# ECON 0150 | Economic Data Analysis

*The economist's data analysis pipeline.*

## *Part 4.3 | Regression Assumptions, Multiple Sample Tests*

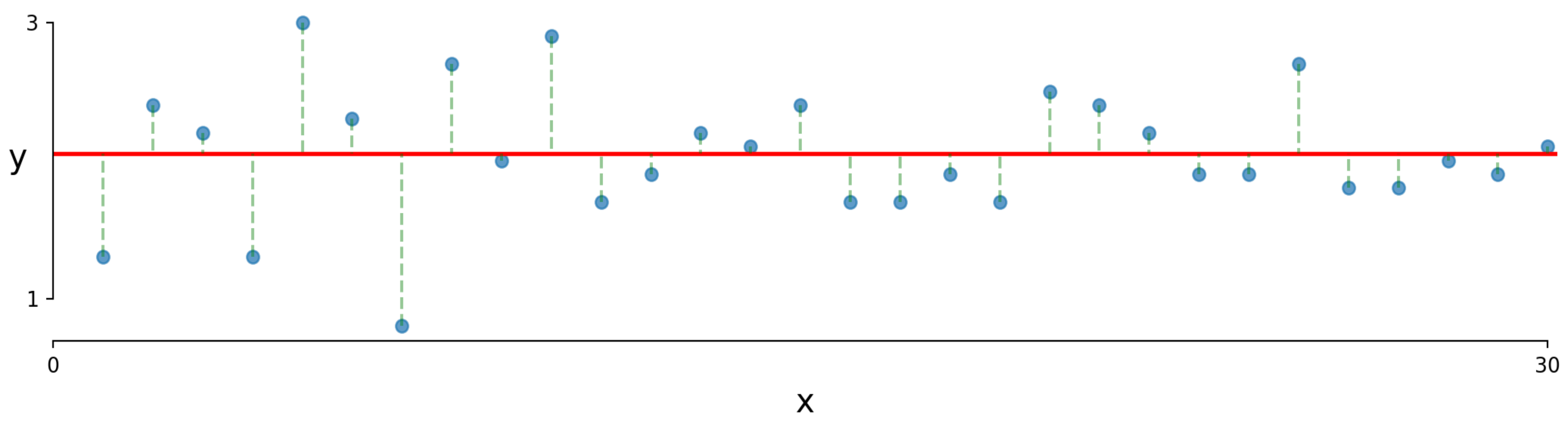
# Regression: Key Concepts

*A regression is a flexible way to run many statistical tests.*

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

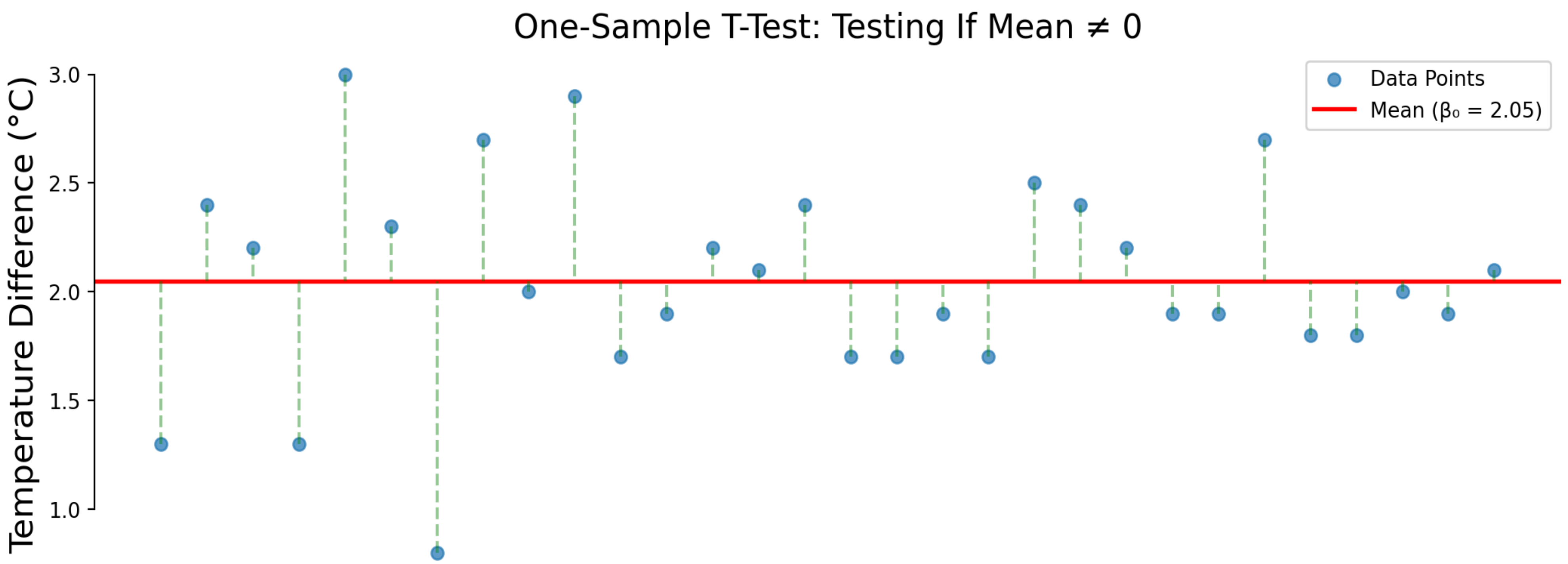
- $\beta_0$ *is the intercept (value of* $\bar{y}$ *when* $x = 0$*)*
- $\beta_1$ *is the slope (change in y per unit change in x)*
- $\varepsilon_i$ *is the error term (random noise around the model)*

