Common Limites and Errors

Professor Ji-Woong Chung
Korea University

Outline

Data Limitations

- The data we use is almost never perfect:
  - Variables are often reported with error.
  - Exit and entry into dataset typically not random.
  - Datasets only cover certain types of firms.

Measurement Error – Examples

- Measurement error occurs when observed values differ from the true values.
- Two main types:
  - **Random (innocent) errors:** Pure noise, not systematically related to other variables.
    E.g. Survey respondents round or misremember income.
    - Leads to greater variance, but no systematic bias if uncorrelated with regressors.
  - **Systematic (nonrandom) errors:** Certain groups misreport in predictable ways.
    E.g. High-GPA teenagers underreport marijuana use; firms understate liabilities.
    - Correlation with covariates $\Rightarrow$ biased estimates.
- **Key question:** How do these errors affect causal inference and estimation?

Measurement Error – Why It Matters

- The impact depends on which variable is measured with error.
- **If the dependent variable (y) is mismeasured:**
  - Random noise: Increases residual variance $\Rightarrow$ larger standard errors.
  - Systematic error: If correlated with x, coefficient estimates become biased.

  E.g. Low-education respondents underreport income $\Rightarrow$ downward bias on education effect.
- **If the independent variable (x) is mismeasured:**
  - Classical error (mean-zero, uncorrelated): Attenuation bias $\Rightarrow$ slope biased toward 0.
  - Non-classical error (correlated): Bias in unpredictable directions; contaminates other coefficients.

  E.g. Noisy education measure $\Rightarrow$ underestimated returns to schooling.
- **Summary:** Random $\Rightarrow$ inefficiency; Systematic $\Rightarrow$ bias.

Measurement Error – Solutions

- Measurement error correction requires knowing its **source and structure**.
- **Common approaches:**
  - **Instrumental Variables (IV):** Find variable correlated with true x but not error.

    E.g. Administrative wage data as instrument for self-reported income.
  - **Validation samples or repeated measures:** Estimate or correct error variance.
  - **Structural modeling:** Explicitly model the measurement process
- **Challenges:**
  - Hard to correct without auxiliary or validation data.
  - Unknown error patterns $\Rightarrow$ unpredictable bias.
- Always scrutinize data accuracy—small errors can distort inference.

Survivorship Issues – Examples

- Observations may be missing or included for **systematic reasons**, not by chance.
- **Example 1:** IPO firms
  - Datasets of public firms exclude private firms.
  - Firms that go public may already differ (e.g., more profitable, faster-growing, or better governed).
- **Example 2:** Distressed or failed firms
  - Firms severely affected by a shock may disappear due to bankruptcy.
  - Remaining sample overrepresents "survivors."
- **Question:** How do such missing or selective exits bias our estimates?

Survivorship Issues – Why It Matters

- **Selection bias** can lead to misleading conclusions.
- **Example 1: IPOs and growth**
  - High post-IPO growth may not be caused by going public.
  - Rather, firms that went public were already high-growth candidates.
- **Example 2: Negative events and exits**
  - If failing firms disappear after a shock, the observed average effect looks smaller (or even positive).
  - Survivors are systematically different from those that dropped out.
- Result: Bias in estimates, especially in causal or panel analyses.

Survivorship Issues – Solutions

- No perfect fix, but several **diagnostic checks** help:
- **1. Check for selective attrition:**
  - In DiD, test whether treatment status predicts dropping from the data.
  - If treatment increases exit probability, estimate may be biased.
- **2. Compare characteristics of dropouts vs. survivors:**
  - Are exiting observations systematically different in key covariates?
  - If yes, assess how their absence might affect estimates.
- **3. Sensitivity checks:**
  - Include censored or imputed outcomes where possible.
  - Use survival models (hazard or selection models) if dropout is endogenous.

Sample is Limited – Examples

- Many widely used datasets cover only a **subset of firms**.
- **Example 1: Compustat**
  - Focuses on large, listed U.S. firms.
  - Excludes small, private, and young firms.
- **Example 2: Execucomp**
  - Covers CEO pay and incentives only for S&P 1500 firms.
  - Omits privately held or smaller listed firms.
- **Question:** How could this limited coverage bias our findings?

