An Analysis of the Impact of Neighbors’ Expectations on Regional Home Values

by

Stephanie Langeland

Final Project Presented to

Professor Elena Krumova

In Partial Fulfillment

of the Requirement for

QMSS 4015: Data Analysis for the Social Sciences

Columbia University

Graduate School

of

Arts and Sciences

April 28, 2017

Table of Contents

1 Introduction 3

1.1 Background 3

1.2 Hypothesis 3

2 Datasets and Variables 4

2.1 SHED 4

2.2 HPA 8

2.3 R Code and Data Manipulation 10

3 Quantitative Analysis 10

3.1 Exploratory Data Analysis 10

3.2 Model #1 12

Multiple Linear Regression Model #1 12

Logistic Regression Model #1 14

3.3 Model #2 15

Multiple Linear Regression Model #2 15

Logistic Regression Model #2 16

4 Conclusion 17

4.1 Limitations & Future Research 17

Appendix 19

Table 1 19

Table 2 19

Distribution of HPA Data 20

Distribution of SHED Data (H1) 22

Distribution of SHED Data (H4) 23

Distribution of Regional Data 24

Multiple Linear Regression Model #1 Diagnostics 25

Multiple Linear Regression Model #1 Summary Statistics 27

Logistic Regression Model #1 Summary Statistics 28

Logistic Regression Model #1 Odds Ratios 28

Logistic Regression Model #1 Predicted Probabilities Table A 28

Logistic Regression Model #1 Predicted Probabilities Table B 29

Logistic Regression Model #1 Predicted Probabilities Table C 29

Logistic Regression Model #1 Confusion Matrix 29

Multiple Linear Regression Model #2 Diagnostics 30

Multiple Linear Regression Model #2 Summary Statistics 32

Logistic Regression Model #2 Summary Statistics 33

Logistic Regression Model #2 Odds Ratios 33

Logistic Regression Model #2 Predicted Probabilities Table A 33

Logistic Regression Model #2 Predicted Probabilities Table B 34

Logistic Regression Model #2 Predicted Probabilities Table C 34

Works Cited 35

# 1 Introduction

## 1.1 Background

Perspective homebuyers should understand the past, current, and expected future state of the neighborhood where they are looking for a home. However, obtaining such summary statistics is not a straightforward effort as such data may not be readily available in a succinct consumable format. This study proposes a residential real estate model that summarizes the past and current as well as forecasts the future state of housing markets for nine regions of the United States. The model uses homeowners’ opinions of the past value of their homes and expected future values of the homes in their neighborhoods to predict whether actual home prices will increase or decrease. Data for homeowners’ opinions about the past value of their homes and expected future values of homes in their neighborhood come from the Survey of Household Economics & Decisionmaking (SHED). These data are used to forecast movements in the Home Price Analyzer (HPA), which reflects actual home prices. All data are split into nine geographic regions of the United States. Therefore, there will be a model for each region.

## 1.2 Hypothesis

*Null Hypothesis: No relationship exists between residents’ opinions about the value of their homes & homes in their region versus the HPA.*

*Alternative Hypothesis: A relationship exists between residents’ opinions about the value of their homes & homes in their region versus the HPA.*

This study operates under the assumption that residents of each neighborhood (respondents of the SHED) are the most reliable source of current information about the socioeconomic condition of their neighborhood. Therefore, residents of a particular neighborhood are expected to provide reliable forecasts of expected home values of homes in their neighborhood. Residents have the most detailed view of current events in their neighborhood as well as potential threats or positive influences to the condition of that area. Therefore their opinions, noted in the SHED, should be reliable predictors of actual home values.

