic, time, education, occupation	−0.191	0.000963	0.563
Full-time, BA	Basic	−0.285	0.00159	0.131
Full-time, BA	Basic, time	−0.230	0.00158	0.177
Full-time, BA	Basic, time, education	−0.233	0.00155	0.216
Full-time, BA	Basic, time, education, occupation	−0.163	0.00158	0.374
All, BA	Basic	−0.384	0.00173	0.119
All, BA	Basic, time	−0.227	0.00151	0.380
All, BA	Basic, time, education	−0.229	0.00148	0.407
All, BA	Basic, time, education, occupation	−0.163	0.00151	0.525

Notes: "Basic" regression is the log of annual earnings regressed on the female dummy, age as a quartic, race, and year. "Time" adds log hours per week and log weeks. "Education" adds dummies for education categories (and those above a BA for the college graduate sample). "Occupation" adds three-digit occupation dummies. "Full-time" is 35 and above hours per week and 40 and above weeks per year. "All" includes workers 25 to 64 years old with positive earnings and positive hours worked during the past year. The "full-time" sample consists of full-time, full-year individuals 25 to 64 years old excluding those in the military using trimmed annual earnings data (exceeding 1,400 hours × 0.5 × 2009 minimum wage). The "BA" sample includes workers with at least a college or university bachelor's degree. The number of observations is 2,603,968 for full-time, 3,291,168 for all, 964,705 for full-time BA or more, and 1,162,638 for all BA or more.

Abel and Fuster (2021)

TABLE 3—REGRESSION ESTIMATES OF THE TREATMENT EFFECT OF REFINANCING ON MONTHLY PROBABILITIES OF MORTGAGE DEFAULT AND NON-MORTGAGE SERIOUS DELINQUENCY

	Panel A. Mortgage default				*Panel B. Non-mtg. serious delinq.*			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
IV								
Basis points	−2.68	−2.46	−2.17	−3.94	−3.48	−3.15	−3.33	−3.35
(SE)	(0.77)	(0.77)	(0.77)	(1.03)	(1.28)	(1.33)	(1.32)	(3.50)
OLS								
Basis points	−4.30	−4.11	−4.10	−4.27	−2.43	−2.32	−2.31	−2.26
(SE)	(0.25)	(0.27)	(0.26)	(0.29)	(0.27)	(0.27)	(0.27)	(0.27)
Quarter FEs	✓	✓	✓	✓	✓	✓	✓	✓
Zip-code FEs	✓	✓	✓	✓	✓	✓	✓	✓
Observables	✓	✓	✓	✓	✓	✓	✓	✓
Q-by-zip FEs		✓	✓	✓		✓	✓	✓
Guar. lag FEs			✓	✓			✓	✓
Cohort FEs				✓				✓
N (mill.)	11.5	11.5	11.5	11.5	13.9	13.9	13.9	13.9

Notes: For borrower i in month t, the refinancing indicator is turned on if she has completed a refinance in some month $\tau \leq t$. The mortgage default indicator is turned on if she is at least 90 days delinquent on her mortgage in month t, and the non-mortgage serious delinquency indicator is turned on if she has had three consecutive months with delinquent balances on non-first-mortgage debt. Borrowers are censored after their first month in default/serious delinquency. The takeup-weighted monthly mortgage default rate is 5.9 bp, while it is 11.7 bp for non-mortgage serious delinquency. IV estimates result from instrumenting for the refinance indicator with HARP eligibility, interacted with a full set of quarter indicators. "Observables" include ten equally-sized bines for each of: CLTV (lagged three months and at origination), credit score (lagged three months and at origination), credit utilization (lagged three months and at origination), initial mortgage rate, initial debt balances, and remaining principal balance. We also include indicators for mortgage "purpose" (e.g., purchase, cash-out refi, etc.). Standard errors (in parentheses) are clustered at the county level.

Standard Errors, Revisited

- We began the course with the assumption of random sampling in which observations' deviations from average (u_i) are i.i.d
  - Independent and Identically-Distributed
- We quickly relaxed the "identically distributed" assumption
  - Heteroskedasticity-robust SEs allow variance of u_i to depend on X_i
- With panel data, not reasonable to assume that the draws are independent
  - E.g. If we have a panel of firms over time, it seems likely that the observations of a single firm may be correlated...
  - ...or maybe observations on a single day are correlated

Clustered Standard Errors

- The solution to this is called "clustering" or "clustered SEs"
- We won't look at the formula, but the idea is:
  - If you split the data into "clusters" (e.g. "firms" or "months" or "firm-months"...
  - ...clustered SEs assume there is no correlation *between* clusters but there may be a lot of correlation *within* clusters
    - Caveat: Clustered SEs are unreliable if you have very few clusters. It's like having very few observations from a cross-section
- Clustering SEs is a big deal
  - As a general observation, homo- vs. heteroskedastic SEs usually doesn't matter much, and the effect of correcting for heteroskedasticity can go either way
  - Failing to cluster SEs can often lead to drastic underestimation of SEs
    - Failing to cluster is like assuming you have more independent draws from the data than you do!

