EC 570
Fall 2021
Stata Take Home Final Exam
Due: December 10 2021 11:59 PM
Total Points: 65

For this question, you will need to use a dataset of your own choosing.

THIS IS A DRAFT: DO NOT GRADE
Collin Campbell, and
Mitch Priestley, and
Brannigan Vogt,
do not offer this document as our submission.
THIS IS A DRAFT: DO NOT GRADE

Use the data to answer the following questions. **Don't forget to copy and paste in only the relevant part of your Stata output for each question.**

1. Upload your dataset in Stata format via D2L's Dropbox function. **(1 points)**

Steps:

1. On the Assessments tab (near the top) click on Dropbox.
2. Click on Student Data Files.
3. Click on Add a File and upload your data file.
4. In the comments section list all group member names, including yourself.
5. Submit.

If you have trouble uploading your dataset, please contact me ASAP.

Obtain `usa_00003.dat` and `usa_00003.do` from IPUMS. In Stata with proper working directory, `File -> Do` run the do file and it will process in the dat file.

Then clean the data with following Stata command sequence,

```
drop if age > 64
drop if age < 18

replace valueh = . if valueh > 9999990
replace valueh = . if valueh == 0035000
replace valueh = . if valueh == 0050000
replace valueh = . if valueh ==  0100000
replace valueh = . if valueh ==  0200000
replace valueh = . if valueh ==  0400000
replace valueh = . if valueh == 1000000

replace hhincome = . if hhincome == 9999999

replace costelec = . if costelec > 9990
replace costelec =. if costelec == 0000

count if educd == 999

gen collegegrad = (educd >= 101)
```

Dropping sample down to just ages 18-64 brings sample size down to $1,921,823$.

2. Write a **brief** description of your dataset (about 1-2 paragraphs). Who collected the data? What kind of data is it (cross sectional, time series, repeated cross-sectional, or panel)? When were the data collected? How was the sample constructed? What is the population of interest? What are the main variables (or types of variables) covered in the data? Include a proper citation of your data (see some instructions and an example here (http://guides.library.ucsc.edu/citedata)). If possible, include a link to where you downloaded your dataset. **Clarity is very important here and throughout this portion of homework assignments. (2 points)**

The US Census Bureau annually attempts to observe around 3.5 million residential addresses through the American Community Survey (ACS) (https://www.census.gov/programs-surveys/acs). This Stata take-home final exam will consider a dataset derived from the Year 2019 ACS (https://www.census.gov/programs-surveys/acs/about/forms-and-instructions/2019-form.html). These data were accessed using Integrated Public Use Microdata Series (IPUMS). The Census Bureau contacted addresses (https://www.census.gov/programs-surveys/acs/respond/other-ways-to-respond.html) by mail requesting that the dwellers of the residence fill out an online form. If dwellers at some address do not respond online promptly than the Census Bureau will mail a paper version of the questionnaire. Once a paper form has been ignored, the addresses may receive an in-person visit for interview from a Census Bureau employee.

All members of the US should be considered as the population of interest. The Census Bureau also wants more local data, towards offering cities counties and states vital household economic social demographic data, making for an overall regionally-representative sample.

When we derived our cross-sectional 2019 ACS data using IPUMS, we included variables with regards to house value and total household income. Our full list of selected variables includes those two as well as ownership of dwelling, annual electricity cost, annual gas cost, annual water cost, annual home heating fuel cost, multigenerational household, sex, age, race, educational attainment, employment status, total personal income, social security income, welfare (public assistance) income, interest dividend and rental income, retirement income, poverty status, and veteran status.

Steven Ruggles, Sarah Flood, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouweiler and Matthew Sobek. *IPUMS USA: Version 11.0* [dataset]. Minneapolis, MN: IPUMS, 2021. https://doi.org/10.18128/D010.V11.0 (https://doi.org/10.18128/D010.V11.0)

“ACS is monthly rolling samples of households that was designed to replace the Census long form. Nationally-representative ACS data has been available each year since 2000”

“include about 3 million households nationwide”

"The public use samples of the ACS and PRCS are extracted from the Census Bureau’s larger internal data files and are thus subject to additional sampling error and further data processing (such as imputation and allocation).

The sampling unit is the household and all persons residing in the household. To protect individual confidentiality, geographic identifiers are currently restricted to the state level, and individual variables, such as income and housing values, are Top coded.

The ACS/PRCS sample design approximates the Census 2000 long form sample design and oversamples areas with smaller populations. Each month a systematic sample is drawn to represent each U.S. county or county equivalent. The selected monthly sample is mailed the ACS/PRCS survey at the beginning of the month. Nonrespondents are contacted via telephone for a computer assisted telephone interview (CATI) one month later. One third of the nonrespondents to the mail or telephone survey are contacted in person for a computer assisted personal interview (CAPI) one month following the CATI attempt. Weights included with the ACS PUMS for the household and person-level data adjust for the mixed geographic sampling rates, nonresponse adjustments, and individual sampling probabilities. Estimates from the ACS IPUMS samples may not be consistent with summary table ACS estimates due to the additional sampling error."

https://usa.ipums.org/usa/chapter2/chapter2.shtml#ACS (https://usa.ipums.org/usa/chapter2/chapter2.shtml#ACS)

The population of interest in our research analysis is subset of US adults: US residents age 18-64 with total housing units value below one million dollars.

3. Choose **2 continuous (i.e. not dummy/binary or categorical)** variables from your dataset. You will be studying the relationship between these variables so one should be considered an outcome variable y and one an explanatory variable x_1. Choose ones that make sense for this type of relationship. Describe both variables here in words, including what they are meant to capture, units of measurement, etc. Be clear on which is the outcome variable and which is the explanatory variable. **(2 points)**

$$y : valueh \text{ i.e., total home value}$$
$$x_1 : hhincome \text{ i.e., total household income}$$

We have chosen to treat house value as outcome variable and household income as explanatory variable.

"HHINCOME reports the total money income of all household members age 15+ during the previous year. The amount should equal the sum of all household members' individual incomes, as recorded in the person-record variable INCTOT. The persons included were those present in the household at the time of the census or survey. People who lived in the household during the previous year but who were no longer present at census time are not included, and members who did not live in the household during the previous year but who had joined the household by the time of the census or survey, are included. For the census, the reference period is the previous calendar year; for the ACS and the PRCS, it is the previous 12 months.

Note that household income differs from family income, which is reported in FTOTINC. The family income variable only reports the incomes of household members related to the head, while HHINCOME includes the incomes of all household members."

“VALUEH reports the value of housing units in contemporary dollars.”

