Log(0)

STEMFORALL 2025

Students: Aurona Wang, Sweeney Luan,
Grace Brandt, Yujia Hu
Supervisor: Curt Signorino

Roadmap

Background

Project question

Past solutions

Our approach

Analytical and MC results

Classical Linear Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Black Voter Registration = β_0 + β_1 Income + β_2 Education + β_3 Industry + β_4 Poll Tax + ϵ
Socioeconomic Predictor

Democracy level = β_0 + β_1 log(GDP) + β_2 Education_PC + β_3 Fractionalization + ϵ
Institutional Indicator

Log models

Linear-Log → $y = \beta_0 + \beta_1 \ln(x) + \epsilon$

Log-Linear → $\ln(y) = \beta_0 + \beta_1 x + \epsilon$

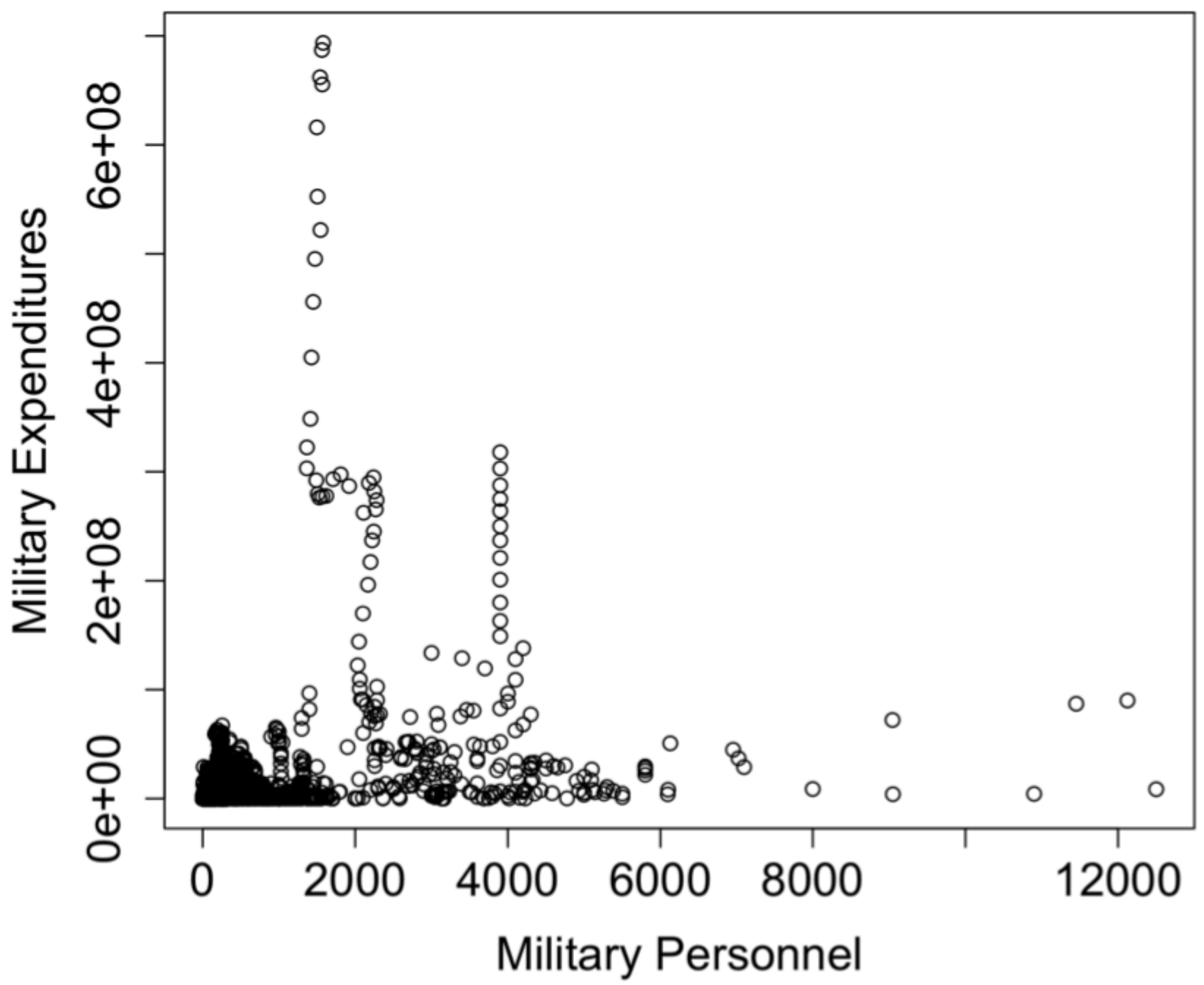
Log-Log → $\ln(y) = \beta_0 + \beta_1 \ln(x) + \epsilon$

