# Measurement Error and Misclassification: Introduction

Sections 2.1 & 2.2

Lillian Li

1

# Why Should We Care About Measurement Error?

# What Happens When We Ignore Measurement Error?

**Consequences:**

- Degraded quality of inference
- Biased parameter estimates
- Incorrect conclusions


**Common in:** Medical studies, epidemiology, survey research, longitudinal studies

# What is Measurement Error?

**Our Definition:** Any setting where the ideal measurement of a variable may differ from the actual value obtained by data collection.


**Types of Error:**

- **Systematic Error (Bias):** Repeatable, from instrument/procedure imperfections
- **Random Error:** Inherent variability, unreproducible
- **Sampling Error:** Due to using sample vs. entire population

# Why Do We Get Measurement Error?

**Common Sources:**

1. **Physical limitations:** Hard to access (e.g., coronary arteries → carotid arteries)
2. **Cost constraints:** Expensive precise measurements
3. **Nature of variable:** Some things can't be measured exactly (radiation exposure)
4. **Temporal variation:** Variables change over time
5. **Ethical considerations:** Intentional alteration for privacy ($X^* = X + e$)

**Real Example:** Cancer risk from radiation exposure - impossible to measure precisely

# Notation and Terminology

**True Variables:**

- Y: True response variable
- X: True covariate(s)
- Z: Error-free covariates

**Measured Variables:**

- X*: Surrogate/measured covariate(s)

# Illustrating Measurement Error Effects

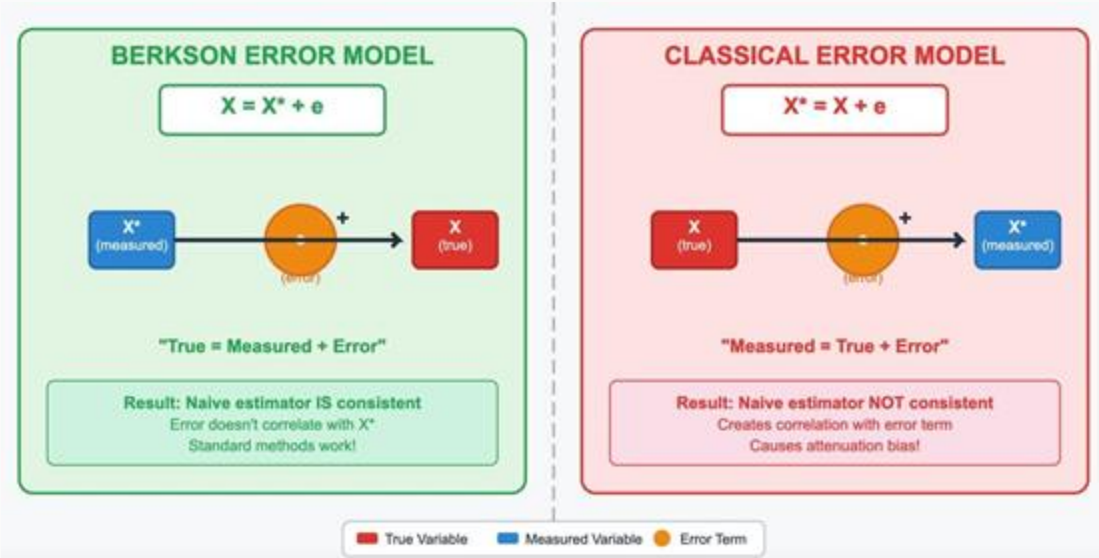
**Setup:** Simple linear regression example

- Model: $Y_i = \beta_0 + \beta_x X_i + \varepsilon_i$
- True covariate $X_i$ not observed
- Surrogate $X_i^*$ available instead

**Key Questions:**

1. What happens when we use $X_i^*$ instead of $X_i$?
2. Does the relationship between $X_i$ and $X_i^*$ matter?
3. Can we still get valid estimates?