4. Provide summary statistics (number of observations, mean, standard deviation, minimum and maximum) for your two variables described in part (3). **Calculate these summary statistics only for the sample that you use in part (6).** This means that your sample size for each variable should match the sample size in part (6). To do this, you first run the regression in part (6). Then immediately after you run the regression, you use the summarize command and an "if" statement as follows:

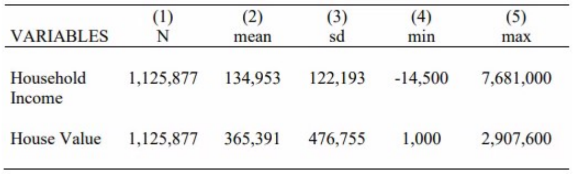
```
regress y x
summarize y x if e(sample)==1
```

Be sure to display the summary statistics **in a neatly formatted table** (i.e. do NOT simply copy and paste Stata output). **(2 points)**

You can refer these:

- Using outreg2 to report regression output, descriptive statistics, frequencies and basic crosstabulations (https://www.princeton.edu/~otorres/Outreg2.pdf)
- In Stata type: `help outreg2`

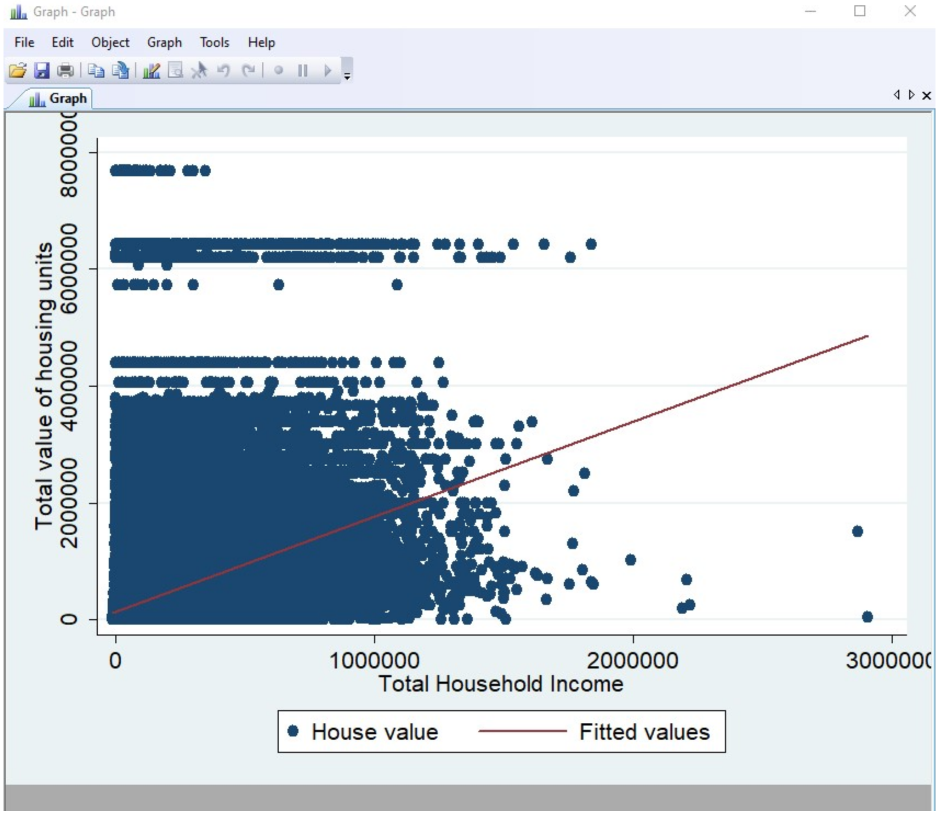
```
summarize valueh hhincome if e(sample)==1
```



	(1)	(2)	(3)	(4)	(5)
VARIABLES	N	mean	sd	min	max
Household Income	1,125,877	134,953	122,193	-14,500	7,681,000
House Value	1,125,877	365,391	476,755	1,000	2,907,600

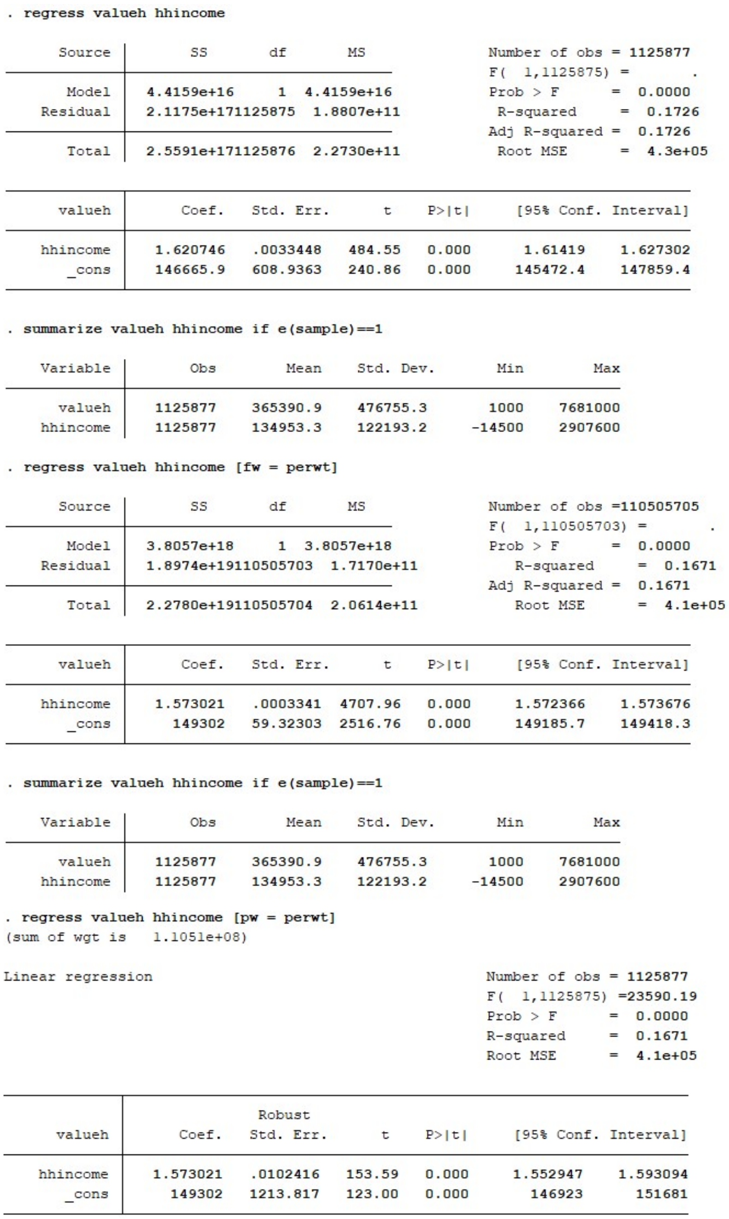
5. Create a scatterplot of you two variables. You may copy and paste your graph from Stata, but please make sure your graph is clearly formatted (e.g. that the axes are labeled clearly with labels that make sense – **not** just nonsensical variable names). **(2 points)**

```
twoway (scatter valueh hhincome) (lfit valueh hhincome) , xtitle("Total Household Income") ytitle("Total value of housing units")
```



