ECON 0150 | Economic Data Analysis

The economist's data analysis pipeline.

Part 4.3 | Regression Assumptions, Multiple Sample Tests

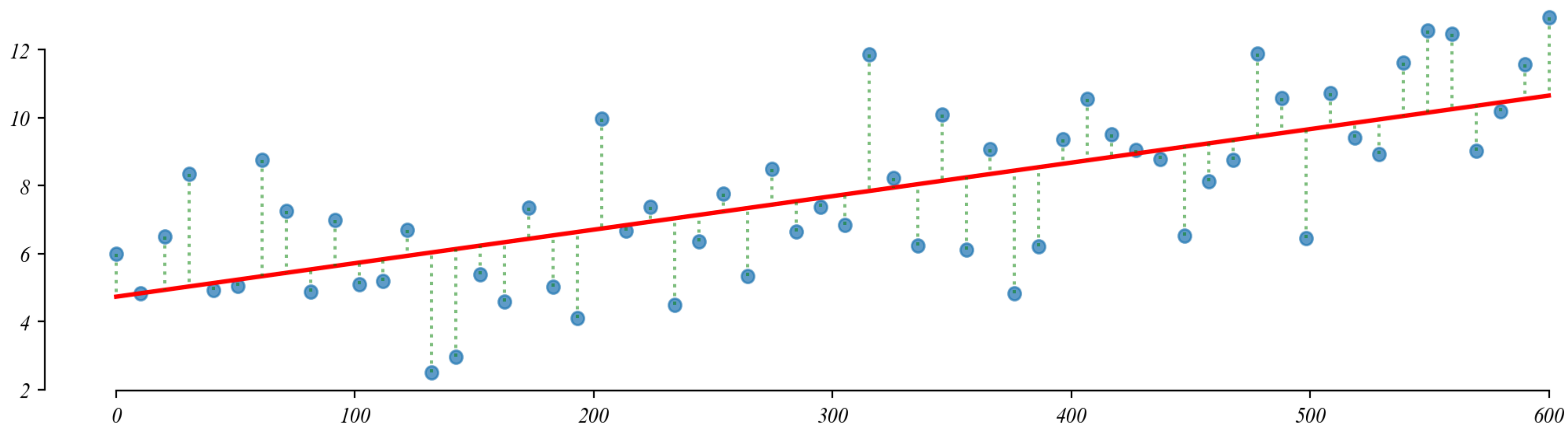
General Linear Model

... a flexible approach to run many statistical tests.

The Linear Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

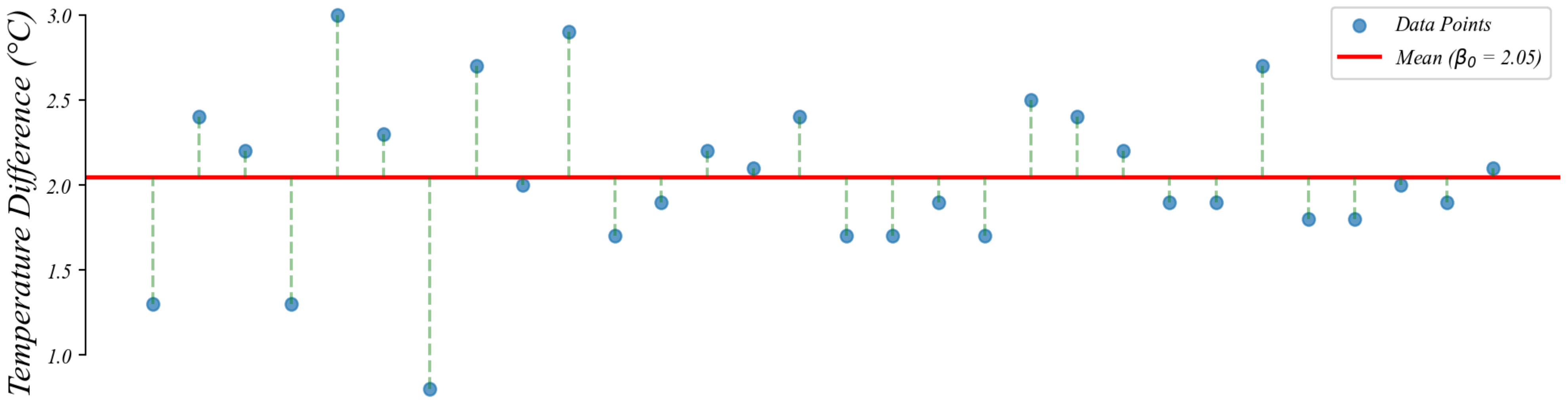
- β_0 *is the intercept (value of $\bar{y}$ when $x = 0$)*
- β_1 *is the slope (change in y per unit change in x)*
- ε_i *is the error term (random noise around the model)*

OLS Estimation: Minimizes $\sum_{i=1}^{n} \varepsilon_i^2$



One-Sample T-Test

A one-sample t-test is a horizontal line model.