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



# T-Tests Using Regression

*One-sample t-test as a horizontal line model*




> *Model:* $Temperature = \beta_0 + \varepsilon$

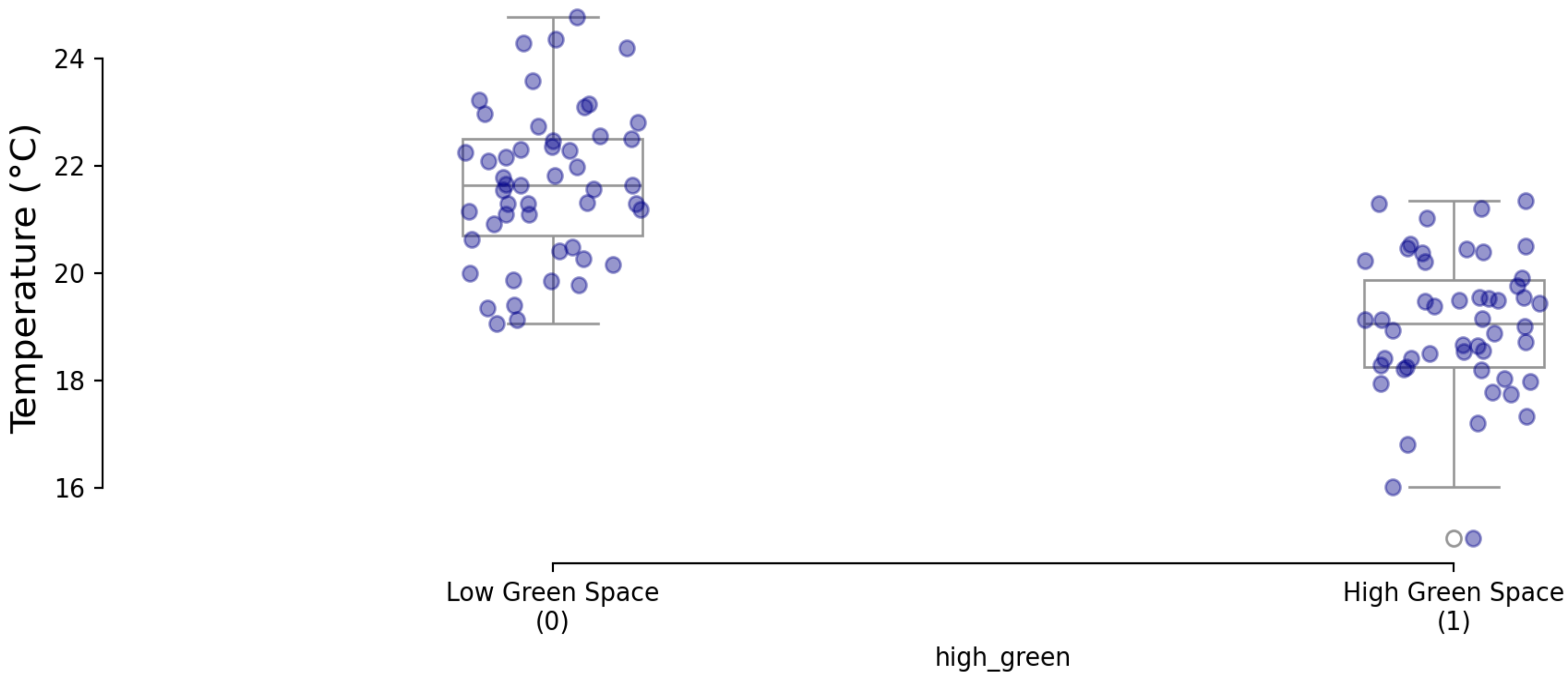
> *Interpretation: The intercept* $\beta_0$ *is the estimated mean temperature*