For example, residents who see an uptick in vandalism in their neighborhood may form a negative outlook of the neighborhood’s prosperity, potentially resulting in a decline in home values if they anticipate that their neighbors may vacate the neighborhood due to increased crime. This expectation, which is based on perception, may cause residents to sell their homes as well. If homeowners expect their home values to decline in the near-term, they may sell their homes quickly rather than waiting for the highest offer. If this occurs on a large scale in a neighborhood, comparable prices of homes in the area will decrease, pushing aggregate home values down. In this situation, negative expectations can result in a decline in aggregate home values for a neighborhood. Alternatively, residents who see positive changes, i.e., improvements in local schools or a decrease in crime, in their community may form an optimistic outlook for their neighborhood. This positive perception of prosperity may cause residents to forecast higher home values in the area. This expectation may lead sellers to price their homes higher, which may result in homes that do not sell if they are overvalued or an overall increase in home values if homes do sell for top dollar. The purpose of using the SHED, which represents these expectations, is to understand how well homeowners forecast their neighborhood home values. To analyze homeowners’ prediction accuracy, the model will use the SHED variables as predictors and actual home values via HPA data as the dependent variable.

# 2 Datasets and Variables

## 2.1 SHED

In September 2013, October 2014, and October – November 2015, the Federal Reserve Board’s Division of Consumer and Community Affairs conducted the Survey of Household Economics and Decisionmaking (SHED). The intention of the SHED is “to capture a snapshot of the financial and economic well-being of U.S. households, as well as to monitor their recovery from the recent recession and identify any risks to their financial stability” (Board of Governors of the Federal Reserve System 5). The survey collected data that were not easily accessible from other sources. The purpose of the SHED is to provide

a nationally representative snapshot of the economic situation of households in the United States at the time of the survey, as well their perspective on financial conditions in the recent past and expectations for conditions in the near future (Board of Governors of the Federal Reserve System 5).

The surveys cover a broad range of topics such as personal finances of households (U.S. adults), housing, living arrangements, economic fragility, emergency savings, income, spending, savings, banking, credit access & behavior, credit usage, retirement, health-related expenses, education, student debt, and auto lending. Notably, the survey asks respondents questions about how the value of their home changed in the past as well as their expectations of the future values of homes in their neighborhood (Board of Governors of the Federal Reserve System 5). The specific SHED questions used in the model will be discussed later in this section.

The surveys utilized samples of adults over the age of 18 years old from KnowledgePanel. KnowledgePanel is a probability-based web panel of over 50,000 randomly selected U.S. households (Board of Governors of the Federal Reserve System 5). The sample is designed to be representative of the U.S. population due to the random selection and sheer size of the sample. For the 2013 SHED, “e-mails were sent to 6,912 randomly selected members of KnowledgePanel resulting in 4,134 completed surveys and yielding a final stage completion rate of 59.8 percent” (Board of Governors of the Federal Reserve System 5). For the 2014 SHED:

e-mails were sent to 2,190 randomly selected respondents from the 2013 SHED (“re-interviewed respondents”) and 4,059 randomly selected respondents from the remaining members of KnowledgePanel (“fresh respondents”). The survey also includes an oversample of lower-income individuals by sending e-mails to 2,726 randomly selected respondents with a household income under $40,000 per year who are not included in the initial sample of re-interviewed respondents or fresh respondents … Overall, of the 8,975 respondents contacted for the survey, 5,896 respondents completed it, yielding an overall final stage completion rate of 65.7 percent (Board of Governors of the Federal Reserve System 5-6).

Additionally, for the 2015 SHED:

e-mails were sent to 2,853 respondents from the 2014 SHED (“re-interviewed respondents”) and 3,332 randomly selected respondents from the remaining members of KnowledgePanel (“fresh respondents”). The survey also includes an oversample of lower-income individuals by sending e-mails to 2,496 randomly selected respondents with a household income under $40,000 per year who are not included in the initial sample of re-interviewed respondents or fresh respondents … Of the 8,681 respondents contacted for the survey, 5,695 respondents completed it, yielding an overall final stage completion rate of 65.5 percent (Board of Governors of the Federal Reserve System 5-6).