Sample is Limited – Why It Matters

- Limited samples threaten **external validity**.
- **Example 1: Treatment effects in Compustat**
  - You may find no effect in large public firms.
  - But the same treatment could strongly affect unobserved small or private firms.
- **Example 2: CEO incentives in Execucomp**
  - Correlation between incentives and risk-taking may reflect large-firm governance structures.
  - May not generalize to smaller or family-controlled firms.
- Key issue: **Selection on observables and unobservables** into the dataset.

Sample is Limited – Solutions

- **1. Be explicit about scope:**
  - Avoid overgeneralization; limit conclusions to the covered population.
  - Emphasize that results apply to large public firms if using Compustat or Execucomp.
- **2. Argue representativeness:**
  - Show your sample captures an economically important segment.
  - E.g., S&P 1500 firms represent majority of U.S. market capitalization.
- **3. Extend the data:**
  - Hand-collect missing data or merge with private firm databases.
  - Building new datasets can yield high-impact, publishable research.

Example

- Ali, Klasa, and Yeung (RFS 2009) provide a striking case of data mismeasurement.
- Many finance theories emphasize **industry concentration** as a key variable:
  - E.g., competition, market power, R&D incentives, and financing constraints.
- Researchers typically measure concentration (Herfindahl index) using **Compustat**.
- **Question:** What's wrong with that approach?

Example [Part 1]

- **Systematic measurement error:**
  - Compustat excludes private firms $\Rightarrow$ distorted Herfindahl index.
  - Ali, Klasa, and Yeung construct an alternative using **U.S. Census data** (which includes all firms).
  - Correlation between the Compustat and Census-based measures = only **13%**.
- The error is not random:
  - Bias is related to observable industry traits—e.g., turnover, entry, exit, and listing propensity.

Example [Part 2]

- Using the Census-based measure, Ali, Klasa, and Yeung (RFS 2009) show that:
  - The mismeasurement meaningfully changes empirical conclusions.
  - Four major published results are overturned.
- **Example:**
  - Previous studies (using Compustat) found a **negative** link between concentration and R&D.
  - With accurate Census data, the relationship becomes **positive**.
- **Lesson:** Measurement error in key variables can fundamentally alter conclusions.

Outline

Hypothesis Testing Mistakes

- Researchers often compare treatment effects across groups by estimating separate DiDs.
- **Example:**
  - Estimate treatment effect for small firms.
  - Estimate treatment effect for large firms.
- Then they conclude: "The effect is stronger for large firms."

Example Inference from Analysis

	Small Firms	Large Firms	Low D/E Firms	High D/E Firms
Treatment×Post	0.031	0.104**	0.056	0.081***
	(0.121)	(0.051)	(0.045)	(0.032)
N	2,334	3,098	2,989	2,876
R^2	0.11	0.15	0.08	0.21
Year FE	✓	✓	✓	✓
Firm FE	✓	✓	✓	✓

- Researchers often conclude:
  - "Treatment effect is larger for big firms."
  - "High D/E firms respond more."
  - *But are those differences statistically significant?*

Be Careful Making Such Claims

- **Problem:** Differences across subsamples may not be statistically significant.
- You can't tell by "eyeballing" coefficients.
  - Statistical significance depends on the **covariance** between estimates.
- **Proper test:** Include an interaction term (triple difference) in a single regression.

Example Triple Interaction Result

	All Firms
Treatment×Post	0.031
	(0.121)
Treatment×Post×Large	0.073
	(0.065)
N	5,432
R^2	0.12
Year FE	✓
Firm FE	✓
Year×Large FE	✓

- Difference between large and small firms is **not statistically significant**.
- Always include interaction with year dummies to match subgroup DiDs.

Practical Advice

- Don't make claims you haven't statistically tested.
- Always report the **p-value for the difference** across groups.
- If the difference isn't significant (e.g., $p = 0.15$), say so — triple differences are noisy.
- Be cautious with phrasing:
  - Instead of: "Large firms respond more,"
  - Say: "We find an effect for large firms but not for small firms."