6. Regress your outcome variable on your explanatory variable. Copy paste Stata output here. Report your results in an equation (like in the lecture notes) AND in a **formatted table** using outreg2. **(2 points)**

```
regress valueh hhincome
summarize valueh hhincome if e(sample)==1
regress valueh hhincome [fw = perwt]
summarize valueh hhincome if e(sample)==1
```



```
. regress valueh hhincome

      Source |       SS       df       MS              Number of obs = 1125877
-------------+------------------------------           F(  1,1125875) =       .
       Model |  4.4159e+16     1  4.4159e+16           Prob > F      =  0.0000
    Residual |  2.1175e+171125875  1.8807e+11           R-squared     =  0.1726
-------------+------------------------------           Adj R-squared =  0.1726
       Total |  2.5591e+171125876  2.2730e+11           Root MSE      =  4.3e+05

------------------------------------------------------------------------------
      valueh |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    hhincome |   1.620746   .0033448   484.55   0.000      1.61419    1.627302
       _cons |   146665.9   608.9363   240.86   0.000     145472.4    147859.4
------------------------------------------------------------------------------

. summarize valueh hhincome if e(sample)==1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      valueh |   1125877    365390.9    476755.3       1000    7681000
    hhincome |   1125877    134953.3    122193.2     -14500    2907600

. regress valueh hhincome [fw = perwt]

      Source |       SS       df       MS              Number of obs =110505705
-------------+------------------------------           F(  1,110505703) =       .
       Model |  3.8057e+18     1  3.8057e+18           Prob > F      =  0.0000
    Residual |  1.8974e+19110505703  1.7170e+11           R-squared     =  0.1671
-------------+------------------------------           Adj R-squared =  0.1671
       Total |  2.2780e+19110505704  2.0614e+11           Root MSE      =  4.1e+05

------------------------------------------------------------------------------
      valueh |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    hhincome |   1.573021   .0003341  4707.96   0.000     1.572366    1.573676
       _cons |     149302   59.32303  2516.76   0.000     149185.7    149418.3
------------------------------------------------------------------------------

. summarize valueh hhincome if e(sample)==1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      valueh |   1125877    365390.9    476755.3       1000    7681000
    hhincome |   1125877    134953.3    122193.2     -14500    2907600

. regress valueh hhincome [pw = perwt]
(sum of wgt is   1.1051e+08)

Linear regression                                      Number of obs = 1125877
                                                       F(  1,1125875) =23590.19
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1671
                                                       Root MSE      =  4.1e+05

------------------------------------------------------------------------------
             |               Robust
      valueh |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    hhincome |   1.573021   .0102416   153.59   0.000     1.552947    1.593094
       _cons |     149302   1213.817   123.00   0.000       146923      151681
------------------------------------------------------------------------------
```

ANSWER 6
REQUIRES outreg2 COMMAND

7. Interpret the slope coefficient and the intercept. **(2 points)**

ANSWER 7

8. With respect to your results in the table in part (6), explain how there will be a omitted variables bias? Do you think the bias will be positive or negative? **Explain. (2 points)**

I think *costelec* is positively correlated with *valueh* and with *hhincome* leading to a positive bias.

Big expensive house, high cost of electricity, higher price, higher income to afford.

ANSWER 8

9. Choose another **continuous** variable to add to your regression from your dataset (x_2). Briefly describe this new variable (including units of measurement). **(1 point)**

$$x_2 : costelec \text{ i.e., Annual household cost of electricity}$$

“COSTELEC gives the annual electricity cost for each housing unit (rented or owned)”

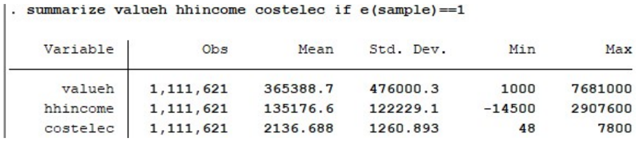
dollars

ANSWER 9

10. Provide a neatly formatted table with summary statistics for **all** three variables: variables you chose in part (3) and the one you chose in part (9): $y, x_1\ x_2$.

Calculate these summary statistics only for the sample that you use in part (14). This means that your sample size for each variable should match the sample size in part (14). Because you are including a new variable, this may change your sample size from part(6) (and part(4)) so you may need to recalculate summary statistics even for y and x_1. **(2 points)**

ANSWER 10



```
. summarize valueh hhincome costelec if e(sample)==1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      valueh |  1,111,621    365388.7    476000.3       1000    7681000
    hhincome |  1,111,621    135176.6    122229.1     -14500    2907600
    costelec |  1,111,621    2136.688    1260.893         48       7800
```

11. Write down your multiple regression equation (including your new variable). See example below; **replace the xs and ys with your variable names. (1 point)**

$$y = \beta_0 + \beta_1 x_1 + \beta_1 x_2 + u$$

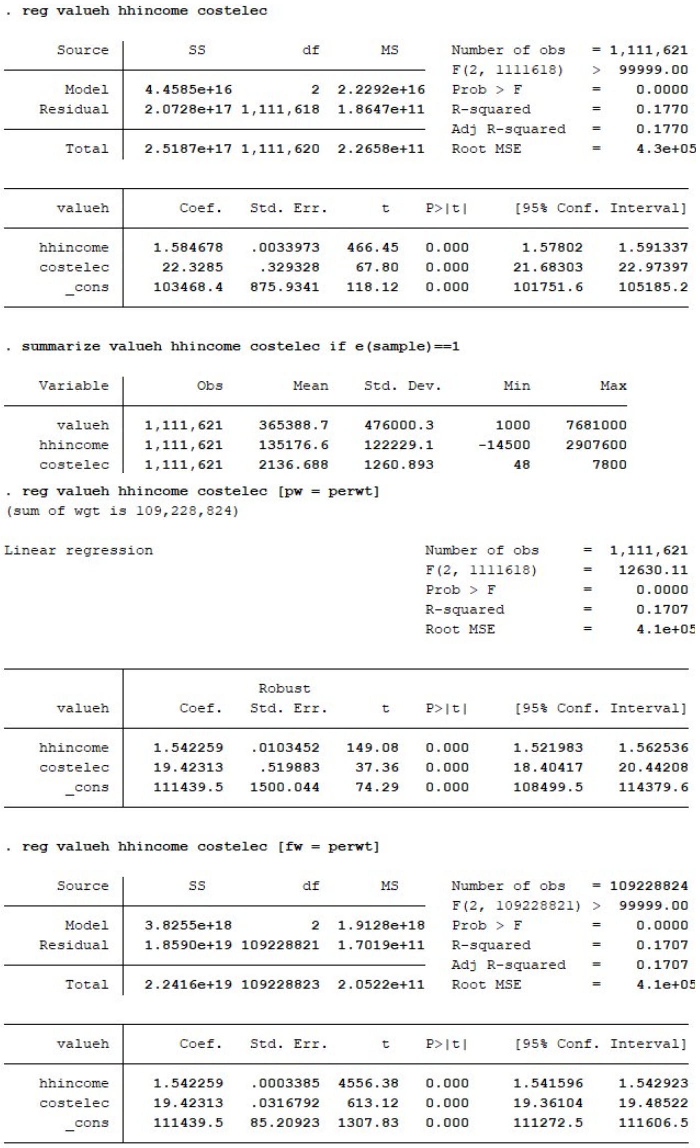
$$valueh = \beta_0 + \beta_1 hhincome + \beta_1 costelec + u$$

12. Does your new variable (x_2) help you reduce some of the omitted variables bias that you discuss in part (8)? **Explain. (2 points)**

ANSWER 12

13. Run the regression in part (11). Copy paste your Stata output here. Report your results in equation format. Interpret each coefficient and indicate whether or not it is statistically significant (for both $\hat{\beta}_1$ and $\hat{\beta}_2$). [You do not need to interpret the intercept, $\hat{\beta}_0$.) **(2 points)**

