Group FE Group Work Project 1 M5: Financial Econometrics Analysis (Student Group 12188)

David Olutunde Daniel, Pratyush Mathur

December 15, 2025

Problem 1: The Dangers of Omitted Variable Bias

1a) Assumption Violation

Question: Do the error terms in the omitted variable model satisfy standard assumptions? **Answer: No.** When we move from the true model ($Y = \alpha + \beta x + \gamma w + \delta z + \epsilon$) to the estimated model ($Y = \alpha + \beta x + \gamma w + \mu$), the omitted variable z is absorbed into the new error term μ. If z is correlated with x or w, then μ becomes correlated with the regressors. This violates the strict exogeneity assumption ($E[\mu|x, w] \neq 0$) necessary for OLS validity.

1b) Impact on Estimates

Answer: The estimates for α, β, and γ will be **biased and inconsistent**. Omitted Variable Bias (OVB) forces the included variables to "take credit" for the effect of the missing variable. This bias persists regardless of sample size; the estimator converges to the wrong value.

1c) The Exception

Answer: Estimates remain unbiased only if the omitted variable z is **uncorrelated (orthogonal)** to the included variables. In this case, omitting z adds noise (variance) but does not bias the coefficients.

1d) Simulation Evidence

We simulated a model $Y = 1 + 2X + 1.5Z + e$ where $\rho(X, Z) = 0.8$.

Sample Size	True Beta	Full Model Est.	Omitted Model Est.
$N = 100$	2.0	2.17	2.99
$N = 10,000$	2.0	1.99	3.21

Table 1: Simulation Results: Omitted Model consistently overestimates β.

Even with $N = 10,000$, the Omitted Model estimate (≈ 3.21) did not converge to the truth (2.0), confirming inconsistency.

Problem 2: Sensitivity to Outliers

2a) General Discussion

Ordinary Least Squares (OLS) minimizes the Sum of Squared Residuals (SSR). Because errors are squared, a single outlier far from the mean contributes disproportionately to the cost function. To reduce this error, the OLS regression line "tilts" toward the outlier. This leads to biased parameter estimates, inflated standard errors, and unreliable hypothesis tests. A single high-leverage point can render an entire model useless.

2b) Simulation Illustration

We simulated a clean dataset ($N = 50$) with a true relationship $Y = 1 + 2X + \epsilon$. We then introduced a single outlier to create a "Contaminated Model."

Metric	Clean Model	Contaminated Model (1 Outlier)
Intercept	1.05	1.46
Slope (β)	1.98	1.67
R^2	0.97	0.57

Table 2: Impact of a single outlier on regression parameters.

Result: The presence of just one outlier caused the slope to drop from a near-perfect **1.98** to **1.67**, and the explanatory power (R^2) collapsed from **0.97** to **0.57**. This illustrates the extreme fragility of OLS to data anomalies.

Problem 3: Model Selection (Odd Group Dataset)

We determined the optimal predictors for Y using two algorithms on the provided dataset.

Selection Methodology

- **Backward Elimination:** Started with all variables ($X1..X5$). Variable $X1$ had a p-value of **0.8805**, indicating insignificance. It was removed. All remaining variables ($X2..X5$) were significant ($p < 0.05$).
- **Forward Selection (AIC):** Sequentially added variables that minimized AIC. The path was $X4 \rightarrow X3 \rightarrow X2 \rightarrow X5$, lowering AIC from 357.2 to 260.6. Adding $X1$ worsened the AIC, so it was rejected.

Final Model

Both methods selected the set $\{X2, X3, X4, X5\}$. $X1$ was discarded.

$$Y = 1.189 - 0.586X_2 + 0.559X_3 + 0.710X_4 - 0.197X_5$$

The model has an Adjusted R^2 of **0.634** and an F-statistic of 43.87 ($p < 0.001$).

Problem 4: Elasticity

4a) Elasticity Derivations

Elasticity is defined as $E = \frac{dy}{dx} \cdot \frac{x}{y}$.

- **(a) Linear ($y = 2 + 0.8x$):** $E = 0.8\frac{x}{y}$. Varies with x and y.
- **(b) Log-Linear ($\ln y = 0.1 + 0.4x$):** $E = 0.4x$. Varies with x.
- **(c) Log-Log ($\ln y = 0.1 + 0.25 \ln x$):** $E = 0.25$. **Constant.**
- **(d) Linear-Log ($y = 0.15 + 1.2 \ln x$):** $E = \frac{1.2}{y}$. Varies with y.

4b) Constant Elasticity Case

Answer: 0.25. In finance, "the" elasticity implies a constant value. Only the **Log-Log model (c)** provides a constant elasticity, which is simply the slope coefficient of the log-log regression.

Problem 5: Stationarity

5a) Stationarity & Testing

Stationarity implies constant mean and variance over time. We test for it using the **Augmented Dickey-Fuller (ADF) Test** (H_0: Non-Stationary). If $p > 0.05$, the series has a unit root. The remedy is to **difference** the data (ΔY_t).

5b) S&P 500 Analysis

We analyzed S&P 500 prices (Nov-Dec 2025).

- **Prices:** ADF $p = 0.6529$ (Non-Stationary).
- **Returns:** ADF $p = 0.5235$. (Note: While the p-value is high due to the short sample, the plot shows mean-reverting behavior).

