AAE637
Applied Econometrics II
Assignment #1

Due: Feb. 6, 2018
Total Points: 115 pts

Answer the following questions using the MATLAB software system. When handing in your assignments submit your MATLAB command file, your output file and any Word document generated to complete the assignment. I have created a DropBox for your assignments located within the *Learn@UW* system accessible from your *MyUW* desktop.

When handing in assignment related material us the following naming convention: **FI_LN**_Assign#**X(q=Y)**.xlsx (doc,m, etc.) where **FI** is your first initial, **LN** is your last name, and **X** is the assignment number and **q** is the question number (if more than 1 file per assignment).

You should remember that under the classical regression model the formula for estimating the multiple regression coefficients, $\beta_S=(X'X)^{-1}X'Y$, is are obtained from the first-order-conditions from the unconstrained minimimization of the sum of squared errors (SSE). In this class, the algorithms you will use will return the estimated coefficients via some type of iterative method where alternative solutions are evaluated. Depending on data and model structure, estimaton of coefficients could take hundreds of iterations. When handing in your assignments **do not show all of the iterations** in the output file you are handing in for a particular question. Saving the first few and last few iteration results should be sufficient.

1. (**15** pts) Suppose you have a random variable x_i, with known mean, μ, and variance, σ^2, where i = 1,…,T, (i.e., x_i~N(μ, σ^2))

   a. (**5** pts) Given the definition of the sample mean, $\hat{\mu}=\frac{\sum_{i=1}^{T} x_i}{T}$ where T is the number of observations in your sample data set, explictly **derive** (not just state) the formula for the ***variance of the sample mean***, Var($\hat{\mu}$), using information from all T observations. Note: For this application we are assuming that the individual values of the x_i's are ***independent*** of one another. What does it mean to say that these observations are independent of one another?.

   b. (**5** pts) The file, RECS_Data_2009_base.xlsx is an Excel file that contains a portion of the 2009 Residential Energy Consumption Survey (RECS) data collected by the Energy Information Administration of the U.S. Department of Energy (DOE). According to the DOE:

> *EIA administers the Residential Energy Consumption Survey (RECS) to a nationally representative sample of housing units. Specially trained interviewers collect energy characteristics on the housing unit, usage patterns, and household demographics. This information is combined with data from energy suppliers to these homes to estimate energy costs and usage for heating, cooling, appliances and other end uses — information critical to meeting future energy demand and improving efficiency and building design.*

There are two worksheets within the above file: (i) the codebook defining a subset of the RECS dataset variables (i.e., Codebook) and (ii) the data obtained from 12,083 surveyed households (i.e., Data). Assume electricity consumption is: Normally distributed and the consumption of electricity by one household is independent of electricity consumption of other households in the sample.

Given the above assumptions and the 2009 RECS data, statistically test the null hypothesis that the ***sample mean*** per capita electricity consumption equals 3,750 KWH.

c. (**5** pts) Given the above assumptions and the RECS data, test the null hypothesis that the **difference in sample means** in electricity consumption is 0 when comparing households with heating and cooling square footage less than 2,750 sq ft. vs. households with 2,750 sq ft or more of heating/cooling space? **Note:** These two subgroups could have distributions with different means ***and*** variances.

2. (**40** pts) Now lets analyze household electricity useage as a function of a number of exogenous household characteristics. To do this, I would like you to estimate a linear (in parameters) regression model where you are attempting to explain electricity useage variability across RECS households. The annual electricity use equation you want to estimate is represented via the following:

$$\mathrm{KWH}_t = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\ln(\mathrm{HDD65}_t) + \boldsymbol{\beta_2}\ln(\mathrm{CDD65}_t) + \boldsymbol{\beta_3}\ln(\mathrm{House\_Age}_t) +$$

$$\boldsymbol{\beta_4}\mathrm{HHSize}_t + \boldsymbol{\beta_5}\ln(\mathrm{Tot\_SqFt\_H/C}_t) + \boldsymbol{\beta_6}\ln(\mathrm{HH\_Inc}_t) + \boldsymbol{\beta_7}\mathrm{Elec\_Pr}_t + \varepsilon_t \quad (2.1)$$

where t=1,…12,083), the error term $\varepsilon_t \sim N(0,\sigma^2)$, the $\boldsymbol{\beta_i}$'s are regression coefficients you are to estimate, ln represents the natural logarithm and variables names are defined in the datafile CodeBook worksheet.[1]

In (2.1) we use the logarithm for some of the explanatory variables so to allow for (i) *marginal effects* being dependant on the exogenous variable value but

[1] As noted in the data codebook, a cooling degree day (CDD) is the number of degrees that a day's average temperature is above 65o Fahrenheit and air conditioning is starting to be used. You then add the number of degree days across all days encompassed by the survey to obtain an estimate of total annual cooling degree days (i.e., CDD65). Similarly, a heating degree day (HDD) is the number of degrees that a day's average temperature is below 65o Fahrenheit, the temperature below which buildings need to be heated. You then add the number of these degree days across all days encompassed by the survey to obtain an estimate of total annual heating degree days (i.e., HDD65).

independent of electricity use and (ii) *elasticity measures* being impacted by electricity use but independent of exogenous variable values.