ANSWER 13



```
. reg valueh hhincome costelec

      Source |       SS           df       MS      Number of obs   = 1,111,621
-------------+----------------------------------   F(2, 1111618)   >  99999.00
       Model |  4.4585e+16         2  2.2292e+16   Prob > F        =    0.0000
    Residual |  2.0728e+17 1,111,618  1.8647e+11   R-squared       =    0.1770
-------------+----------------------------------   Adj R-squared   =    0.1770
       Total |  2.5187e+17 1,111,620  2.2658e+11   Root MSE        =    4.3e+05

------------------------------------------------------------------------------
      valueh |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    hhincome |   1.584678   .0033973   466.45   0.000      1.57802    1.591337
    costelec |    22.3285    .329328    67.80   0.000     21.68303    22.97397
       _cons |   103468.4   875.9341   118.12   0.000     101751.6    105185.2
------------------------------------------------------------------------------

. summarize valueh hhincome costelec if e(sample)==1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      valueh |  1,111,621    365388.7    476000.3       1000    7681000
    hhincome |  1,111,621    135176.6    122229.1     -14500    2907600
    costelec |  1,111,621    2136.688    1260.893         48       7800
. reg valueh hhincome costelec [pw = perwt]
(sum of wgt is 109,228,824)

Linear regression                               Number of obs     =  1,111,621
                                                F(2, 1111618)     =   12630.11
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1707
                                                Root MSE          =     4.1e+05

------------------------------------------------------------------------------
             |               Robust
      valueh |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    hhincome |   1.542259   .0103452   149.08   0.000     1.521983    1.562536
    costelec |   19.42313    .519883    37.36   0.000     18.40417    20.44208
       _cons |   111439.5   1500.044    74.29   0.000     108499.5    114379.6
------------------------------------------------------------------------------

. reg valueh hhincome costelec [fw = perwt]

      Source |       SS           df       MS      Number of obs   = 109228824
-------------+----------------------------------   F(2, 109228821) >  99999.00
       Model |  3.8255e+18         2  1.9128e+18   Prob > F        =    0.0000
    Residual |  1.8590e+19 109228821  1.7019e+11   R-squared       =    0.1707
-------------+----------------------------------   Adj R-squared   =    0.1707
       Total |  2.2416e+19 109228823  2.0522e+11   Root MSE        =    4.1e+05

------------------------------------------------------------------------------
      valueh |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    hhincome |   1.542259   .0003385  4556.38   0.000     1.541596    1.542923
    costelec |   19.42313   .0316792   613.12   0.000     19.36104    19.48522
       _cons |   111439.5   85.20923  1307.83   0.000     111272.5    111606.5
------------------------------------------------------------------------------
```

14. Now run a new regression where you characterize y in natural log form rather than levels, i.e.

$$\ln(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

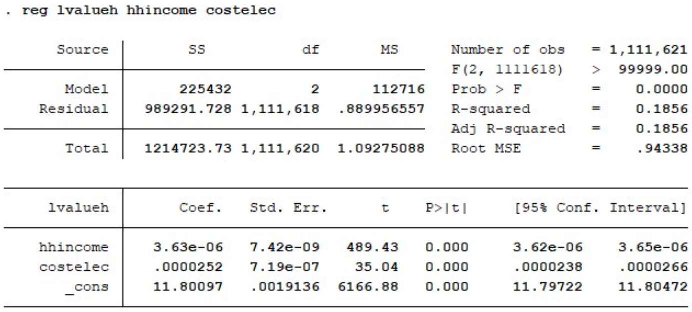
[*Hint: to do this, you will need to generate a new variable that is the natural log of y. For example, you might call this new variable lny.*] Copy paste your Stata output here. Report your results in equation format. Interpret the coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$. If there is no meaningful interpretation of the coefficients, please explain. **(2 points)**

```
gen lvalueh = ln(valueh)
```

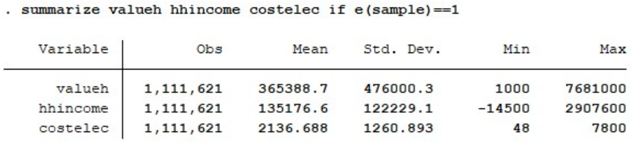
$$\ln(valueh) = \beta_0 + \beta_1 hhincome + \beta_2 costelec + u$$

Additional $\$100$ total household income at a constant costeleclevel will be associated with 0.0363% increase in home value.

Additional $\$100$ annual electricity cost at a constant hhincome level will be associated with 0.252% increase in home value.



```
. reg lvalueh hhincome costelec

      Source |       SS           df       MS      Number of obs   = 1,111,621
-------------+----------------------------------   F(2, 1111618)   >  99999.00
       Model |      225432         2      112716   Prob > F        =    0.0000
    Residual |  989291.728 1,111,618  .889956557   R-squared       =    0.1856
-------------+----------------------------------   Adj R-squared   =    0.1856
       Total |  1214723.73 1,111,620  1.09275088   Root MSE        =    .94338

------------------------------------------------------------------------------
     lvalueh |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    hhincome |   3.63e-06   7.42e-09   489.43   0.000     3.62e-06    3.65e-06
    costelec |   .0000252   7.19e-07    35.04   0.000     .0000238    .0000266
       _cons |   11.80097   .0019136  6166.88   0.000     11.79722    11.80472
------------------------------------------------------------------------------
```



```
. summarize valueh hhincome costelec if e(sample)==1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      valueh |  1,111,621    365388.7    476000.3       1000    7681000
    hhincome |  1,111,621    135176.6    122229.1     -14500    2907600
    costelec |  1,111,621    2136.688    1260.893         48       7800
```

15. Now run a regression where you include a quadratic term in either x_1 or x_2 (but with y **in levels**, as in part (11)). For example, if you choose to include a quadratic term in x_2, your regression will be

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + u$$

[*Hint: To do this, you will need to generate a new variable that is the square of your explanatory variable of choice. In this example, you might call this new variable x2sq.*] Copy paste your Stata output here. Report your results in equation format. What is the association between your outcome and your variable that includes the quadratic term, **at the sample mean of that variable**? Explain this association in words. **(2 points)**

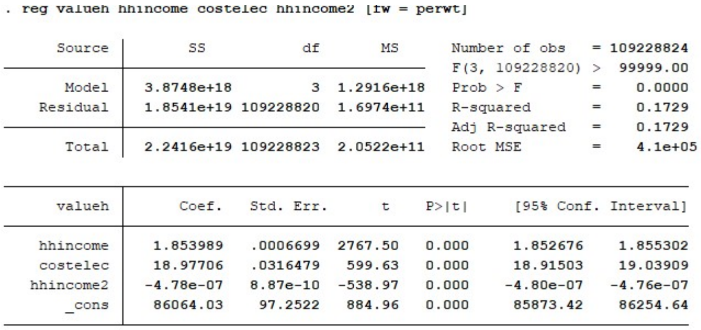
```
gen hhincome2 = hhincome^2
```

$$valueh = \beta_0 + \beta_1 hhincome + \beta_2 costelec + \beta_3 hhincome^2 + u$$
$$\overline{hhincome} = 135176.6$$

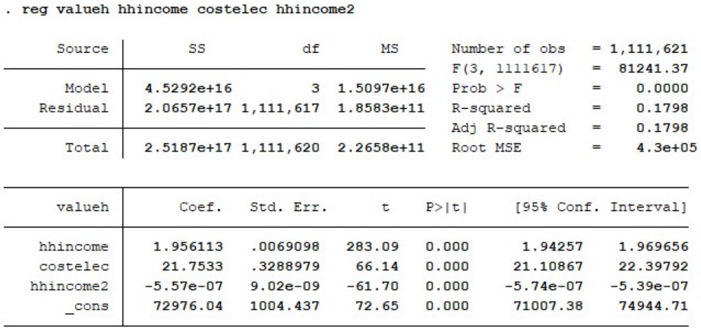
At this sample mean of hhincome, a 1 dollar increase in hhincome would be associated with an 1.81 dollar increase in valueh. $(1.956113 + 135176.6 * -5.57e - 07 * 2)$

In FW weight it is $(1.853989 + 132840.8 * -4.78e - 07 * 2) = 1.73$ increase in *valueh* per unit increase *hhincome* at *hhincome* mean.