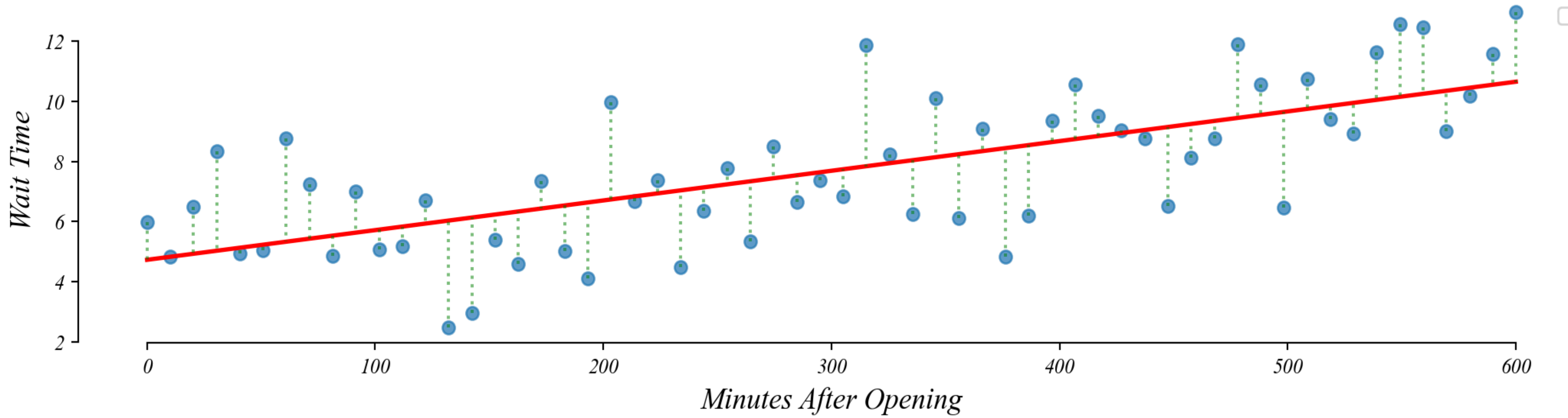
$$Temperature = \beta_0 + \varepsilon$$

> *the intercept β_0 is the estimated mean temperature*

> *the p-value is the probability of seeing β_0 if the null is true*

Relationships Between Variables

A test of relationships is a slope model.



$$\text{WaitTime} = \beta_0 + \beta_1 \text{MinutesAfterOpening} + \epsilon$$

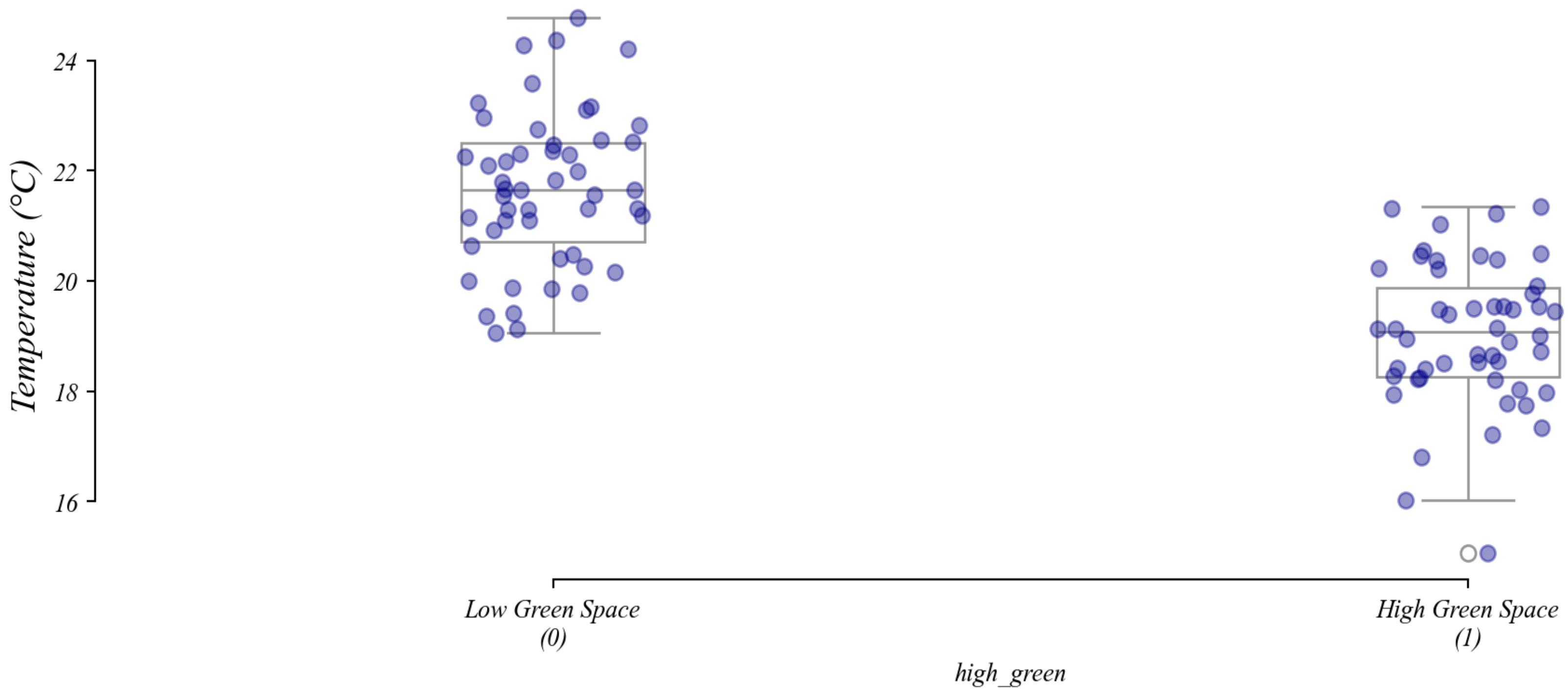
> *the intercept parameter β_0 is the estimated temperature at 0 on the horizontal*

> *the slope parameter β_1 is the estimated change in y for a 1 unit change in x*

> *the p-value is the probability of seeing parameter (β_0 or β_1) if the null is true*

New Setting: Two Samples

Is temperature lower with more green space?