log-log, log-linear, linear-log models

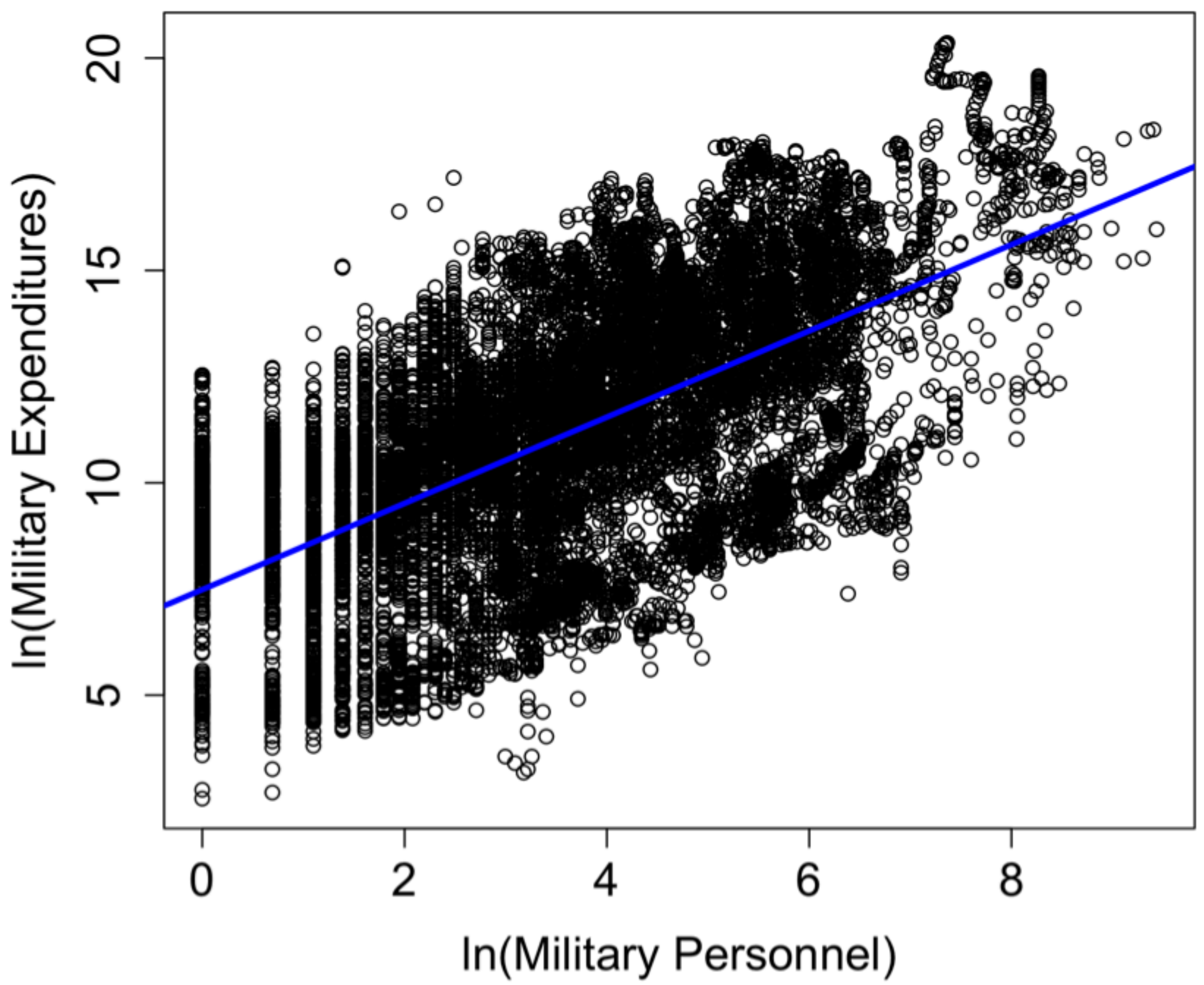
Can arise for theoretical reasons — e.g., Economics

Log transforms make right-skewed variables more symmetric empirically and stabilize variance and allow interpretation of corfficients as approximate percentage changes

Raw data



Log-transformed



Problem: What if our variable has zeros?

ln(0) = undefined

R: log(0) = -Inf

Log models

$$y = \beta_0 + \beta_1 \ln(x) + \epsilon$$ ← Linear-Log Regression

$$\ln(y) = \beta_0 + \beta_1 x + \epsilon$$ ← Log-Linear Regression

$$\ln(y) = \beta_0 + \beta_1 \ln(x) + \epsilon$$ ← Log-Log Regression

Past "solutions"

Delete all observations with a ln(0)?

