ECON 0150 | Economic Data Analysis

The economist's data analysis pipeline.

Part 5.4 | Using GLM Appropriately

Selecting the appropriate test

Matching research questions with appropriate statistical approaches

Focus:

- *Translating research questions into appropriate models*
- *Visualizing data to inform modeling choices*
- *Selecting and interpreting the right regression approach*
- *Examples across various economic contexts*

Example 1: Impact of Microlending

Do microloans improve income in low-income communities?

Research Question: *Does a microlending program increase average monthly income in a low-income community?*

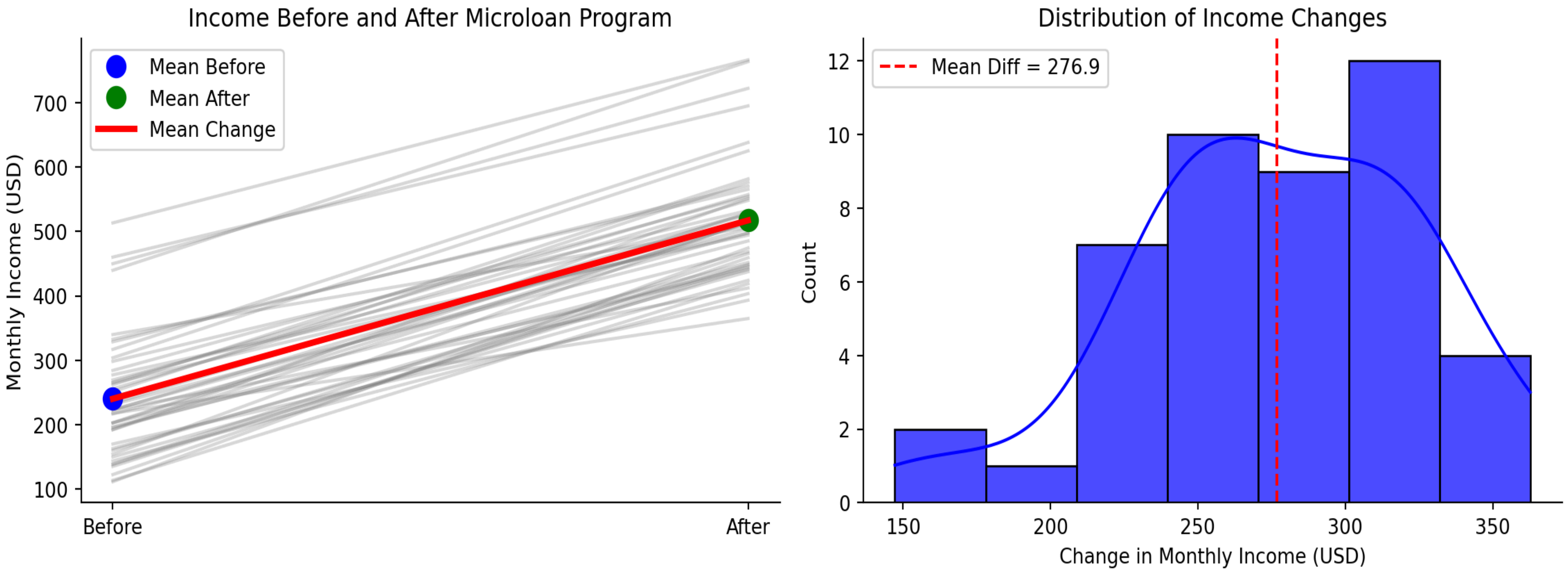
The Data:

- *Monthly income (in USD) for 45 participants before and after receiving microloans*
- *Data structure: Panel Data (paired observations)*

Example 1: Impact of Microlending

Do microloans improve income in low-income communities?

Visualization: jittered scatter, boxplot, line graph, histogram



Model: Paired Sample t-test

$$\text{income\_change} = \beta_0$$

Example 1: Impact of Microlending

Do microloans improve income in low-income communities?