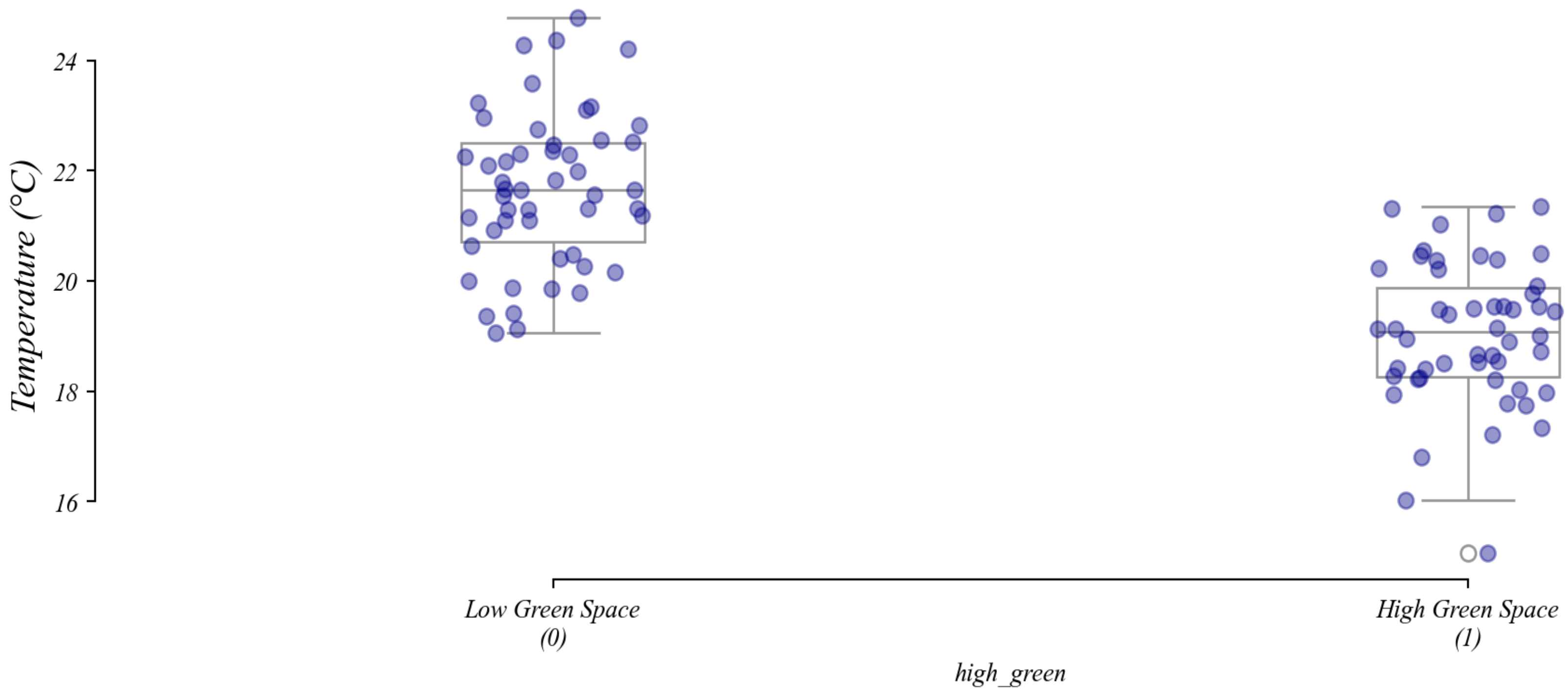
$$Temperature = \beta_0 + \beta_1 \cdot HighGreen + \varepsilon$$

> *how would we interpret* β_0 *here?*

> *the average temperature at* $x = 0$*, which is Low Green Space locations*

New Setting: Two Samples

Is temperature lower with more green space?



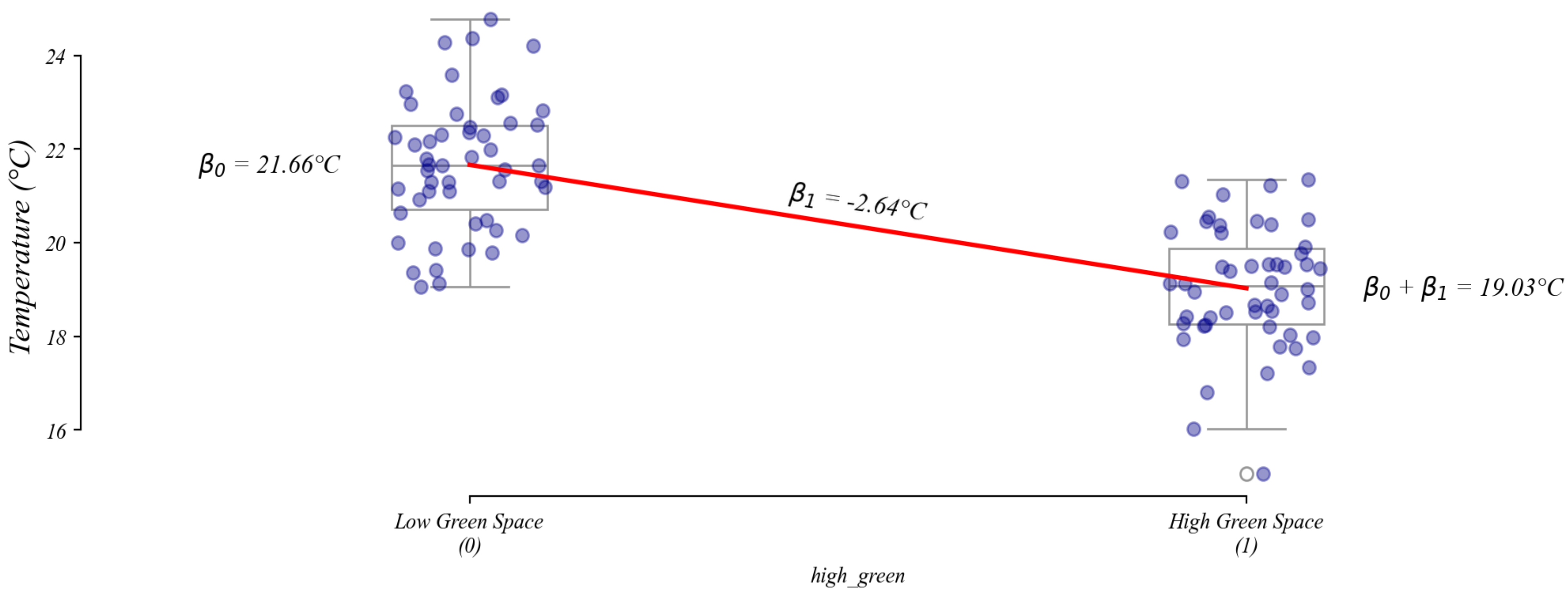
$$Temperature = \beta_0 + \beta_1 \cdot HighGreen + \varepsilon$$

