ECON 453
Fall 2023
Problem Set 2 – 38 points

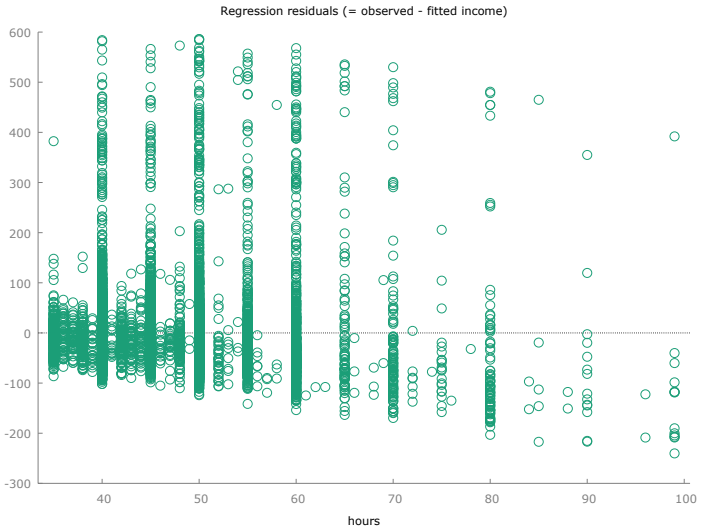
ANSWER KEY

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).
   - 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.

```
Q1: OLS, using observations 1-12704
Dependent variable: income

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const      -182.967      6.59109      -27.76    1.04e-164 ***
  female      -17.9316     1.44087      -12.44    2.39e-035 ***
  age           4.27699    0.154869      27.62    4.36e-163 ***
  hours         3.07145    0.0908645     33.80    6.49e-240 ***
  white        10.7550     1.73394        6.203   5.72e-010 ***

Mean dependent var   93.98083   S.D. dependent var   85.88394
Sum squared resid    79193995   S.E. of regression   78.96986
R-squared            0.154795   Adjusted R-squared   0.154529
F(4, 12699)          581.4392   P-value(F)           0.000000
Log-likelihood      -73528.31   Akaike criterion     147066.6
Schwarz criterion    147103.9   Hannan-Quinn         147079.1

White's test for heteroskedasticity -
  Null hypothesis: heteroskedasticity not present
  Test statistic: LM = 858.042
  with p-value = P(Chi-square(12) > 858.042) = 5.84563e-176
```

We learn from this model that there is a significant gender gap (around $18,000 less for females), there is a significant "white" pay gap (about $10,755 lower income for those that are not white), and that income increases with age and hours worked per week. To me, none of this is particularly surprising. The gender gap is a well-known reality. The racial wage gap is also something that is not new in my findings. Age is positive which probably reflects the fact that work experience increases with age, and it makes sense that working more hours (on average) leads to more income.

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



This one is slightly tricky to interpret. As we discussed in class, this is why we have the formal test that we conduct in the next part of the problem. One reason this is a little bit difficult to work with is that hours worked is a variable people tend to report on "round" numbers (40, 45, etc.). This means we have some significant bunching in our data. To me, I am troubled by what looks like a slight negative trendline to the residuals, but a trendline does not necessarily signal a heteroskedasticity problem. I also think the variance of the residuals starts off fairly narrow, then widens, then narrows again. This type of changing variance is a concern.

- 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).

```
White's test for heteroskedasticity
OLS, using observations 1-12704
Dependent variable: uhat^2

              coefficient    std. error    t-ratio    p-value
  ------------------------------------------------------------
  const        16285.2       18282.4        0.8908   0.3731
  female       25729.4        4704.09       5.470    4.60e-08  ***
  age          -1171.41        892.343     -1.313    0.1893
  hours         -822.494       329.407     -2.497    0.0125    **
  white       -12481.4        5486.64      -2.275    0.0229    **
  X2_X3          -493.893       109.362     -4.516    6.35e-06  ***
  X2_X4          -249.889        66.6571    -3.749    0.0002    ***
  X2_X5         -1942.16       1194.30      -1.626    0.1039
  sq_age           -4.55362      12.8565    -0.3542   0.7232
  X3_X4            49.6546        6.84395    7.255    4.24e-013 ***
  X3_X5           206.236       131.967       1.563    0.1181
  sq_hours         -1.41286       1.96966   -0.7173   0.4732
  X4_X5           173.268        76.9097      2.253    0.0243    **

  Unadjusted R-squared = 0.067541

Test statistic: TR^2 = 858.042030,
with p-value = P(Chi-square(12) > 858.042030) = 0.000000
```