Model: Paired Sample t-test

$$\text{income\_change} = \beta_0 + \epsilon$$

```python
# Create a dataframe with the differences
data = pd.DataFrame({'income_change': income_change})

# Run a one-sample t-test as regression
model = smf.ols('income_change ~ 1', data=data).fit()
print(model.summary().tables[1])
```

Interpretation:

- *The average monthly income increased by β_0 after the microloan program.*
- *We expect to see a result this extreme only p percent of the time.*
- *Or more standard: we reject the null ($\beta_0 = 0$) if $p < 0.05$.*

Example 2: Online Learning Effectiveness

Does online learning yield different outcomes than in-person learning?

Research Question: *Does learning format (online vs. in-person) affect student performance in economics courses?*

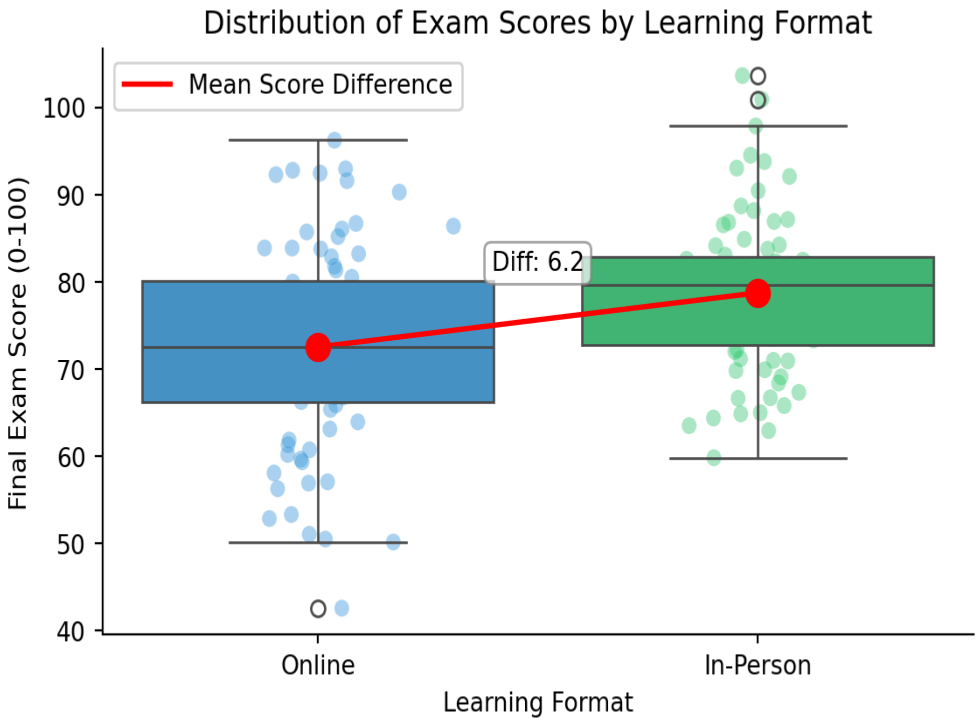
The Data:

