Business Statistics and Analysis Capstone

Week-01

Data for the Capstone Project

The aim of this project is to study housing affordability relative to Median income, Poverty level income, and Fair market rent.

- ➢ Data is made available every two years. We will use 2005, 2007, 2009, 2011, and 2013 data.
- ➢ Housing-level variables include,
  - Number of rooms in the housing unit.
  - The year it was built.
  - Occupied or vacant, Rented or Owned
  - Single-family or multi-family structure.
  - Number of units in the building.
  - Market value, Housing costs.
  - Number of people living, Household income.
  - Type of residential area.

The URL to download data for the Capstone Project:

https://www.huduser.gov/portal/datasets/hads/hads.html

We downloaded the ASCII version of files for the “HADS(Housing Affordability Data System) Data derived from AHS(American Housing Survey) National Data” for 2005, 2007, 2009, 20011 and 2013.

- ➢ Now we have five text files.
- ➢ Conversion of Excel. Reduction of variables.
- ➢ A document file with variable definitions.
- ➢ Now, we have five Excel files with a reduced number of variables.

Merging data files in Excel:

We used the “CONTROL” variable in the data. Merging is helpful to do a longitudinal analysis.

Week-02

Assignment-01

Q1) Are there differences in the Market Values of occupied versus vacant housing units?

Q2) Is there a pattern in these differences over the period 2005 through 2013?

VALUE: Current Market Value of the Housing Unit

STATUS: Whether the Housing Unit is “Occupied” or “Vacant”

My analysis falls under these four categories:

1) Summarizing the data for the VALUE and STATUS variables
    a) Basic descriptive statistics.
    b) Graphical summary.
2) Test for differences in the variable VALUE between ‘occupied’ and ‘vacant’ housing units. (Did Differences in Means Hypothesis Test)
3) Did the above analysis separately for all five years, 2005 through 2013.
4) Prepared a brief summary report, which includes the above categories. (Excel file related to this uploaded in the GitHub)

Suspect data:

1) Many housing units may have a negative or very low ‘Current Market Value’ (VALUE).
2) For our analysis, we deleted all housing units which have a market value of less than $1,000.

Q1) Ans: The difference in the Market Values is significant only for the years 2005 and 2011. In these years the market value of 'Occupied' units was greater than 'Not-Occupied' units.

For the remaining years, there is no significant difference in the market value across 'Occupied' and 'Not-Occupied' units.

Q2) Ans: The pattern discernable is that the Market value of 'Occupied' units is never less than that for 'Not-Occupied' units. It is either greater (as in the years 2005 and 2011) or equal (as in the remaining years).

Assignment-02

FMR: Fair Market Rent for the Housing Unit.
Compared Fair Market Rent (FMR) across different years:

- ➢ Pairwise comparisons across years.
- ➢ Pairing the data help control for differences across Housing Units.
- ➢ Merged all five data files using the CONTROL variable.
- ➢ Kept only those housing units that have data on FMR for all five years.

My analysis falls under the following categories:

1) Prepared the data by merging across all five years.
    a) Kept only the CONTROL and FMR variables.
    b) After merging I have the CONTROL variable and FMR values for all five years.
    c) Deleted the entire Housing Unit if FMR is ‘missing’ or ‘negative’.
2) Provided the descriptive statistics on FMR. Numerical as well as graphical.
3) Did appropriate analysis to compare the pairwise differences in Fair Market Rents (FMR) across the five years, 2005 through 2013. Paired t-test is preferable here. Like the paired t-test for 2005 and 2007, 2007 and 2009….
4) Prepared a brief summary report, which is accessible in my GitHub repository.

Analysis of the differences in Fair Market Rents across the various years.

As seen by the various statistical tests (t-tests for differences in means) it can be seen that the Fair Market Rents continuously rose across these various years. Further, if we calculate the percentage increases across years, the highest increase was observed from 2007 to 2009, the period overlapping the subprime mortgage crisis.