# Two Measurement Error Models




**Model 1 (Berkson Error):** $X_i = X_i^* + e_i$

- True value = Measured value + Error
- Naive estimator IS consistent

**Model 2 (Classical Error):** $X_i^* = X_i + e_i$

- Measured value = True value + Error
- Naive estimator IS NOT consistent

# The Attenuation Effect

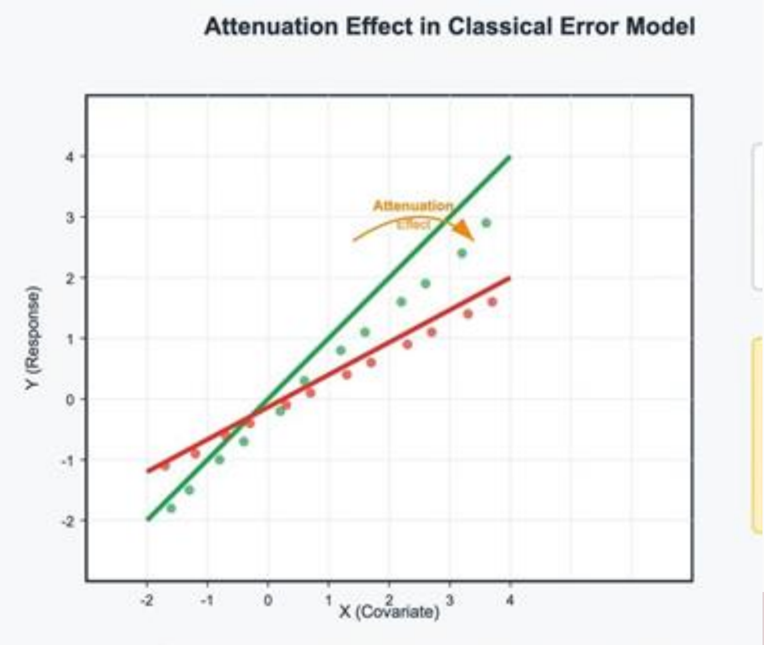
**Under Classical Error Model ($X_i^* = X_i + e_i$):**

**Naive estimator converges to:** $\beta^*_x = \omega\beta_x$

**Where reliability ratio:** $\omega = \sigma^2_x/(\sigma^2_x + \sigma^2_e) \leq 1$

**Result: Attenuation bias** - estimates are shrunk toward zero

**Why this happens:** Error in $X_i^*$ creates correlation between predictor and error term in the regression




Classical measurement error causes the slope to be attenuated (shrunk toward zero)

The stronger the measurement error (smaller ω), the flatter the observed relationship

# Simulation Results - Classical Error

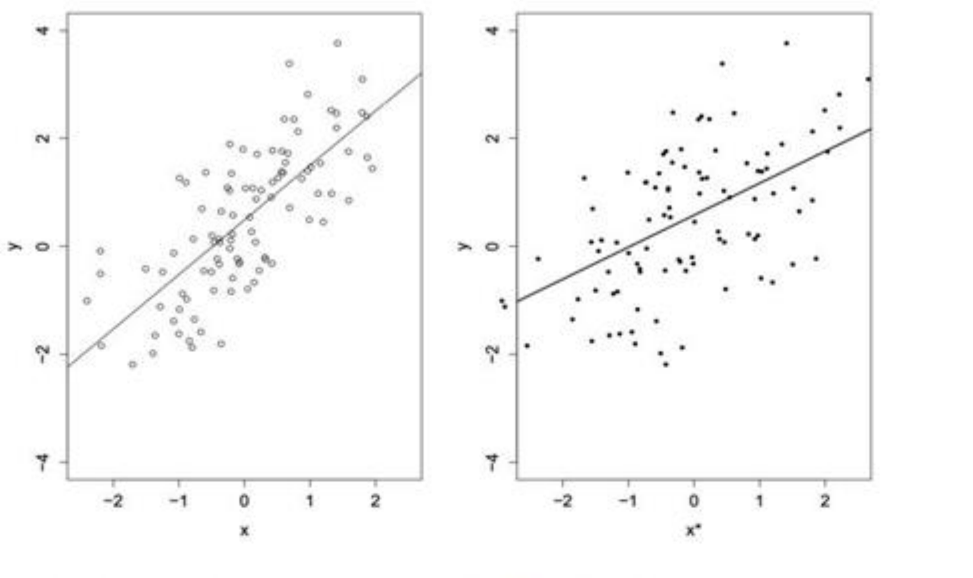



**Fig. 2.1.** *Effects of Measurement Error Model (2.3) on Simple Linear Regression*

**Simulation Setup:**

- $n = 100$, $\beta_0 = 0.5$, $\beta_x = 1.0$
- $X_i \sim N(0,1)$, $e_i \sim N(0,1)$
- Model: $X_i^* = X_i + e_i$

