Andrew Plum

**ECON 453**

Fall 2023

Problem Set 2 – 38 points

Submit by end of day Monday October 9th

Please download the gretl session file “PS2 Session” from Canvas. This is a dataset from the American Community Survey in 2019 (this in an annual (large) survey conducted by the Census Bureau). The data contain information at the individual level regarding income and other variables. The following is how I selected the sample:

* The individual is between 25 and 40 years of age
* The individual is employed, reports working at least 35 hours per week, and reports making at least $20,000 in annual income
* The individual has a bachelor’s degree and no degree beyond that (only has a bachelor’s)
* The individual majored in Accounting, Economics, Finance, or Marketing

You will submit one document (Word or PDF) for this problem set. Please copy/paste the relevant regression results or graphs into your document, then add your discussions.

1. (7 points) Let’s begin by looking at some of the factors that influence income.

* Run a regression using income as the dependent variable and the following regressors: female, age, hours, and a dummy for whether the person’s race is white (1 if Yes, 0 if no).

Done

* Report/copy your results. Summarize what we learn from the model. Interpret the coefficients on the dummy variables specifically. Overall, do the estimated coefficients match your expectations? Explain briefly.

A white sheet with numbers and lines

Description automatically generated with medium confidence

All regressors in the regression have statistical significance and the adjusted R^2 is low. The coefficient for the female regressor is -17.9316; this means that being female is associated with a decrease in income. The coefficient for the age variable is 4.27699, indicating that as age increases, income also tends to increase. The hours variable coefficient is 3.07145, suggesting that increasing the number of hours worked is associated with an increase in income. The white regressor’s coefficient is 10.7550, indicating that being white is associated with a higher income compared to other races. From these regression results we would conclude that this is a significant income gap between being male and being female as well as there also being a significant income gap between being white and all the other races. We would also predict you would have more income if you are older and/or if you work more hours. I was not surprised by the income gaps in the results because I’ve heard about their existence before. I also wasn’t surprised by older people and people who work more hours making more money because it makes intuitive sense that more senior workers are valued more for the work experience they likely have compared to new employees and most people probably only want to work longer hours if it means more pay.

* Let’s check for heteroskedasticity. Present and briefly discuss a residual plot using hours as the x-axis. Is this troubling?

A graph of green and white lines

Description automatically generated

This graph is difficult to interpret if there is heteroskedasticity. I am inclined to say that there is heteroskedasticity because the variance of the points on the residual plot does not seem to be uniform throughout the graph. The variance seems to be small at the beginning and then the variance becomes large around 40 and seems to stay large until 70 where is the variance seems to become small again. It also seems like the trend line of the residual plot has a slightly negative slope which could indicate a problem. One other thing to note is that points are grouped on the graph in lines, and I think this is due to people rounding their hours work when reporting it.

* Run White’s test for heteroskedasticity. Report the results of this test and discuss what that means. What would you recommend doing based on these results? (you do not need to implement, just discuss what the next steps should be).

A screenshot of a computer

Description automatically generated

The p-value at the bottom of the test indicates we have a problem with heteroskedasticity as it is less than 0.05. There are several things we could possibly do to fix the model. We could restrict the sample so that the variance of the residuals of the restricted sample is more uniform. So, we could restrict from 0 to before 40 hours, and 40 hours to 70 or 80 hours, and after 70 or 80 hours to 100 hours. Another thing that I could possibly do to fix the model is to change the kind of regression (so maybe we have a squared or logged term), and another solution that should work to fix the model is using robust standard errors to adjust the model to account for the heteroskedasticity.

1. (5 points) Run the same regression again but use **logged income** as the dependent variable. Use the same set of regressors as in Question 1.

* Report/copy your results

A screenshot of a computer

Description automatically generated

* Summarize what we learn from this version of the regression and interpret the coefficients on the dummy variables specifically.

The units for income are now in percentage changes because the income variable has been logged. All the regressors are still statistically significant, and it should be noted the adjusted R^2 of this model is slightly higher than the model where income wasn’t logged indicating a slightly better fit. From the results of this regression, we learn females make about 14.6% less than males; white people make about 10.7% more on average than other races; each year a worker’s age increases, their income increases by about 3.6%; and for each additional hour a worker works, their income increases 2.3%.

* Has this changed our heteroskedasticity situation as compared to part 1?

A green and white chart

Description automatically generated

A screenshot of a computer

Description automatically generated

According to White’s test the p-value was still 0 which would indicate we would still have heteroskedasticity; so no the heteroskedasticity situation has not changed from the model in part 1.

* In your (humble) opinion, which should be our preferred version of the model (the one in question 1 or question 2)? Explain your reasoning.

