ECON 453
In-Class Exercise 5
October 5, 2023

ANSWER KEY

1. Fun with Omitted Variable Bias (OVB)
    a. Regress income on the following regressors: female and hours worked. Report and briefly interpret the female coefficient.

Here are my results:

```
Model 1: OLS, using observations 1-4492
Dependent variable: income

              coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       -59.2011      9.03446     -6.553    6.28e-011 ***
  female      -17.2232      3.67296     -4.689    2.82e-06  ***
  hours         4.12609     0.189941    21.72     1.42e-099 ***

Mean dependent var   123.5052   S.D. dependent var   119.0908
Sum squared resid    56963743   S.E. of regression   112.6483
R-squared            0.105668   Adjusted R-squared   0.105269
F(2, 4489)           265.1939   P-value(F)           1.4e-109
Log-likelihood      -27593.79   Akaike criterion     55193.59
Schwarz criterion    55212.82   Hannan-Quinn         55200.36
```

The female coefficient is -17.223. This means we expect a female to make about $17,223 less in annual income then a male. This is our estimate of the gender gap in earnings.

    b. How biased might our estimate of the gender gap be if we do not know the hours of work for an individual (we were forced to omit it from our model)? To find this, you first need to run a regression where hours is the dependent variable and female is the independent variable. Report this coefficient and discuss what it tells us.

```
Model 2: OLS, using observations 1-4492
Dependent variable: hours

              coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       46.3597       0.158761    292.0     0.0000    ***
  female      -2.57359      0.286019     -8.998   3.33e-019 ***

Mean dependent var   45.56679   S.D. dependent var   8.929260
Sum squared resid    351732.5   S.E. of regression   8.850812
R-squared            0.017713   Adjusted R-squared   0.017494
F(1, 4490)           80.96359   P-value(F)           3.33e-19
Log-likelihood      -16167.72   Akaike criterion     32339.44
Schwarz criterion    32352.26   Hannan-Quinn         32343.96
```

The dependent variable in this one is the usual hours worked in a week. The results tell us that females work (on average) about 2.6 hours less per week. These are self-reported hours work, so it might be more appropriate to say that females report working about 2.6 hours fewer per week than males.

    c. Use the estimated coefficients on hours from part a and female from part b to determine how much bias we would have if we omitted hours of work from our model estimating the gender gap. What do you expect the estimated gender gap would be after the bias?

Determining the omitted variable bias comes from knowing the relationship between the omitted variable (hours worked) and the dependent variable (income) as well as the relationship between the omitted and the regressor of interest (female). We now have both of these from parts a and b. We can find the amount of bias by multiplying

together. If we omit hours, the bias on the female coefficient would be = 4.12609 * -2.57359 = -10.6189. We should expect the female coefficient will be 10.6189 lower if we omit the hours variable (there will be a negative bias).

d. Run a regression where you regress income on (only) the female dummy variable. Compare to your estimate from part c and be impressed (or sad).

```
Model 3: OLS, using observations 1-4492
Dependent variable: income

            coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       132.083       2.12393     62.19     0.0000    ***
  female      -27.8421      3.82642     -7.276    4.03e-013 ***

Mean dependent var   123.5052   S.D. dependent var   119.0908
Sum squared resid    62951864   S.E. of regression   118.4080
R-squared            0.011654   Adjusted R-squared   0.011434
F(1, 4490)           52.94429   P-value(F)           4.03e-13
Log-likelihood      -27818.29   Akaike criterion     55640.59
Schwarz criterion    55653.41   Hannan-Quinn         55645.10
```

In part a, our estimated female coefficient was -17.2232. We just learned that omitting the hours variable should cause a negative bias of 10.6189 units. This means the estimate gender gap would be -17.2232 + -10.6189 = -27.8421. OMG, that is what we found in our regression. Super cool.

The main idea here is that knowing hours worked is important in predicting income. If we do not have that information, it may cause us to overstate the amount of the gender gap that exists. Females tend to work fewer hours, so this explains part (but not nearly all) of why female incomes are lower.