This association is increasing with decreasing returns.



```
. reg valueh hhincome costelec hhincome2 [fw = perwt]

      Source |       SS           df       MS      Number of obs   = 109228824
-------------+----------------------------------   F(3, 109228820) >  99999.00
       Model |  3.8748e+18         3  1.2916e+18   Prob > F        =    0.0000
    Residual |  1.8541e+19 109228820  1.6974e+11   R-squared       =    0.1729
-------------+----------------------------------   Adj R-squared   =    0.1729
       Total |  2.2416e+19 109228823  2.0522e+11   Root MSE        =    4.1e+05

------------------------------------------------------------------------------
      valueh |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    hhincome |   1.853989   .0006699  2767.50   0.000     1.852676    1.855302
    costelec |   18.97706   .0316479   599.63   0.000     18.91503    19.03909
   hhincome2 |  -4.78e-07   8.87e-10  -538.97   0.000    -4.80e-07   -4.76e-07
       _cons |   86064.03    97.2522   884.96   0.000     85873.42    86254.64
------------------------------------------------------------------------------
```



```
. reg valueh hhincome costelec hhincome2

      Source |       SS           df       MS      Number of obs   = 1,111,621
-------------+----------------------------------   F(3, 1111617)   =  81241.37
       Model |  4.5292e+16         3  1.5097e+16   Prob > F        =    0.0000
    Residual |  2.0657e+17 1,111,617  1.8583e+11   R-squared       =    0.1798
-------------+----------------------------------   Adj R-squared   =    0.1798
       Total |  2.5187e+17 1,111,620  2.2658e+11   Root MSE        =    4.3e+05

------------------------------------------------------------------------------
      valueh |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    hhincome |   1.956113   .0069098   283.09   0.000      1.94257    1.969656
    costelec |    21.7533   .3288979    66.14   0.000     21.10867    22.39792
   hhincome2 |  -5.57e-07   9.02e-09   -61.70   0.000    -5.74e-07   -5.39e-07
       _cons |   72976.04   1004.437    72.65   0.000     71007.38    74944.71
------------------------------------------------------------------------------
```

16. Create a new formatted table using outreg2 that includes all of your regression results from parts (6), (11), (14), and (15) and paste it here. The final product should be a table with 4 columns:

1. Column 1 has only one explanatory variable, x_1 (from question 6) 2. Column 2 has two explanatory variables, x_1 and x_2 (from question 11) 3. Column 3 has two explanatory variables, x_1 and x_2 and the dependent variable is ln(y) (from question 14) 4. Column 4 has three explanatory variables, x_1, x_2 and x_2^2. (The dependent variable is y.) (from question 15)

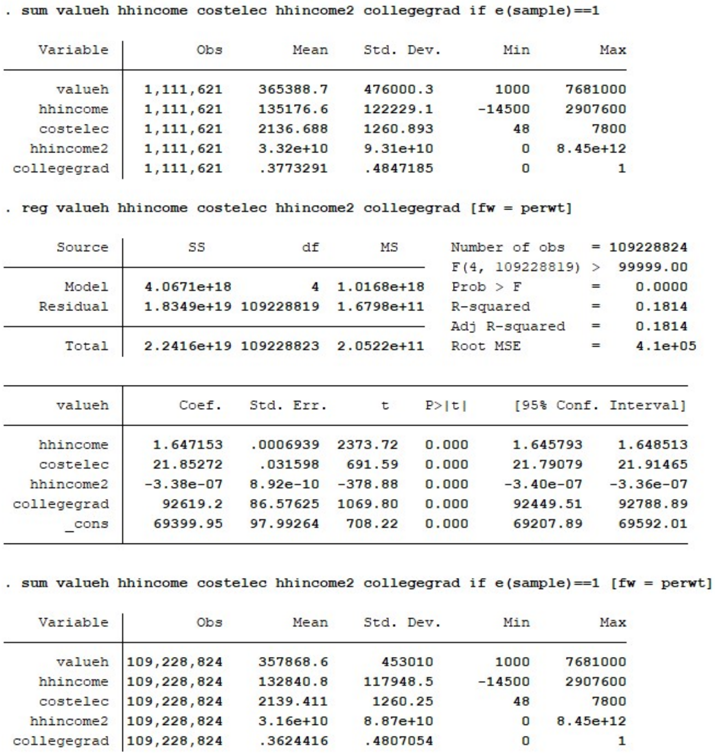
Be sure that your table is formatted clearly, with column numbers and easily identifiable entries (e.g. no mysterious variable names). (2 points)

ANSWER 16

Choose a **binary explanatory** variable to add to your regression from your dataset. If you do not have a binary variable in your dataset, create one from a continuous or categorical variable. For example, if you have data on income (continuous), use it to create a dummy variable (e.g. high or low income). Or, if you have data on school type (private, public, religious), you might create a dummy variable that is 1 for private or religious and 0 for public. If you want, you can also have a binary dependent variable, but this is NOT required. However, if you use a binary dependent variable, you must be careful when interpreting coefficients – specifically, you'll need to interpret everything in terms of percentage point changes in the probability that your dependent variable is equal to 1.

17. Copy and paste your variables description from previous parts and add the description of this new variable. **(2 points)**

collegegrad is a binary variable which we generated using *educd* which provided detailed information on educational attainment, and we set the *collegegrad* to true when bachelors or higher is achieved and to false when the observation has no bachelor, professional, or doctoral degree.

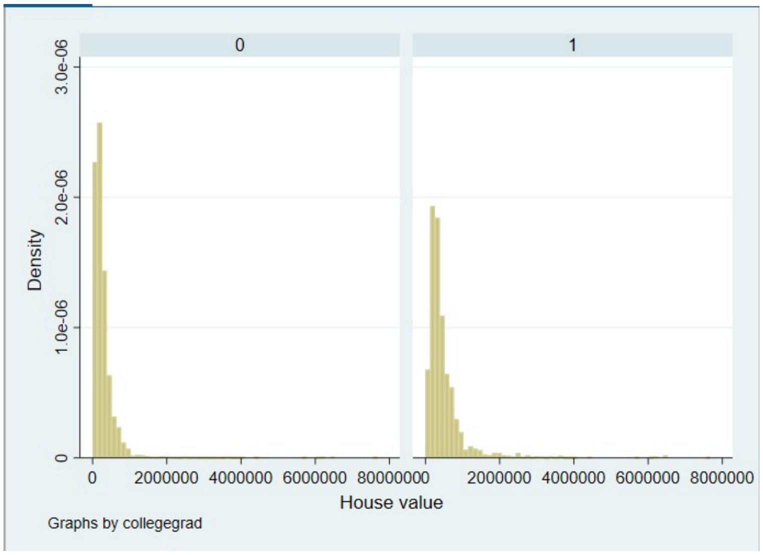


```
. sum valueh hhincome costelec hhincome2 collegegrad if e(sample)==1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      valueh |  1,111,621    365388.7    476000.3       1000    7681000
    hhincome |  1,111,621    135176.6    122229.1     -14500    2907600
    costelec |  1,111,621    2136.688    1260.893         48       7800
   hhincome2 |  1,111,621    3.32e+10    9.31e+10          0   8.45e+12
 collegegrad |  1,111,621    .3773291    .4847185          0          1

. reg valueh hhincome costelec hhincome2 collegegrad [fw = perwt]

      Source |       SS           df       MS      Number of obs   = 109228824
-------------+----------------------------------   F(4, 109228819) >  99999.00
       Model |  4.0671e+18         4  1.0168e+18   Prob > F        =    0.0000
    Residual |  1.8349e+19 109228819  1.6798e+11   R-squared       =    0.1814
-------------+----------------------------------   Adj R-squared   =    0.1814
       Total |  2.2416e+19 109228823  2.0522e+11   Root MSE        =    4.1e+05

------------------------------------------------------------------------------
      valueh |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    hhincome |   1.647153   .0006939  2373.72   0.000     1.645793    1.648513
    costelec |   21.85272    .031598   691.59   0.000     21.79079    21.91465
   hhincome2 |  -3.38e-07   8.92e-10  -378.88   0.000    -3.40e-07   -3.36e-07
 collegegrad |    92619.2   86.57625  1069.80   0.000     92449.51    92788.89
       _cons |   69399.95   97.99264   708.22   0.000     69207.89    69592.01
------------------------------------------------------------------------------

. sum valueh hhincome costelec hhincome2 collegegrad if e(sample)==1 [fw = perwt]

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      valueh |109,228,824    357868.6      453010       1000    7681000
    hhincome |109,228,824    132840.8    117948.5     -14500    2907600
    costelec |109,228,824    2139.411     1260.25         48       7800
   hhincome2 |109,228,824    3.16e+10    8.87e+10          0   8.45e+12
 collegegrad |109,228,824    .3624416    .4807054          0          1
```