This model has a slightly higher adjusted R^2 which would indicate the logged income does have a slightly better fit. Another thing to consider is that putting income in percentages by logging income might make it easier to comprehend the changes in income for each of the regressors. There also might be an advantage to using percentages here because it doesn’t require a base line number to compare the percentages to whereas you would need a base line number if the units weren’t in percentages because knowing how much more or less income one would have based on their situation would not be as useful without the context of a base line number to compare to.

1. (6 points) Interactions! Create an interaction term between the female and white variables. Run a regression where we use income (**not logged**) as the dependent variable, and the following regressors: female, age, hours, white, and the interaction of female and white.

* Report/copy your results.

A screenshot of a computer

Description automatically generated

* Predict the annual income of 4 types of individuals based on gender and race (white/nonwhite). For each, assume they are 33 years old and work 40 hours per week.

Female & nonwhite = -185.909 + (1 \* -11.0236) + (33 \* 4.27003) + (40 \* 3.06540) + (0 \* 14.9482) + (0 \* -8.83401) = -185.909 - 11.0236 + 140.91099 + 122.616 + 0 + 0 = 66.59439

Female & white = -185.909 + (1 \* -11.0236) + (33 \* 4.27003) + (40 \* 3.06540) + (1 \* 14.9482) + (1 \* -8.83401) = -185.909 - 11.0236 + 140.91099 + 122.616 + 14.9482 - 8.83401 = 72.70858

Male & nonwhite = -185.909 + (0 \* -11.0236) + (33 \* 4.27003) + (40 \* 3.06540) + (0 \* 14.9482) + (0 \* -8.83401) = -185.909 + 0 + 140.91099 + 122.616 + 0 + 0 = 77.61799

Male & white = -185.909 + (0 \* -11.0236) + (33 \* 4.27003) + (40 \* 3.06540) + (1 \* 14.9482) + (0 \* -8.83401) = -185.909 + 0 + 140.91099 + 122.616 + 14.9482 + 0 = 92.56619

* Summarize what this tells us about the gender and racial wage gaps. What should we conclude, for example, about whether the gender gap is worse for racial minorities?

The results here indicate that the gender income gap is actually more present for white people than it is for racial minorities as the gender income gap is larger for white people than it is for racial minorities and this relationship indicated by the interaction is statistically significant. The gender income gap is about $8,834 dollars wider for white people than it is for racial minorities. This surprised me initially. It should be noted, White people still earn more income than nonwhite people of the same gender.

1. (5 points) Run your own model that uses income as the dependent variable and includes an interaction term. You can use whatever variables you would like for this (as long as there is some logic to it).

* Discuss your idea/hypothesis, what is the question you have in mind, what do you expect you will find?

I am testing to see if the impact of having children is different for people who are doing remote work and those not doing remote work. I predict that those who have children and do remote work will make less money than those without children who do remote work because the remote workers with children will be distracted by their children at home which results in lower quality work which results in lower income earned. This would be very interesting if I am correct because I believe people with children tend to earn more income on average than those without children likely because they feel like they have an active responsibility to provide for their family.

* Summarize any steps you took to create/adjust variables or the sample (if necessary). The idea here is that I would be able to replicate your results if I wanted to.

What I did was I first created a variable called “child\_true” which was equal to “nchild != 0”. So the value of “child\_true” will be equal to 1 if the person has children. Then I created an interaction term called “remote\_X\_child\_true” and it was set equal to “remote\*child\_true”.

* Report/copy your results.

A screenshot of a computer

Description automatically generated

* Summarize what we learned from your regression. Do the results seem reasonable to you/match your expectations?

The results of the regression indicate that the interaction term is not statistically significant. This is because the p-value of the interaction term is 0.1101 which is greater than 0.05. This means that we can conclude that there is no statistical difference in the income gap between a person working remotely with kids and a person working remotely without kids. The adjusted R^2 of the regression is also very low indicating very low explanatory power.

1. (7 points) For this model, we will use **logged income** as the dependent variable. Run a regression that includes female, age, hours, and dummy variables for the different majors. Use Economics as your reference category.

* Report/copy your results

A screenshot of a computer

Description automatically generated

* Test the equality of coefficients. For each of the tests, state the null hypothesis, report your p-value, and state your conclusion.
  + Is there a significant difference in income between Economics and Accounting majors?

Because economics is the reference category is economics, the null hypothesis would be Beta(Accounting)=0 which translates to the linear restriction “b[ACCT] = 0” in gretl.

A screenshot of a computer

Description automatically generated

The p-value we get from the test is 6.32075e-026 which is less than 0.05 which means I can reject the null hypothesis. What this means is I can now conclude from the results of the original regression that accounting majors make 15.6% less income than economics majors.