2. Run a regression where you use income as the dependent variable and the following regressors: female, age, hours, masters.

```
Model 4: OLS, using observations 1-4492
Dependent variable: income

            coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const      -215.932      12.9449     -16.68    1.17e-060 ***
  female      -19.3392      3.56434     -5.426   6.07e-08  ***
  age           4.54663     0.287161    15.83    5.54e-055 ***
  hours         3.97721     0.183714    21.65    6.12e-099 ***
  masters      19.4530      3.55116      5.478   4.54e-08  ***

Mean dependent var   123.5052   S.D. dependent var   119.0908
Sum squared resid    53128386   S.E. of regression   108.8141
R-squared            0.165883   Adjusted R-squared   0.165139
F(4, 4487)           223.0853   P-value(F)           7.0e-175
Log-likelihood      -27437.24   Akaike criterion     54884.48
Schwarz criterion    54916.53   Hannan-Quinn         54895.77
```

a. Briefly discuss the findings and whether these match your expectations.

We find that the gender gap is around $19,339, each year of age increases earnings by about $4,500, each additional hour worked increases earnings by about $4k, and having a master's degree increases earnings by about $19,450. These are all statistically significant and make sense to me. It seems odd that income would increase just because of age, but this is probably just a proxy for work experience in this case.

b. Find/calculate the standardized coefficients for the hours and age variables from the regression in part a. Comment briefly on what these tell us.

I prefer to use the "sols" function package in gretl. When I do that, I get these results:

Independent variables:

	std. dev.	standardized coeff.
const	0	-1.8132
female	0.46176	-0.074986
age	5.7576	0.21981
masters	0.4676	0.07638
hours	8.9293	0.29821

The standardized coefficients on age and hours tell us that hours worked tend to have a larger impact on income than a person's age. A one standard deviation increase in age is expected to increase income by 0.22 standard deviations, while a one standard deviation increase in hours is expected to increase income by about 0.30 standard deviations. As a sidenote, we don't really want/need the standardized coefficients on dummy variables (what does a one standard deviation increase in "female" mean?). You need to have them included to ensure the model/standardized values of the other variables are correct.

c. Run the same model (as in part b) but add dummy variables for race categories. Think carefully about how you create categories/adjust the sample. You should include at least two dummy variables for race/ethnicity. Briefly explain your thought process on how to add race to the model. Briefly summarize what you learned about race and income.

Race categories are tricky to deal with in Census data. There are many categories, and while this is a pretty large sample (4492 individuals), there are only 6 Native Americans. Some of the other categories also have relatively low numbers. I chose to restrict the sample to remove Native Americans (too small of a sample), as well as the "other", and multiple race categories. This will hopefully give a clearer picture of how income impacts race. I will include dummies for black and "Asian" (combine categories 4, 5, and 6), and Hispanic. "White" will be my omitted, "reference" category.

Model 5: OLS, using observations 1-4285
Dependent variable: income

	coefficient	std. error	t-ratio	p-value	
const	-209.942	13.4183	-15.65	1.06e-053	***
female	-17.2349	3.71053	-4.645	3.51e-06	***
age	4.54146	0.295047	15.39	4.48e-052	***
hours	3.95389	0.189952	20.82	9.68e-092	***
masters	19.3173	3.64198	5.304	1.19e-07	***
Black	-37.9181	7.85523	-4.827	1.43e-06	***
Asian	-7.60260	4.39662	-1.729	0.0838	*
hispanic	-24.1469	7.21218	-3.348	0.0008	***

Mean dependent var	124.5645	S.D. dependent var	119.8263
Sum squared resid	50898064	S.E. of regression	109.0890
R-squared	0.172539	Adjusted R-squared	0.171185
F(7, 4277)	127.4032	P-value(F)	1.0e-170
Log-likelihood	-26182.07	Akaike criterion	52380.14
Schwarz criterion	52431.05	Hannan-Quinn	52398.12

Notice in my model table that the sample size (n) is lower in my Model 3. This is because of my restricted sample. Here I learn that all three categories are lower than white, with black having the largest estimated magnitude. The Asian coefficient is negative, but is only weakly significant (p-value of .0838).

d. Use the results in your model from part c. Test the equality of two (or more!) of the coefficients Report what you tested and your conclusion (briefly).