The only thing we really care about in the White's test results is the p-value at the bottom. We go into this with the null hypothesis that there is no problem (homoskedastic errors). Since our p-value is so low, we have no choice but to reject the null and finally admit that we have a problem. In terms of what to do next, we have a couple of options. We could try to adjust our specification (restrict the sample, try logging the income variable, etc.) or we could correct the standard errors. This would mean we would use what are generally called "robust" standard errors.

2. (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

```
Q2: OLS, using observations 1-12704
Dependent variable: l_income

              coefficient    std. error    t-ratio    p-value
  ------------------------------------------------------------
  const        2.09972       0.0462656     45.38     0.0000    ***
  female      -0.146326      0.0101141    -14.47     4.59e-047 ***
  age          0.0358274     0.00108709    32.96     1.22e-228 ***
  hours        0.0231502     0.000637816   36.30     1.51e-274 ***
  white        0.106632      0.0121713      8.761    2.18e-018 ***

Mean dependent var   4.319004   S.D. dependent var   0.615883
Sum squared resid    3902.062   S.E. of regression   0.554322
R-squared            0.190175   Adjusted R-squared   0.189920
F(4, 12699)          745.5416   P-value(F)           0.000000
Log-likelihood      -10528.22   Akaike criterion     21066.44
Schwarz criterion    21103.69   Hannan-Quinn         21078.90

Log-likelihood for income = -65396.8
```

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

The key thing to remember when we log the dependent variable is that we should be interpreting the coefficients as approximating percentage changes. For example, the female coefficient is now telling us that females make about 14.6% less income than males, all else equal. The white/nonwhite gap is estimated to be about 10.7%. Each year of age increases income by about 3.6%, and each additional hour worked per week is estimated to increase income by about 2.3%.

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

```
White's test for heteroskedasticity
OLS, using observations 1-12704
Dependent variable: uhat^2

             coefficient    std. error    t-ratio    p-value
  ------------------------------------------------------------
  const      -0.0578975     0.354283      -0.1634    0.8702
  female      0.159795      0.0911579      1.753     0.0796    *
  age         0.0388468     0.0172922      2.246     0.0247    **
  hours      -0.0312049     0.00638339    -4.888     1.03e-06  ***
  white      -0.0821056     0.106323      -0.7722    0.4400
  X2_X3      -0.00567012    0.00211927    -2.676     0.0075    ***
  X2_X4      -0.000194162   0.00129171    -0.1503    0.8805
  X2_X5      -0.0265350     0.0231437     -1.147     0.2516
  sq_age     -0.000578536   0.000249139   -2.322     0.0202    **
  X3_X4       0.000336098   0.000132625    2.534     0.0113    **
  X3_X5       0.00426307    0.00255731     1.667     0.0955    *
  sq_hours    0.000316945   3.81690e-05    8.304     1.11e-016 ***
  X4_X5      -0.00116224    0.00149039    -0.7798    0.4355

  Unadjusted R-squared = 0.066291

Test statistic: TR^2 = 842.159093,
with p-value = P(Chi-square(12) > 842.159093) = 0.000000
```

No. I ran White's Test on our model from question 2, and we still have a significant heteroskedasticity problem. The p-value on our test is still 0.00000000. Biscuits!

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

I like the second one (logged dependent variable) more than the first. Part of this is that I have been around these types of models for a long time at this point, and it is very much the standard to log income in these kinds of estimates. A major reason for that is annual income tends to be a very skewed variable. As we discussed, taking logarithmic transformations is one way to help deal with skew that might otherwise cause bias in our estimates. The other thing is I prefer to discuss things like the gender gap in percentage, rather than dollar, terms. I think it makes it easier to discuss and communicate. If we say females makes about 18,000 less than men, the obvious follow-up question is "what is the base we are taking that 18,000 from?". If we say females make about 15% less than males, then we do not need to have another number to compare to. We might also note that the adjusted-R^2 value is higher in the version of the model with the logged income variable.

3. (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.

