

Problem Set 5: Education and Income Across States
BUS 2160 · Statistics for Business & Economics | Methodist University
Submit via Canvas | *Individual assignment*

This assignment requires the use of the provided state panel dataset, the supplemental Chicago Fed article, and concepts from the textbook and lectures.

- i. You will estimate a bivariate regression relating educational attainment to median household income across the 50 states.
- ii. You will interpret the regression coefficients, statistical significance, R^2 , and standard error of the regression.
- iii. You will use predicted values and residuals to examine where the model performs well and where important differences remain unexplained.
- iv. You will use the Chicago Fed article to evaluate how fitted relationships should be interpreted and why a strong statistical association does not, by itself, establish causation.

Your written responses should demonstrate command of bivariate regression, coefficient interpretation, statistical significance, model fit, predicted values, residuals, and the limits of observational evidence.

Overview

Regression analysis estimates how one variable changes, on average, as another variable changes. In a bivariate regression, the dependent variable is modeled as a function of one explanatory variable.

For this assignment, the dependent variable is real median household income, measured in 2024 dollars. The explanatory variable is the percentage of adults age 25 and older who hold a bachelor's degree or higher.

The slope measures the estimated difference in median household income associated with a one-percentage-point difference in bachelor's degree attainment. The intercept gives the model's predicted income when the bachelor's degree share equals zero.

The fitted regression line produces a predicted income for each state. The residual is calculated as:

$$\text{Residual} = \text{Actual income} - \text{Predicted income}$$

A positive residual means a state's actual income is higher than the model predicts. A negative residual means its actual income is lower than predicted.

R^2 measures the share of cross-state variation in income accounted for by the fitted linear relationship. The standard error of the regression measures the typical size of the residuals in the units of the dependent variable.

Supplemental Reading

Read the required supplemental source, **“Do Higher Education Levels Matter for Regional Growth?”** by Ila Gupta and Thomas Walstrum, published in *Chicago Fed Insights* in March 2026 and available through the link provided in Canvas.

The article examines the relationship between bachelor's degree attainment and subsequent employment growth across states and regional areas. Pay particular attention to:

- i. How the authors interpret the slope of a fitted line
- ii. The variation in outcomes among places with similar education levels
- iii. The factors the authors identify as possible explanations for that variation

- iv. The difference between the descriptive relationship shown in the article's scatterplot and the separate evidence used to make a causal argument
- v. The difference between the outcome examined in the article and the outcome examined in this assignment

The reading is required and directly supports your responses to Questions 4 and 5. Do not submit a separate summary. Use the article to interpret the regression evidence you produce from the state dataset.

Data Preparation

Complete the following work in Excel. Include the completed scatterplot, regression output, and requested calculations in your submission.

1. Open the provided dataset, ps5_6_state_panel.csv.
2. Filter the dataset to include only observations from 2024. Your cross-section should contain one observation for each of the 50 states.
3. Create a scatterplot using:
 - a. **bachelors_pct** on the horizontal axis
 - b. **median_hh_income_real** on the vertical axis

Add a linear trendline. Display the regression equation and R^2 on the chart. Label both axes clearly.

4. Use Excel's Regression tool to estimate the following model:

Real median household income = $\alpha + \beta \times$ bachelor's degree share, where:

- a. **median_hh_income_real** is the dependent variable
- b. **bachelors_pct** is the independent variable

Include the complete regression output in your submission.

5. Create two new columns in the 2024 dataset:
 - a. **predicted_income**
 - i. Calculate each state's predicted income using the estimated intercept and slope.
 - b. **residual**
 - i. Calculate each state's residual as:

Actual real median household income – Predicted real median household income

6. Confirm that the residuals sum to approximately zero. Minor differences caused by rounding are acceptable.
7. Identify:
 - a. The two states with the largest positive residuals
 - b. The two states with the largest negative residuals
 - c. Two states with bachelor's degree shares within 2 percentage points of one another but noticeably different median household incomes

Label all calculations, figures, and regression output clearly.

Assignment

Answer each question in complete sentences. There is no minimum length for answers, but minimal effort will not result in maximum credit. Refer to the grading philosophy in the syllabus if further clarification is needed.

Use statistical reasoning and terminology from class. Responses that merely copy values from the regression output without interpreting them will not earn full credit.

1. Interpret the estimated regression model.

- i. Report the estimated regression equation.
- ii. Interpret the slope in dollars per percentage point. State clearly which two variables are being related.
- iii. Interpret the intercept and evaluate whether it has a useful substantive meaning in this model. Refer to the range of bachelor's degree attainment observed in the dataset.
- iv. Report the p-value for the slope coefficient. At the 5 percent significance level, determine whether the estimated relationship is statistically distinguishable from zero.

2. Evaluate the fit and precision of the model.

- i. Report and interpret R^2 . Explain what variation the statistic describes and what it does not mean.
- ii. Report the standard error of the regression and interpret it in dollars.
- iii. Explain why a statistically significant slope does not necessarily mean that the model predicts every state's income closely.
- iv. Based on R^2 and the standard error of the regression, evaluate whether the model provides a useful but incomplete description of cross-state income differences.

3. Use predicted values and residuals to examine the model's errors.

- i. Report the two largest positive residuals and the two largest negative residuals. Identify the states and report their actual income, predicted income, and residual.
- ii. Explain what the sign and magnitude of each residual indicate.
- iii. Compare the two states you selected with similar bachelor's degree shares but different incomes. Explain how their results demonstrate a limitation of predictions based on education alone.
- iv. Explain why a residual should be interpreted as an unexplained component of the model rather than as proof of any particular omitted cause.

4. Compare your regression with the Chicago Fed analysis.

- i. Identify the explanatory variable and outcome variable used in the article's fitted relationship. Explain how that model differs from the regression estimated in this assignment.
- ii. Explain what the fitted line and the dispersion around it represent in each analysis. Why should the amount of dispersion in the two figures not be compared directly?
- iii. Use the article and your residual analysis to explain why places with similar education levels may experience different outcomes.
- iv. Explain why the positive relationship in your 2024 regression does not, by itself, establish that increasing a state's bachelor's degree share would cause median household income to rise by the estimated slope.

5. How much does educational attainment explain differences in median household income across states?
- i. Take a clear position on whether the bivariate regression provides a weak, moderate, or strong explanation of cross-state income differences.
 - ii. Support your conclusion using the estimated slope, statistical significance, R^2 , standard error of the regression, and specific residuals.
 - iii. Use the Chicago Fed article to evaluate how the fitted relationship should be interpreted when substantial differences remain among places with similar education levels.
 - iv. Explain which claims the regression supports and which predictive or causal claims it does not support.
 - v. Address the strongest evidence that complicates your conclusion and explain why it does or does not change your position.

The strongest responses will distinguish among the size of the estimated relationship, statistical significance, model fit, prediction accuracy, and causal interpretation.

Full credit requires accurate regression calculations, precise interpretation of the model, explicit use of the state data and Chicago Fed article, and a conclusion consistent with the evidence presented in the earlier questions. Responses that do not properly assess the limits of model fit measures such as R^2 or residuals will receive partial credit at best.