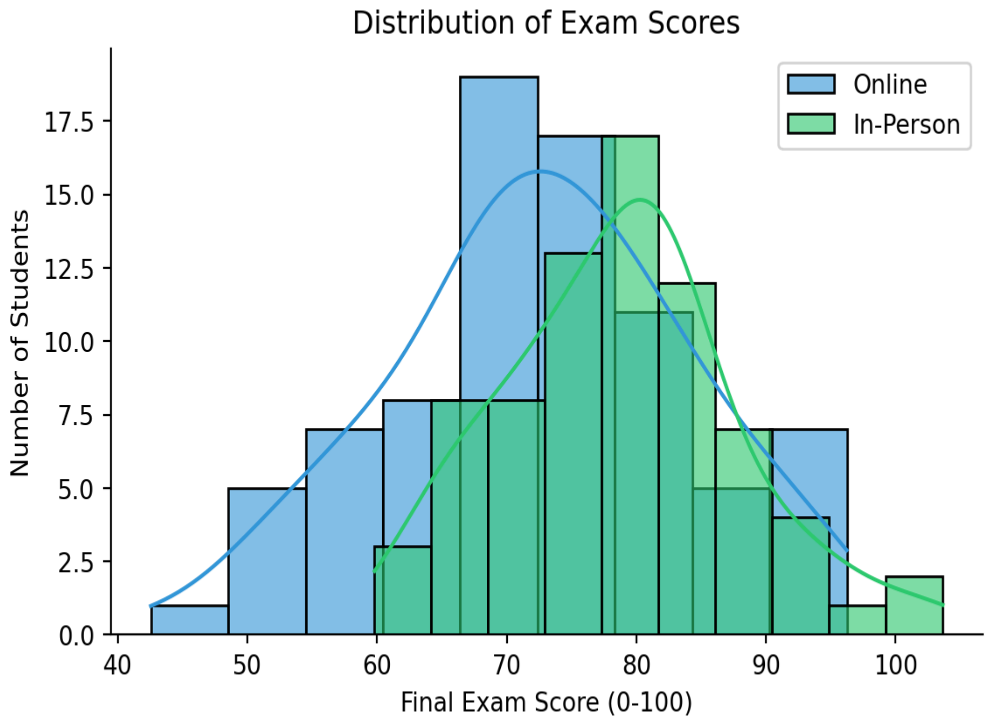
- *Final exam scores (0-100) for students in two different course sections*
- *One section taught online, the other in-person*
- *Random assignment of students to sections*

Example 2: Online Learning Effectiveness

Does online learning yield different outcomes than in-person learning?

Visualization: boxplot, histogram





Model 2: Two-Sample t-test (comparing groups)

$$\text{Score} = \beta_0 + \beta_1 \text{Online} + \epsilon$$

Example 2: Online Learning Effectiveness

Does online learning yield different outcomes than in-person learning?

Model 2: Two-Sample t-test (comparing groups)

$$\text{Score} = \beta_0 + \beta_1 \text{Online} + \epsilon$$

```python
# Create dummy variable
data['online'] = (data['Format'] == 'Online').astype(int)

# Run regression
model = smf.ols('Score ~ online', data=data).fit()
print(model.summary().tables[1])
```

Interpretation:

- *In-person students scored β_1 points higher on average than online students*
- *We expect to see a result this extreme only p percent of the time.*
- *The difference is statistically significant ($p < 0.05$)*

Example 3: Renewable Energy Investment Impact

How does investment in renewable energy affect carbon emissions?

Research Question: *What is the relationship between renewable energy investment and carbon emissions across countries?*