In the 2014 and 2015 SHEDs, the oversample of lower income households improve the precision of estimates among the low-income population and allow for a sufficient sample size to reliably compare results for certain questions of interest across segments of the population.

All SHED data were downloaded from the Board of Governors of the Federal Reserve System’s website. There is one CSV file available for each SHED with all of the survey questions coded as variables in columns and each observation as a row. The individual datasets were merged in R to create one master dataset (refer to [section 2.3](#_2.3_R_Code) for a description of how the data were cleaned and merged). The final SHED dataset contains 15,675 observations.

For this study, only relevant variables were extracted from the original individual datasets to create a master SHED dataset. The master SHED dataset includes the following variables:

1. Year of the survey (“year”).
2. The 2013 (Board of Governors of the Federal Reserve System 19), 2014 (Board of Governors of the Federal Reserve System 10), and 2015 (Board of Governors of the Federal Reserve System 19) SHED asked “in the next 12 months, how much, if at all, do you think that home prices in your neighborhood will change?” The responses to this question represent the future expectations of respondents’ neighborhood home values. This variable is named “H4”. The responses to this question are coded as such:
   1. “Go down by more than 5 percent”, “go down by a lot”, “go down by 5 percent or less”, and “go down by a little” are coded as “0”.
   2. “Go up by 5 percent or less”, “go up by a little”, “go up by more than 5 percent”, and “go up by a lot” are coded as “1”.
   3. “Refused” is coded as “2”.
   4. “Stay about the same” is coded as “3”.
   5. “Don't know” is coded as “4”.

*Note: This study only focuses on increases or decreases in expectations rather than the unit/percentage change in H4. Hence the coding methodology for “0” and “1” above.*

1. In 2013, the SHED asked, “compared to five years ago (since 2008), do you think the value of your home today is higher, lower or stayed the same?” (Board of Governors of the Federal Reserve System 16). Whereas the 2014 (Board of Governors of the Federal Reserve System 9) and 2015 (Board of Governors of the Federal Reserve System 19) SHED asked, “compared to 12 months ago, do you think the value of your home today is higher, lower or stayed the same?” Since these questions all pertain to the past value of each respondent’s home value, they are coded under the same variable named “H1” to represent the past state of the housing market in that region. The responses to these questions are coded as such:
   1. “Lower value”, “value is a lot lower”, and “value is a little lower” are coded as “0”.
   2. “Higher value”, “value is a lot higher”, and “value is a little higher” are coded as “1”.
   3. “Refused” is coded as “2”.
   4. “Value has stayed the same” is coded as “3”.
   5. “Don’t know” is coded as “4”.

*Note: This study only focuses on increases or decreases in values rather than the incremental change in H1. Hence the coding methodology for “0” and “1” above.*

1. In order to merge the HPA and SHED data, the state of each respondent’s residence was used to create a new variable named “state\_region”. This variable converted each respondent’s state into a numeric label that represents the nine Census Regions of the United States (U.S. Census Bureau) that are coded as such:
   1. “1” is the New England region, which is inclusive of Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont.
   2. “2” is the Mid-Atlantic region, which is inclusive of New Jersey, New York, and Pennsylvania.
   3. “3” is the East-North Central region, which is inclusive of Indiana, Illinois, Michigan, Ohio, and Wisconsin.
   4. “4” is the West-North Central region, which is inclusive of Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, and South Dakota.
   5. “5” is the South Atlantic region, which is inclusive of Delaware, District of Columbia, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, and West Virginia.
   6. “6” is the East-South Central region, which is inclusive of Alabama, Kentucky, Mississippi, and Tennessee.
   7. “7” is the West-South Central region, which is inclusive of Arkansas, Louisiana, Oklahoma, and Texas.
   8. “8” is the Mountain region, which is inclusive of Arizona, Colorado, Idaho, New Mexico, Montana, Utah, Nevada, and Wyoming.
   9. “9” is the Pacific region, which is inclusive of Alaska, California, Hawaii, Oregon, and Washington.