18. Add the summary statistics for your new variable to the summary statistics table from before and display here. You must provide a neatly formatted table. **(2 points)**

Collegegrad is average $.4847185$ illustrating a 48% bachelors or higher attainment of the regression sample.

19. Create a separate histogram of your outcome variable for each value of your binary variable using Stata's histogram command with the "by" option. For example, your code might be: `histogram y, by (x3)` where y is your dependent variable and $x3$ is your binary variable. **(2 points)**



ANSWER 19

20. Write down your multiple regression equation (including your new binary variable x 3).
Note: you may either add your new binary variable to your regression from (11) OR you can replace one of your continuous variables with your new binary variable (if, for example, you used one of your previous continuous variables to create a new binary variable). **Make sure your regression contains at least one continuous variable and one binary variable. (2 points)**

ANSWER 20

21. Run the regression in part (20). Copy paste your Stata output here. Report your results in equation format. Interpret the coefficient on the **binary variable** and explain whether or not it is statistically significant. **(2 points)**

$$valueh = \beta_0 + \beta_1 hhincome + \beta_2 costelec + \beta_3 collegegrad + u$$

collegegrad is unsurprisingly statistically significant as this variable is well known to positively correlate with house value. What the output suggests is that for some level of household income and some level of annual electricity cost, being a *collegegrad* is associated with about a $\$92,000$ increase in house value.

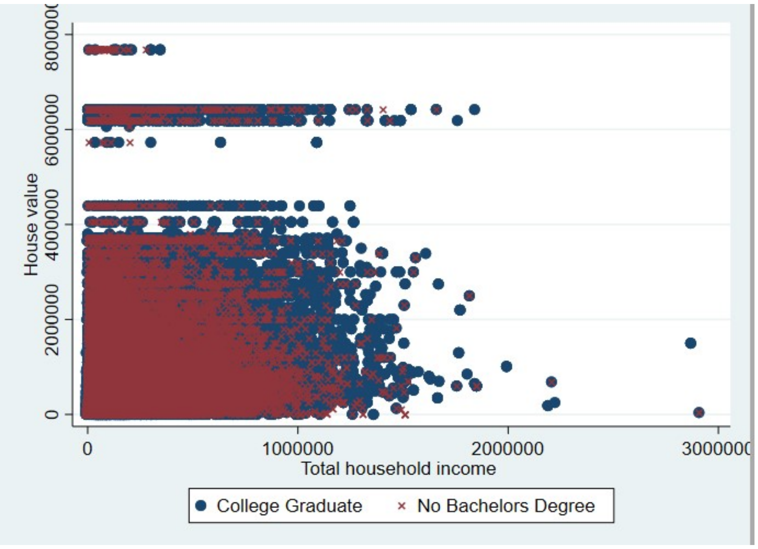
22. Create a scatterplot that shows the relationship between your dependent variable and one continuous explanatory variable, separately for when your binary variable is equal to 1 and when it is equal to 0. **Make sure your scatterplot is very clear and easy to read – it should have different markers for when your binary variable is equal to 1 and when it is equal to 0 and it should have a legend. (2 points)**

Hint: The Stata code might be helpful:

```
twoway (scatter y x1 if x3==1) (scatter y x1 if x3==0, ms(X)), legend(label(1 "High Income") label(2 "Low Incom
e"))
```

where y is your dependent variable, $x1$ is your continuous variable and $x3$ is your binary variable.

```
twoway (scatter valueh hhincome if collegegrad==1) (scatter valueh hhincome if collegegrad==0, ms(X)), legend(label(1 "College Graduate") label(2 '
```



23. Create two linear fit plots that shows the relationship between your dependent variable and your continuous variable (the same one from part (22)), separately for when your binary variable is equal to 1 and when it is equal to 0 (but on the SAME graph). Make sure your linear fit plots are very clear and easy to read – it should have different line styles for when your binary variable is equal to 1 and when it is equal to 0 and it should have a legend.

For example, your code might be `twoway (lfit y x1 if x3==1) (lfit y x1 if x3==0, lp(-)), legend(label(1 "High Income") label(2 "Low Income"))` where y is your dependent variable, x1 is your continuous variable and x3 is your binary variable.

Based on your plot, does it look like the relationship between your dependent variable and your continuous explanatory variable differs, depending on your binary variable? **(2 points)**

```
twoway (lfit valueh hhincome if collegegrad==1) (lfit valueh hhincome if collegegrad==0, lp(-)), legend(label(1 "Bachelor Degree") label(2 "No Bach
```

Based on the plot, the relationship differs as the slope for a degree holder is steeper than the slope for people without a bachelor or higher degree. At some household income level, if the observation has a degree then they will on average have a higher home value than an observation at the same household income level but no bachelors degree.

24. Now run a new regression where you allow the effect of the continuous explanatory variable (the same one from parts (22) and (23)) to differ, depending on the value of your binary variable. Hint: You might have to use an interaction term. Copy paste your Stata output here. Report your results in equation format. **(2 points)**

```
 reg valueh hhincome hhincomeXcollegegrad [fw = perwt]
```

This new regression allows the effect of *hhincome* to differ with *collegegrad*. That is, the effect of *hhincome* will need to be interpreted depending on if the observation has a college degree or not.

25. Is the interaction term in part (24) statistically significant? What does this tell you about the relationship between your dependent variable and your continuous explanatory variable? **(2 points)**

The interaction term in part (24) is statistically significant, this tells me that the relation between house value and household income depends on the degree status of a person. Is different for different bachelor degree status.

26. According to your regression results in part (24), what is the association between your dependent variable and your continuous explanatory variable when your binary variable is equal to 1? What about when your binary variable is equal to 0? Are these associations statistically significant? [Hint: you may need to use Stata's lincom command.] **(2 points)**

For college graduates (*collegegrad* == 1) the association between home value and household income is that for each additional dollar of household income there is an associated 1.6872341 dollar increase in home value for collegegraduates in our sample on average. For non-college graduates (*collegegrad* ==1) the association between home value and household income is that for each additional dollar of household income there is an associated 1.189297 dollar increase in home value for non-collegegraduates in our sample on average. The stata regression output contains the results of hypothesis tests for each coefficient individually, and we can use the command `lincom hhincome + hhincomeXcollegegread` to test the hypothesis that these two are simultaneously significant. We find that in all cases the coefficients are significant and they are practically significant with their magnitudes.