> *Green lines: Residuals (difference between data and mean)*

> *The t-test checks if* $\beta_0$ *is significantly different from zero*

# Example: Two-Sample t-Test Using Regression

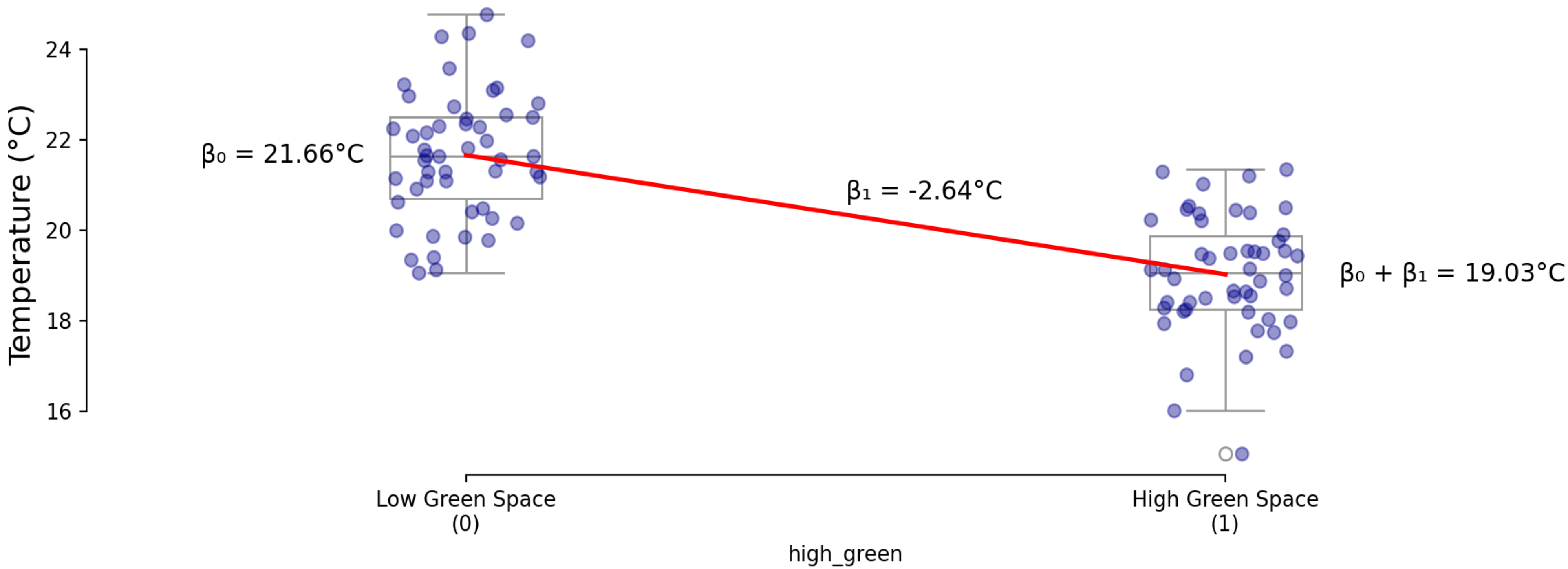
*Is temperature lower with more green space?*




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

# Example: Two-Sample t-Test Using Regression

*Model:* $Temperature = \beta_0 + \beta_1 \cdot HighGreen + \varepsilon$



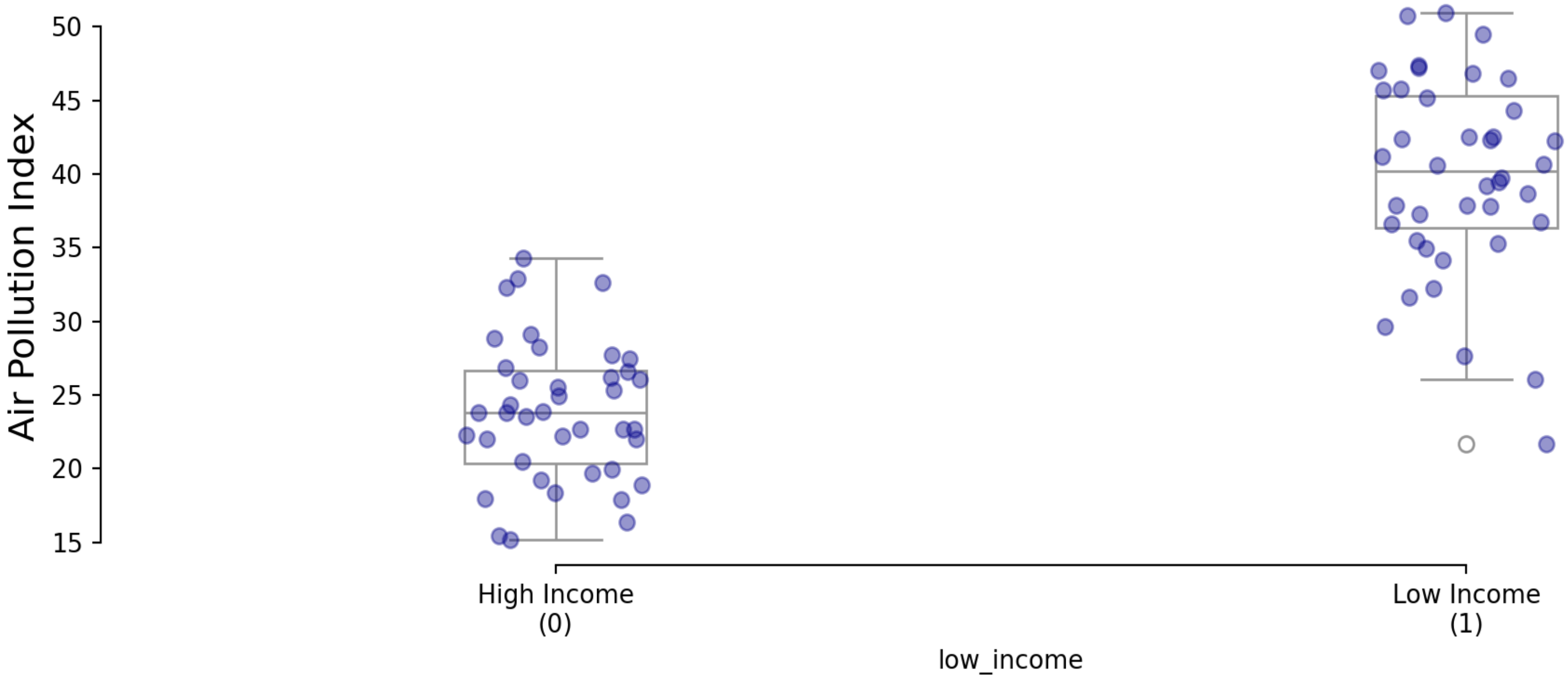

> *The t-test on* $\beta_1$ *tests if this difference is significant*

$\beta_0$ = Mean temperature in low green space cities (22.03°C)

$\beta_1$ = Temperature difference in high green space cities (-3.02°C)