> *how would we interpret* β_1 *here?*

> *one unit increase in* x*, which puts us in High Green Space*

New Setting: Two Samples

Is temperature lower with more green space?



$$Temperature = \beta_0 + \beta_1 \cdot HighGreen + \varepsilon$$

> β_0 *is the mean temperature in low green space cities (22.03°C)*

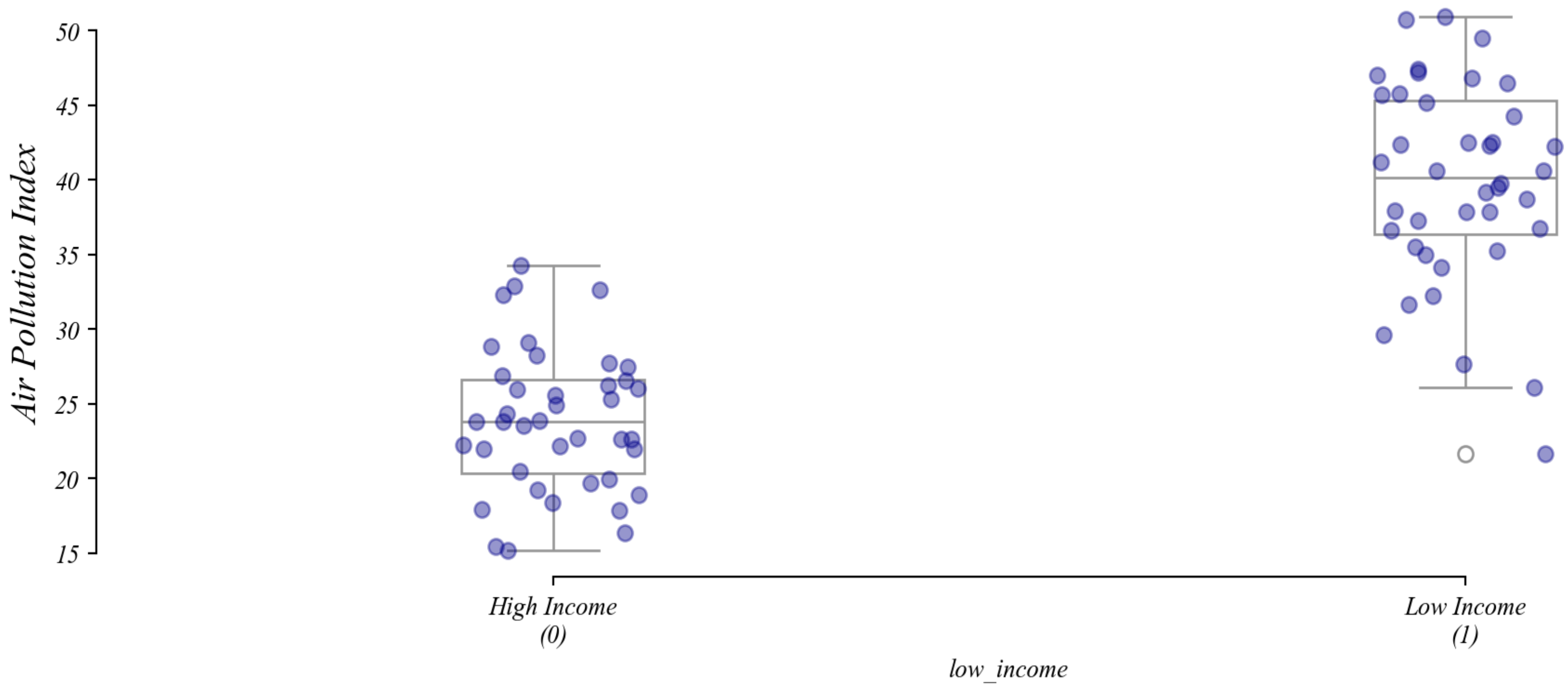
> β_1 *is the temperature difference in high green space cities (-3.02°C)*

> *the t-test on β_1 tests if this difference is significant*

Example: Neighborhood Income and Pollution

Do low-income neighborhoods face higher pollution levels?