**Results:**

- Left panel: True data $(X_i, Y_i)$ - steeper slope
- Right panel: Surrogate data $(X_i^*, Y_i)$ - flatter slope
- Clear attenuation effect visible

# Berkson Error Results

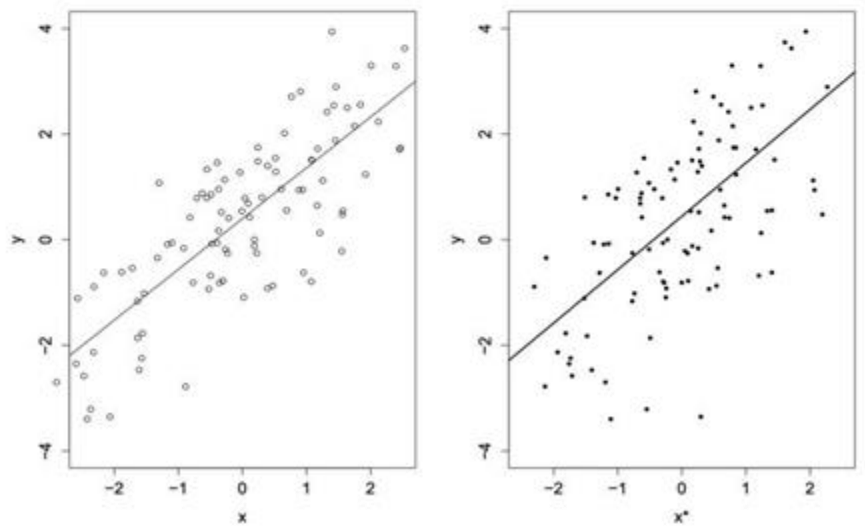



Fig. 2.2. *Effects of Measurement Error Model (2.2) on Simple Linear Regression*

**Simulation with Berkson Error:**

- Model: $X_i = X_i^* + e_i$
- Same simulation parameters

**Results:**

- Both panels show parallel regression lines
- Naive estimator remains consistent
- Only variance of error term changes

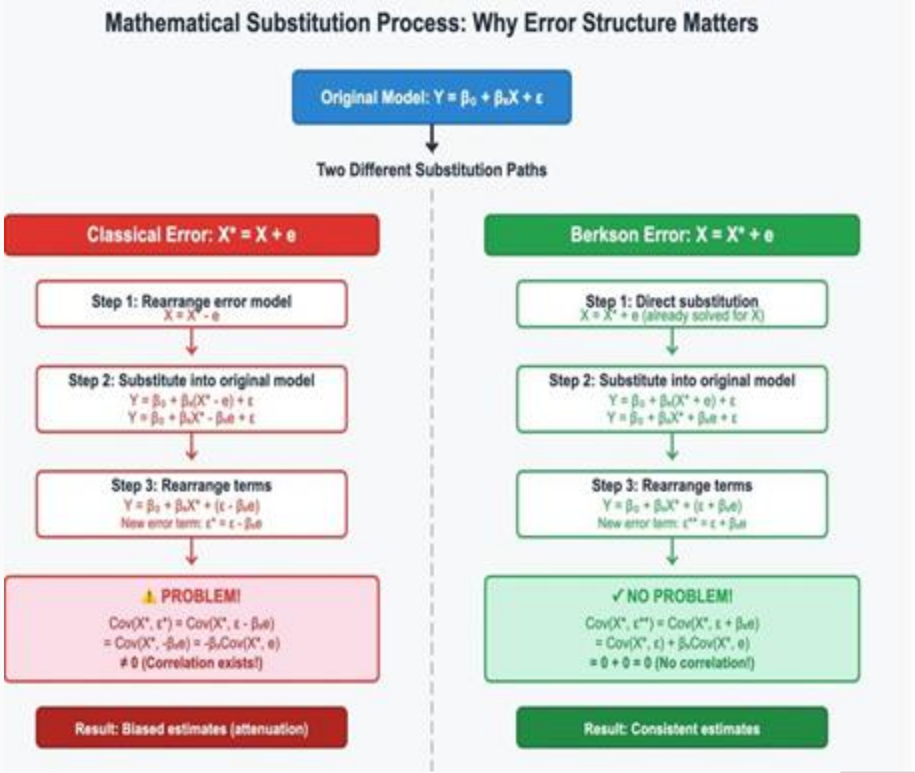
# Why the Difference?