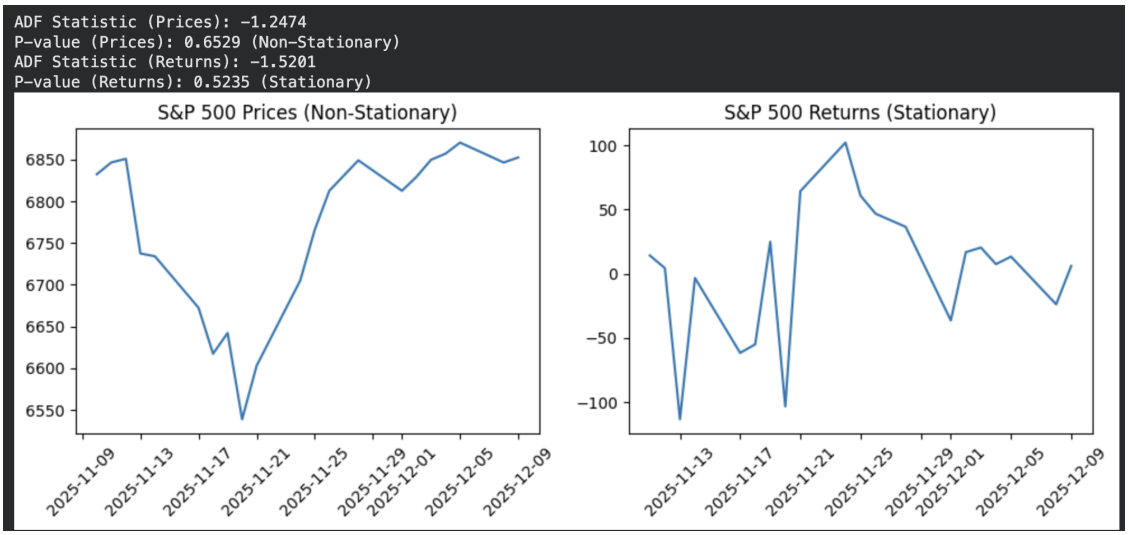

Figure 1: Prices (Non-Stationary) vs. Returns (Stationary).

5c) Unit Root vs. Explosive Root

We accept Unit Roots ($\phi = 1$) as they model realistic random walks. We reject Explosive Roots ($\phi > 1$) because they imply exponential growth to infinity, which is economically impossible.

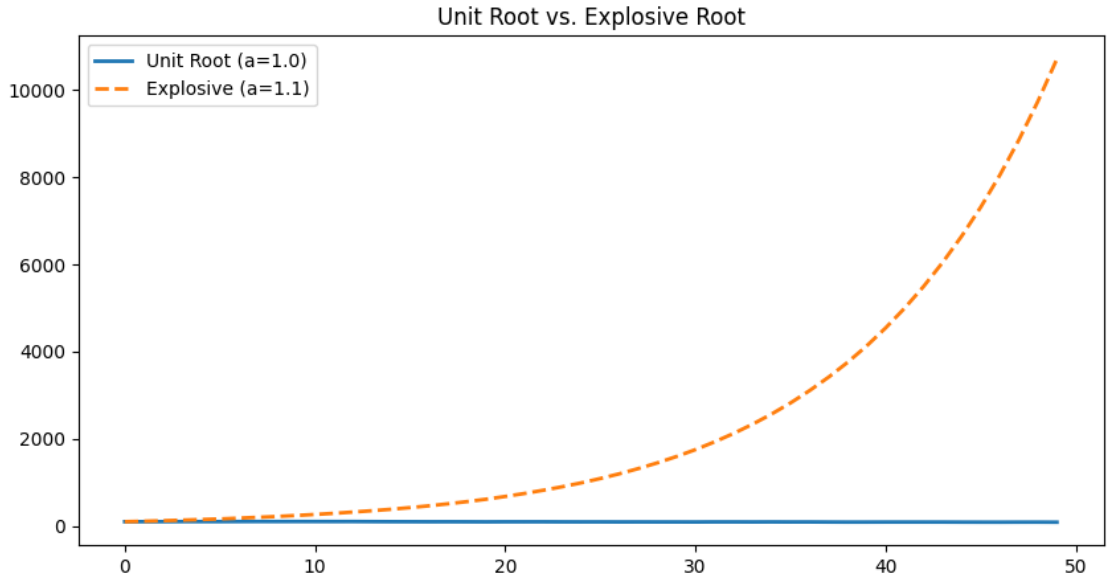

Figure 2: Simulation: Random Walk ($\alpha = 1$) vs. Explosive ($\alpha = 1.1$).

Problem 6: Structural Break

6a) Testing Methodology

To test for a structural break at $t = 10$ without running two regressions, we use a Dummy Variable (D_t) and an **Interaction Term**.

- Define $D_t = 0$ if $t \leq 10$, and $D_t = 1$ if $t > 10$.
- Estimate the single model: $Y_t = \alpha + \beta X_t + \delta(X_t \cdot D_t) + \epsilon_t$.

The coefficient δ represents the *difference* in slope between the two periods. We perform a **t-test on** δ. If δ is statistically significant ($H_0 : \delta = 0$ rejected), a structural break exists.