Step 1: Summarize the data



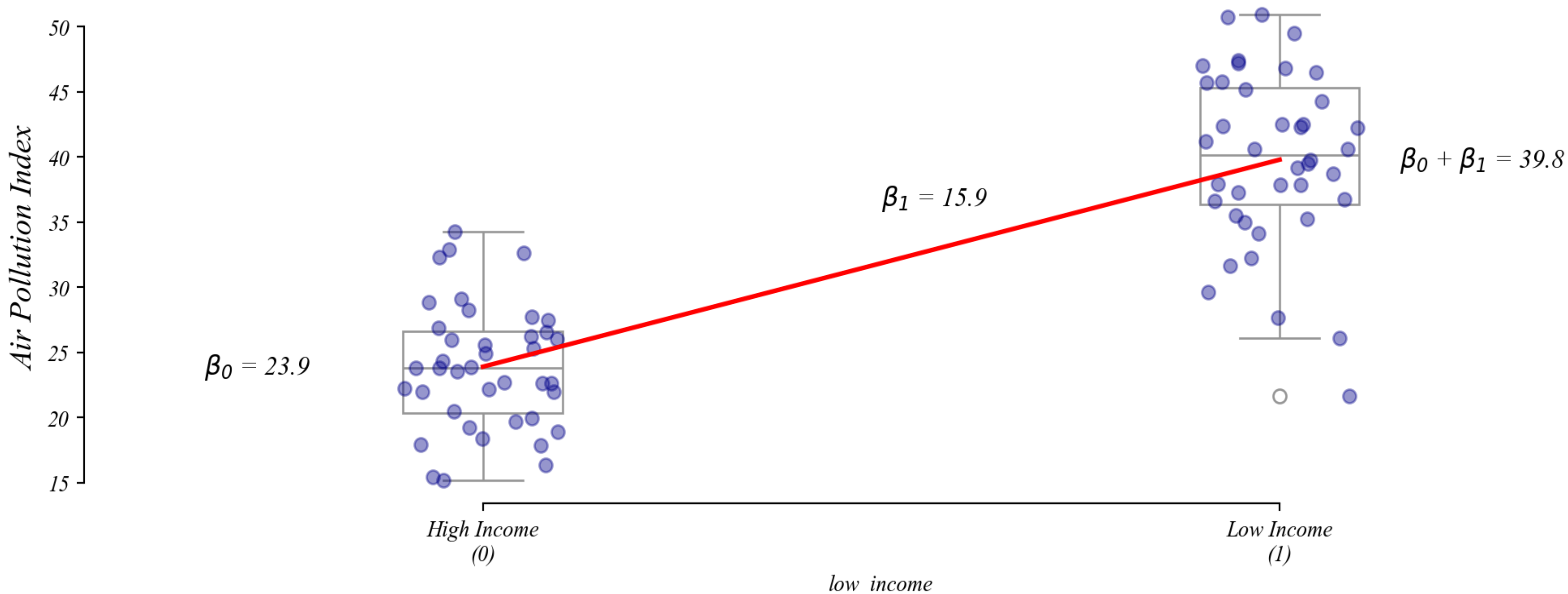
Step 2: Build a model

$$Pollution = \beta_0 + \beta_1 \cdot LowIncome + \varepsilon$$

Example: Neighborhood Income and Pollution

Do low-income neighborhoods face higher pollution levels?