27. Think about the multiple regression equations from before [i.e from parts 11, 14, 15, 21 and 24] and write down the one that makes most sense to you. It can include binary variables, but it does not have to. Other than including multiple explanatory variables, there is no restriction on what type of variables you include. **[1 point]**

$$valueh = \beta_0 hhincome + \beta_1 hhincome^2 + \beta_2 costelec + \beta_3 collegegrad + u$$

28. Estimate the multiple regression equation in part (27) and report the results here. Don’t forget to specify robust standard errors if your sample size is large enough. **[1 point]**

ANSWER 28

29. Do you think the **dependent** variable in your regression from part (28) is reported with error? Explain briefly. If you think it is measured with error, explain whether you think that measurement error will lead to bias and whether that bias will lead you to under- or overestimate the true effect. *Regardless of whether you think the measurement error causes bias*, what effect does measurement error in your dependent variable have on the precision of your estimates? **[2 points]**

If measurement error in valueh is correlated with any of explanatory covariates then we have bias.

Yes *valueh* is reported with error. Many entries are of the midpoint of some range, i.e., we do not have exact home values rather the most reasonable estimate (midpoint) from some range of values recorded. If home values tend to be higher in the range or tend to be lower than its clear how using the midpoint of the range would not exactly work to represent the true home values. Measurement error in an outcome variable causes the estimates to be less precise.

The measurement error will not bias the estimated coefficients as the error basically gets crammed into unobserved/unexplainable. The measurement error WILL cause the estimation to be less precise.

30. Do you think the **explanatory** variables in your regression from part (28) are reported with error? Explain briefly. If some or all variables in your regression are measured with error, explain whether you think that measurement error will lead to bias and whether that bias will lead you to under- or overstimate the true effect. **[2 points]**

Yes, measurement error exists in costelec and hhincome. Costelec is clearly often recorded by taking some estimated monthly electric cost and factoring by twelve to get estimated annual electric cost, this leads to many values such as 600 (50*12) and 1200 (100*12). These annual electric cost values were clearly formed by taking some monthly value times 12 rather than by taking the annual cost itself directly. hhincome has often been rounded to trailing 0s. hhincome intends to display total household income but the survey respondents may have neglected to include some household member or left out some income, if thats the case then. educd seems most likely to be recorded accurately, with the primary matter being the accuracy of self-reports, which to my knowledge are quite fine for educational attainment. so collegegrad may be considered to not have measurement error.

31. Do you think your regression estimates suffer from sample selection bias? Explain. (For this part, do not discuss missing data.) **[2 points]**

Yes some sample selection bias.

32. Do you have missing data for any of the observations in your regression sample? Do you think the data are missing at random? Will the missing data cause bias? **[2 points]**

To help see whether the data are missing at random (and thus whether or not the missing data are likely to cause bias), create a dummy variable that is equal to 1 when data is missing and that is 0 when the data are not missing (for each variable in your regression). Then regress that dummy variable on other explanatory variables.

For example, if you have data on income, age, and race, you would create the following dummy variables:

```
gen missing_inc = (income==.)
gen missing_age = (age==.)
gen missing_race = (race==.)
```

and then you would regress each of those on the other variables, e.g.

```
reg missing_inc age i.race
```

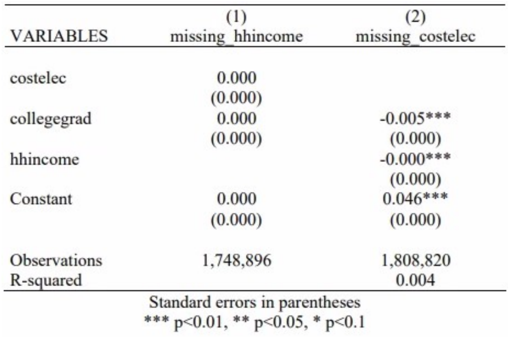
If income is (statistically significantly) more likely to be missing for older (or younger) people or for people of a specific race, then the data are not missing at random and will likely cause bias. (Clearly this will only be helpful for observations where you have missing values for some but not all variables.)

Note: in order to show your results concisely, use the outreg2 command. Continuing the example above, this would mean doing the following:

```
reg missing_inc age i.race
outreg2 using missingdata, bdec(3) word excel replace
reg missing_age i.race income
outreg2 using missingdata, bdec(3) word excel append
reg missing_race age income
outreg2 using missingdata, bdec(3) word excel append
```

Copy and paste only the final formatted table here.

```
gen missing_costelec = (costelec == .)
gen missing_hhincome = (hhincome == .)
gen missing_valueh = (valueh == .)
reg missing_hhincome costelec collegegrad
outreg2 using missingdata, bdec(3) word excel replace
reg missing_costelec hhincome collegegrad
outreg2 using missingdata, bdec(3) word excel append
```



	(1)	(2)
VARIABLES	missing_hhincome	missing_costelec
costelec	0.000	
	(0.000)	
collegegrad	0.000	-0.005***
	(0.000)	(0.000)
hhincome		-0.000***
		(0.000)
Constant	0.000	0.046***
	(0.000)	(0.000)
Observations	1,748,896	1,808,820
R-squared		0.004

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

33. Does your regression sample contain outliers and/or influential observations? In order to answer this question, you should provide some supporting evidence (copy and paste here). This might include things like scatterplots, histograms, detailed summary statistics, or regression results with and without certain observations. Which regression results – with or without the outliers/influential points – do you think are more believable? Explain. **[2 points]**

---

valueh has large groupings at those midpoint values I mentioned earlier.

34. Use the Breusch-Pagan Test to test for heteroscedasticity. Paste the Stata output for the regression you use to perform the test and state your conclusion here. Hint: when you run the regression to calculate the residuals, you should NOT specify robust standard errors. **[2 points]**

ANSWER 34

35. Use the two versions of the White test to test for heteroscedasticity. You should use the built-in Stata command to perform these tests. Paste the Stata output for the tests and state your conclusions here. **[2 points]**

ANSWER 35

Appendix containing regression analysis but weighted specification.

Relevant IPUMS Data Dictionary ———

Variable: “COSTELEC”

Name: COSTELEC

Label: Annual electricity cost

Variable Text: COSTELEC for 1970 reports each rented housing unit’s annual electricity cost, excluding amounts included with contract rent payments. For later years, COSTELEC gives the annual electricity cost for each housing unit (rented or owned), again excluding amounts included in contract rent or other types of payments. For 1970 and 1980, units within the universe that used no electricity can be identified. Beginning in 1990, the form combines the categories “no charge” and “no electricity used.”

COSTELEC amounts for renters are part of RENTGRS. Census Bureau research comparing respondents’ reported costs with utility company records indicates that respondents tend to overstate their costs.

In 1970, the universe for the U.S. Census samples specifies renter-occupied units rented for cash rent, not one-family houses on 10+ acres and not group quarters; however in the Puerto Rican census of 1970, this specification is for renter-occupied units rented for cash rent, not one-family houses on 3+ cuerdas, and not group quarters.