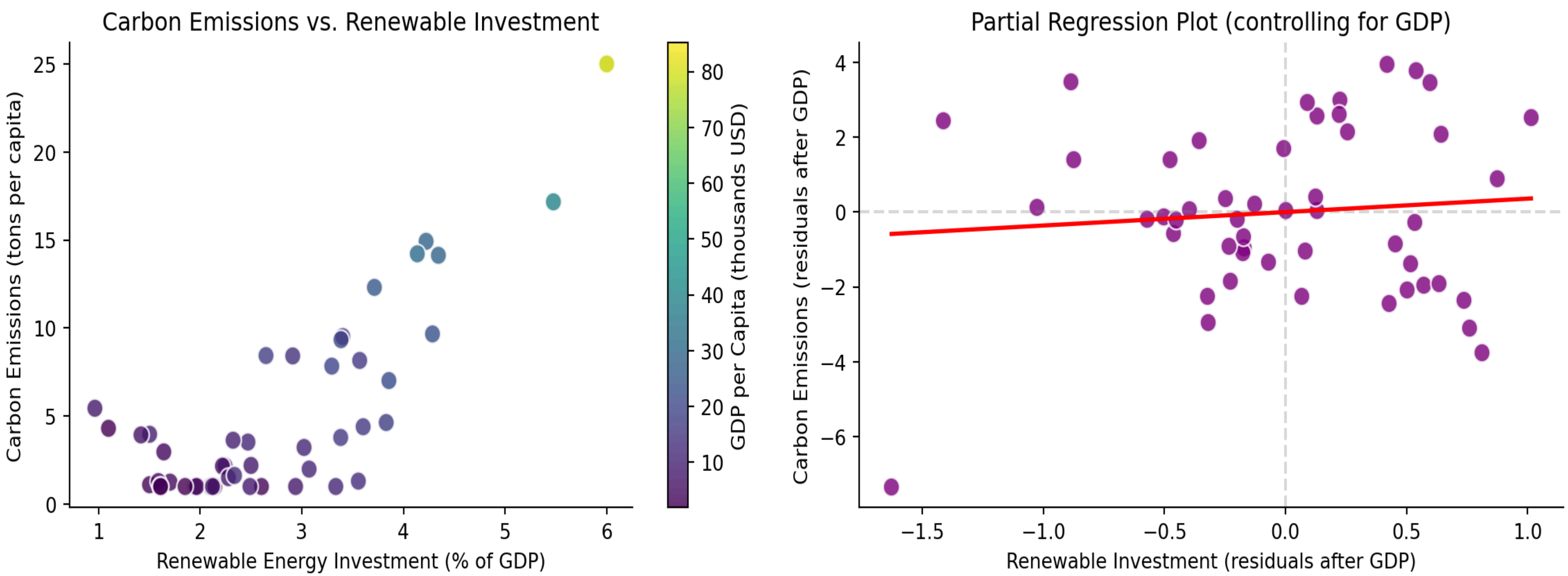
The Data:

- *Cross-sectional data for 50 countries*
- *Renewable energy investment (% of GDP)*
- *Carbon emissions (tons per capita)*
- *GDP per capita (USD)*

Example 3: Renewable Energy Investment Impact

How does investment in renewable energy affect carbon emissions?

Visualization: scatterplot



Model 3: Multiple Regression with Control Variable

$$\text{Carbon\_Emissions} = \beta_0 + \beta_1 \text{Renewable\_Investment} + \beta_2 \text{GDP\_per\_capita} + \epsilon$$

Example 3: Renewable Energy Investment Impact

How does investment in renewable energy affect carbon emissions?

Model 3: Multiple Regression with Control Variable

$$\text{Carbon\_Emissions} = \beta_0 + \beta_1 \text{Renewable\_Investment} + \beta_2 \text{GDP\_per\_capita} + \epsilon$$

```python
# Run multiple regression
model = smf.ols('Carbon_Emissions ~ Renewable_Investment + GDP_per_capita',
                data=countries).fit()
print(model.summary().tables[1])
```

Interpretation:

- *Each 1 percentage point increase in renewable investment is associated with a β_1 change in carbon emissions, controlling for GDP*
- *Effect is statistically significant ($p < 0.05$)*
- *GDP per capita has a separate positive relationship with emissions (β_2)*

Example 4: Gender Wage Gap by Education

Does the gender wage gap vary with education level?