Step 3: Estimate the model

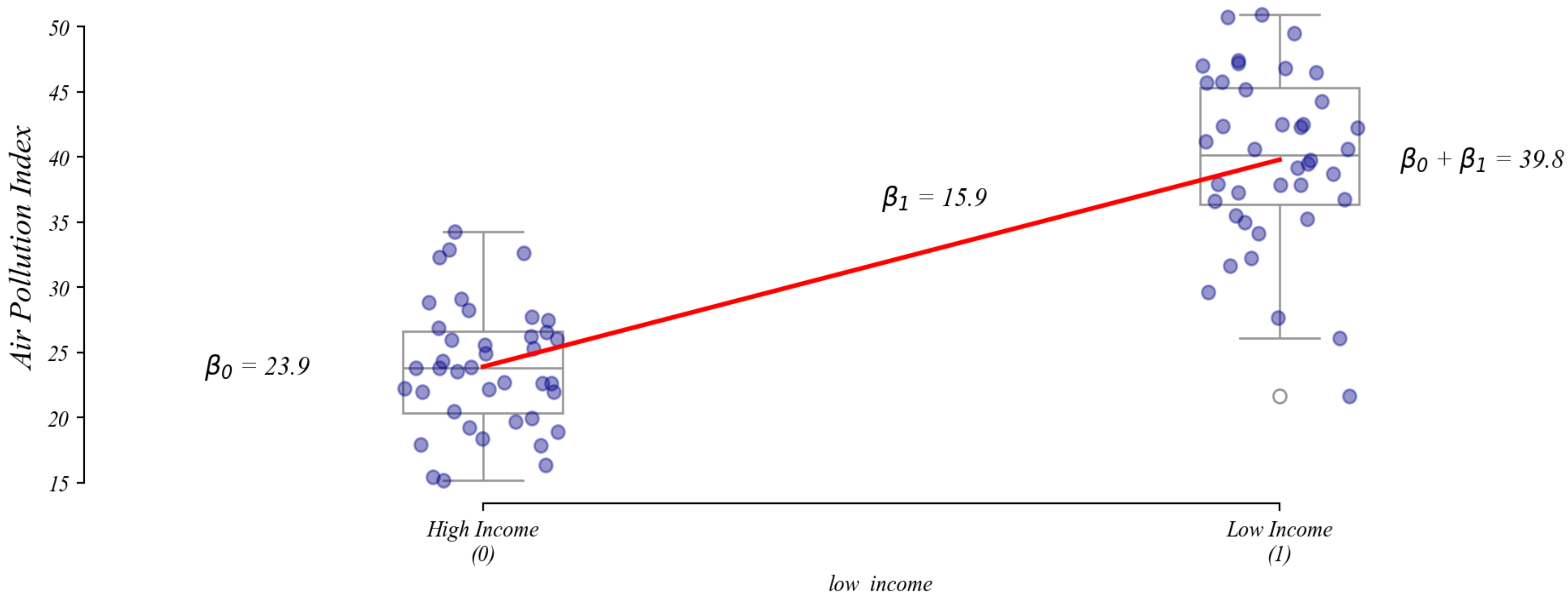


- β_0 = *Mean pollution in high-income areas (24.8)*
- β_1 = *Additional pollution in low-income areas (+15.0)*

Example: Neighborhood Income and Pollution

Do low-income neighborhoods face higher pollution levels?

Step 4: Interpret and communicate the findings



> A significant positive β_1 suggests environmental quality differences between neighborhoods

OLS Assumptions

Our test results are only valid when the model assumptions are valid.

1. ***Linearity***: *The relationship between X and Y is linear*
2. ***Independence***: *Observations are independent from each other*
3. ***Homoskedasticity***: *Equal error variance across all values of X*
4. ***Normality***: *Errors are normally distributed*

Model Diagnostics: Why Check Assumptions?

Assumption violations affect our inferences

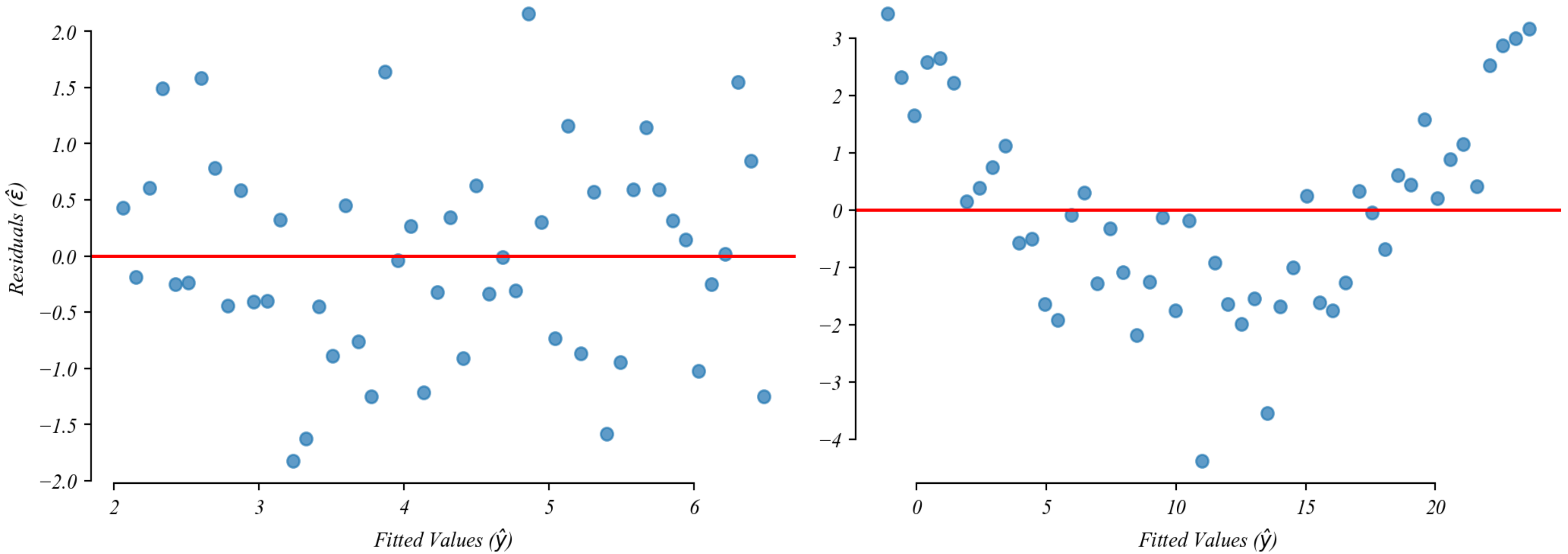
If assumptions are violated:

- *Coefficient estimates may be biased*
- *Standard errors may be wrong*
- *p-values may be misleading*
- *Predictions may be unreliable*

Checking for Linearity

The error term should be unrelated to the fitted value.