Reminder – we create the interaction term by multiplying the two variables together. In this case, that means multiplying the female dummy by the white dummy. Here is what I get when I include that in my regression:

```
Q3: OLS, using observations 1-12704
Dependent variable: income

              coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       -185.909      6.69000     -27.79     4.83e-165 ***
  female       -11.0236     3.07007      -3.591    0.0003    ***
  age            4.27003    0.154859     27.57     1.35e-162 ***
  hours          3.06540    0.0908759    33.73     5.84e-239 ***
  white         14.9482     2.39028       6.254    4.14e-010 ***
  fem_white     -8.83401    3.46703      -2.548    0.0108    **

Mean dependent var   93.98083   S.D. dependent var   85.88394
Sum squared resid    79153524   S.E. of regression   78.95279
R-squared            0.155227   Adjusted R-squared   0.154894
F(5, 12698)          466.6510   P-value(F)           0.000000
Log-likelihood      -73525.07   Akaike criterion     147062.1
Schwarz criterion    147106.8   Hannan-Quinn         147077.1
```

   - 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.

Male, non-white = -185.909 + (4.270*33) + (3.065*40) = 77.601

Male, white = -185.909 + (4.270*33) + (3.065*40) + (14.948*1) = 92.549

Female, non-white = -185.909 – (11.024*1) + (4.270*33) + (3.065*40) = 66.577

Female, white = -185.909 – (11.024*1) + (4.270*33) + (3.065*40) + (14.948*1) – (8.834*1) = 72.691

   - 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 interaction term is statistically significant, so we first learn that there is a difference in how the gender gap affects different races, as well as how the racial gap affects different genders. Thinking about what the estimates above tell us can help you think through what the interaction term tell us more clearly. For example, for this type of person (33 years old, 40 hours per week), the gender gap for non-white individuals is estimated to be about $11,024. You can find this by subtracting the non-white male and non-white female (77,601 – 66,577). We also know this from the "female" coefficient in the regression. For white individuals, this gender gap will be adjusted down by the interaction coefficient. The gap there should be -11.024 – 8.834 = -19.858. This is what we get if we subtract the white female estimated income from the white male (92,549 – 72,691 = 19,858). So, the gender gap is larger for white individuals. We could also look at how the gap between white/non-white differs by gender. The estimated gap for males is $14, 948. For females, the gap between white and non-white is lower, $6,114.

4. (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 going to test whether the impact of marriage on income is different between males and females. My hypothesis is that marriage will have a positive impact on income for males. This could be because a worker will tend to take on more tasks/go for promotions etc., if they are attempting to "provide" for a family. For females, I am not as sure. The same logic may also apply, but there is also the issue of marriage leading to the potential for children, and children tend to impact female incomes in a negative way.

- 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.

I created a dummy variable for "married" that I defined as marst<3. For this one, I counted "separated" people as not married, though that is debatable (they are technically still married, in a legal sense). I then created the interaction variable between married and female.

- Report/copy your results.

```
Model 6: OLS, using observations 1-12704
Dependent variable: income

              coefficient   std. error   t-ratio    p-value
  ------------------------------------------------------------
  const       -162.538      6.55706     -24.79     1.65e-132 ***
  female        -8.78785    2.14846      -4.090    4.33e-05  ***
  age            3.50672    0.166768     21.03     1.61e-096 ***
  hours          3.06995    0.0903690    33.97     3.41e-242 ***
  married       24.2230     1.94400      12.46     1.98e-035 ***
  married_fem  -17.9917     2.86310      -6.284    3.41e-010 ***

Mean dependent var   93.98083   S.D. dependent var   85.88394
Sum squared resid    78451775   S.E. of regression   78.60202
R-squared            0.162716   Adjusted R-squared   0.162387
F(5, 12698)          493.5418   P-value(F)           0.000000
Log-likelihood      -73468.50   Akaike criterion     146949.0
Schwarz criterion    146993.7   Hannan-Quinn         146964.0
```

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

What we learn from my regression is that married males do indeed make more money than non-married males. The estimated coefficient tells us I expect married males to make about $24,223 more per year than non-married males. The interaction term tells us that the impact of marriage on income for females is significantly lower than it is for males. Females that are married are estimated to make about 6,231 more than females that aren't married (24.223-17.992 = 6.231). This comes from taking the "married" coefficient and adjusting it based on the interaction term. Marriage is still estimated to increase income for females, but only by about ¼ the amount that it impacts income for males. As always, my hypotheses have been proven correct.

