Project Assignment #2

Answer the following questions using your ACS data set. To complete this assignment, you will upload three things to canvas: (1) a Word or PDF document containing your complete answers to the questions below; (2) your Stata "do" program; and (3) your Stata "log" file. The Stata "do" program (and therefore your "log" file) must include comments stating which question you are answering with the code you write.

1. Create the following variables (these commands will appear in your Stata do and log files. Label every variable. You do not need to write an explanation of these commands in your written answers):

   a. "wage". Create the wage variable described in Project Assignment #1 (include the same drop commands).

   b. "female". A dummy variable equal to one if the person is female; zero otherwise.

   c. "schoolyr". Years of schooling. Same as in Project Assignment #1.

   d. "exp". Potential labor market experience. "exp" equals the person's age minus their years of schooling minus six.

   e. "private", "govt" and "othemp". Three mutually exclusive and exhaustive dummy variables. "private" equals one if the person is employed by private for profit company, zero otherwise. "govt" equals one if the person is employed by any level of government, zero otherwise. "othemp" is everyone else.

      Note: use the variable "classwkrd" to create the new variables. Check that your three new variables are mutually exclusive and exhaustive by executing the following commands:

      ```
      assert private+govt+othemp==1
      ```

   f. "married". A dummy variable equal to one if the person is married.

   g. "foreign" dummy variable equal to one if the person was born outside the U.S. (use the variable "bpl").

i. “hispan”, “black”, “asian”, “white”, “othrace”. Use the following Stata code to create 5 race categories:

```
   gen byte hisp           = (hispand>=100 & hispand<=499)
   label variable hisp     "Hispanic"

   gen byte white             = (raced==100 | raced==110) & hisp==0
   gen byte black             = raced==200 & hisp==0
   gen byte indian            = inrange(raced,300,399)==1 & hisp==0
   gen byte asian             = inrange(raced,400,699)==1 & hisp==0
   gen byte other_race        = inrange(raced,700,996)    & hisp==0
   label variable white       "White"
   label variable black       "Black"
   label variable indian      "Indian"
   label variable asian       "Asian"
   label variable other_race  "Other Race"

   assert white+hisp+black+asian+indian+other_race==1

   gen byte race_cat6       = 0
   replace  race_cat6       = 1 if white==1
   replace  race_cat6       = 2 if hisp==1
   replace  race_cat6       = 3 if black==1
   replace  race_cat6       = 4 if asian==1
   replace  race_cat6       = 5 if indian==1
   replace  race_cat6       = 6 if other_race==1
   label variable race_cat6  "Race Category (race is non-Hispanic)"

#delimit ;
   label define race_cat6_lbl
                            0   "ERROR!"
                            1   "White"
                            2   "Hispanic"
                            3   "Black"
                            4   "Asian"
                            5   "Indian"
                            6   "Other race";
#delimit cr

   label values race_cat6 race_cat6_lbl
   tab race_cat6, missing
   assert inlist(race_cat6,1,2,3,4,5,6)==1 & r(r)==6
```

j. Create the following education categories:

Dummy Variable	**Definition (highest education level)**
$elem_i$ =	Less than a high school education.
hs_i =	High school diploma.
$college_i$ =	College degree.
ma_i =	Masters degree.
phd_i =	Ph.D.

Use the variable “educd” to create the new variables. You may (and should) tailor the specific categories to fit with your specific occupation. What is the mean of each category? Check that your new variables are mutually exclusive and exhaustive by executing the following command:

```
assert elem+hs+college+ma+phd==1
```

Create the variable “educ_cat5” that is analogous to the “race_cat6” variable, but for the educations categories.

2. In a word processing program (i.e., Microsoft Word), create a table called Table 1. In Table 1, list all of the variables you have created along with a brief description of each. For example:

Table 1: Variable Definitions

Variable	Description
Wage	Hourly wage.
Female	Dummy variable equal to one for females, zero for males.
Age	Age in years.
Experience	Potential labor market experience (age – years of school – 6).

3. In a word processing program (i.e., Microsoft Word), create a table called Table 2. In Table 2, list all of the variables you described in Table 1 along with their means and standard errors. Also list the means and standard errors of each variable for men and women separately. Describe any interesting features of your version of Table 2. How do the characteristics men and women differ? How are they similar?