Week-03

Assignment-03

A Model for Market Value of Housing Unit:

VALUE: Current Market Value of the Housing Unit.

$$VALUE = ß_0 + ß_1X_1 + ß_2X_2 + \ldots + ß_KX_K$$

I used only the 2013 data and considered only 'Single Family Housing'.
TYPE = 1 and STRUCTURE TYPE = 1

Some more cleansing steps of the data include

- ➢ The VALUE variable may have a negative or very low value.
- ➢ For our analysis, we deleted all housing units which have a market value of less than 1,000$.
- ➢ This would also delete all data on housing units which are rental units, OWNRENT = '2'
- ➢ Chose only those housing units that have a market value that is 1000 dollars or greater, that is, VALUE ≥ $1000.
- ➢ Chose only the 'Single Family Units', that is STRUCTURE TYPE = 1 and TYPE = 1.

After data cleaning:

- ➢ Calculated various descriptive statistics for the VALUE variable.
- ➢ Visualized the empirical distribution of VALUE.
  - ○ Plotted a histogram of VALUE.
  - ○ Plotted a histogram of Ln(VALUE).

Among these two plots, the plot of the histogram of LN(VALUE) more closely resembled a Bell curve.

- ➢ Selected my set of independent variables or the X variables.

I used some tips from my mentor for selecting "X" variables, which include

1) The number of "X" variables needs to be less than 16.
2) Need to appropriately code the categorical variables.

i) Collapsing categories.
METRO3: '1' → Central city area
'2', '3', '4' or '5' → Not a Central city area

3) We may try out the natural logarithmic or other transformation of variables.
4) Thinking about the appropriateness of using the various "X" variables in the model.

My report will cover the following (Uploaded my report in GitHub named Assignment-3_analysis)

1) The set of variables used in the regression model and a brief justification for their use.
2) The estimated regression model along with an explanation of any variable transformations you have done.
3) Interpretation of the impact of various variables included in my model.

Week-04

Assignment-04

In the previous assignment, We used Y and X variables belonging to the same year. Using the regression model to predict market value two years from now may be problematic.

To build a predictive model for the market value of the Housing Unit, We used the Y variable from 2013 and the X variables from 2011.

- ➢ Merged the VALUE variable from the 2013 data into the 2011 data. (VLOOKUP command using the CONTROL variable to match the data files)
- ➢ Data cleaning
  - Deleted all rows for housing units that are not common across the two years.
  - Considered only 'Single Family Housing' TYPE = 1 and STRUCTURE TYPE = 1
  - Deleted all Housing units which have a market value of less than $1000, that is, deleted units with VALUE < $1000.

- ➢ Estimated a regression model using the 2013 VALUE variable and the X variables from the 2011 data.

$$VALUE = ß_0 + ß_1X_1^{(2011)} + ß_2X_2^{(2011)} + \ldots + ß_KX_K^{(2011)}$$

How well does the model predict future Market values? It can be answered by using the following methods.

- ➢ **R-square measure.**
- ➢ **'Holdout Analysis'.**
  - Hold out some data and not include it in the regression.
  - Use the estimated regression model to predict the held-out data.

Holdout Analysis

- ➢ From the data that you create, select 1000 Housing Units at random. This is your 'Holdout Sample'.
- ➢ To select a random sample
  - Use the =RANDBETWEEN() function in Excel.
  - Sort the data on this random value and select the top 1000 Housing Units for 'Holdout Sample'.
  - Estimate regression model on the remaining set of Housing Units.

- ➢ Using 'β' coefficients from the regression model and X variables from 'Holdout Sample', predict VALUE for each Housing Unit in the Holdout Sample.

Compared these thousand predictions with the actual market value of those housing units.

- ➢ Create a 'Mean Absolute Deviation' measure.
- ➢ It is the average of the absolute difference between the actual and the predicted value.

$$\text{Mean Absolute Deviation} = \left(\sum_{i=1}^{1000} |\text{Actual}_i - \text{Predicted}_i| \right) / 1000$$

My report covers the following:

1) The regression output.
2) The calculation of ‘Mean Absolute Deviation’.

Mean Absolute Deviation = $ 78,958.36