- Throws out a lot of data
- Often those are very interesting observations, can lose out on potential patterns in zeros
- Produces a truncated dataset, which requires a different estimator

Add a small constant to all observations: ln(x+c) or ln(y+c)

- Existing research shows this can bias estimates
- Option to estimate constant as its own parameter for less bias, but more complicated and not well-studied

Create another transformation: Inverse Hyperbolic Sine

Are we even modeling the process correctly?

- Solution may differ depending on what we think DGP is
- Normality/skew isn't always the best indicator for the true relationship

Our approach

Rethink how 0's appear in log(x)

DGP1: $y_i = \beta_0 + \beta_1 \log\left[x_i + D\, z_i\right] + \epsilon_i$

DGP2: $y_i = \beta_0 + \beta_1 \log(x_i)\,(1 - z_i) + \beta_2\, z_i + \epsilon_i$

$$z_i = \begin{cases} 0, & \text{if } x \neq 0 \\ 1, & \text{if } x = 0 \end{cases}$$

Claim 1: From an estimation perspective, these DGP's are observationally equivalent.

Our approach

Rethink how 0's appear in log(x)

DGP1: $y_i = \beta_0 + \beta_1 \log\left[x_i + D\, z_i\right] + \epsilon_i$

DGP2: $y_i = \beta_0 + \beta_1 \log(x_i)\,(1 - z_i) + \beta_2\, z_i + \epsilon_i$

Claim 2: Using OLS with the estimating equation

Est1: $y_i = B_0 + B_1 \log(x_i + d\, z_i) + B_2 z_i + \epsilon_i$

$\hat{B}_1$ **is an unbiased estimate of** β_1 **in either DGP above.**

Est1: $y_i = B_0 + B_1 \log(x_i + d\, z_i) + B_2 z_i + \epsilon_i$

DGP1: $y_i = \beta_0 + \beta_1 \log\left[x_i + D\, z_i\right] + \epsilon_i$

$$E(\hat{B}_0) = \beta_0$$
$$E(\hat{B}_1) = \beta_1$$
$$E(\hat{B}_2) = \beta_1 \log(D/d)$$