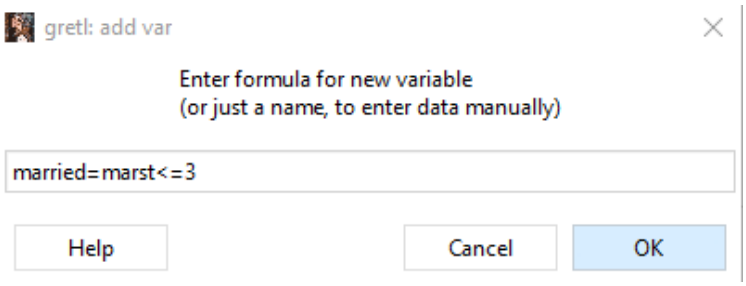
I will test whether the black and Hispanic coefficients are significantly different from each other. To do this, I go to my Model 3 results window and select Tests → Linear restrictions. In the window that popped up I typed: b[6]-b[8]=0. The results came up as:

```
Restriction:
 b[Black] - b[hispanic] = 0

Test statistic: F(1, 4277) = 1.78087, with p-value = 0.182113

Restricted estimates:

              coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       -210.601      13.4105      -15.70    4.47e-054 ***
  female       -17.3520      3.70983      -4.677   3.00e-06  ***
  age            4.54271     0.295073     15.40    4.29e-052 ***
  hours          3.96795     0.189677     20.92    1.33e-092 ***
  masters       19.2823      3.64221       5.294   1.26e-07  ***
  Black        -30.4060      5.47928      -5.549   3.04e-08  ***
  Asian         -7.51975     4.39658      -1.710   0.0873    *
  hispanic     -30.4060      5.47928      -5.549   3.04e-08  ***

  Standard error of the regression = 109.099
```

The null hypothesis here is that the two coefficients are equal. Since the p-value is 0.18, I will fail to reject the null. While both black and Hispanic are significantly lower than white incomes, they are not significantly different from each other.

3. For this set of models, we will use hours worked per week (**hours**) as the dependent variable.
   a. Create two dummy variables, one for "married", and one for "kids". The "married" should be a simple yes/no regarding whether the person is currently married. The "kids" variable should be 0 if the person has no children in the household and 1 if there are any. Explain briefly how you created these.

I included "Married, spouse present", "Married, spouse absent", and "Separated" as "Married". The separated group is debatable, but I tend to be an optimist and I am still holding out hope for these people's relationships until it becomes official.



b. Run a regression using hours as the dependent variable where regressors are female, age, married, kids. Briefly comment on your findings.

```
Model 6: OLS, using observations 1-4492
Dependent variable: hours

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       44.2023       0.887699     49.79     0.0000    ***
  female      -2.54273      0.286098     -8.888    8.88e-019 ***
  age          0.0562092    0.0273512     2.055    0.0399    **
  married      0.789496     0.339573      2.325    0.0201    **
  kids        -0.634644     0.358615     -1.770    0.0768    *

Mean dependent var   45.56679   S.D. dependent var   8.929260
Sum squared resid    350819.1   S.E. of regression   8.842266
R-squared            0.020264   Adjusted R-squared   0.019390
F(4, 4487)           23.20074   P-value(F)           5.26e-19
Log-likelihood      -16161.88   Akaike criterion     32333.76
Schwarz criterion    32365.81   Hannan-Quinn         32345.05
```

I find that females work an average of about 2.5 hours less per week, hours increase (slightly) with age. Married people work a little bit more (I get it) and people with kids in the household work a little bit less. Female, age, and married all have p-values below 0.05, while the kids variable is only weakly significant.

c. Create an interaction term between married and female. Add this to the model from part b. Discuss what this tells us about gender, marriage, and hours worked.