Research Question: *Does the gender wage gap differ across education levels?*

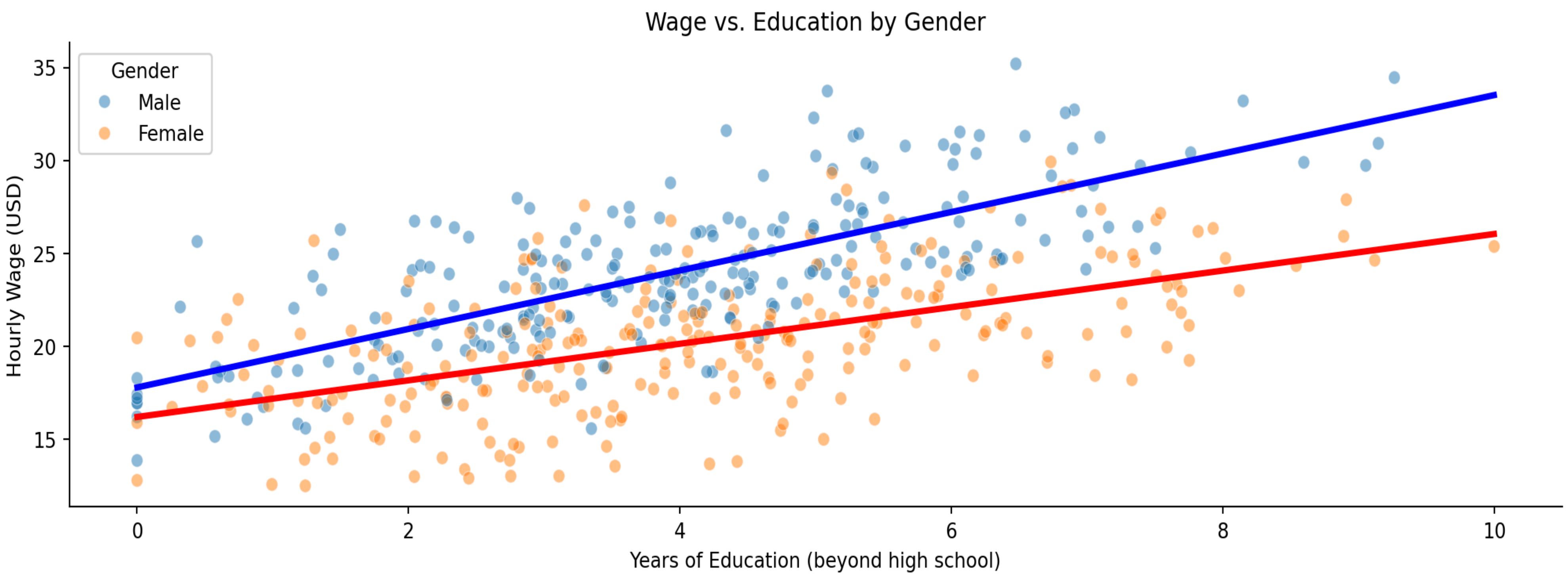
The Data:

- *Survey of 500 full-time workers*
- *Hourly wage*
- *Gender*
- *Years of education*
- *Experience*

Example 4: Gender Wage Gap by Education

Does the gender wage gap vary with education level?

Visualization: scatterplot, multiple groups



Model 4: Interaction Model

$$wage = \beta_0 + \beta_1 female + \beta_2 edu \beta_3 female \cdot edu + \beta_3 exp + \epsilon$$

Example 4: Gender Wage Gap by Education

Does the gender wage gap vary with education level?

Model 4: Interaction Model

$$wage = \beta_0 + \beta_1 female + \beta_2 edu \beta_3 female \cdot edu + \beta_3 exp + \epsilon$$

```python
# Create female dummy variable
workers['female'] = workers['gender']

# Run interaction model
model = smf.ols('wage ~ female + education + female:education + experience',
                data=workers).fit()
print(model.summary().tables[1])
```

Interpretation:

- *β_1: Base gender gap for workers with no education beyond high school*
- *β_2: Return to education for male workers*
- *β_3: Additional effect of education for female workers (beyond the male return)*

Example 5: Housing Prices and Transit

How does transit access and green space affect housing prices?