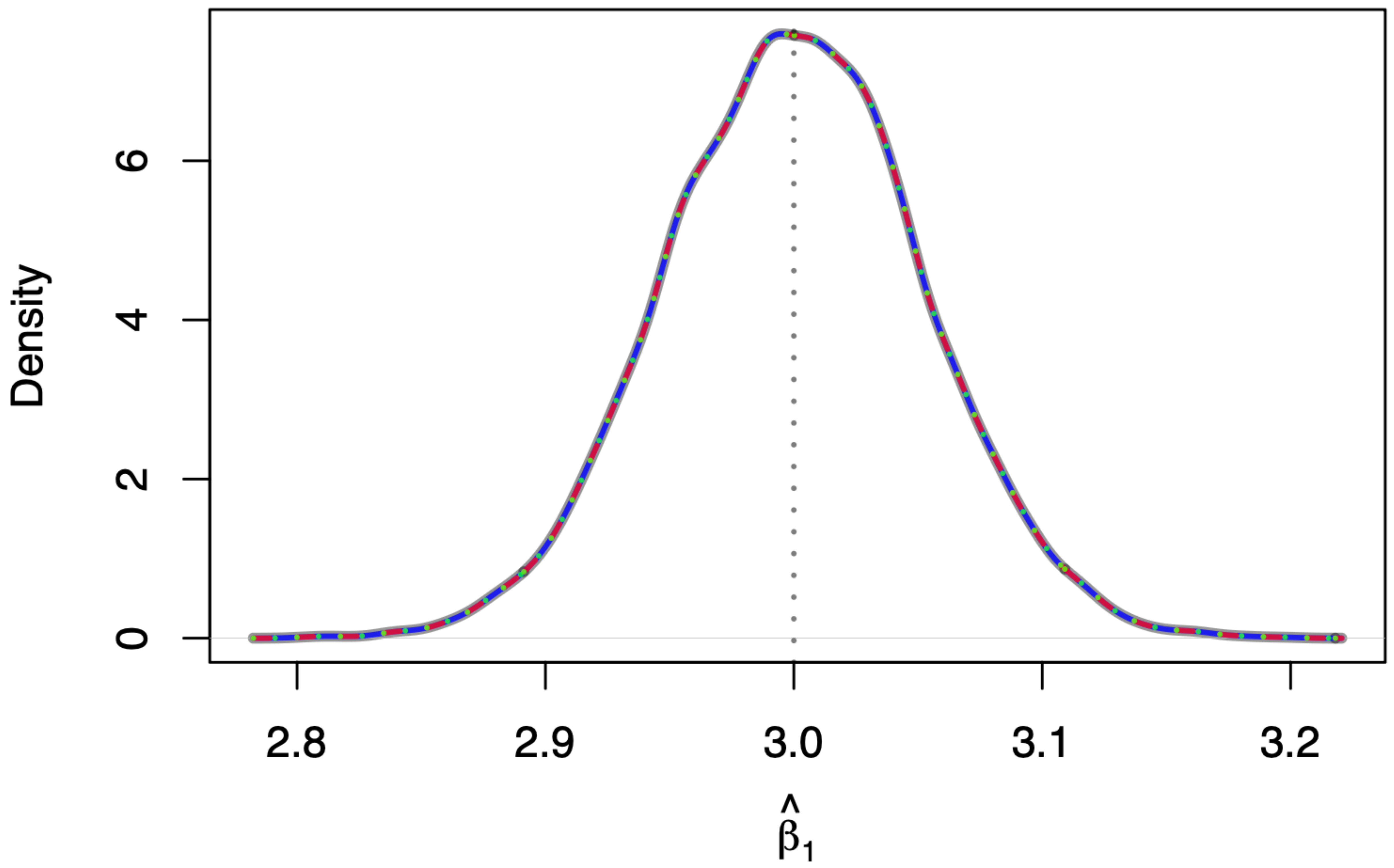
DGP2: $y_i = \beta_0 + \beta_1 \log(x_i)\,(1 - z_i) + \beta_2\, z_i + \epsilon_i$

$$E(\hat{B}_0) = \beta_0$$
$$E(\hat{B}_1) = \beta_1$$
$$E(\hat{B}_2) = \beta_2 - \beta_1 \log(d)$$

DGP1: $y_i = \beta_0 + \beta_1 \log \left[x_i + D\, z_i\right] + \epsilon_i$

DGP2: $y_i = \beta_0 + \beta_1 \log(x_i)\,(1 - z_i) + \beta_2\, z_i + \epsilon_i$