Outline

Outline

Unobserved Heterogeneity – Motivation

- Controlling for **unobserved heterogeneity** is a fundamental challenge in empirical finance.
- **Why?** Many important factors cannot be directly measured or included in data:
  - Managerial talent, corporate culture, or risk appetite.
  - Local demand or regulatory conditions.
  - Investor sentiment or regional economic trends.
- These unobservables can be **correlated** with key explanatory variables:
  - $\Rightarrow$ Leads to **omitted variable bias**.
- Important sources of heterogeneity are often shared across **groups**:
  - Industry-level demand shocks.
  - Region-specific economic or policy environments.
  - Time-period shocks common to all firms.

Many Different Strategies Are Used

- As discussed earlier, **Fixed Effects (FE)** can control for unobserved heterogeneity and yield consistent estimates when the unobservables are time-invariant.
- But researchers use several alternative or complementary strategies to remove **group-level heterogeneity**:
  - **Adjusted-Y (AdjY):** Demean the dependent variable within groups (e.g., subtract the industry-year average: "industry-adjusted" outcomes).
  - **Average Effects (AvgE):** Include group-level averages of outcomes as controls (e.g., add the mean of y for a given state-year or industry-year).
- Each method aims to remove variation driven by shared shocks or persistent group differences.
  - FE fully removes group-level heterogeneity (e.g., via industry–year dummies).
  - AdjY and AvgE are simplified approximations useful in small samples, but only FE yields consistent estimates when unobservables correlate with regressors.

The Underlying Model [Part 1]

- Start with a simple data-generating process:

$$y_{i,j} = \beta X_{i,j} + f_i + \epsilon_{i,j}$$

  - i: Group index (e.g., industry, state, bank, or fund family)
  - j: Observation within group (e.g., firm, branch, fund)

- Model components:
  - $y_{i,j}$: outcome (e.g., investment, leverage, return)
  - $X_{i,j}$: explanatory variable of interest (e.g., policy, treatment)
  - f_i: unobserved group-level factor (e.g., industry demand, regulation)
  - $\epsilon_{i,j}$: idiosyncratic error term
- The key question: what happens if we try to control for f_i without properly including a fixed effect?

The Underlying Model [Part 2]

- Standard assumptions about the data structure:
  - N: Number of groups is large; J: Observations per group is small.
  - $\text{Var}(f_i) = \sigma_f^2,\ \mathbb{E}[f_i] = 0$
  - $\text{Var}(X_{i,j}) = \sigma_X^2,\ \mathbb{E}[X_{i,j}] = 0$
  - $\text{Var}(\epsilon_{i,j}) = \sigma_\epsilon^2,\ \mathbb{E}[\epsilon_{i,j}] = 0$
- X and ϵ are i.i.d. across groups, but may be correlated **within** groups:
  - Within-group correlation $\Rightarrow$ common shocks.
  - Across-group independence ensures valid asymptotics.

The Underlying Model [Part 3]

- Additional assumptions:
  - $\text{Cov}(f_i, \epsilon_{i,j}) = 0$ — group factors are uncorrelated with idiosyncratic errors.
  - $\text{Cov}(X_{i,j}, \epsilon_{i,j}) = \text{Cov}(X_{i,j}, \epsilon_{i,-j}) = 0$ — exogeneity of X.
    - $X_{i,j}$ is exogenous with respect to both its own error term and the error terms of other group members, enabling unbiased and consistent estimation of β in the fixed effects model.
  - $\text{Cov}(X_{i,j}, f_i) = \sigma_{Xf} \neq 0$ — regressor correlated with group unobservables.
- Implication:
  - If we omit f_i, OLS suffers from classic **omitted variable bias**.
- FE removes f_i through within-group demeaning. AdjY and AvgE only partially do so, leaving residual correlation with f_i. This incomplete adjustment can amplify—or even reverse—the bias.