Research Question: *How does public transit access and green space proximity affect residential property values?*

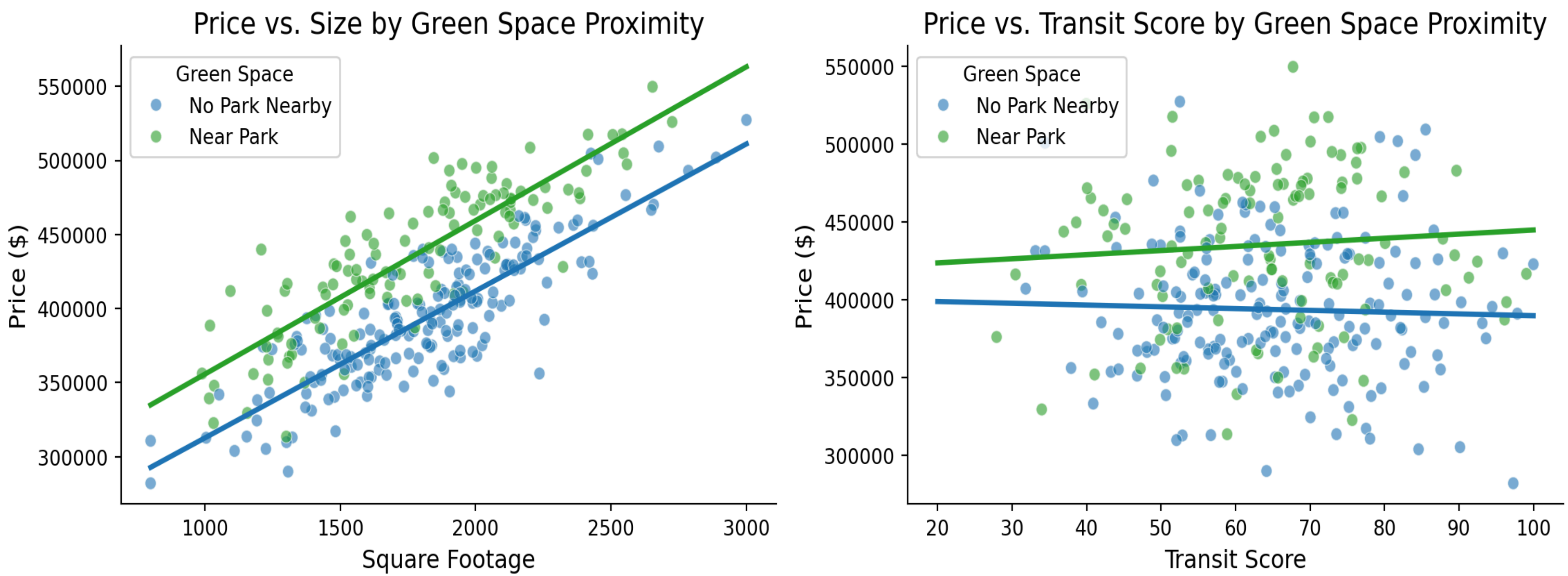
The Data:

- *Housing transactions in a metropolitan area (n = 300)*
- *Sale price*
- *Square footage*
- *Transit score (0-100)*
- *Green space proximity (binary)*

Example 5: Housing Prices and Transit

How does transit access and green space affect housing prices?

Visualization: scatterplot, multiple groups



Model 5: Multiple Regression with Categorical and Continuous Predictors

$$\text{price} = \beta_0 + \beta_1 \text{sq\_ft} + \beta_2 \text{transit\_score} + \beta_3 \text{green\_space} + \epsilon$$

Example 5: Housing Prices and Transit

How does transit access and green space affect housing prices?

Model 5: Multiple Regression with Categorical and Continuous Predictors

$$\text{price} = \beta_0 + \beta_1 \text{sq\_ft} + \beta_2 \text{transit\_score} + \beta_3 \text{green\_space} + \epsilon$$

```python
# Normalize square footage for easier interpretation
housing['sq_footage_100s'] = housing['sq_footage'] / 100
housing['transit_score_10s'] = housing['transit_score'] / 10

# Run multiple regression model
model = smf.ols('price ~ sq_footage_100s + transit_score_10s + green_space',
                data=housing).fit()
print(model.summary().tables[1])
```

Interpretation:

- *β_1: Price increase for each additional 100 square feet*
- *β_2: Price increase for each 10-point improvement in transit score*
- *β_3: Price premium for homes near green spaces*

Example 6: Seasonal Effects on Retail Sales

How do seasonal patterns impact retail sales?