# Example: Neighborhood Income and Pollution

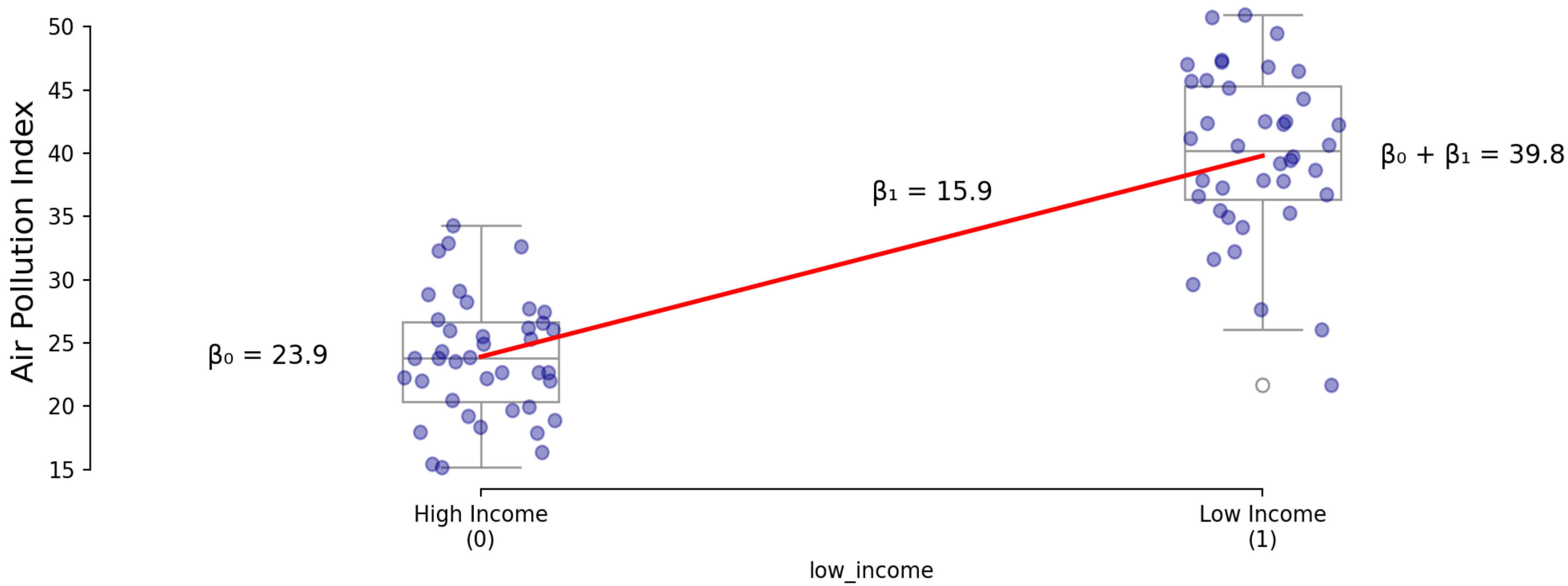
*Do low-income neighborhoods face higher pollution levels?*