We Already Know OLS Is Biased

$$\text{True model: } y_{i,j} = \beta X_{i,j} + f_i + \epsilon_{i,j}$$

$$\text{But OLS estimates: } y_{i,j} = \beta^{OLS} X_{i,j} + u_{i,j}^{OLS}$$

- By omitting the group effect f_i, OLS suffers from standard omitted variable bias:

$$\hat{\beta}_{OLS} = \beta + \frac{\sigma_{Xf}}{\sigma_X^2}$$

- Direction and size of bias depend on the covariance between X and f_i.

Adjusted-Y ($AdjY$) Estimation

- **Idea:** Remove unobserved group effects by demeaning the dependent variable within groups:

$$y_{i,j} - \bar{y}_i = \beta^{AdjY} X_{i,j} + u_{i,j}^{AdjY}$$

- Group mean:

$$\bar{y}_i = \frac{1}{J} \sum_{k \in i} y_{i,k} = \frac{1}{J} \sum_{k \in i} (\beta X_{i,k} + f_i + \epsilon_{i,k})$$

$$\Rightarrow \bar{y}_i = \beta \bar{X}_i + f_i + \bar{\epsilon}_i$$

- Some researchers exclude the observation itself or use medians, but bias remains.

Example: *AdjY* Estimation in Practice

- Example regression:

$$Q_{i,j,t} - \overline{Q}_{i,t} = \alpha + \beta X_{i,j,t} + \epsilon_{i,j,t}$$

- Variables:
  - $Q_{i,j,t}$: Tobin's Q for firm j in industry i, year t
  - $\overline{Q}_{i,t}$: industry-year mean Q ("industry-adjusted Q")
  - $X_{i,j,t}$: explanatory variables (e.g., governance, leverage)
  - Often combined with firm or year fixed effects
- **Question:** Why is *AdjY* still inconsistent?

Why *AdjY* Is Inconsistent

- Substitute the group mean:

$$\bar{y}_i = \beta \bar{X}_i + f_i + \bar{\epsilon}_i$$

$$y_{i,j} - \bar{y}_i = \beta(X_{i,j} - \bar{X}_i) + (\epsilon_{i,j} - \bar{\epsilon}_i)$$

- The transformation removes f_i in the mean but not in the regressor.
- When we regress $(y_{i,j} - \bar{y}_i)$ on $X_{i,j}$ instead of $(X_{i,j} - \bar{X}_i)$, the omitted $\bar{X}_i$ induces bias.
- Hence, *AdjY* omits a relevant group-level term.

AdjY and Omitted Variable Bias

- The true transformed model is:

$$y_{i,j} - \bar{y}_i = \beta X_{i,j} - \beta \bar{X}_i + (\epsilon_{i,j} - \bar{\epsilon}_i)$$

- But *AdjY* estimates:

$$y_{i,j} - \bar{y}_i = \beta^{AdjY} X_{i,j} + u_{i,j}^{AdjY}$$

- Because it omits $\bar{X}_i$, the estimator is biased:

$$\hat{\beta}_{AdjY} = \beta - \frac{\sigma^2_{X\bar{X}}}{\sigma^2_X}$$

- With positive $\text{Cov}(X, \bar{X})$—common under shared industry shocks—bias is typically negative.

Adding a Second Variable, Z

- Suppose the true model has two regressors:

$$y_{i,j} = \beta X_{i,j} + \gamma Z_{i,j} + f_i + \epsilon_{i,j}$$

- Maintain previous assumptions and add:
  - $\text{Cov}(Z_{i,j}, \epsilon_{i,j}) = 0$, $\text{Var}(Z) = \sigma_Z^2$
  - $\text{Cov}(X, Z) = \sigma_{XZ}$
  - $\text{Cov}(Z, f_i) = \sigma_{Zf}$
- AdjY still omits group-level means $(\bar{X}_i, \bar{Z}_i)$, creating intertwined biases.

AdjY Estimates with Two Variables

- The biases are now complex:

$$\hat{\beta}_{AdjY} = \beta + \triangle, \quad \hat{\gamma}_{AdjY} = \gamma + \Diamond$$

- Biases depend on correlations among X, Z, f_i.
- As Gormley and Matsa (2014) show:
  - Both coefficients can move in unpredictable directions.
  - Even sign reversals are possible.

