EC285, Winter 2018. Assignment 3

Select one of the following data sets from the online data repository, ODESI. You will be using this data set for Assignment 3. The objective this time, in contrast to Assignment 2, is to choose a data set and variables that you believe are interesting and to try to explain a relationship between variables using data and the tools you have learned so far. You will also hopefully gain experience obtaining your own data without me having to download and upload it to MyLS for you.

Some suggested data sets from ODESI
Labour Force Survey, August 2017
Survey of Labour and Income Dynamics (person file), 2010
Census of Population public use microdata, 2006
National Graduates Survey, 2013
Aboriginal Peoples' Survey, 2012
Survey of Financial Security, 2012
Survey of Household Spending, 2009
Programme for International Assessment of Adult Competencies (PIAAC), 2012
Canadian Community Health Survey, 2014
Canadian Travel Survey (person level file), 2004
Canada Survey of Giving, Volunteering and Participating (main file), 2010

You can also find a data set on ODESI on your own that is not on the above list and as long as it has the variables you need and you clearly indicate which data set you are using, that is also acceptable. If you have chosen a data set from above and you do not find that it suits your needs or interests, I recommend switching to another before continuing with the assignment. **Do not use the Food Expenditure Survey as your data set; that is the example data set and using it will result in a zero.**

Do:

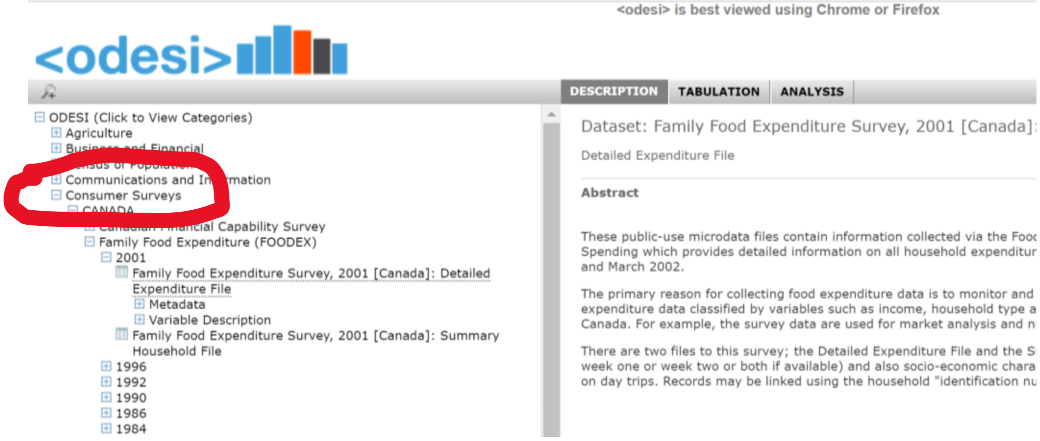
For the data set you selected from above:

1. Find your data set in ODESI (see instructions below for more info): http://odesi2.scholarsportal.info/webview/
2. Use the Metadata / study description (from the menus to the left) to help you figure out the context of the data set. The documentation can also be downloaded at the same time as the data (by checking a box); this documentation will also contain much useful information about the variables and data set.
3. In ODESI, using the menus to the left:
    a. Find one categorical variable, and screenshot its variable description page (paste it into a document)
    b. Find one quantitative variable, and screenshot its variable description page (paste into a document).
4. Download the data set in Stata format (the menu may describe it as Stata v. 8 format, which is fine; Stata can always open data files from earlier versions)
5. Open a log file using the log using command, as in earlier assignments. **Note: this assignment will require you to print off the log file after you are done and attach it to your hard copy of your assignment.**
6. Open the data set with Stata
7. Set up random number generator in Stata by "setting seed" equal to your student number, so that different student numbers should get different results (**set seed 12345678**)
Very important: do not skip this step, or the next 3 steps. If you do not do these steps, or fail to change 12345678 to your student number, you will be very heavily penalized.