These variables were compiled into one master SHED dataset that was then merged with the CoreLogic HPA data to create a final master dataset.

## 2.2 HPA

CoreLogic prepares the Home Price Analyzer (HPA), which represents actual home values for the purpose of this study. CoreLogic is a California based corporation that is a

leading provider of consumer, financial and property information, analytics and services to business and government. The company combines public, contributory and proprietary data to develop predictive decision analytics and provide business services that bring dynamic insight and transparency to the markets it serves. CoreLogic has built the largest U.S. real estate, mortgage application, fraud, and loan performance databases and is a recognized leading provider of mortgage and automotive credit reporting, property tax, valuation, flood determination, and geospatial analytics and services. More than one million users rely on CoreLogic to assess risk, support underwriting, investment and marketing decisions, prevent fraud, and improve business performance in their daily operations (Allen 5).

Based on CoreLogic’s credentials and reputation in the real estate industry, this study operates under the assumption that the underlying data driving the HPA accurately reflect the U.S. residential real estate market. The CoreLogic data used in this study were obtained through a corporate subscription.

The HPA is an automated valuation model (AVM) “that searches from more than 200 million historical residential sales records using three independent methods: hedonic, location-based and index-based” (CoreLogic). An AVM is “a computerized system that analyzes data to provide an estimate of market value for a property at a given point in time. AVMs can also be used to provide a broader analysis of a real estate market based on the AVM-determined values of homes in a defined area, of a particular type, or in a specific price tier” (CoreLogic 1). Additionally, CoreLogic AVM data cover 98 percent of all U.S. zip codes as well as 3,100 counties in all 50 states and Washington DC. This is representative of 99 percent of the U.S. population, 97 percent of all 145 million properties, over 50 million active mortgages, and 96 percent of loan-level non-agency mortgage securities (CoreLogic 1). In addition to CoreLogic AVMs’ representativeness of the U.S. real estate market, these AVMs also undergo rigorous testing to maintain their status as leading performers. These AVMs were used by eighteen out of the nation’s twenty largest mortgage lenders for property valuation (CoreLogic 1). All models are also tested on a daily basis. Since CoreLogic produces approximately 100 percent of U.S. residential housing stock on a rolling basis, actual purchase transactions are pulled (within three to five days of the transaction) and compared against the AVM to benchmark the AVM’s accuracy. These results are immediately analyzed and used to address any deviations from the performance standards of the models. In addition to the in house testing performed by CoreLogic, clients that utilize CoreLogic AVMs also perform independent testing whose results are sent back to CoreLogic for immediate analysis (CoreLogic 2). Therefore, the HPA is a representative measure of residential home values for each of the nine U.S. Census Regions. The HPA, as the dependent variable, should be a reliable measure to understand whether the SHED data, as independent variables, are effective predictors of actual home values.

The raw HPA data were collected at the city level rather than the state level like the SHED data. Therefore, there is an individual HPA observation for each city. In order to merge these data with the SHED data, which are recorded on the state level, the following two procedures were carried out. (1) For each HPA observation, the associated state was noted with the city name. Each HPA location at the city level observation was converted to a state value, then to one of the nine U.S. Census Bureau Regions and coded under the “state\_region” variable (refer to the [description of the “state\_region” variable in section 2.1](#stateregion) for details). (2) Additionally, each HPA observation is unique as it is based on a city. In order to homogenize the data at the regional level, two variables were created to represent the mean and median HPA values for each region. Both the mean and median variables for the HPA will be used in the models to understand which is a more reliable dependent variable. Therefore, the SHED respondents will be predicting home values for their region rather than their neighborhood as asked in the shed question. Refer to [section 4.1](#_4.1_Limitations_&) for more details about the implications of using regional values for the HPA and SHED data.