**Classical Error ($X_i^* = X_i + e_i$):**

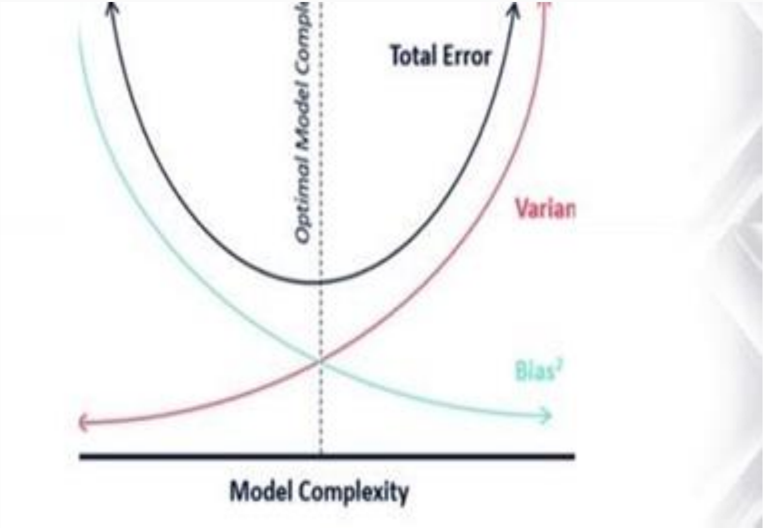
- Substituting into $Y_i = \beta_0 + \beta_x X_i + \varepsilon_i$ gives:
- $Y_i = \beta_0 + \beta_x X_i^* + \varepsilon_i^*$ where $\varepsilon_i^* = \varepsilon_i - \beta_x e_i$
- Problem: biased estimates; $\varepsilon_i$ is correlated with $X_i$!

**Berkson Error ($X_i = X_i^* + e_i$):**

- Substituting gives: $Y_i = \beta_0 + \beta_x X_i^* + \varepsilon_i^{**}$ where $\boldsymbol{\varepsilon}_i^{**} = \varepsilon_i + \beta_x e_i$
- Good news: consistent estimates; $\varepsilon_i^{**}$ is uncorrelated with $X_i^*$



# Practical Implications




**Key Lessons:**

1. **Measurement error matters** - can severely bias results
2. **Error structure is crucial** - determines if naive analysis works
3. **Attenuation is common** - classical error shrinks estimates toward zero
4. **Problem-specific analysis needed** - no universal solution

**Trade-offs:**

- Naive estimators: Biased but lower variance
- Corrected estimators: Unbiased but higher variance
- Need sufficient sample size for correction to be worthwhile

# Looking Ahead

**This was just the beginning:**

- Simple linear regression with one covariate
- More complex models have additional complications
- Multiple covariates create multidimensional problems
- Nonlinear models can reverse signs, not just attenuate

**General Principle:** The nature and degree of measurement error effects depend on:

- Response model form
- Measurement error model structure
- Variable relationships and variability

**Next steps:** Explore correction methods and more complex scenarios

# Summary

**Main Points:**

1. Measurement error is ubiquitous and consequential
2. The relationship between true and measured variables determines bias
3. Classical error ($X^* = X + e$) causes attenuation bias
4. Berkson error ($X = X^* + e$) preserves consistency
5. Ignoring measurement error can lead to wrong conclusions

**Bottom Line:** Understanding measurement error structure is essential for valid statistical inference.

# Thanks!