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

# Example: Neighborhood Income and Pollution

*Model:* $Pollution = \beta_0 + \beta_1 \cdot LowIncome + \varepsilon$

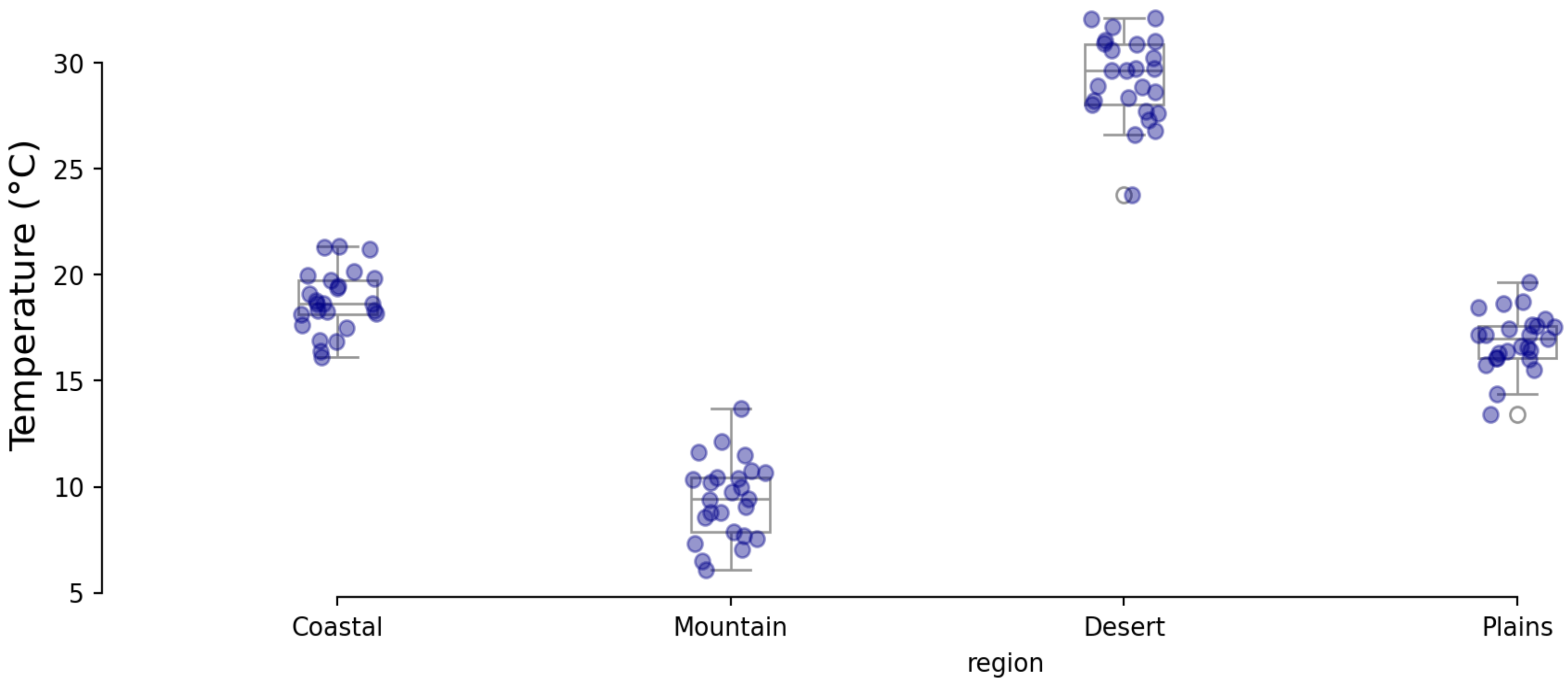



*> A significant positive $\beta_1$ suggests environmental quality differences between neighborhoods*

- $\beta_0$ = *Mean pollution in high-income areas (24.8)*
- $\beta_1$ = *Additional pollution in low-income areas (+15.0)*

# Example: Comparing Many Regions

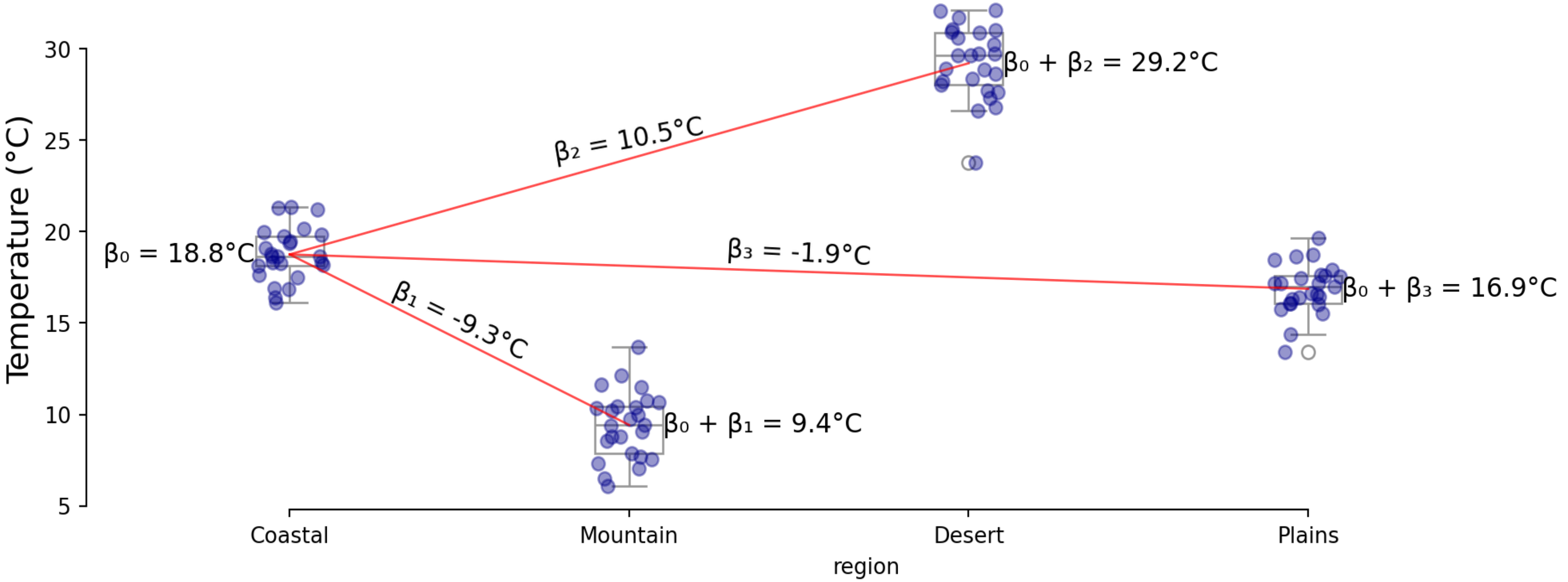
*How does temperature differ across climate regions?*




$$Temperature = \beta_0 + \beta_1 \cdot Mountain + \beta_2 \cdot Desert + \beta_3 \cdot Plains + \varepsilon$$