> *which one of these figures shows linearity?*



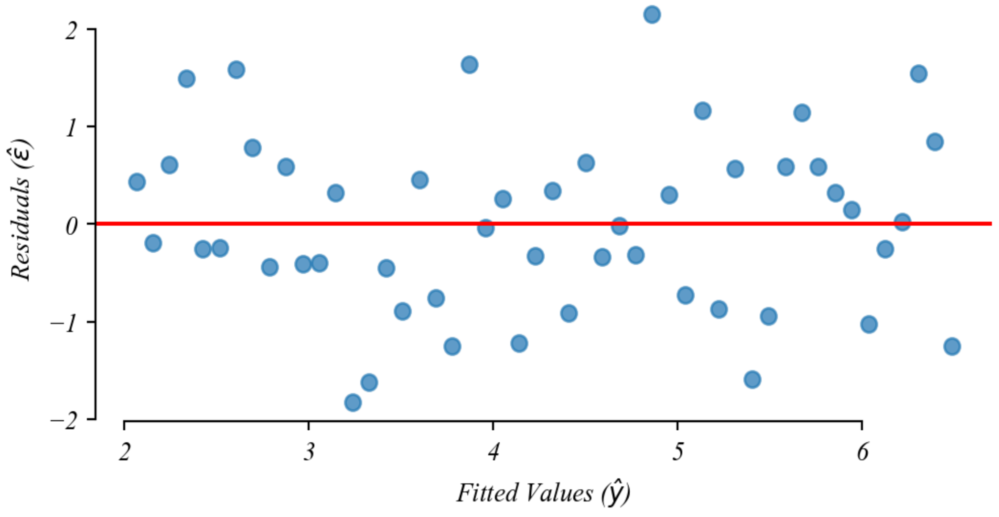
> *the left one is what we want to see*

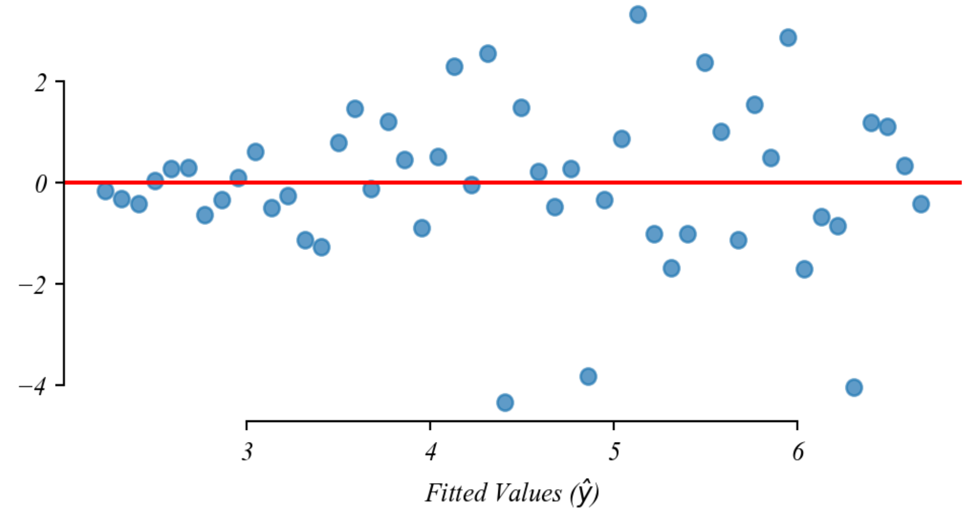
> *residual plots should show that the model is equally wrong everywhere*

Checking for Homoskedasticity

Residuals should be spread out the same everywhere.

> *which one of these figures shows homoskedasticity?*





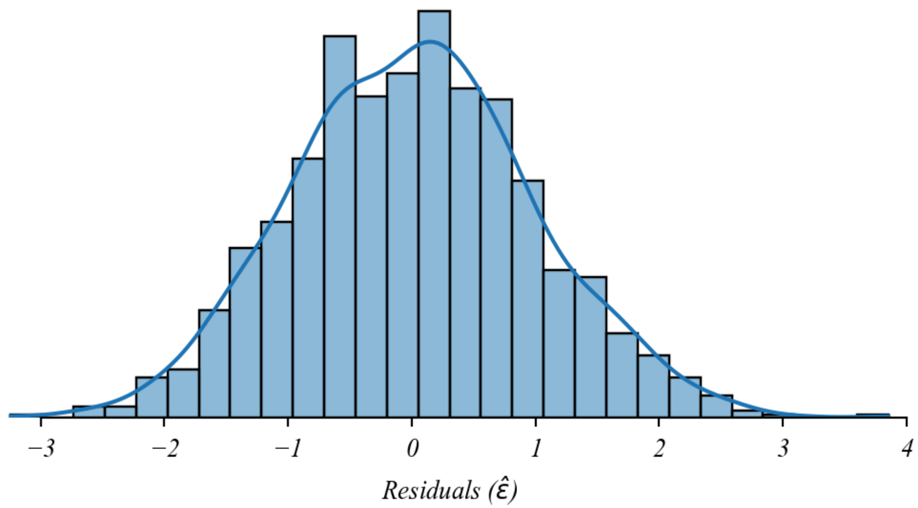
> *the left figure shows constant variability (homoskedasticity)*

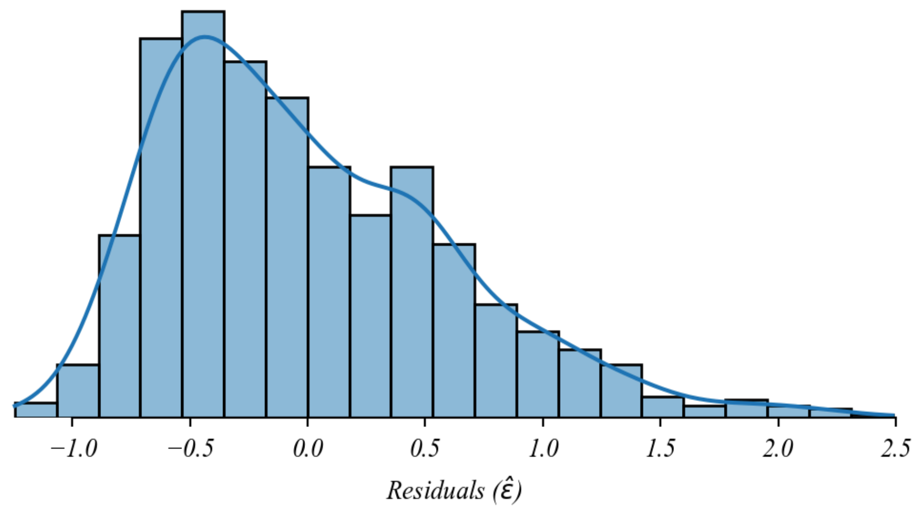
> *the right one has increasing variability (heteroskedasticity)*