a. (**20 pts**) Write your own **native** MATLAB code to read the data, estimate the classical regresssion model (i.e., the β_i's) shown in (2.1) via the use of the equation: $\beta_s=(X'X)^{-1}X'Y$ and the estimation of the associated paramater vector covariance matrix (Σ_β) via the following: $\Sigma_\beta=\sigma^2(X'X)^{-1}$. That is, **do not use** the MATLAB regression function but write your own code. Develop your code so that it generates an output file that has a **nicely organized** results table (regardless of the number of parameters estimated) showing variables names and associated:

   i. Estimated coefficients;
   ii. Estimated coefficient standard errors;
   iii. T-statistic values;
   iv. P-values of the above t-values;
   v. Overall equation F-statistic;
   vi. Equation R^2 and adjusted R^2 statistics;
   vii. Estimate of the error variance; and
   viii. The Parameter covariance matrix, Σ_β .

   Note: After transforming the variables, where required, exclude data that has missing values otherwise MATLAB will generate an error message. Given the above results, what percent of the total variation in electric energy consumption about the mean is explained by your estimated regression?

   In designing your software think about the structure of output tables that you prefer whether it was generated by Stata, Excel or other regression systems. When writing your MATLAB code you may want to use the MATLAB function found in the file table_bwg.m . See if can figure out what this file is doing. Having your regression function call out this table creating function will greatly simplify your coding.

b. (**5 pts**) Do the results with respect to the heating and cooling degree day variable coefficients make sense? Why or why not? Statistically test the following hypothesis: At the mean of the **raw data**, the ***marginal effect*** of HDD's equals the ***marginal effect*** of CDD's. ***Hint***: How do you calculate the variance of the sum of two random variables? How do you calculated the variance of the sum of two weighted random variables?

c. (**5 pts**) At the mean of the raw data do you obtain the same results as above with respect to equality of estimated CDD and HDD ***elasticity impacts*** on electricity use?[2] Provide statistical results to support your statement.

[2] I refer to Raw Data here so that you do not take the mean of the natural logarithm of a particular but instead the logarithm of a variable at its mean value.

d. (**5 pts**) What is the *price elasticity* of demand, Γ, when evaluated at the **mean** values of the raw data? Provide statistical evidence justifying your answer as to whether this price elasticity is :

   i. Statistically different from 0?

   ii. Elastic, i.e., more negative than -1?

e. (**5** pts) Test the following hypothesis: Including both the CDD and HDD variables results in a statistically significant increase in explanatory power of the regression model. Explain how you undertook this statististical test.

3. (**35 pts**) To answer question #2 above, you developed native MATLAB code to estimate your regression model.

   a. (**5 pts**) Use whatever drawing software you are familiar with to develop a flowchart of a *function* that would be used to estimate the classical regression model similar to what you did in #2 above by being called out by a command file. If you are not familiar with a drawing program, you may want to use the drawing capability of PowerPoint. In this flowchart indicate the flow of information from the command file, model inputs, regression outputs, relationship between model components, etc. Save this flowchart to an image file and place in the assignment dropbox.

   b. (**30 pts**) Using the above flowchart as a base, develop a MATLAB function that estimates the classical regression model. When developing this code make sure it is written in a general way so it can handle data with **any number of observations and any number of coefficients to estimate**. That is, you want matrix dimensions to be dynamically determined where appropriate. Make sure this function returns all regression related statistics you estimated in #2 above.

   Once you create the function have a 2nd file, i.e., the commnd file, call out the function and use it to estimate equation (2.1). The role of the command file is to pass to the function those items that change across application of the function. That is, any user inputs are set in the command file and then passed to the function. These could include the name of raw data being used in the analysis, the data actually being used, associated variable names and any type of program constants that control model design. As an example, sometimes we do not want to include an intercept in a regression model. Therefore I would like you to create a global variable defined via the following: $\text{Intercept\_D} = \begin{cases} 1 \text{ if intercept in the model} \\ 0, \text{otherwise} \end{cases}$. The command file will pass this value to the function. (Hint: What regression statistic change depending on whether there is an intercept in the model?)

4. (**25** pts) Re-estimate the regression model represented in equation (2.1) using the function you developed in #3 above except at this time add to the previous

exogenous variables, the variable (*HHINC* × *Price*) which is the product of the variables *HHINC* and *Elec_Pr*.

a. (**5** pts) Show your regression results obtained from estimating this second regression model using your newly created OLS function. From a statistical perspective, does the *marginal impact* of a change in *electricity price* on *electricity consumption change* depending on the level of *income*? Why or why not?

b. (**5** pts) At the mean of the raw data, what is the *elasticity* of a change in price on electricity purchase? Is this elasticity statistically different then -1.0? What is the justification for your answer?

c. (**5** pts) At the mean of the raw data what is the income elasticity of electricity use? Is this elasticity statistically different from 0.50? Explain how you obtained your results.

d. (**10** pts) Assume the raw data is set at sample means *except* for price, does the **income** elasticity of electricity consumption, when price is 75% of the sample mean equal the elasticity when price is 150% of the sample mean, ceteris paribus? Provide statistical evidence to support your assertion.