Milk Industry: Effect of wholesale price on store prices

- Have daily data on $\sim 5,000$ stores across 9 years
- Have the wholesale price of milk (same for whole region) and store-level prices each day

	Dependent variable: Store Price			
Constant	2.01	2.01	2.01	2.01
(SE)	(0.00031)***	(0.00032)***	(0.00544)***	(0.01693)***
Wholesale Price	0.67	0.67	0.67	0.67
	(0.00013)***	(0.00012)***	(0.00154)***	(0.00643)***
SE type	i.i.d.	Hetero	Clust: Store	Clust: Day
Stata SE Option		robust	vce(cluster Store)	vce(cluster Day)

- SE on Wholesale Price assuming iid is $\sim 1/46$ of clustered SE!
- Intuition: Store prices move together, but column 1 views those all as 5,000 independent observations, but they are not independent

Time Series Data

- Time Series analysis is incredibly important (e.g. all of macroeconomics, essentially)
- It is also extremely technical and challenging
  - Ironically, Panel Data analysis simpler than Time Series analysis.
  - While Panel Data has a time series component, the cross-sectional component saves us from some headaches
- In this course, we will not cover Time Series econometrics in detail. Rather, I want to flag 2 common mistakes people often make that you should beware of whenever doing or consuming Time Series analysis
  1. It is often wrong to run a regression in "levels," and you should be extremely careful if doing so
  2. You need to adjust SEs, but even then you likely should not trust them too much

Beware Time Series Regressions in Levels

- Suppose you have two time series, Y_t and X_t
- It is almost always a bad idea to run a regression in "levels:"

$$Y_t = \hat{\beta}_0^{Lev} + \hat{\beta}_1^{Lev} \cdot X_t + \hat{u}_t^{Lev} + \hat{u}_t^{Lev}$$

- Usually much better to run in First Differences

$$\Delta Y_t = \hat{\beta}_0^{FD} + \hat{\beta}_1^{FD} \cdot \Delta X_t + \hat{u}_t^{FD}$$

- Many time series have trends (think GDP and Consumption)
  - The "Levels" regression might find a strong relationship, but it's just picking up the common trend.
  - The "First Differences" regression eliminates a linear trend and so is safer

Do the Yankees and Red Sox Help or Hurt Each Other?

- Yankees and Red Sox are rival baseball teams
- They account for roughly 10% of each others' games
- They compete for the same bids into postseason
- Seems likely that Yankees' success is bad for Red Sox and vice versa
- Looking at annual team revenue from 2001-2019:

$$Y_t = \underbrace{\hat{\beta}_0^{Lev}}_{} + \underbrace{\hat{\beta}_1^{Lev}}_{} \cdot RS_t$$