Table 2: Schoolteacher Characteristics, by Gender

	Men	Women	All
	(1)	(2)	(3)
Wage	15.59	13.57	14.11
	(.15)	(.09)	(.08)
Age	42.16	41.17	41.44
	(.20)	(.12)	(.11)
Years of schooling	14.35	13.95	14.05
	(.03)	(.02)	(.01)
Potential years of experience	19.34	18.88	19.00
	(.20)	(.12)	(.11)
Married	.78	0.73	.75
	(.01)	(.01)	(.002)
Race/Ethnicity:			
White	.88	.85	.86
	(.01)	(.001)	(.001)
Hispanic	.03	.03	.03
	(.002)	(.003)	(.001)
Black	.06	.09	.08
	(.01)	(.002)	(.001)
Asian	.01	.01	.01
	(.002)	(.003)	(.001)
Native American	.01	.01	.01
	(.001)	(.001)	(.002)
Other Race	.01	.01	.01
	(.001)	(.002)	(.001)
Sample Size	2,256	6,133	8,389

Source: 2010-2020 American Community Survey (ACS) data. The samples include elementary and secondary schoolteachers aged 18-65.

Notes—Standard error in parenthesis. White, black, Asian, Native American, and other race are non-Hispanic. Other race includes two or more races.

4. Estimate the OLS regression: $wage_i = \alpha + \beta \exp_i + \varepsilon_i$.

   a. Interpret the OLS coefficients $\hat{\alpha}$ and $\hat{\beta}$.

   b. In the regression output, Stata reports a t-statistic. Show exactly how Stata calculates the reported t-statistics for $\hat{\beta}$.

   c. Is the OLS coefficients $\hat{\beta}$ statistically different from zero at the 90, 95, and 99% confidence levels? Show your work.

   d. Use Stata's "predict" command to generate the predicted wage (type "help predict"). What is the average predicted wage? How does the average predicted wage compare to the average observed wage? Explain your result.

5. Estimate the OLS regression: $wage_i = \alpha + \beta \exp_i + \varepsilon_i$ twice, once for men and once for women.

   a. Plot the sample regression functions for men and women using the following Stata command. After you create the graph, cut-and-paste it into your Word document (the graph will not appear in your log file).

```
twoway lfit wage exp, by(sex)
```

   b. Are the returns to experience the same for men and women?

   c. Estimate the regression: $wage_i = \alpha + \beta_1 \exp_i + \beta_2 f_i + \beta_3 (\exp_i \times f_i) + \varepsilon_i$, where $f_i = female_i$. How does the output from this single regression compare to the output from the two regressions above? Explain why this is the way it is.

6. Create a "centered" age variable called "*age_centered*" equal to $age_i - \overline{age}$, and a standardized age variable called "*age_std*" equal to $(age_i - \overline{age}) / std.dev(age)$. Create the standardized age variable two ways: the first by using a single "egen" command, and the second using a single "generate" command immediately following a "summarize" command. Use the "assert" command to show that the two variables are the same.

Estimate the following models:

i. $wage_i = \alpha^a + \beta_1^a female_i + \beta_2^a age_i + \beta_3^a (age_i \times female_i)$.

ii. $wage_i = \alpha^b + \beta_1^b female_i + \beta_2^b (age\_centered_i) + \beta_3^b (age\_centered_i \times female_i)$.

iii. $wage_i = \alpha^c + \beta_1^c female_i + \beta_2^c (age\_std_i) + \beta_3^c (age\_std_i \times female_i)$.

