ECON 0150 | Spring 2025 | Homework 09

Due: Tuesday, March 25, 5PM

Homework is designed to both test your knowledge and challenge you to apply familiar concepts in new applications. Answer clearly and completely. You are welcomed and encouraged to work in groups so long as your work is your own. Use the provided datasets to answer the following questions. Then submit your figures and answers to Gradescope.

1. A labor economist is analyzing the economic impacts of remote work policies by comparing job satisfaction scores for employees at companies with different work arrangements. The dataset `work_satisfaction.csv` shows satisfaction scores (0-100 scale) for workers at traditional office-based companies versus those with flexible remote work policies.

a) Create a box plot and/or scatter plot that visualizes this data for the question.

b) Calculate the mean satisfaction score for each group. What is the difference in means between employees at office-based companies and those with remote-flexible policies? We'll compare this difference in means to the regression result in d).

```
# Create both samples
office_based = data[data.remote_flexible == 0].satisfaction
remote_flexible = data[data.remote_flexible == 1].satisfaction

# Means
print(office_based.mean())
print(remote_flexible.mean())
```

c) Perform a two-sample t-test to determine the probability of an event as extreme as the sample satisfaction difference between the work arrangement groups. We'll compare this test to the regression result in d).

```
# Two-sample t-test
from scipy import stats
t_stat, p_value = stats.ttest_ind(office_based, remote_flexible, equal_var=False)
print('t-statistic:', t_stat)
print('p-value:', p_value)
```

d) Fit a regression model to test for a satisfaction difference between work arrangement groups.

```
# Run the regression model
import statsmodels.formula.api as smf
model = smf.ols('satisfaction ~ remote_flexible', data=data).fit()
print(model.summary().tables[1])
```

e) Interpret the slope coefficient. Be sure to mention the meaning, magnitude, and p-value for the slope.

2. A climate researcher is studying the relationship between community wealth disparities and heat island intensity across metropolitan areas. The data is available in the file `wealth_heat.csv` and includes: `community` (name of the neighborhood); `wealth_gap` (an index with higher values indicateinggreater inequality); `heat_intensity` (°C above surrounding areas).

a) Create a visualization of heat island intensity versus wealth gap index. Describe any apparent relationship.

b) Fit a linear regression model to predict heat island intensity based on wealth gap index.

```
# Fit the model
import statsmodels.formula.api as smf
model = smf.ols('heat_intensity ~ wealth_gap', data=wealth_heat).fit()

# Print summary statistics
print(model.summary().tables[1])
```

c) Interpret the slope coefficient. Be sure to mention them meaning, magnitude, and p-value for the slope.

d) Plot the model's residuals against the model predictions.

```
# Get predicted values and residuals
wealth_heat['predicted'] = model.predict()
wealth_heat['residuals'] = model.resid

# Plot residuals (y) against predictions (x)
plt.scatter(wealth_heat['predicted'], wealth_heat['residuals'])
plt.axhline(y=0, color='r', linestyle='-')
```

e) Based on the residual plot, identify any potential violations of regression assumptions. Do you see evidence of:

- Non-linearity
- Heteroskedasticity

f) Create a histogram of the residuals. Do they appear to be normally distributed?