Est1: $y_i = B_0 + B_1 \log(x_i + d\, z_i) + B_2 z_i + \epsilon_i$

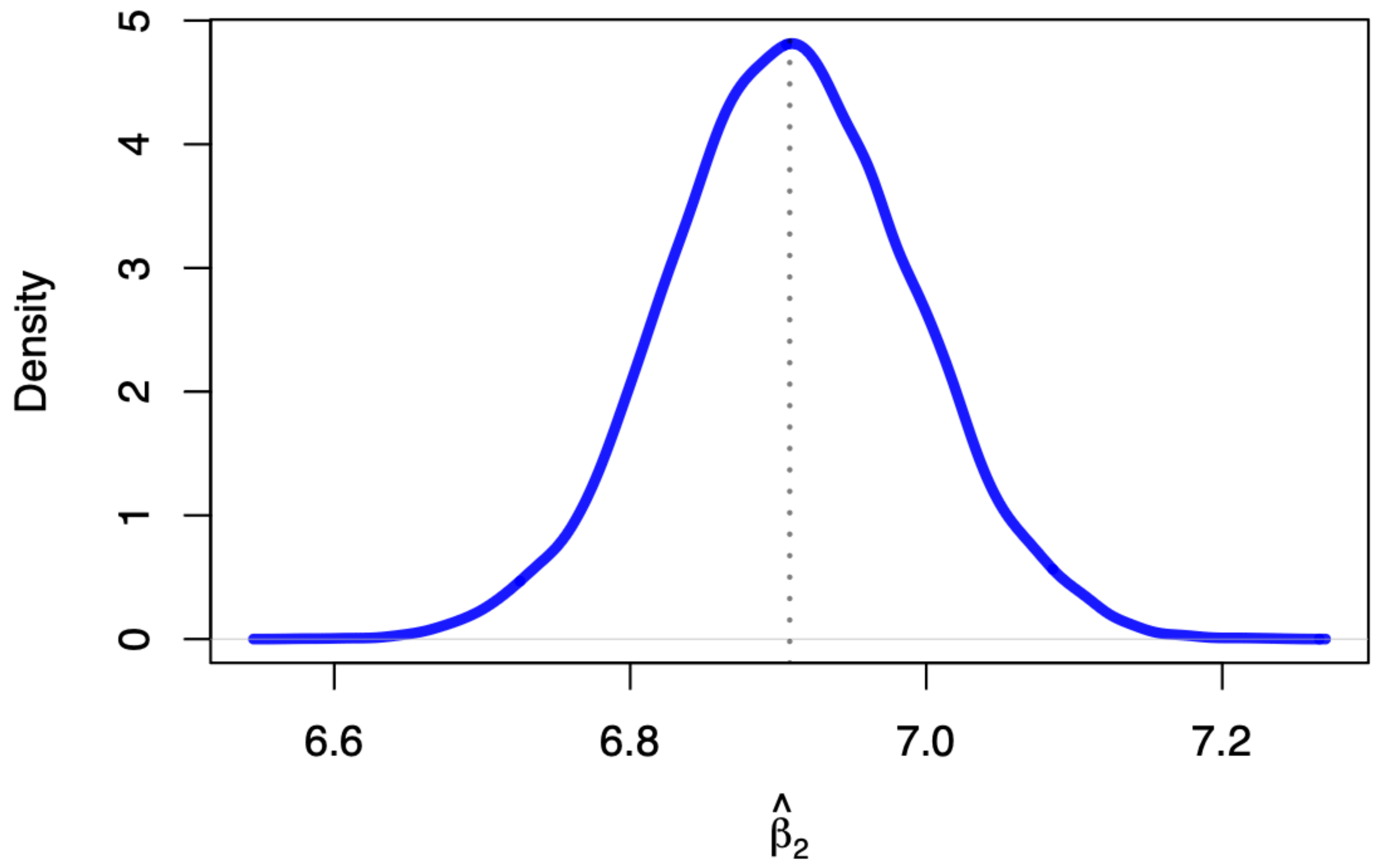


DGP1: $y_i = \beta_0 + \beta_1 \log\left[x_i + D\, z_i\right] + \epsilon_i$

Est1: $y_i = B_0 + B_1 \log(x_i + d\, z_i) + B_2 z_i + \epsilon_i$

Does the predicted value of β_2 match the estimated value?

$E(\hat{B}_2) = \beta_1 \log(D/d)$



Analysis of misspecified model

DGP1: $y_i = \beta_0 + \beta_1 \log\left[x_i + D\, z_i\right] + \epsilon_i \quad \longrightarrow$

$$y_i = \beta_0 + \beta_1 \log\left[x_i + dz_i\right] + \beta_2 z_i + \epsilon_i$$

Est1: $y_i = B_0 + B_1 \log(x_i + d\, z_i) + \epsilon_i$

Claim 3: Given the DGP above, omitting the dummy variable z induces omitted variable bias in $\hat{B}_1$.

Based on the formula for omitted variable bias, the estimated value of $\hat{B}_1$ should be

$$\hat{B}_1 = \hat{\beta}_1 + \hat{\beta}_2 \frac{\text{cov}(\log(x+dz), z)}{\text{var}(\log(x+dz))}$$