The interaction term is calculated by simply multiplying the two variables together, (married_female =married*female).

File Edit Tests Save Graphs Analysis LaTeX

```
Model 7: OLS, using observations 1-4492
Dependent variable: hours

                 coefficient   std. error   t-ratio   p-value
  ------------------------------------------------------------
  const           43.8679       0.889099     49.34     0.0000   ***
  female          -1.03204      0.445396     -2.317    0.0205   **
  age              0.0510193    0.0273202     1.867    0.0619   *
  married          1.65467      0.391363      4.228    2.41e-05 ***
  kids            -0.681447     0.358034     -1.903    0.0571   *
  married_female  -2.56785      0.581074     -4.419    1.01e-05 ***

Mean dependent var   45.56679   S.D. dependent var   8.929260
Sum squared resid    349298.5   S.E. of regression   8.824066
R-squared            0.024510   Adjusted R-squared   0.023423
F(5, 4486)           22.54301   P-value(F)           2.11e-22
Log-likelihood      -16152.12   Akaike criterion     32316.24
Schwarz criterion    32354.70   Hannan-Quinn         32329.80
```

The interaction term between married and female is statistically significant. This tells us that the impact of marriage on hours worked differs by gender. The coefficient on married (1.655) tells us that married males tend to work 1.65 hours more per week than non-married males. The interaction term tells us how to adjust that for females. For females, married decreases hours worked by about 0.9 hours (1.655 – 2.568 = -0.913).

d. Now create a second interaction term between kids and female. Add this to the model from part c (yes, you will have two interaction terms!). Create two separate prediction equations, one for males and one for females. The equations should tell us how to predict hours worked based on age, marital status, and child ownership. Summarize what we learn.

I mean, how much fun is this? Right? Two interaction terms in one regression. Here are my results:

```
Model 8: OLS, using observations 1-4492
Dependent variable: hours

                  coefficient   std. error   t-ratio   p-value
  ------------------------------------------------------------
  const            43.8521      0.888880     49.33     0.0000   ***
  female           -0.938707    0.447975     -2.095    0.0362   **
  age               0.0539136   0.0273548     1.971    0.0488   **
  married           1.37459     0.418164      3.287    0.0010   ***
  kids             -2.05067     0.805458     -2.546    0.0109   **
  married_female   -2.65521     0.582727     -4.557    5.34e-06 ***
  married_kids      1.64481     0.866788      1.898    0.0578   *

Mean dependent var   45.56679   S.D. dependent var   8.929260
Sum squared resid    349018.3   S.E. of regression   8.821509
R-squared            0.025293   Adjusted R-squared   0.023989
F(6, 4485)           19.39687   P-value(F)           1.87e-22
Log-likelihood      -16150.32   Akaike criterion     32314.64
Schwarz criterion    32359.51   Hannan-Quinn         32330.45
```

The coefficient on "kids" tells us that males with kids and males without kids work the same amount of hours (on average). For females, however, the interaction term is -2.14. Females with kids work about 2 hours fewer per week as compared to females without children in the household. This is consistent with research showing females tend to still take on more household duties (even in cases where the female makes more money). The married coefficient says married males work about an hour more than non-married males (I get it). When we adjust this with the interaction term, there is really no impact of marriage on hours for females. This tells us our finding in part c comes from the fact that marriage and having children tend to be correlated. The real reduction of hours for females come from the kids, not the marriage.

To estimate these equations:

For males (this is easy, since interaction terms and female values will all be 0)

Predicted hours for males = 43.82 + .05*age + 1.2*married + 0.04*kids.

For females, we will plug in 1 for the female variable

Predicted hours = 43.82 – (0.81*1) +0.05*age+1.2*married+0.04*kids-1.34(married*1) -2.14(kids*1)

Predicted hours for females = 43.01 + 0.05*age – 0.14*married – 2.10*kids