Research Question: *How do retail sales vary by season when accounting for overall trends?*

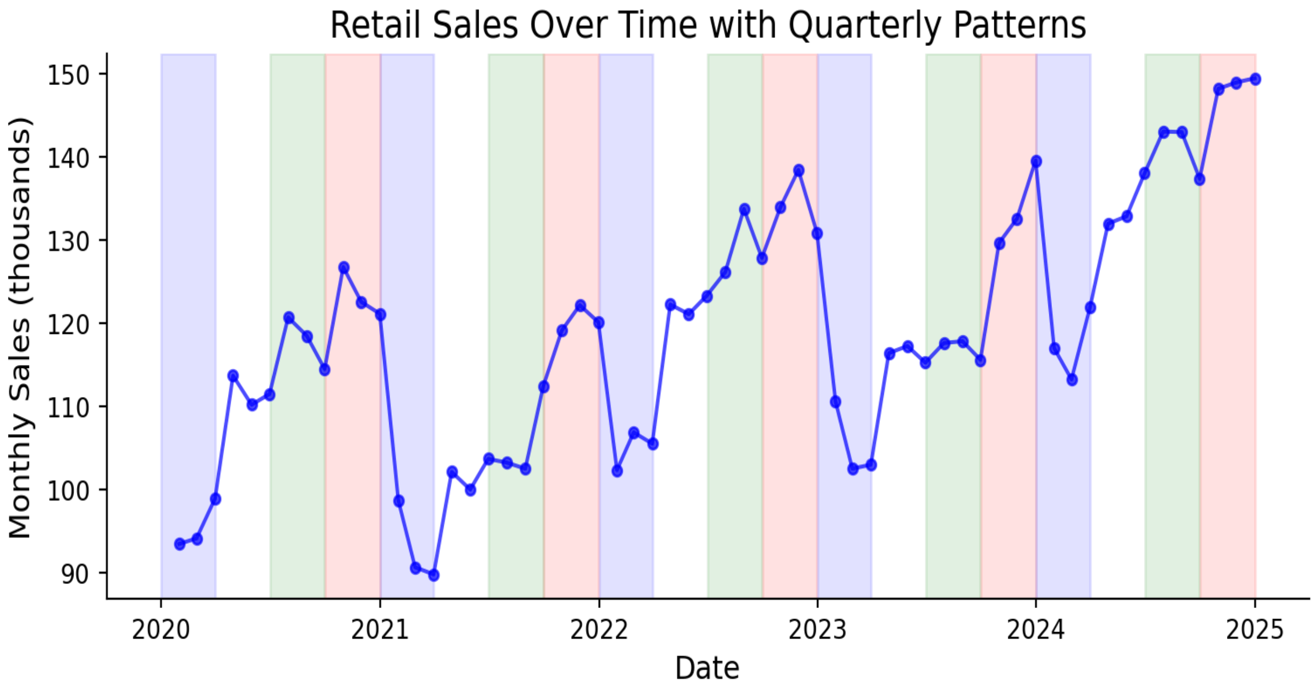
The Data:

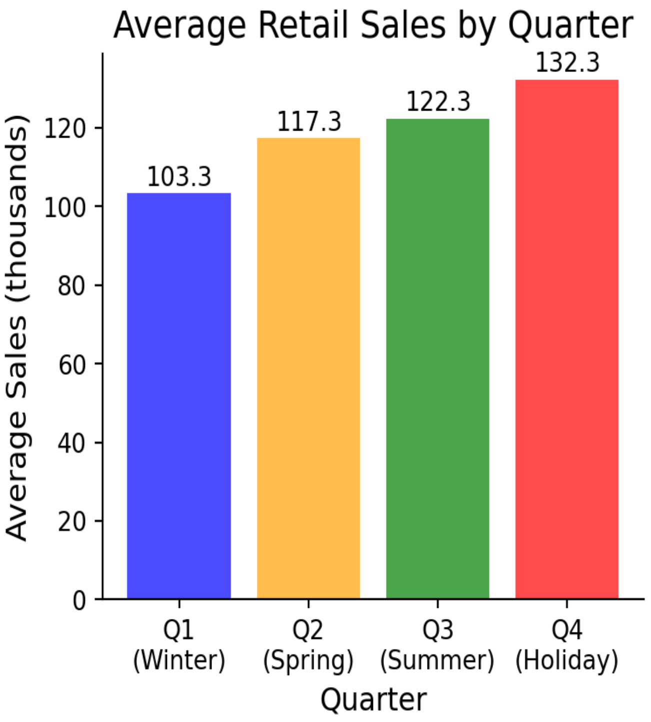
- *Monthly retail sales data over 5 years (n = 60 months)*
- *Variables: Sales (in thousands), time trend, seasonal indicators*

Example 6: Seasonal Effects on Retail Sales

How do seasonal patterns impact retail sales?

Visualization: line graph, timeseries, bar graph by season





Example 6: Seasonal Effects on Retail Sales

How do seasonal patterns impact retail sales?

Model 6: Time Series with Seasonal Fixed Effects

$$\text{sales} = \beta_0 + \beta_1 \text{time} + \beta_2 \text{Q2} + \beta_3 \text{Q3} + \beta_4 \text{Q4} + \epsilon$$

```python
# Run model with time trend and seasonal dummies
model = smf.ols('sales ~ time + Q2 + Q3 + Q4', data=retail_data).fit()
print(model.summary().tables[1])
```

Interpretation:

- *β_1: Underlying monthly trend in sales (growth rate per month)*
- *β_2, β_3, β_4: Seasonal effects for Q2, Q3, and Q4 relative to Q1 (the reference quarter)*
- *The model captures both the long-term trend and seasonal patterns*

Key Insight: Seasonal fixed effects allow us to quantify and test the significance of seasonal patterns while controlling for the underlying trend

Model Selection Framework

Matching research questions to statistical approaches

Question Type	**Model**
Change in single group	$y = \beta_0 + \varepsilon$ (*One-sample t-test*)
Differences between groups	$y = \beta_0 + \beta_1 Group + \varepsilon$ (*Two-sample t-test*)
Relationship between vars	$y = \beta_0 + \beta_1 x + \varepsilon$ (*Simple regression*)
Multiple factors	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ (*Multiple reg*)
Group-specific relationships	$y = \beta_0 + \beta_1 x + \beta_2 Group + \beta_3 x \times Group + \varepsilon$ (*Interactions*)
Temporal patterns	$y_t = \beta_0 + \beta_1 t + \beta_2 Season + \varepsilon_t$ (*Time series with fixed effects*)
Many more!	(*You can construct your own*)

Key Takeaways

Connecting economic questions to appropriate statistical models

1. ***Start with the research question***
    - *The nature of the question guides model selection*
    - *Consider what parameters would directly answer your question*
2. ***Visualize data first***
    - *Plots reveal patterns that inform model specification*
    - *Helps identify potential non-linearities or interactions*

Key Takeaways

Connecting economic questions to appropriate statistical models

3. ***Match the model to the data structure***
    - *Paired observations call for paired tests*
    - *Categorical predictors often require dummy variables*
    - *Time series data usually needs detrending or seasonal adjustment*
4. ***Interpret coefficients carefully***
    - *Connect each coefficient back to your research question*
    - *Consider both statistical and practical significance*
    - *Always think about what's being held constant ```*