> *residual plots should show that the model is equally wrong everywhere*

Checking for Normality

Residuals should be normally distributed





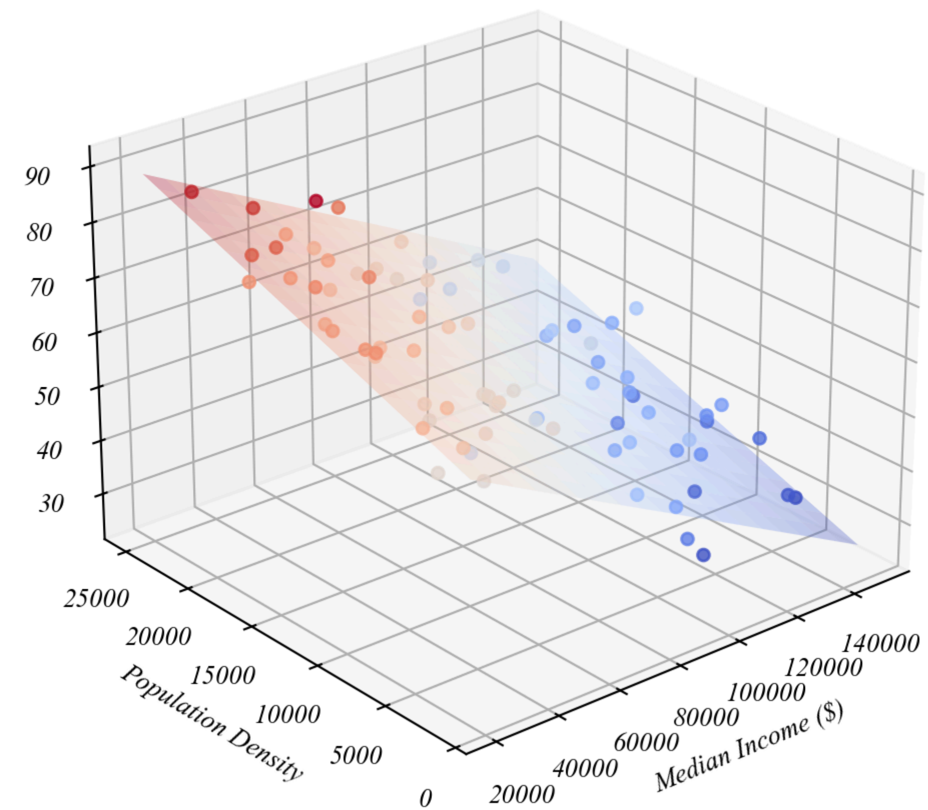
> *left shows a nice bell shape (roughly normally distributed)*

> *right shows a skewed distribution (not normally distributed)*

> *by the CLT we can still use regression without this if the sample is large*

Extending to Multiple Regression

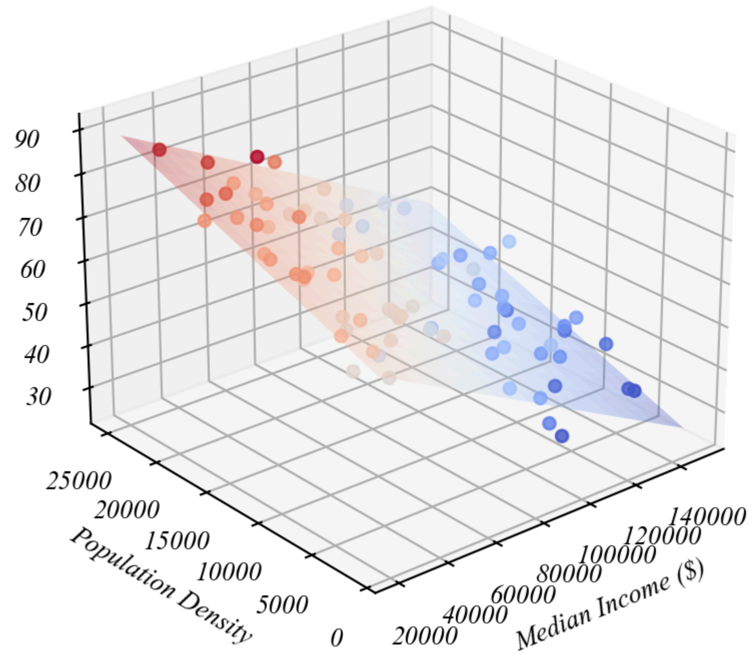
Adding control variables to isolate relationships



$$Pollution = \beta_0 + \beta_1 \cdot Income + \beta_2 \cdot Density + \varepsilon$$

Extending to Multiple Regression

Adding control variables to isolate relationships



- β_0 = *Baseline pollution level (70.0)*
- β_1 = *Effect of income on pollution, holding density constant (-0.0003)*
- β_2 = *Effect of density on pollution, holding income constant (+0.001)*

Key Takeaways

Regression provides a unified framework for statistical testing

One-Sample T-Test: Continuous outcome variable (y) with only an intercept

$$y = \beta_0 + \varepsilon$$

Relationships: Continuous outcome variable (y) with a continuous predictor (x)

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Two-Sample T-Test: Continuous outcome variable (y) with a dummy ($Group$)

$$y = \beta_0 + \beta_1 \cdot Group + \varepsilon$$

Multiple Regression: Adding control variables to isolate relationships

> all use the same OLS framework and interpretation of coefficients and p-values