5. (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

```
Q5: OLS, using observations 1-12704
Dependent variable: l_income

            coefficient   std. error    t-ratio     p-value
  ------------------------------------------------------------
  const       2.24574      0.0461893      48.62     0.0000    ***
  female     -0.121357     0.0103022     -11.78     7.27e-032 ***
  age         0.0364003    0.00108359     33.59     4.35e-237 ***
  hours       0.0228666    0.000634969    36.01     1.68e-270 ***
  MKTG       -0.129867     0.0148314      -8.756    2.27e-018 ***
  FIN         0.00330367   0.0147639       0.2238   0.8229
  ACCT       -0.155654     0.0147502     -10.55     6.32e-026 ***

Mean dependent var   4.319004   S.D. dependent var   0.615883
Sum squared resid    3862.500   S.E. of regression   0.551548
R-squared            0.198386   Adjusted R-squared   0.198007
F(6, 12697)          523.7144   P-value(F)           0.000000
Log-likelihood      -10463.49   Akaike criterion     20940.98
Schwarz criterion    20993.13   Hannan-Quinn         20958.42
```

   - 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?

To test this one, we simply need to look at the coefficient on Accounting. Since Economics is our omitted (reference) category, this is what the Accounting coefficient is comparing. This means my null hypothesis is really just: $\beta_{ACCT} = 0$. The fact that the p-value indicates this is a statistically significant coefficient tells us that we can reject the null, Economics majors make more than Accounting majors, based on our sample. Suck it, accounting.

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

Oh, now this one is more fun. This will involve us testing the equality of two of our estimated coefficients. My null in this case is that $\beta_{ACCT} = \beta_{MKTG}$. To test this, I will use gretl's test of linear restrictions. My results are below.

```
Restriction:
 b[MKTG] - b[ACCT] = 0

Test statistic: F(1, 12697) = 3.81058, with p-value = 0.0509516

Restricted estimates:

            coefficient   std. error    t-ratio     p-value
  ------------------------------------------------------------
  const       2.24770      0.0461835      48.67     0.0000    ***
  female     -0.120508     0.0102941     -11.71     1.71e-031 ***
  age         0.0363052    0.00108261     33.53     2.56e-236 ***
  hours       0.0228854    0.000634966    36.04     6.34e-271 ***
  MKTG       -0.142938     0.0132353     -10.80     4.53e-027 ***
  FIN         0.00329905   0.0147656       0.2234   0.8232
  ACCT       -0.142938     0.0132353     -10.80     4.53e-027 ***

  Standard error of the regression = 0.55161
```

The p-value on this test is about 0.051, which is right near the standard threshold for determining statistical significance. There is some evidence that Marketing makes more than Accounting, but it is not definitive.

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

Overall, we learn that Economics and Finance majors make about the same income, and Accounting and Marketing majors make about the same income, with Marketing possible earning slightly more than Accounting. We learn that the Econ/Finance people make significantly more than the Accounting/Marketing people. These results are a little bit surprising to me, but I think at least a part of this is due to the fact that Econ/Finance people are more likely to be located in larger cities. This would mean they might make more income because of a higher cost of living where they are located. There might also be something about Marketing majors taking on more varied job types as compared to Econ/Finance, but I do not know this for sure, just speculating.

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

Based on the results we have so far, we can probably make our model more efficient by comparing just two categories for major (instead of 4). We can definitely combine ECON/FIN and can probably combine ACCT/MKTG. The other things we could think about would be what other kinds of factors we can account for. Could we find data, for example, to account for where the individuals live? Could we account for occupation? There are many factors that would predict income that we haven't included in the model to this point.

6. (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?
   - Summarize any steps you took to create/adjust variables or the sample (if necessary).
   - Report/copy your results.
   - Summarize what we learned from your regression. Do the results seem reasonable to you/match your expectations?
   - Assess the validity of your model regarding the issues of multicollinearity and heteroskedasticity. Are these problematic in your model? Present your evidence for/against.
   - What would you do to improve this analysis going forward?

These wildcard questions are really just a gift from me to you. It is the gift of getting to investigate the data on your own and try to succinctly summarize what we learn from your analysis. As before, the things I am looking for on these problems are that the analysis you perform matches the question you asked, and that you are interpreting the results correctly.