8. Assign a random number to each observation (**gen rnumber = runiform()*1000000**)
9. Sort on the random number (**sort rnumber**)
10. Keep only 250 randomly selected observations (**keep if _n<=250**)
11. Generate a dummy variable using the categorical variable you have chosen (see below for more details on how to do this).
12. Use the **tabulate** command on your categorical variable. (**tabulate catvar1**). Note: some categorical variables may have lots of categories. You can either try choosing a categorical variable with only several (10-15 at most) categories, or you can show/print off only the first page of tabulated results to save space. Screenshot the results and include it in your assignment.
13. Use the **tabulate** command on your dummy variable. (**tabulate dummyvar**). Screenshot the results and include it in your assignment.
14. Run a regression using the **regress** command, with your quantitative variable as the dependent (left hand side) variable, and the dummy variable to be the independent (right hand side variable). (**regress var1 dummyvar**) Screenshot the regression output from Stata and include it in your assignment.
15. Generate the residuals from your regression using the **predict** command with the "**residuals**" option. Save the residuals in your data set as a new variable called **ehat**.
(predict ehat, residuals).
16. Generate the fitted values (the y hats) from your regression using the **predict** command. Save the fitted values in your data set as a new variable called **yhat**.
(predict yhat)
17. Use the **summarize** command to generate statistics on the residuals, fitted values, your dependent variable, and the dummy variable (**summarize ehat yhat var1 dummyvar**). Screenshot the table and include it in your assignment.
18. Save the commands you used to do this in sequence as a .do file. Ideally you should be able to run the entire do file to get all the results requested.

Downloading a Data Set from ODESI:

1. Go to http://odesi2.scholarsportal.info/webview/
2. Depending on what data set you choose above, navigate to the relevant drop down menu on the left. For example, the Food Expenditures Survey would be held in the "Consumer Surveys" section of ODESI:


<odesi> is best viewed using Chrome or Firefox
<odesi>
DESCRIPTION
TABULATION
ANALYSIS
⊟ ODESI (Click to View Categories)
⊞ Agriculture
⊞ Business and Financial
Census of Population
⊞ Communications and Information
⊟ Consumer Surveys
⊟ CANADA
Canadian Financial Capability Survey
⊟ Family Food Expenditure (FOODEX)
⊟ 2001
Family Food Expenditure Survey, 2001 [Canada]: Detailed Expenditure File
⊞ Metadata
⊞ Variable Description
Family Food Expenditure Survey, 2001 [Canada]: Summary Household File
⊞ 1996
⊞ 1992
⊞ 1990
⊞ 1986
⊞ 1984

Dataset: Family Food Expenditure Survey, 2001 [Canada]:

Detailed Expenditure File

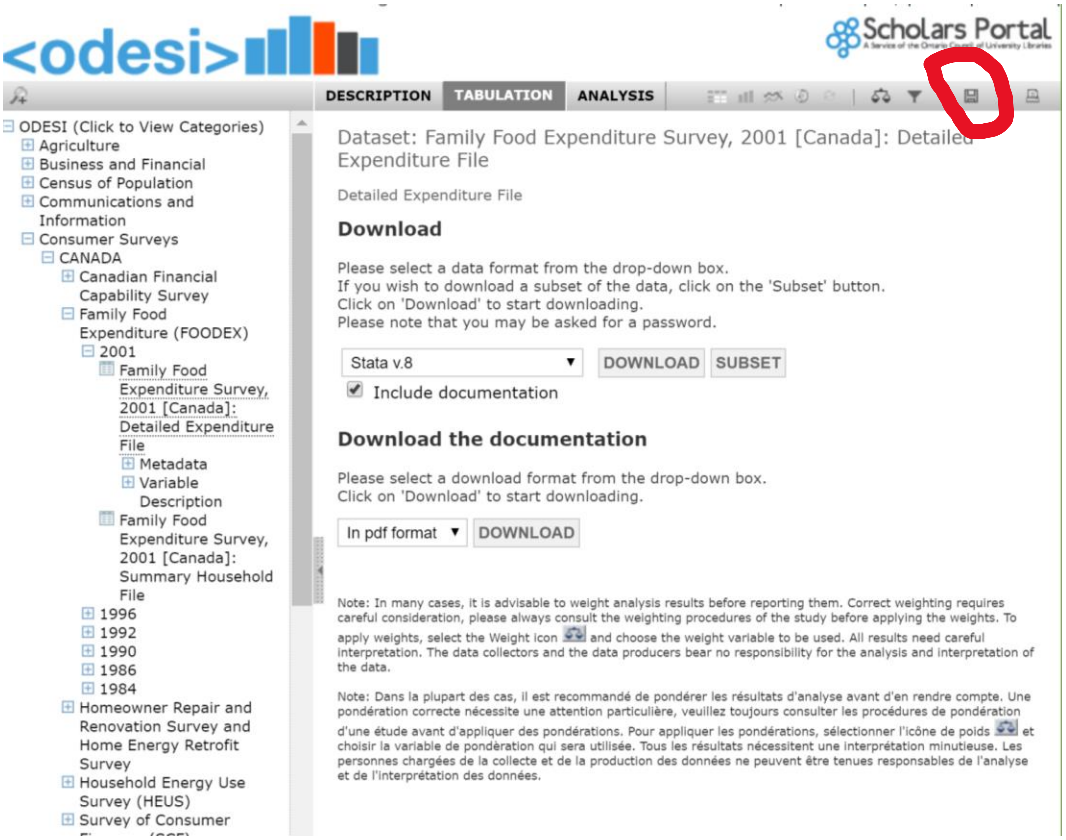
Abstract

These public-use microdata files contain information collected via the Fooc Spending which provides detailed information on all household expenditur and March 2002.

The primary reason for collecting food expenditure data is to monitor and expenditure data classified by variables such as income, household type a Canada. For example, the survey data are used for market analysis and n

There are two files to this survey; the Detailed Expenditure File and the S week one or week two or both if available) and also socio-economic chara on day trips. Records may be linked using the household "identification nu

3. Go to the download option in the top right portion of your screen (the floppy disk, circled below). Once you do so, select Stata (or Stata 8) as your preferred download option:


<odesi>
Scholars Portal
DESCRIPTION
TABULATION
ANALYSIS
ODESI (Click to View Categories)
Agriculture
Business and Financial
Census of Population
Communications and Information
Consumer Surveys
CANADA
Canadian Financial Capability Survey
Family Food Expenditure (FOODEX)
2001
Family Food Expenditure Survey, 2001 [Canada]: Detailed Expenditure File
Metadata
Variable Description
Family Food Expenditure Survey, 2001 [Canada]: Summary Household File
1996
1992
1990
1986
1984
Homeowner Repair and Renovation Survey and Home Energy Retrofit Survey
Household Energy Use Survey (HEUS)
Survey of Consumer

Dataset: Family Food Expenditure Survey, 2001 [Canada]: Detailed Expenditure File

Detailed Expenditure File

Download

Please select a data format from the drop-down box.
If you wish to download a subset of the data, click on the 'Subset' button.
Click on 'Download' to start downloading.
Please note that you may be asked for a password.

Stata v.8 ▼ DOWNLOAD SUBSET

☑ Include documentation

Download the documentation

Please select a download format from the drop-down box.
Click on 'Download' to start downloading.

In pdf format ▼ DOWNLOAD

Note: In many cases, it is advisable to weight analysis results before reporting them. Correct weighting requires careful consideration, please always consult the weighting procedures of the study before applying the weights. To apply weights, select the Weight icon and choose the weight variable to be used. All results need careful interpretation. The data collectors and the data producers bear no responsibility for the analysis and interpretation of the data.

Note: Dans la plupart des cas, il est recommandé de pondérer les résultats d'analyse avant d'en rendre compte. Une pondération correcte nécessite une attention particulière, veuillez toujours consulter les procédures de pondération d'une étude avant d'appliquer des pondérations. Pour appliquer les pondérations, sélectionner l'icône de poids et choisir la variable de pondèration qui sera utilisée. Tous les résultats nécessitent une interprétation minutieuse. Les personnes chargées de la collecte et de la production des données ne peuvent être tenues responsables de l'analyse et de l'interprétation des données.

4. **Please check to see that you can download your data before doing the rest of your assignment**; you don't want to waste time filling out information for variables/a data set that you can't end up using in the end!

Note: you will need to either access ODESI on campus or input your Laurier login information to download the data. **Please check this ahead of time, in case there is an issue for you with downloading.**

Generating a Dummy Variable:

In the sample code attached to this assignment, I have chosen to use the categorical variable marstat to create a dummy variable equal to 1 if a respondent was ever married, which I have called evermarried. I first create a variable equal to zero for all observations:

```
gen evermarried=0
```

I next replace values of evermarried with 1 if an observation registers a value of 1 for marstat (married) OR 3 (married but divorced/widowed/separated). Note that Stata reads the vertical slash | as an "or" when specifying conditions:

```
replace evermarried=1 if marstat==1|marstat==3
```

If you have a categorical variable where the responses are numbers, then using the above code and tweaking it (replacing variable names, adding more or fewer categories, replace category values) will work for you.

