AAEC 4804/5804G, STAT 4804: Fundamentals of Econometrics

Your Name Here

Spring 2025 – Homework #1

Instructions

This homework is intended to help you review the material covered in Lecture 2. There is a joint emphasis on both the theoretical and practical aspects of the material. **You are strongly encouraged to work with your classmates, but you must submit your own answers.**

For Questions 1 & 2, I will allow you to submit your written answers on a separate sheet of paper. This should lower your anxiety about typing out the equations in `R Markdown`. However, you are free to type your answers if you prefer. *This will be rewarded with extra credit.*

For Questions 3 & 4, you are required to type your answers in `R Markdown`. You should also include the relevant `R` code used to answer the questions. See the Homework Solution Template for an example of how to structure your answers.

Q1: Modifying the OLS Model

Your friend, Peter, is always experimenting with things. His latest victim is the classical simple linear regression model. He is considering the modified regression model:

$$y_i = \kappa + \beta_0^* + \beta_1^* x_i + u_i,\ i = 1, \ldots, n \tag{1}$$

where κ is a known, non-zero constant. He is interested in estimating the parameters β_0^* and β_1^*.

a. As per our class notes, derive the OLS estimators of β_0^* and β_1^* **by minimizing the sum of squared residuals**.

b. Do these estimators differ from the ones where $\kappa = 0$? Briefly explain.

c. You and Peter are in a heated argument about whether the least squares residuals of this model, $\hat{u}_i$, necessarily sum to zero. He claims that they do, while you are adamant that they do not. Who is correct after all? Briefly explain.

d. From this regression model, will $\hat{u}_i$ continue to be uncorrelated with x_i? Briefly explain.

Q2: Rescaling the data.

You decided to rescale your data before running a regression. You rescaled your dependent variable by multiplying by η such that $y^* = \eta y$, and your independent variable by multiplying by τ such that $x^* = \tau x$.

You then ran the regression:

$$y_i^* = \beta_0 + \beta_1 x^* + u_i \tag{2}$$

a. **Show** and **discuss** the impact of these rescalings on the OLS estimators. (**Hint: You are not required to re-derive them but could instead manipulate the standard OLS estimators derived in class.**)

Q3: Estimating the OLS Estimators via `lm()`

A popular dataset in econometrics is `bwght` from the `wooldridge` package. The dataset contains information on births to women in the U.S. If we were to consider two variables in the dataset, `cigs` and `faminc`, where `cigs` is the number of cigarettes smoked per day by the mother while pregnant and `faminc` is the family income in thousands of dollars:

a. Which variable would be the dependent variable, and which would be the independent variable? Briefly explain.

b. Write down the OLS regression model that captures the causal relationship you described in part (a).

$$cig = \beta_0 + \beta_1 faminc + u$$

c. Report the average number of cigarettes smoked per day by the mother while pregnant as well as the average family income.

d. Report the OLS estimates of the model you wrote in part (b) using the `lm()` and `coef()` functions in `R`.

e. Interpret the estimated coefficients.

f. What proportion of the variation in your y variable is explained by the x variable? **Be sure to explain in plain English and explicitly state the variables you are referring to.**

g. Using your results from part (d), what is the income elasticity of cigarette consumption **calculated at the average of the variables**?

Q4: Functional Forms

(C1 JW, 7th Ed.) The data in 401K are a subset of data analyzed by Papke (1995) to study the relationship between participation in a 401(k) pension plan and the generosity of the plan. The variable *prate* is the percentage of eligible workers with an active account; this is the variable we would like to explain. The measure of generosity is the plan match rate, *mrate*. This variable gives the average amount the firm contributes to each worker's plan for each \$1 contribution by the worker. For example, if $mrate = 0.50$, then a \$1 contribution by the worker is matched by a 50¢ contribution by the firm.

Note: Your dataset of interest is `k401k` from the `wooldridge` package.

a. What is the average participation rate and the average match rate? What about the min and max values for each variable? Index the `summary()` function to extract the relevant information.

b. Now estimate the following simple regression model:

$$prate = \beta_0 + \beta_1 mrate + u$$

Store the model results as `m1`. **You are not required to report the full results of the model here, just to store the results.**

c. Now estimate the following regression model:

$$prate = \beta_0 + \beta_1 \log(mrate) + u$$

Store the model results in this step as `m2`.

d. Using the `stargazer` package, report the full results of both models. Use the `digits` argument to report your results to three decimal places.

e. Interpret the intercept, slope, and coefficient of determination of **both** models.

f. Using `m1`, compute the elasticity of the participation rate with respect to the match rate **at the sample means**. Does a one-percent change in the match rate have a large or small effect on the participation rate? Briefly explain.

g. Using the `predict()` function, compute the predicted values of *prate* for both models starting at $mrate = 0.05$ and increasing by 0.05 until $mrate = 4.5$. I would like you to use the `seq()` function in `R` to generate the sequence of values for *mrate*. (**Hint: A quick Google search, or using the help function in your consoles should help with this.**)

Next, plot the predicted values of *prate* against *mrate* for both models on the same graph. Make sure to label the axes and include a legend to distinguish between the two models. Which model appears to fit the data better? Briefly explain. (**Hint: It would help you visualize the fits if you were to add a scatter plot of the (partially transparent) original data to the graph as well. Also, please restrict your y-axis to range between 50 and 110.**)