# Example: Comparing Many Regions

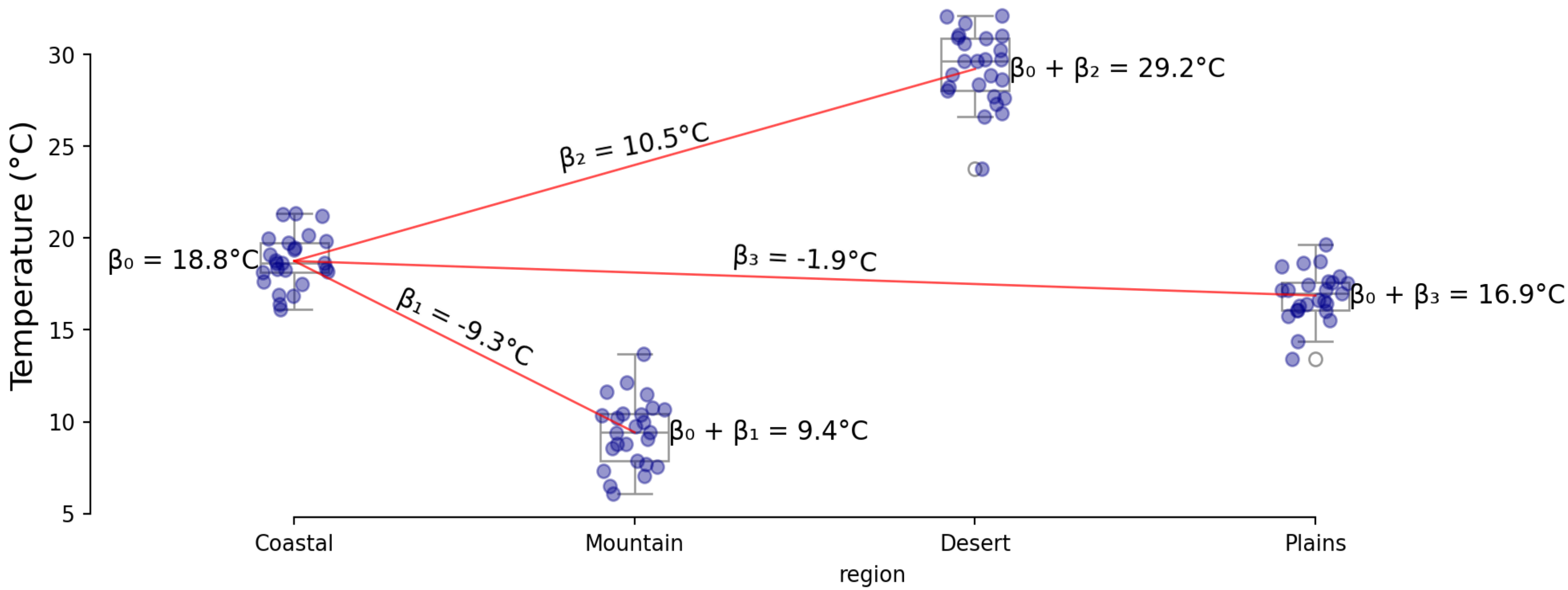
*Model:* $Temperature = \beta_0 + \beta_1 \cdot Mountain + \beta_2 \cdot Desert + \beta_3 \cdot Plains + \varepsilon$



- $\beta_0$ = *Mean temperature in Coastal areas (18.8°C)*
- $\beta_1$ = *Difference between Mountain and Coastal (-9.3°C)*
- $\beta_2$ = *Difference between Desert and Coastal (+10.5°C)*
- $\beta_3$ = *Difference between Plains and Coastal (-1.9°C)*

# Example: Comparing Many Regions

*Model:* $Temperature = \beta_0 + \beta_1 \cdot Mountain + \beta_2 \cdot Desert + \beta_3 \cdot Plains + \varepsilon$