If you instead have a categorical variable which is saved as a "string" variable in your data set, you will need to use slightly different Stata language to generate your dummy. For example, if I wanted to generate a dummy variable called east that was equal to 1 if the respondent lived in Ontario or a province east of Ontario, but the province variable's values were all in the actual province names (not numbers, I would code it in the following way:

```
gen east=0
```

***note: below should be all in one line in Stata; if it cannot fit, let Stata overflow everything onto the next line. Do not manually include a line break/space between Nova and Scotia.

```
replace east=1 if province=="Ontario"|province=="Quebec"|province=="Nova
Scotia"|province=="Newfoundland"|province=="PEI"|province=="New Brunswick"
```

Constructing your own dummy variable will be very different across every student, depending on the variable that is chosen. It is therefore a little more difficult to come up with a set of instructions that will apply to everyone. This is also a little closer to truly using data and transforming it to conduct your own independent analysis.

Discuss:

1. Describe your data set (discuss the who what when were how and why). If you find your own data set off of ODESI, be especially careful to describe it here.
2. Using the information from ODESI and from your tabulation, describe the categorical variable.
3. Describe the dummy variable that you constructed using the categorical variable, with the help of your tabulation results. Why is your dummy variable interesting and what does the sample mean of your dummy variable tell you in your particular case?
4. Using the information from ODESI, describe the quantitative variable. Be sure to include its definition and units of measurement (if applicable).
5. Use the regression results describe the relationship between the two variables, making sure to interpret the slope coefficient using what we know of dummy variables. Discuss also whether you think this is a causal relationship.
6. Using your regression results and the statistics of the residual, fitted values, the dependent variable, and the dummy variable in your data set (from the **summarize** table), show that:

$$\bar{y} = \widehat{b_0} + \widehat{b}_1\bar{x} + \bar{e}$$

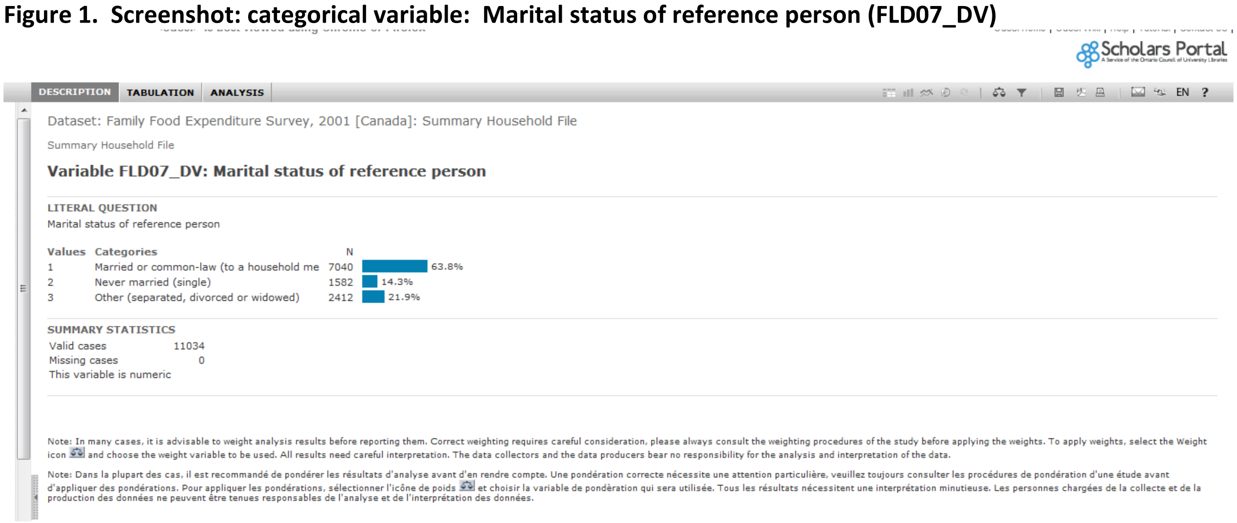
What to hand in:

1. A document along the lines of the example provided, including the code you used to get the Stata output (either your printed do-file or a list of all the commands you used, as in previous assignments). A cover page is not necessary.
2. A printed copy of your log file containing all of the commands and output that you used for your assignment (and was included in your do file). This will need to be included in your assignment when handing it in. You can open and print your log file directly from Stata using the log drop-down menu under "File", selecting "view", choosing your saved log file, and printing it when it opens up. You may also open it using a text file editor, like notepad, and printing it from there, although the formatting might be a little different than the Stata log viewer.