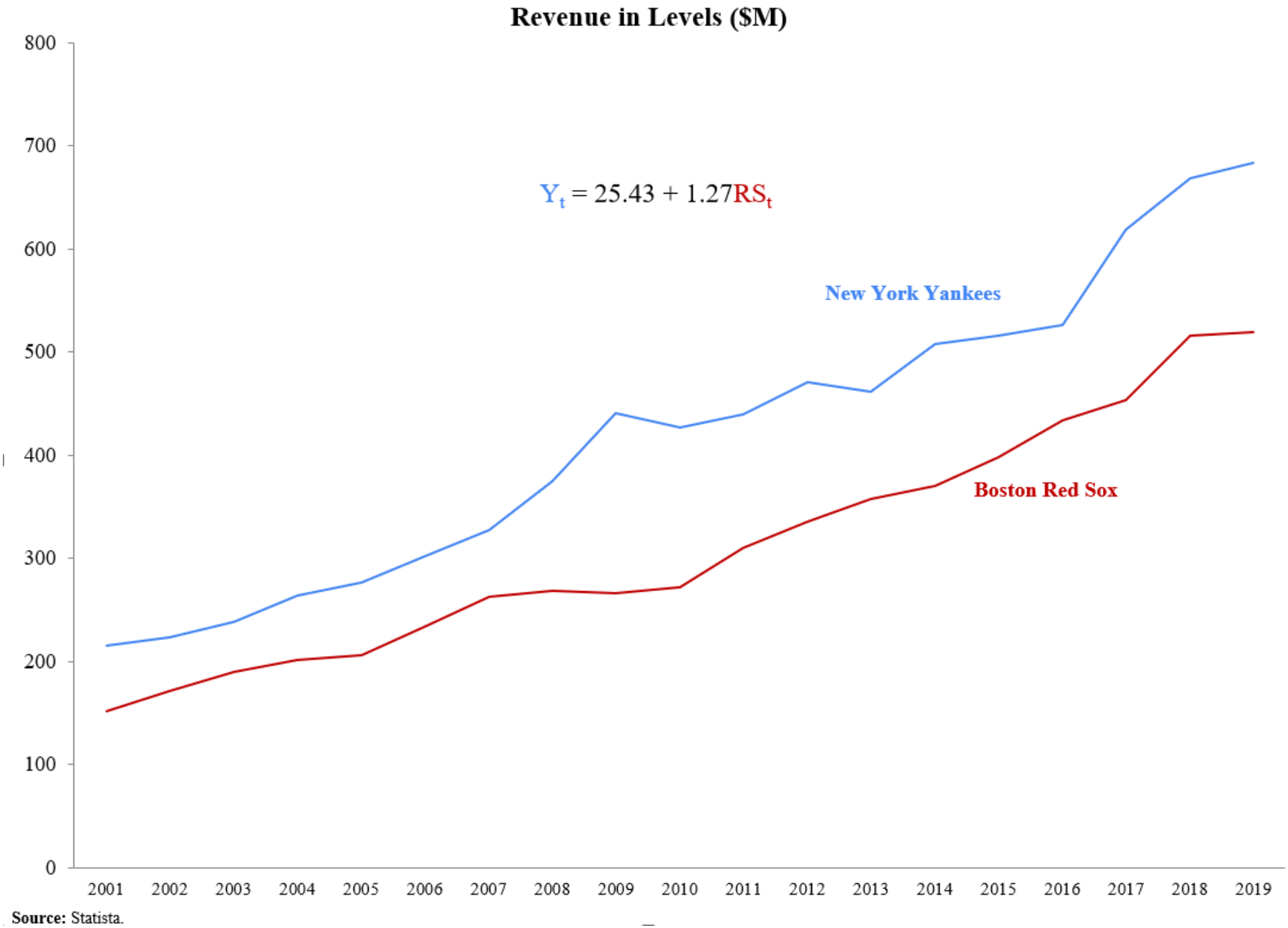
Do the Yankees and Red Sox Help or Hurt Each Other?

- Yankees and Red Sox are rival baseball teams
- They account for roughly 10% of each others' games
- They compete for the same bids into postseason
- Seems likely that Yankees' success is bad for Red Sox and vice versa
- Looking at annual team revenue from 2001-2019:

$$Y_t = \underbrace{\hat{\beta}_0^{Lev}}_{25.43} + \underbrace{\hat{\beta}_1^{Lev}}_{1.27} \cdot RS_t$$