> *This performs the same analysis as ANOVA but gives specific comparisons*

# 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*

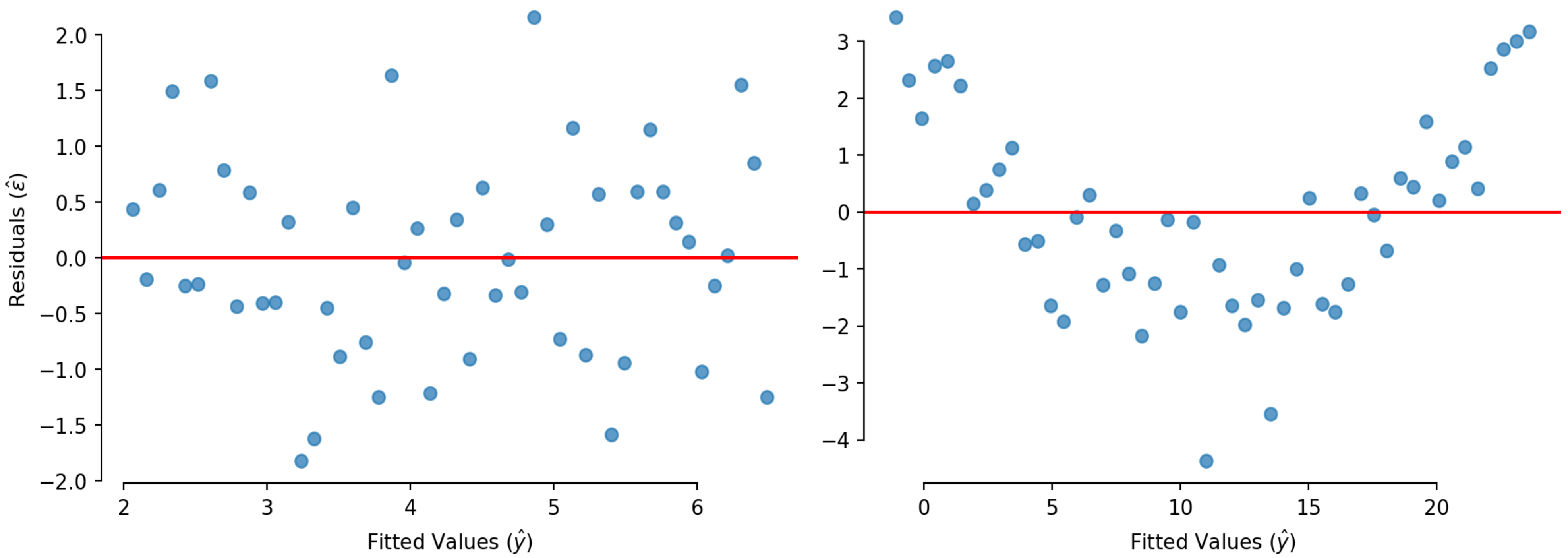
**If assumptions are violated:**

- *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?*



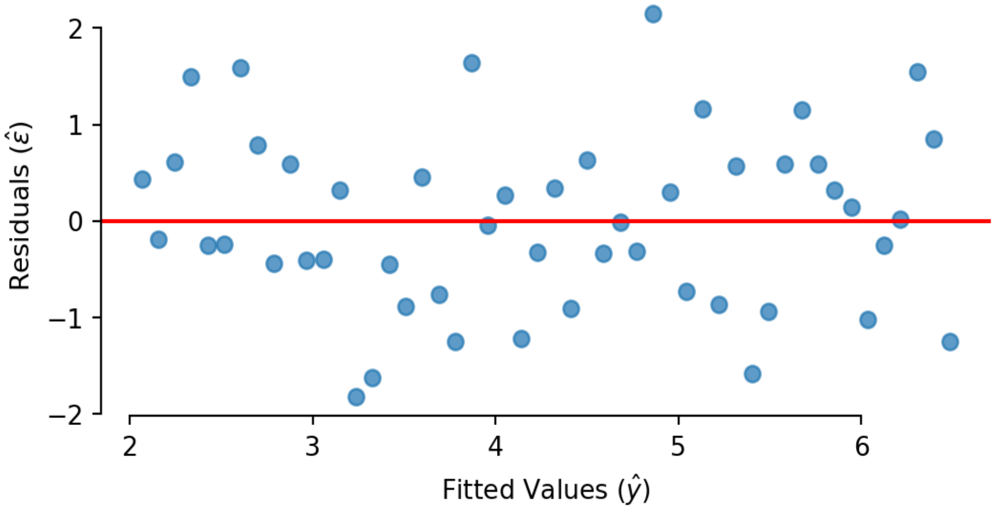

> *the left one is what we want to see*

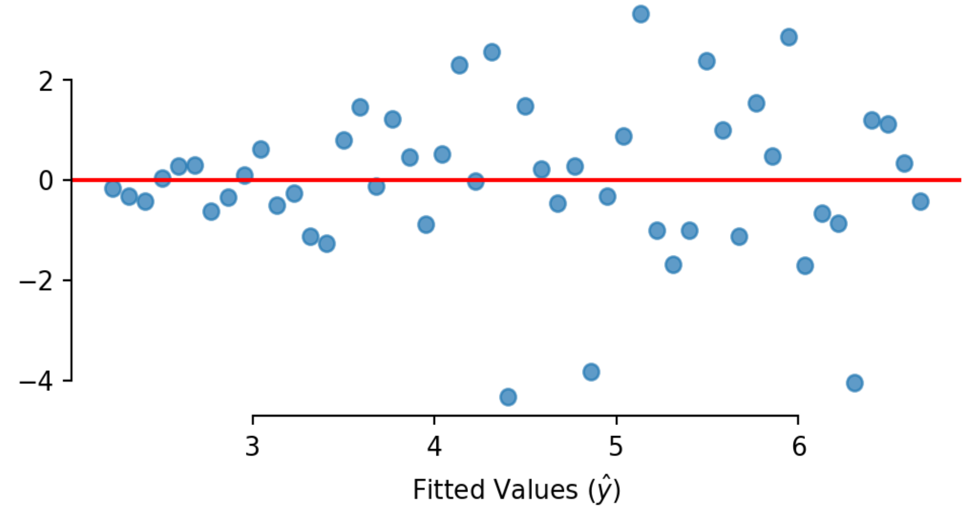
> *residual plots should show that the model is equally wrong everywhere*