EXAMPLE FOR EC285: Assignment 3, 2018

Using the Family Food Expenditure Survey (FOODEX) 2001

Figure 1. Screenshot: categorical variable: Marital status of reference person (FLD07_DV)



Dataset: Family Food Expenditure Survey, 2001 [Canada]: Summary Household File

Summary Household File

Variable FLD07_DV: Marital status of reference person

LITERAL QUESTION

Marital status of reference person

Values	Categories	N	
1	Married or common-law (to a household me	7040	63.8%
2	Never married (single)	1582	14.3%
3	Other (separated, divorced or widowed)	2412	21.9%

SUMMARY STATISTICS

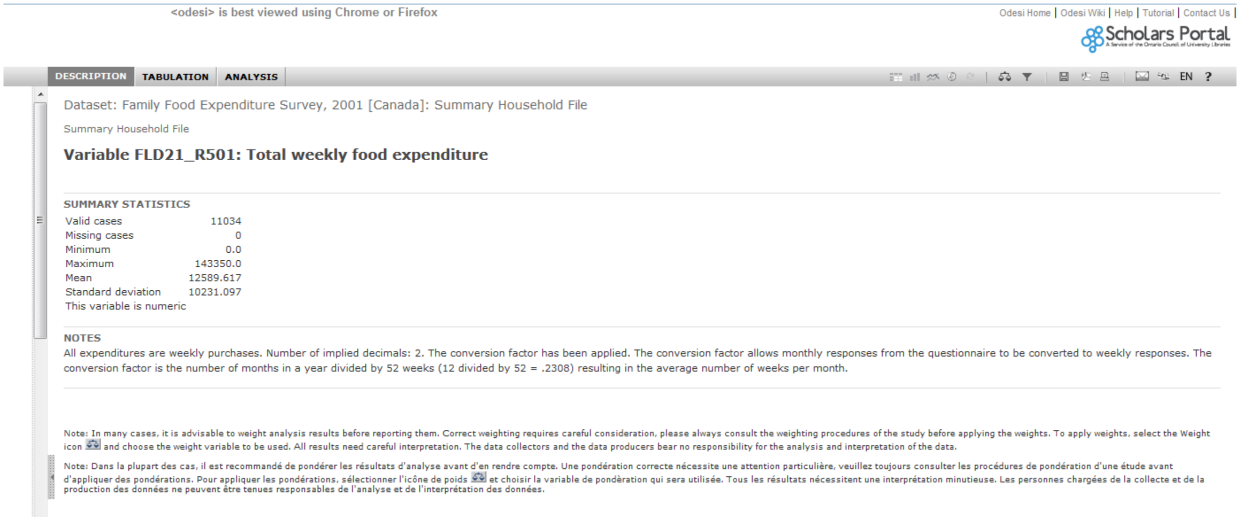
Valid cases	11034
Missing cases	0

This variable is numeric

Note: In many cases, it is advisable to weight analysis results before reporting them. Correct weighting requires careful consideration, please always consult the weighting procedures of the study before applying the weights. To apply weights, select the Weight icon ⚖ and choose the weight variable to be used. All results need careful interpretation. The data collectors and the data producers bear no responsibility for the analysis and interpretation of the data.

Note: Dans la plupart des cas, il est recommandé de pondérer les résultats d'analyse avant d'en rendre compte. Une pondération correcte nécessite une attention particulière, veuillez toujours consulter les procédures de pondération d'une étude avant d'appliquer des pondérations. Pour appliquer les pondérations, sélectionner l'icône de poids ⚖ et choisir la variable de pondèration qui sera utilisée. Tous les résultats nécessitent une interprétation minutieuse. Les personnes chargées de la collecte et de la production des données ne peuvent être tenues responsables de l'analyse et de l'interprétation des données.