a. Explain how you can calculate $\hat{\alpha}^b$ and $\hat{\beta}^b$ from regression (*ii*) using only the Stata output for regression (*i*) and $\overline{age}$. Modify the Stata code below to verify that your calculations are correct using the Stata output for regressions (*i*) and (*ii*).

```
* Regression (i)
   regress wage female age i.female#c.age
   di _b[_cons]
   di _b[female]
   di _b[age]
   di _b[1.female#c.age]
   sum age if e(sample)==1
   di r(mean)

* Calculate coefficients from (ii) using output from (i)
   local alpha_b = {your equation using the _b vars}
   local beta1_b = {your equation using the _b vars}
   di  "`alpha_b'"
   di  "`beta1_b'"

* Regression (ii)
   regress {your regression for (ii)}
   assert float(_b[_cons])==float(`alpha_b')
   assert float(_b[female])==float(`beta1_b')
```

b. Compare the interpretation of β_1^a to the interpretation of β_1^b. With this interpretation in mind, in what way does testing the null hypothesis $H_0 : \beta_1^a = 0$ have a different interpretation than testing the null hypothesis $H_0 : \beta_1^b = 0$? (hint: draw a graph to help explain your answer).

c. In what way does testing the null hypothesis $H_0 : \beta_1^b = 0$ have a different interpretation than testing the null hypothesis $H_0 : \beta_1^c = 0$? Explain why?

d. Compare all of the coefficients, standard errors, and t-stats between models (*ii*) and (*iii*). Explain the similarities and/or differences in the interpretation of each coefficient. Explain why the similarities and differences exist in the numbers contained in the output of each regression.

e. If you were to choose between models (*ii*) and (*iii*), which would you choose and why? There is not a right or wrong choice here, so explain why you prefer a centered or standardized variable in this case (i.e., age), and why you might choose the other option for a different variable.

7. Estimate the following four models:

i. $\ln(wage_i) = \beta_1 + \beta_2 schoolyr_i + \beta_3 \ln(\exp_i) + \varepsilon_i$

ii. $\ln(wage_i) = \beta_1 + \beta_2 schoolyr_i + \beta_3 \exp_i + \varepsilon_i$

iii. $wage_i = \beta_1 + \beta_2 schoolyr_i + \beta_3 \ln(\exp_i) + \varepsilon_i$

iv. $wage_i = \beta_1 + \beta_2 schoolyr_i + \beta_3 \exp_i + \beta_4 \exp_i^2 + \varepsilon_i$

a. For each model, interpret β_3 using simple non-technical language. For each model, calculate the marginal effect experience has on hourly wages. Do these calculations by had using the regression output. When necessary, calculate the marginal effect and elasticity at the mean of the data. Comment on how the return to experience differs in each regression. For example, which regression suggests the biggest return to experience?

b. Does each of your four regressions have the same number of observations? If not, explain why.

8. Consider the following regression models:

a.
$$\begin{aligned} wage_i = {} & \alpha^a + \beta_1^a \exp_i + \beta_2^a female_i + \\ & \beta_3^a elem_i + \beta_4^a college_i + \beta_5^a ma_i + \beta_6^a phd_i + \varepsilon_i \end{aligned}$$

b.
$$\begin{aligned} wage_i = {} & \alpha^b + \beta_1^b \exp_i + \beta_2^b female_i + \\ & \beta_3^b elem_i + \beta_4^b hs_i + \beta_5^b ma_i + \beta_6^b phd_i + \varepsilon_i \end{aligned}$$

Estimate each model two ways in Stata:

(*i*) Using the education dummy variables that you created earlier.
(ii) Using the variable “educ_cat5” (note: use Stata “factor variables”).

Explain how regressions (a) and (b) differ. The OLS coefficients from regressions (a) and (b) are mathematically related to each other. Explain how. Use your regression results to demonstrate your answer.

9. Estimate the model:

$$\begin{aligned} wage_i = \alpha + \beta_1 \exp_i + \beta_2 \exp_i^2 + \beta_3 \, female_i \, + \\ \beta_4 elem_i + \beta_5 college_i + \beta_6 ma_i + \beta_7 \, phd_i \, + \\ \beta_8 (elem_i)(female_i) + \beta_9 (college_i)(female_i) + \\ \beta_{10}(ma_i)(female_i) + \beta_{11}(phd_i)(female_i) + \varepsilon_i \end{aligned}$$

Using the regression above, compare the wages of the following groups:

a. Men without a high school education to men with a high school education.
b. Women with a masters degree to men with a masters degree.
c. Women with a college degree to men with a college degree.
d. Women with a masters degree to men with a college degree.