Average Effects (*AvgE*) — Idea

- Researchers often want to control for unobserved group-level factors (f_i) when fixed effects are unavailable or costly.
- **Idea:** Include the group mean of the dependent variable as a proxy for f_i:

$$y_{i,j} = \beta^{AvgE} X_{i,j} + \gamma^{AvgE} \bar{y}_i + u_{i,j}^{AvgE}$$

- Example – Firm profitability regression:

$$\text{ROA}_{i,s,t} = \alpha + \beta X_{i,s,t} + \gamma \overline{\text{ROA}}_{s,t} + u_{i,s,t}$$

  - $\overline{\text{ROA}}_{s,t}$: Average ROA among firms in state s, year t
  - $X_{i,s,t}$: Firm-level controls (e.g., leverage, size, market share)
- **Goal:** Use $\bar{y}_i$ to soak up unobserved shocks (f_i) that affect all group members.

Why *AvgE* Is Problematic

- The true model:

$$y_{i,j} = \beta X_{i,j} + f_i + \epsilon_{i,j}$$

- *AvgE* substitutes a proxy for f_i:

$$\bar{y}_i = \beta \bar{X}_i + f_i + \bar{\epsilon}_i$$

- Substituting this into the regression gives:

$$y_{i,j} = \beta X_{i,j} + \gamma(\beta \bar{X}_i + f_i + \bar{\epsilon}_i) + u_{i,j}$$

- **Two problems arise:**
  1. **Measurement error:** $\bar{y}_i$ is an imperfect proxy for f_i — it includes $\beta \bar{X}_i + \bar{\epsilon}_i$.
  2. **Endogeneity:** $X_{i,j}$ and $\bar{y}_i$ share common components $(\bar{X}_i, f_i)$, creating correlation between $X_{i,j}$ and the regression error.

Measurement Error Bias

- Since

$$\bar{y}_i = f_i + \underbrace{(\beta \bar{X}_i + \bar{\epsilon}_i)}_{\text{measurement error}} ,$$

  $\bar{y}_i$ measures f_i with noise.
- This creates **measurement error bias**:
  - As is well known, even classical measurement error causes all estimated coefficients to be inconsistent
- Bias here is complicated because error can be correlated with both mismeasured variable, f_i, and with $X_{i,j}$.

Summary of OLS, *AdjY*, and *AvgE*

$$\text{True model: } y_{i,j} = \beta X_{i,j} + f_i + \epsilon_{i,j}$$

$$\text{True model: } y_{i,j} - \bar{y}_i = \beta X_{i,j} - \beta \boxed{\bar{X}_i} + \epsilon_{i,j} - \bar{\epsilon}_i$$

$$\textit{AdjY} \text{ estimates: } y_{i,j} - \bar{y}_i = \beta^{AdjY} X_{i,j} + u_{i,j}^{AdjY}$$

$$\textit{AvgE} \text{ estimates: } y_{i,j} = \beta^{AvgE} X_{i,j} + \gamma^{AvgE} \bar{y}_i + u_{i,j}^{AvgE}$$

- All three estimators are inconsistent in the presence of unobserved group heterogeneity.
- *AdjY* and *AvgE* are not necessarily an improvement over OLS.
- *AdjY* and *AvgE* can yield estimates with opposite signs to the true coefficient.

The Differences Will Matter! Example 1 — Capital Structure

- Regression model:

$$(D/A)_{i,t} = \alpha + \beta X_{i,t} + f_i + \epsilon_{i,t}$$

  - $(D/A)_{i,t}$: Book leverage for firm i, year t
  - $X_{i,t}$: Variables affecting leverage (e.g., tangibility, size, profitability)
  - f_i: Firm fixed effect capturing unobserved, time-invariant factors
- Data: U.S. firms, 1950–2010, winsorized at 1% tails
- Goal: Compare how different estimators handle unobserved heterogeneity (f_i)