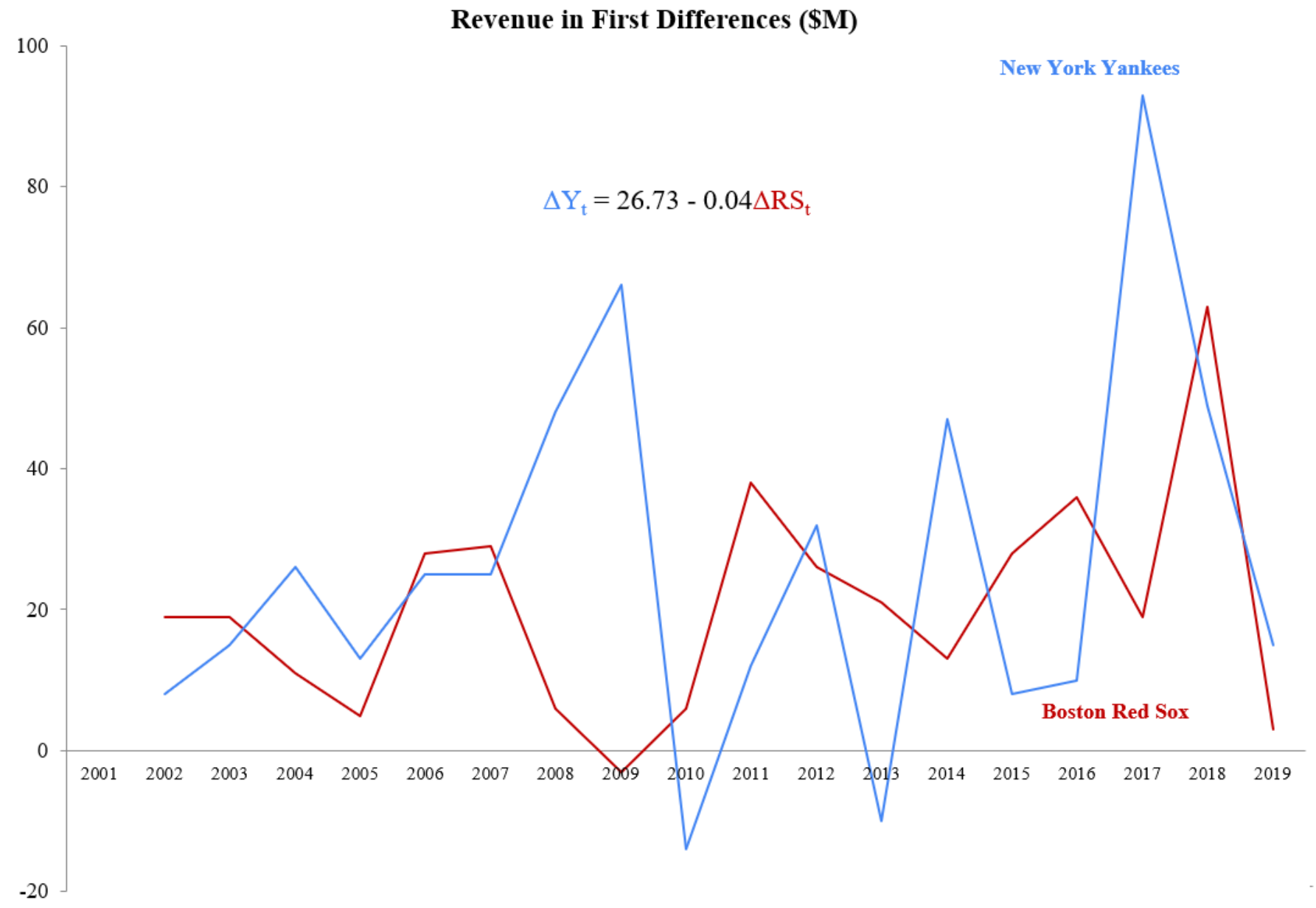
Should we trust this?

Yankees-Red Sox Results



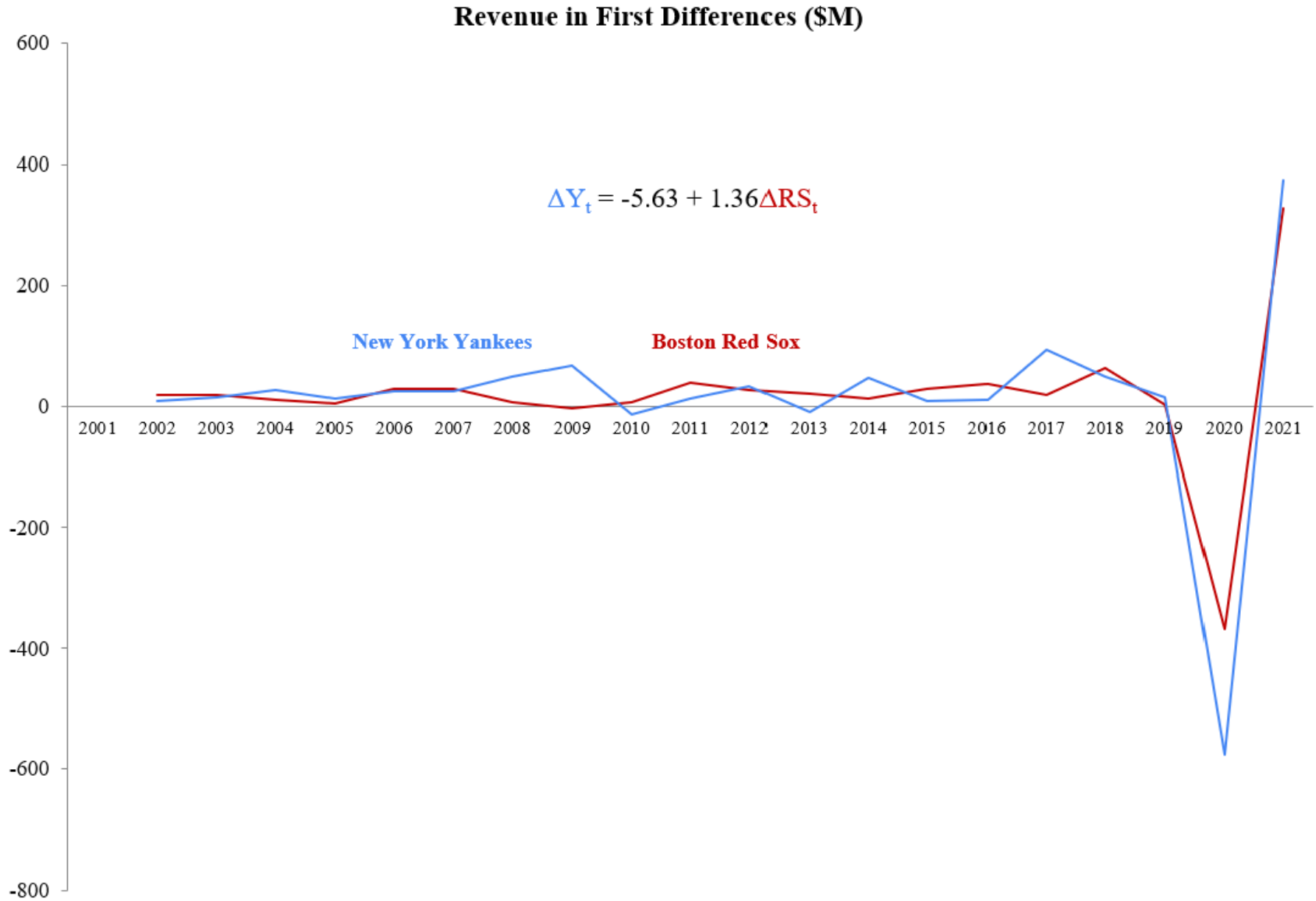
Source: Statista.