In order to understand whether SHED respondents correctly predicted the home value changes in their regions within the 12 months following the SHED, a variable (“hpa\_change”) was created, which is the change in the HPA. For example, for each 2013 SHED observation, the corresponding “hpa\_change” observation is the difference between the 2014 and 2013 HPA value for that region. Additionally, a variable named “hpa\_direction” was created to capture the direction of the change recorded in the “hpa\_chnge” i.e. the sign of the hpa\_change observation: “0” for a negative change or “1” for a positive change. Continuing with the previous example, if the HPA value was “2” in 2013 and “5” in 2014, then the “hpa\_change” observation would be “3” (5 – 2 = 3) and the “hpa\_direction” observation would be “1”.

The final master dataset includes all SHED and HPA data discussed throughout this paper thus far. Using this dataset, the models in [section 3](#section3) will be built to understand whether SHED respondents can predict actual changes in home prices for each U.S. Census Bureau Region.

## 2.3 R Code and Data Manipulation

A zip file was submitted alongside this file containing the raw datasets and R code used to clean and prepare the datasets and models discussed in this paper. Refer to [table 1](#_Table_1) in the appendix, which details the purpose of each file included in the submission folder.

# 3 Quantitative Analysis

*Null Hypothesis: No relationship exists between residents’ opinions about the value of their homes & homes in their region versus the HPA.*

*Alternative Hypothesis: A relationship exists between residents’ opinions about the value of their homes & homes in their region versus the HPA.*

## 3.1 Exploratory Data Analysis

The following models will be built using all data for all regions:

*Model #1:*

*Model #2:*

The model that holds true to the regression assumptions and fits the data best will be further broken down by each region. The individual regional models will help us understand how well the SHED respondents predict the HPA for each region. First, preliminary exploratory data analysis will be conducted in this section to present the distribution of the data and hypothesize the expected outcomes. The next section will discuss the models.

The distribution of the HPA data (refer to the [Distribution of HPA Data](#_Distribution_of_HPA) in the appendix for visualizations) for the HPA as the median and mean over all years and regions reveal that the data does not suffer from missingness but rather some skew. Overall, the data are skewed with more cases showing that regional housing markets experienced more downturns than improvements. This skew may be a result of the housing market recovery that has been ongoing since the Great Recession.

To contextualize the HPA data distribution, which represents the reality of the housing market, the SHED data must be assessed to understand the distribution of the data measuring changes in respondents’ home values since the recession. The distribution of the H1 variable data, capturing the change in SHED respondents’ home values in the past, show case imbalances, which may affect the prediction accuracy of the models (refer to the [Distribution of SHED Data (H1)](#_Distribution_of_SHED_3) in the appendix for visualizations). The expectation was that the SHED respondents would have clearly answered whether their home values increased, decreased, or remained the same. However, missing values, respondents who did not know whether their home values changed, and respondents who refused to answer this SHED question may alter the regression results. Furthermore, the data show that more respondents believed that their home values increased rather than decreased, worsening the skew of the data. A limitation of the SHED data is that respondents have the option to select “refuse” or “don’t know” as a response to this question. If respondents were only given increased, decreased, or unchanged as response options, the data may have been more equally distributed.

The limitations seen in the H1 SHED data are also reflected in the H4 SHED data (refer to the [Distribution of SHED Data (H4)](#_Distribution_of_SHED_4) in the appendix for visualizations). The H4 variable represents the SHED respondents’ expectations of the future values of homes in their neighborhood for all years and regions. Unlike the data for past home values (H1), these data show that more SHED respondents believe that home values in their neighborhood will decrease within the next twelve months, suggesting a negative outlook. These cases are severely skewed towards the negative outlook, which is inline with the HPA data. However, this skewness may adversely impact the ability of the model to predict positive changes in the HPA if there are not enough data due to severe case imbalances. Although no one refused to answer this question, the data suffers from missingness as well as respondents who answered that they did not know if future neighborhood home values would change.