Figure 2. Screenshot: quantitative variable: Total weekly food expenditure (FLD21_R501)


<odesi> is best viewed using Chrome or Firefox
Odesi Home | Odesi Wiki | Help | Tutorial | Contact Us |
Scholars Portal
A Service of the Ontario Council of University Libraries
DESCRIPTION
TABULATION
ANALYSIS
EN
?

Dataset: Family Food Expenditure Survey, 2001 [Canada]: Summary Household File

Summary Household File

Variable FLD21_R501: Total weekly food expenditure

SUMMARY STATISTICS

Valid cases	11034
Missing cases	0
Minimum	0.0
Maximum	143350.0
Mean	12589.617
Standard deviation	10231.097

This variable is numeric

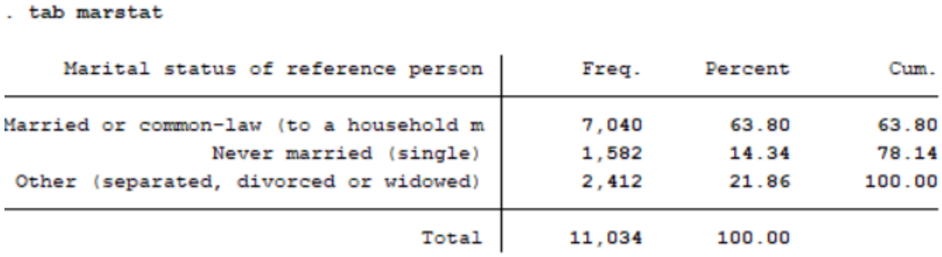
NOTES

All expenditures are weekly purchases. Number of implied decimals: 2. The conversion factor has been applied. The conversion factor allows monthly responses from the questionnaire to be converted to weekly responses. The conversion factor is the number of months in a year divided by 52 weeks (12 divided by 52 = .2308) resulting in the average number of weeks per month.

Note: In many cases, it is advisable to weight analysis results before reporting them. Correct weighting requires careful consideration, please always consult the weighting procedures of the study before applying the weights. To apply weights, select the Weight icon ⚖ and choose the weight variable to be used. All results need careful interpretation. The data collectors and the data producers bear no responsibility for the analysis and interpretation of the data.

Note: Dans la plupart des cas, il est recommandé de pondérer les résultats d'analyse avant d'en rendre compte. Une pondération correcte nécessite une attention particulière, veuillez toujours consulter les procédures de pondération d'une étude avant d'appliquer des pondérations. Pour appliquer les pondérations, sélectionner l'icône de poids ⚖ et choisir la variable de pondèration qui sera utilisée. Tous les résultats nécessitent une interprétation minutieuse. Les personnes chargées de la collecte et de la production des données ne peuvent être tenues responsables de l'analyse et de l'interprétation des données.

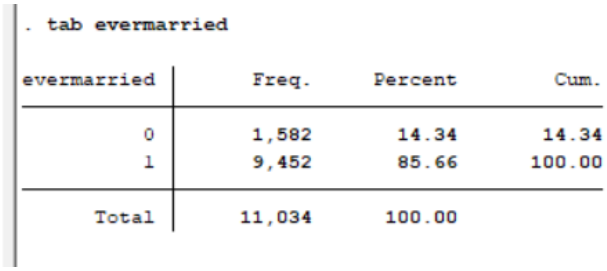
Table 1. Tabulation of marital status variable (marstat)



```
. tab marstat

   Marital status of reference person |      Freq.     Percent        Cum.
--------------------------------------+-----------------------------------
Married or common-law (to a household m |      7,040       63.80       63.80
                Never married (single) |      1,582       14.34       78.14
 Other (separated, divorced or widowed) |      2,412       21.86      100.00
--------------------------------------+-----------------------------------
                                 Total |     11,034      100.00
```