Yankees-Red Sox Results



Source: Statista.

Yankees-Red Sox Results



Source: Statista.

Non-Stationarity

- When data has a trend, that is an example of "non-stationarity"
  - Essentially it means that data from different time periods are not being pulled from the same distribution
    - E.g. Due to rising incomes, the mean of the distribution of Red Sox revenue was higher in 2020 than 2002
- Non-stationarity wreaks havoc on interpretation of coefficients and calculation of SEs
- Regressions of non-stationary variables can only be done with care, with techniques we will not cover
- If data has a linear trend, First Differences eliminates it
  - Makes it more likely the series are stationary
  - But no guarantee :(

The Fundamental Nightmare of Time Series Data

- In cross-sectional data, there is no "order" of the observations.
  - It is reasonable to think they are uncorrelated, as they are random draws from a common distribution
- In Time Series data, there is a clear ordering
  - 2000 came before 2001, which came before 2002, ...
  - 2000 may have *affected* 2001, and 2002, ...
  - They are *not* independent
- Even in stationary data, annual data from 2001 to 2020 is probably *not* 20 independent draws of annual data
  - In the extreme, we might think of it as 1 observation that happens to be 20 years long!
  - In this view, it would be impossible to quantify uncertainty (i.e. compute SEs), because it's just 1 observation

Standard Errors in Time Series Regressions

- As in panel data, we no longer have the independence assumption in time series data
- In panel regressions, we were okay if we used clustered SEs, because the cross-sectional component could still be used
  - But with a Time Series, there is only 1 "cluster," so these can't be used
- There are techniques to try to deal with this
  - You need to use one of them, and be skeptical of anyone who doesn't
  - But even with these techniques, you should generally be less confident that your estimated SEs correctly quantified uncertainty than in the cross-sectional (or even panel) case

Newey-West Standard Errors

- One common approach is to use Newey-West SEs
- This methodology provides an estimate of SEs conditional on an assumption about how long correlation lasts in the data
  - E.g. What happened in 2001 affects 2002 and 2003, but nothing after

GDP, Gov't Spending Regression, Revisited

Outcome: Δ GDP				
Constant	67.81	67.81	67.81	67.81
(SE)	(9.05)***	(10.85)***	(12.14)***	(13.36)***
Δ G	-0.87	-0.87	-0.87	-0.87
	(0.39)**	(0.89)	(1.23)	(1.22)
SE type	i.i.d.	Hetero	Corr: 4 qtrs	Corr: 8 qtrs
Stata Command	regress	regress	newey	newey
Stata SE option		robust	lag(4)	lag(7)