* + Is there a significant difference in income between Marketing and Accounting majors?

The null hypothesis would be Beta(Marketing)=Beta(Accounting) which translates to the linear restriction “b[MKTG] – b[ACCT] = 0” in gretl.

A screenshot of a computer

Description automatically generated

The p-value we get from the test is 0.0509516 which is slightly greater than 0.05 which means I do not reject the null hypothesis. What this means is in the results of the original regression that there is no significant difference in the percentage of income earned between marketing majors and accounting majors.

* Summarize what we learn from the regression results/tests. Do these results match what you would expect?

From the regression results, we learn that difference in the percentage of income earned between economics majors and marketing majors as well as the difference in the percentage of income earned between economics majors and accounting majors are both statistically significant whereas the difference in the percentage of income earned between economics majors and finance majors is not statistically significant. The difference in the percentage of income earned between marketing majors and accounting majors is also not statistically significant. From here, we then see that economics majors and finance majors (because they are not statistically different from one another) earn about 14.3% more income than marketing and accounting majors (grouped together because they are not statistically different from one another). This doesn’t necessarily surprise me because I would imagine the top earners in economics and finance make quite a lot more than the top earners in marketing and accounting because of the nature of their respective industries and how lucrative they can potentially be. I wonder what it would look like if you removed the top earners from each field in the dataset.

* What would you recommend doing to change/improve the model examining the differences in income between majors?

Because the lack of significant difference in the percentage of income earned between economics and finance and between accounting and marketing, in our model we could combine economics and finance majors and we could combine marketing and accounting majors. This would reduce issues of multicollinearity in our model. The adjusted R^2 is still quite low in terms of explanatory power. If we want to predict our dependent variable more accurately, we will want to increase this. One way we could do this is by adding new variables to our regression that increase the adjusted R^2.

1. (8 points) Wildcard! Time to have some fun with the data. Run a regression using any (reasonable) variable as the dependent variable (could be income or logged income, but does not have to be), and at least 3 explanatory variables.

* Discuss your idea/hypothesis, what is the question you have in mind, what do you expect you will find?

I am wondering how age, hours, being an immigrant, being a veteran, and travel time affect the amount of income a person earns. In the regression I will run, I will have income as the dependent variable and for the regressors I will have age of the person, the number of hours the person works, if the person is an immigrant, if the person is a veteran, and the amount of time the person travels to work. Age and hours will continue to have a positive effect on income as they did in the regressions before. I predict being an immigrant will mean the individual will learn less money because to my knowledge they tend to be more willing to work lower paying jobs than native born Americans. I expect veterans to make less money because they sacrifice time serving the country while most people their age are pursuing their careers. There are also disorders they may potentially gain from war ranging from new disabilities that they didn’t have before to mental illnesses; these factors could all impact their income. I think travel time will positively increase income as it increases.

* Summarize any steps you took to create/adjust variables or the sample (if necessary).

I added income as the dependent variable, and I added “age”, “hours”, “immig”, “veteran”, and “trantime” to the regression.

* Report/copy your results.

A screenshot of a computer

Description automatically generated

* Summarize what we learned from your regression. Do the results seem reasonable to you/match your expectations?

All of the regressors from the regression are statistically significant. All of the predictions I made were in line with the results. The variable “veteran” had quite a higher p-value when compared to the rest of the regressors. I didn’t expect that being a veteran would have a greater negative affect on a person’s income than being an immigrant. It was also a bit unexpected to me for travel time to have as great of an effect on income as it did because I would’ve have assumed it would be less impactful because a lot of companies with white-collar are offering more and more remote job positions; I wonder if the coefficient for it is lower than in previous time periods.

* Assess the validity of your model regarding the issues of multicollinearity and heteroskedasticity. Are these problematic in your model? Present your evidence for/against.

A screenshot of a computer

Description automatically generated

I ran White’s test (results above), and the p-value at the bottom of the test indicates we have a problem with heteroskedasticity as it is 0 which is less than 0.05. This means I must reject the null hypothesis of homoskedasticity meaning heteroskedasticity is present. This is problematic for the model because it can make it more difficult to correctly interpret the coefficients. If possible, I should try to adjust the model to account for the heteroskedasticity.

* What would you do to improve this analysis going forward?

There are several things I could possibly do to fix the model. I could restrict the sample so that the variance of the residuals of the restricted sample is more uniform. I would need to examine the data to see where the data sample could be restricted the best. Another thing that I could possibly do to fix the model is to change the kind of regression (so maybe we have a squared or logged term) to see if there is a better fit. One more solution I could try to fix the model is using robust standard errors to adjust the model to account for the heteroskedasticity.