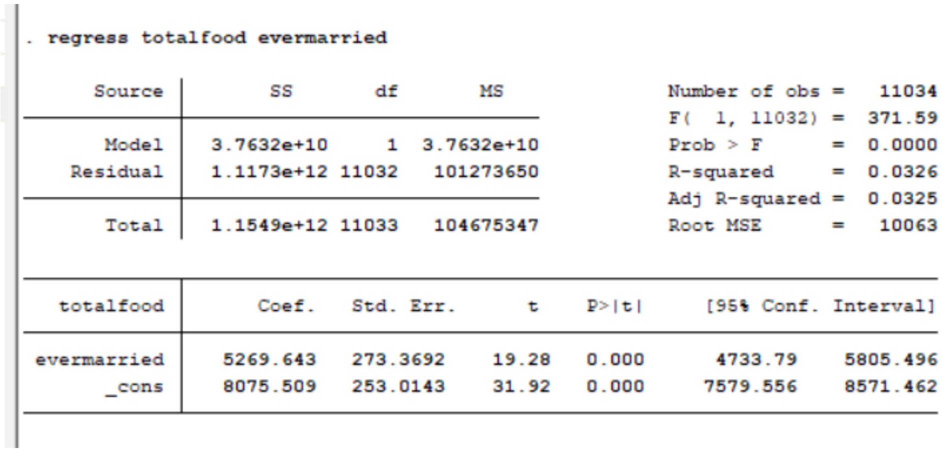
Table 2. Tabulation of ever married dummy variable (evermarried)



```
. tab evermarried

evermarried |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,582       14.34       14.34
          1 |      9,452       85.66      100.00
------------+-----------------------------------
      Total |     11,034      100.00
```

Table 3. Regression output from a regression with totalfood as the dependent variable and evermarried as the independent variable

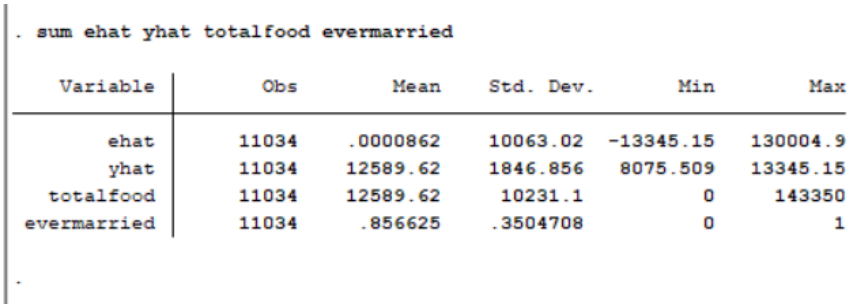


```
. regress totalfood evermarried

      Source |       SS       df       MS              Number of obs =   11034
-------------+------------------------------           F(  1, 11032) =  371.59
       Model |  3.7632e+10     1  3.7632e+10           Prob > F      =  0.0000
    Residual |  1.1173e+12 11032   101273650           R-squared     =  0.0326
-------------+------------------------------           Adj R-squared =  0.0325
       Total |  1.1549e+12 11033   104675347           Root MSE      =   10063

------------------------------------------------------------------------------
   totalfood |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 evermarried |   5269.643   273.3692    19.28   0.000      4733.79    5805.496
       _cons |   8075.509   253.0143    31.92   0.000     7579.556    8571.462
------------------------------------------------------------------------------
```

Table 4. Summary statistics of the residual, fitted values, the dependent variable, and the dummy variable



```
. sum ehat yhat totalfood evermarried

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        ehat |     11034    .0000862    10063.02  -13345.15   130004.9
        yhat |     11034    12589.62    1846.856   8075.509   13345.15
   totalfood |     11034    12589.62     10231.1          0     143350
 evermarried |     11034     .856625    .3504708          0          1

.
```

Stata code used to generate statistics and graphs from commands are also available as a do-file on MyLS called assignment_3_example.do (you should be able to run this using the posted data set foodex.dta

```
cap log close

***note: change C:\yourdirectory to the location where you keep your data for EC285
cd "C:\yourdirectory"
log using EC285assignment3.log, replace

***note: replace foodex.dta with the file name of the ODESI data you are using
use "foodex.dta", clear

***note: change 12345678 below to your student number
set seed 12345678
gen rnumber=runiform()*1000000
sort rnumber
keep if _n<=250

*****generating your dummy variable
***note: in this example, I am generating a dummy variable called evermarried
***which is equal to 1 if a survey respondent indicates they were either married
***or married and then divorced/separated/widowed
gen evermarried=0
replace evermarried=1 if marstat==1|marstat==3
***note: the next line sets evermarried to “missing”, which is denoted by a period in Stata
***if marstat is also a missing value for that observation
replace evermarried=. if marstat==.

tabulate marstat
tabulate evermarried

***note: below, change totalfood to your quant variable, and evermarried to your dummy
reg totalfood evermarried
predict ehat, residuals
predict yhat
summarize ehat yhat totalfood evermarried
log close
```