# Checking for Homoskedasticity

*Residuals should be spread out the same everywhere.*

> *which one of these figures shows homoskedasticity?*



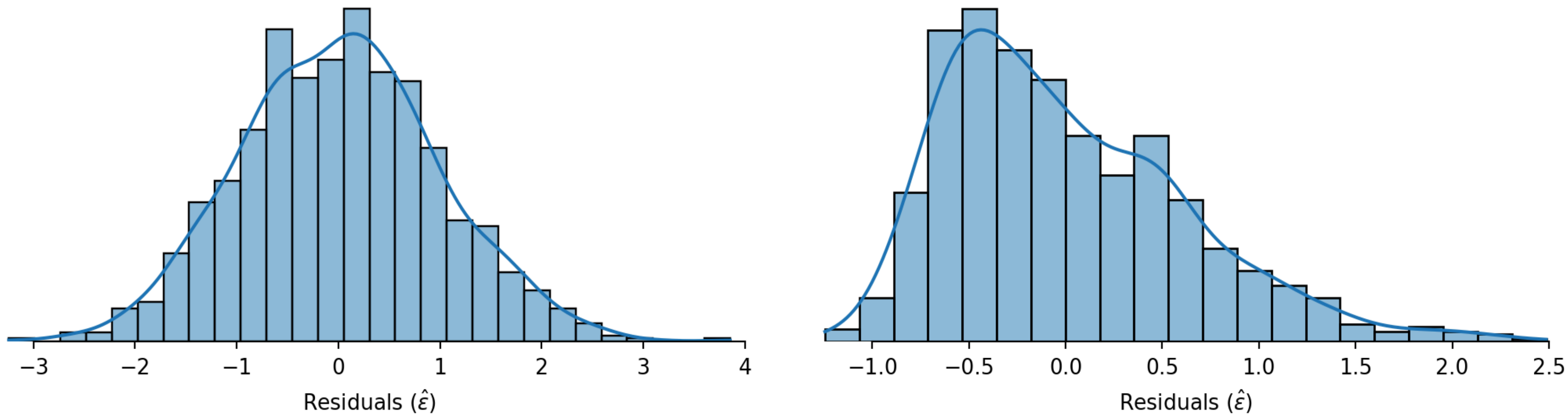

> *the left figure shows constant variability (homoskedasticity)*

> *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*



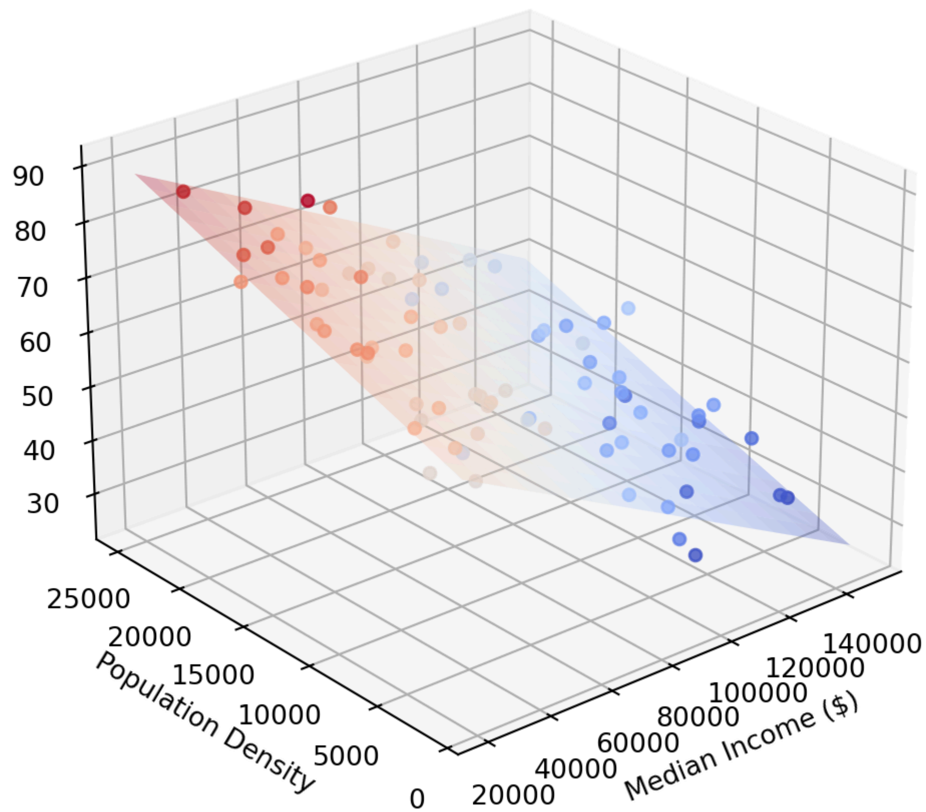

*> left shows a nice bell shape (roughly normally distributed)*

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

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

# Extending to Multiple Regression

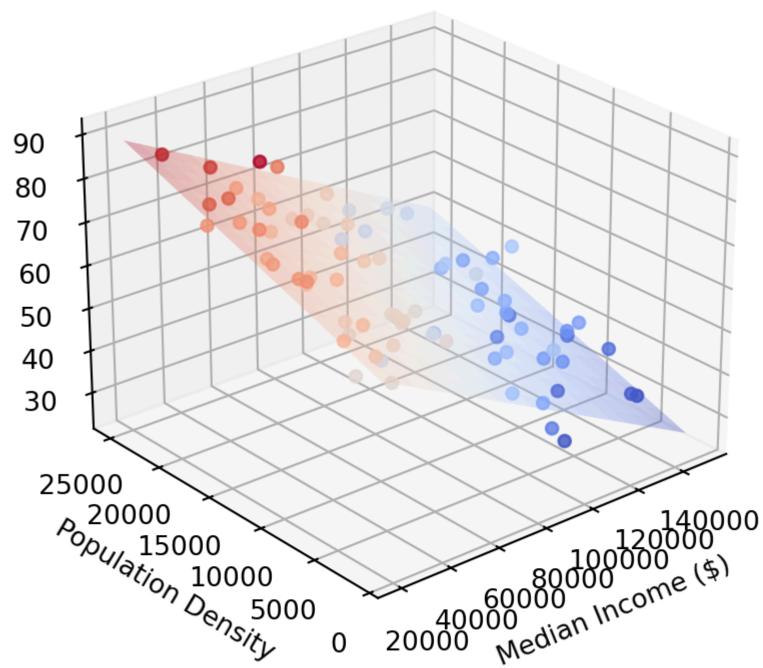
*Adding control variables to isolate relationships*




> *Model:* $Pollution = \beta_0 + \beta_1 \cdot Income + \beta_2 \cdot Density + \varepsilon$

# Extending to Multiple Regression

*Adding control variables to isolate relationships*




- $\beta_0$ = *Baseline pollution level (70.0)*
- $\beta_1$ = *Effect of income on pollution, holding density constant (-0.0003)*
- $\beta_2$ = *Effect of density on pollution, holding income constant (+0.001)*

# Key Takeaways

*Regression provides a unified framework for statistical testing*

**One-Sample T-Test**: Regression with only an intercept ($y = \beta_0 + \varepsilon$)

**Two-Sample T-Test**: Regression with a dummy variable ($y = \beta_0 + \beta_1 \cdot Group + \varepsilon$)

**ANOVA**: Regression with multiple dummy variables for groups

**Multiple Regression**: Adding control variables to isolate relationships

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