Amounts are expressed in contemporary dollars, and users studying change over time must adjust for inflation (See INCTOT for Consumer Price Index adjustment factors). The exception is the ACS/PRCS multi-year files, where all dollar amounts have been standardized to dollars as valued in the final year of data included in the file (e.g., 2007 dollars for the 2005-2007 3-year file). Additionally, more detail may be available than exists in the original ACS samples.

User Note:

The traditional unit of land area in Puerto Rico is the cuerda. The cuerda is equal to about 3930 square meters, 4700 square yards, or 0.971 acres. Because the cuerda and the acre are so close to being equal, they are often treated informally as being equal. Mainlanders sometimes call the unit the “Spanish Acre.” The IPUMS has preserved the units for the mainland U.S. as acres and Puerto Rico as cuerdas.

Concept: Economic Characteristic Variables – HOUSEHOLD

Start Position: 71

End Position: 74

Width: 4

Variable Format: numeric

Implied Decimal Places: 0

Variable: “HHINCOME”

Name: HHINCOME

Label: Total household income

Variable Text: HHINCOME reports the total money income of all household members age 15+ during the previous year. The amount should equal the sum of all household members’ individual incomes, as recorded in the person-record variable INCTOT. The persons included were those present in the household at the time of the census or survey. People who lived in the household during the previous year but who were no longer present at census time are not included, and members who did not live in the household during the previous year but who had joined the household by the time of the census or survey, are included. For the census, the reference period is the previous calendar year; for the ACS and the PRCS, it is the previous 12 months.

Note that household income differs from family income, which is reported in FTOTINC. The family income variable only reports the incomes of household members related to the head, while HHINCOME includes the incomes of all household members.

Amounts are expressed in contemporary dollars, and users studying change over time must adjust for inflation (See INCTOT for Consumer Price Index adjustment factors). The exception is the ACS/PRCS multi-year files, where all dollar amounts have been standardized to dollars as valued in the final year of data included in the file (e.g., 2007 dollars for the 2005-2007 3-year file). Additionally, more detail may be available than exists in the original ACS samples.

User Note: ACS respondents are surveyed throughout the year, and amounts do not reflect calendar year dollars. While the Census Bureau provides an adjustment factor (available in ADJUST), this is an imperfect solution. See the ACS income variables note for further details.

Concept: Economic Characteristic Variables – HOUSEHOLD Start Position: 87

End Position: 93

Width: 7

Variable Format: numeric

Implied Decimal Places: 0

Coder

Variable: “VALUEH”

Name: VALUEH

Label: House value

Variable Text: VALUEH reports the value of housing units in contemporary dollars. For 1930, 1940, and from 2008 onward, VALUEH is a continuous variable. The other years report the midpoint of an interval; see codes and frequencies for intervals.

User Note: Universe shifts and changing methods of determining value complicate use of this variable for comparisons across years. Furthermore, dollar amounts were intervalled differently for each year, and the top codes changed. Users must adjust for the effects of inflation; see INCTOT for Consumer Price Index adjustment factors.

User Note: The traditional unit of land area in Puerto Rico is the cuerda. The cuerda is equal to about 3930 square meters, 4700 square yards, or 0.971 acres. Because the cuerda and the acre are so close to being equal, they are often treated informally as being equal. Mainlanders sometimes call the unit the “Spanish Acre.” The IPUMS has preserved the units for the mainland U.S. as acres and Puerto Rico as cuerdas.

Concept: Economic Characteristic Variables – HOUSEHOLD Start Position: 94

End Position: 100

Width: 7

Variable Format: numeric

Implied Decimal Places: 0

Categories

Value	Label
0000000	$0 (1940)
0000250	Less than $500
0000500	Less than $999
0001000	Less than $2,000
0001500	$2,000- 1,999
0002500	Less than $5,000
0003500	$3,000- 3,999
0004000	$3,000- 4,999
0004500	$4,000- 4,999
0005000	Less than $10,000
0006250	$5,000 - 7,499
0008750	$7,500 - 9,999
0012500	$10,000 - 14,999
0011250	$10,000 - 12,499
0013750	$12,500 - 14,999
0017500	$15,000 - 19,999
0016250	$15,000 - 17,499
0018750	$17,500 - 19,999
0025000	$20,000- 29,999
0022500	$20,000 - 24,999
0021250	$20,000 - 22,499
0023750	$22,500 - 24,999
0030000	$25,000 - 34,999
0026250	$25,000 - 27,499
0027500	$25,000 - 29,999
0028750	$27,500 - 29,999
0032500	$30,000 - 34,999
0031250	$30,000- 32,499
0033750	$32,500- 34,999
0035000	$35,000+
0042500	$35,000 - 49,999
0037500	$35,000 - 39,999
0036250	$35,000- 37,499
0038750	$37,500- 39,999
0045000	$40,000 - 49,999
0042499	$40,000 - 44,999
0047500	$45,000 - 49,999
0050000	$50,000+
0055000	$50,000 - 59,999
0052500	$50,000 - 54,999
0057500	$55,000 - 59,999
0065000	$60,000 - 69,999
0062500	$60,000 - 64,999
0067500	$65,000 - 69,999
0075000	$70,000 - 79,999
0072500	$70,000 - 74,999
0077500	$75,000 - 79,999
0087500	$75,000 - 99,999
0085000	$80,000 - 89,999
0095000	$90,000 - 99,999
0100000	$100,000+
0112500	$100,000 - 124,999
0137500	$125,000 - 149,999
0175000	$150,000 - 199,999
0162500	$150,000 - 174,999
0187500	$175,000 - 199,999

0200000 $200,000+

0225000 $200,000 - 249,999

0275000 $250,000 - 299,999

0350000 $300,000 - 399,999

0400000 $400,000+

0450000 $400,000 - 499,999

0625000 $500,000 - 749,999

0875000 $750,000 - 999,999

1000000 $1,000,000+

9999998 Missing

9999999 N/A

Variable: “EDUCD”

Name: EDUCD

Label: Educational attainment [detailed version]

Variable Text: EDUC indicates respondents’ educational attainment, as measured by the highest year of school or degree completed. Note that completion differs from the highest year of school attendance; for example, respondents who attended 10th grade but did not finish were classified in EDUC as having completed 9th grade. For additional detail on grade attendance, see GRADEATT as well as the detailed version of HIGRADE.

Concept: Education Variables – PERSON

Start Position: 128

End Position: 130

Width: 3

Variable Format: numeric

Implied Decimal Places: 0

Categories

Value Label

000 N/A or no schooling

001 N/A

002 No schooling completed

010 Nursery school to grade 4

011 Nursery school, preschool

012 Kindergarten

013 Grade 1, 2, 3, or 4

014 Grade 1

015 Grade 2

016 Grade 3

017 Grade 4

020 Grade 5, 6, 7, or 8

021 Grade 5 or 6

022 Grade 5

023 Grade 6

024 Grade 7 or 8

025 Grade 7

026 Grade 8

030 Grade 9

040 Grade 10

050 Grade 11

060 Grade 12

061 12th grade, no diploma

062 High school graduate or GED

063 Regular high school diploma

064 GED or alternative credential

065 Some college, but less than 1 year

070 1 year of college

071 1 or more years of college credit, no degree

080 2 years of college

081 Associate’s degree, type not specified

082 Associate’s degree, occupational program

083 Associate’s degree, academic program

090 3 years of college

100 4 years of college

101 Bachelor’s degree

110 5+ years of college

111 6 years of college (6+ in 1960-1970)

112 7 years of college

113 8+ years of college

114 Master’s degree

115 Professional degree beyond a bachelor’s degree

116 Doctoral degree

999 Missing