Capital Structure Regression Results

Dependent Variable: Book Leverage

Variable	OLS	*AdjY*	*AvgE*	FE
Fixed Assets / Total Assets	0.270***	0.066***	0.103***	0.248***
	(0.008)	(0.004)	(0.004)	(0.014)
Ln(Sales)	0.011***	0.011***	0.011***	0.017***
	(0.001)	(0.000)	(0.000)	(0.001)
Return on Assets	-0.015***	0.051***	0.039***	-0.028***
	(0.005)	(0.004)	(0.004)	(0.005)
Z-score	-0.017***	-0.010***	-0.011***	-0.017***
	(0.000)	(0.000)	(0.000)	(0.001)
Market-to-Book Ratio	-0.006***	-0.004***	-0.004***	-0.003***
	(0.000)	(0.000)	(0.000)	(0.000)
Observations	166,974	166,974	166,974	166,974
R^2	0.29	0.14	0.56	0.66

- Notice how *AdjY* and *AvgE* estimates differ sharply from both OLS and FE.
- For example, the profitability (ROA) coefficient flips sign under *AdjY* / *AvgE*.
- Partial corrections for heterogeneity distort inference — bias can even reverse direction.

The Differences Will Matter! Example 2 — Firm Value

- Regression model:

$$Q_{i,j,t} = \alpha + \beta X_{i,j,t} + f_{j,t} + \epsilon_{i,j,t}$$

  - $Q_{i,j,t}$: Tobin's Q for firm i, industry j, year t
  - $X_{i,j,t}$: Firm-level determinants of value (e.g., size, R&D, profitability)
  - $f_{j,t}$: Industry-year fixed effect (controls for sectoral conditions)
- Data: U.S. manufacturing firms
- Question: Do OLS, AdjY, AvgE, and FE produce consistent results?

Firm Value Regression Results

Dependent Variable: Tobin's Q

Variable	OLS	*AdjY*	*AvgE*	FE
Delaware Incorporation	0.100***	0.019	0.040	0.086**
	(0.036)	(0.032)	(0.032)	(0.039)
Ln(Sales)	-0.125***	-0.054***	-0.072***	-0.131***
	(0.009)	(0.008)	(0.008)	(0.011)
R&D Expenses / Assets	6.724***	3.022***	3.968***	5.541***
	(0.260)	(0.242)	(0.256)	(0.318)
Return on Assets	-0.559***	-0.526***	-0.535***	-0.436***
	(0.108)	(0.095)	(0.097)	(0.117)
Observations	55,792	55,792	55,792	55,792
R^2	0.22	0.08	0.34	0.37

- *AdjY* and *AvgE* substantially understate the R&D coefficient compared to FE (3.0 vs 5.5).
- Their partial corrections remove part of the true within-group variation.
- Overall fit (R^2) confirms this — FE explains far more variation, capturing persistent unobserved factors.

General Implications of the Framework

- The same logic applies well beyond firm-level panel regressions:
  - Many "adjusted" estimators implicitly assume the group mean or median removes unobserved heterogeneity.
- However, any estimator that subtracts off a noisy or endogenous benchmark still suffers from omitted-variable or measurement-error bias.
- **Examples of biased AdjY-type estimators:**
  - Subtracting the group **median** or **value-weighted mean** instead of the unobserved fixed effect.
  - Subtracting the mean outcome of a **matched control sample** (as in diversification-discount studies).
  - Comparing "adjusted" outcomes before vs. after an event (as in M&A announcement studies).
  - Using **characteristically adjusted returns** in asset pricing.
- These adjustments remove some noise but not the unobserved heterogeneity that actually drives bias.

AdjY-Type Estimators in Asset Pricing

- In empirical asset pricing, researchers often compare returns across portfolios sorted by firm characteristics.
- Returns are typically "**characteristically adjusted**":
  - Subtract the mean return of a benchmark portfolio with similar size, book-to-market, or R&D intensity.
  - $r_{i,t} - \bar{r}_{\text{benchmark},t}$ is regressed on firm characteristics.
- **Problem:** This is mathematically equivalent to *AdjY*.
  - The benchmark mean ($\bar{r}_{\text{benchmark},t}$) is a noisy, endogenous proxy for the unobserved common component (e.g., systematic factor, industry effect).
  - It does not hold constant the variation in the independent variable across benchmark portfolios.
- Hence, the adjustment does not eliminate unobserved co-movement — it may even exaggerate it.