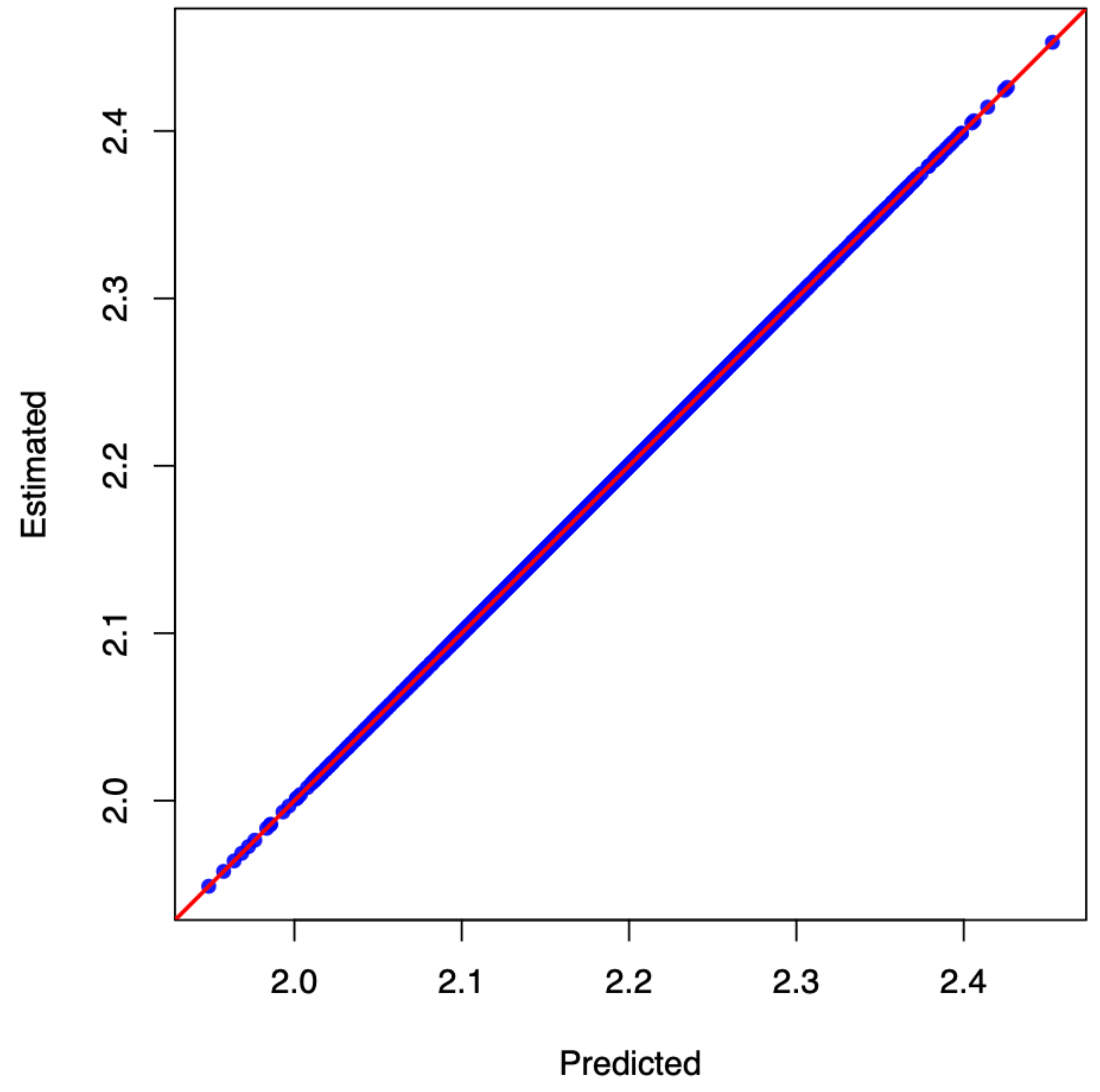
Analysis of misspecified model

DGP1: $y_i = \beta_0 + \beta_1 \log[x_i + D z_i] + \epsilon_i \quad \longrightarrow \quad y_i = \beta_0 + \beta_1 \log[x_i + dz_i] + \beta_2 z_i + \epsilon_i$

Est1: $y_i = B_0 + B_1 \log(x_i + d z_i) + \epsilon_i$

Based on the formula for omitted variable bias, the estimated value of $\widehat{B}_1$ should be

$$\widehat{B}_1 = \widehat{\beta}_1 + \widehat{\beta}_2 \frac{\text{cov}(\log(x+dz),z)}{\text{var}(\log(x+dz))}$$



Future Work

- Applications
  - Implementation on real datasets
  - Replication and comparison of studies using other solutions, like log(x+1)

- Interpretation of x in different fields
  - Biomedicine
  - Economics
  - Political Science

- Optimization of picking D

- Generalization

Summary

Approach: rethink how 0's are generated in log(x)

Two DGP's that are observationally equivalent

Estimation technique that recovers coefficient on log(x) term

Can express problem as a form of omitted variable bias