Asset Pricing *AdjY* Example — R&D and Stock Returns

- Example: Firms sorted into quintiles by R&D intensity (R&D/MVE).
- Researchers compute "characteristically adjusted" yearly returns by subtracting industry-size benchmark means (i.e., an *AdjY* transformation).

Missing	Q1	Q2	Q3	Q4	Q5
-0.012***	-0.033***	-0.023***	-0.002	0.008	0.020***
(0.003)	(0.009)	(0.008)	(0.007)	(0.013)	(0.006)

- Benchmark portfolios: industry–size matched means of returns.
- Difference between Q5 and Q1 = 5.3 percentage points.
- Appears to suggest "high R&D firms outperform."
- But since benchmark returns correlate with firm characteristics and unobserved shocks, this inference may be spurious.

Regression Comparison: *AdjY* vs Fixed Effects

Dependent Variable: Yearly Stock Return

R&D Quintile	*AdjY* Estimate	FE Estimate
Missing	0.021**	0.030***
	(0.009)	(0.010)
Quintile 2	0.010	0.019
	(0.013)	(0.019)
Quintile 3	0.032***	0.051***
	(0.012)	(0.014)
Quintile 4	0.041***	0.068***
	(0.015)	(0.018)
Quintile 5	0.053***	0.094***
	(0.011)	(0.020)
Observations	144,592	144,592
R^2	0.00	0.40

- Regression equivalent of the previous "sorts" result.
- The FE version applies benchmark-period fixed effects to both returns and R&D — conceptually a cleaner "within" estimator.
- *AdjY* coefficients are consistently smaller in magnitude.
- R^2 is near zero under *AdjY* but large under FE — indicating that benchmark adjustment misses systematic variation.

What If *AdjY* or *AvgE* Is the True Model?

- Suppose the data truly follow the *AvgE* structure—where the **group mean outcome** directly affects each member's outcome:

$$y_{i,j} = \beta X_{i,j} + \gamma \bar{y}_i + u_{i,j}$$

  - Then $\bar{y}_i$ itself depends on $\bar{u}_i$, which includes $u_{i,j}$.
  - This creates a simultaneous relationship: individuals affect the group mean and the group mean affects individuals.
  - This is the classic **reflection problem** (Manski, 1993) — identifying peer or group effects becomes impossible without extra structure or instruments.
  - In this case, **none** of the estimators (OLS, *AdjY*, *AvgE*, or FE) recover the true β. [See Leary and Roberts (2010) for a finance application.]

What If *AdjY* or *AvgE* Is the True Model?

- Even if the researcher is interested in deviations from the group mean: $(y_{i,j} - \bar{y}_i)$, the *AdjY* estimator is only consistent if $X_{i,j}$ has no effect on $y_{i,j}$.
  - If $X_{i,j}$ influences $y_{i,j}$, then it must also influence others in the same group $(y_{i,-j})$ through correlated behavior or shared shocks.
  - Therefore, $\bar{X}_i$ also affects $(y_{i,j} - \bar{y}_i)$, implying:

$$\text{Cov}(X_{i,j}, (y_{i,j} - \bar{y}_i)) \neq 0.$$

  - In short, it is impossible for $X_{i,j}$ to affect $y_{i,j}$ but not the group deviation $(y_{i,j} - \bar{y}_i)$.
- When true interdependence exists among group members, simple "adjusted" or "demeaned" models confound cause and reflection. Identifying the effect of $X_{i,j}$ requires either instrumental variables or structural modeling of peer interactions.

Summary of Today [Part 1]

- Our data isn't perfect:
  - Watch for measurement error.
  - Watch for survivorship bias.
  - Be careful about external validity claims.
- Test that estimates across subsamples are statistically different (if you plan to claim differences).

Summary of Today [Part 2]

- Don't use *AdjY* or *AvgE*!
- Do use fixed effects:
  - Use benchmark portfolio-period FE in asset pricing rather than characteristically adjusted returns.