Since the goal is to build a model for each region, severe case imbalances by region may forbid such granular models. The distributions of all observations show that the data are not equally distributed by region (refer to the [Distribution of Regional Data](#_Distribution_of_Regional) in the appendix for visualizations). There are concentrations of cases in the South Atlantic region versus very few cases, comparatively, in the New England region.

To summarize, this exploratory data analysis is supported by [table 2](#_Table_2) in the appendix, which provides descriptive statistics used in the study. The mean, standard deviation, minimum, and maximum values are listed in the table. On average, the change in the mean of the HPA (“hpa\_comp\_mean\_direction”) is 0.484, which is inline with the visualization in the appendix that shows that the data are more skewed towards downturns in the housing market. Also, on average, the change in the median of the HPA (“hpa\_comp\_med\_direction”) is 0.446, which is also skewed to the left like the change in the mean of the HPA. Respondents reported that their home values changed in the past (H1) on average by 1.251. Since the mean is between 1 for increased and 2 for refused, again the data are skewed. Additionally, respondents reported that they expect home values in their neighborhood (H4) to change by 2.104 on average, within the next twelve months. The mean is between 2 for refused and 3 for stayed the same. However, 1 is increased and no respondents refused to answer the question, so the mean is actually between 1 and 3. Finally, we see that on average there is a heavy concentration of observations (state\_region) in the South Atlantic region as the mean is 5.077.

The HPA data are complete as opposed to the SHED data that are inundated by missingness and observations that may not be helpful to the analysis such as “don’t know” and “refuse” responses. These limitations may be compounded by case imbalances by region, which may forbid the construction of models that are able to predict the HPA for every region.

## 3.2 Model #1

*Null Hypothesis: No relationship exists between residents’ opinions about the value of their homes & homes in their region versus the median HPA.*

*Alternative Hypothesis: A relationship exists between residents’ opinions about the value of their homes & homes in their region versus the median HPA.*

### Multiple Linear Regression Model #1

*Model #1:*

Model #1 is a multiple linear regression model for all regions and years using the median HPA as the dependent variable versus past home values (H1) and future neighborhood home values (H4) as the independent variables. An analysis of the Ordinary Least Squares (OLS) assumptions will be followed by an examination of the summary statistics if the OLS assumptions hold true.

The analysis of the regression diagnostic plots includes the “Residuals vs Fitted Plot”, “Normal Q-Q Plot”, “Distribution of Studentized Residuals”, “Scale-Location Plot”, and “Residuals vs Leverage Plot”. The “Residuals vs Fitted Plot” will display equally spread residual points around a horizontal line if a linear relationship exists between the independent and dependent variables. The “Normal Q-Q Plot” will show residuals that fall along the dotted line if they are normally distributed. To support this plot, the “Distribution of Studentized Residuals” shows the data in histogram form to clarify the type of distribution. The “Scale-Location Plot” tests whether there is constant variance (homoscedasticity) for each one of the residuals. Heteroscedasticity manifests when the residuals are not equally spread along the ranges of predictors. The “Residuals vs Leverage Plot” measures whether extreme outliers exist, which may alter the regression results if included or excluded from the model. Such extreme outliers have a large Cook’s distance and are typically located in the upper or lower right corners of the “Residuals vs Leverage Plot”.

For Multiple Linear Regression Model #1, the “Residuals vs Fitted Plot” shows that the model violates the OLS linearity assumption as the residuals are not equally spread around the horizontal regression line, so there is a nonlinear relationship between the predictor variables and outcome variable. But rather, there are two distinct parallel lines, neither of which overlap with the regression line. Next, the “Normal Q-Q Plot” evidences that the residuals are not normally distributed since they severely deviate from the straight dashed line. To support this claim, the histogram of the “Distribution of Studentized Residuals” shows that the data are severely skewed rather then normally distributed. However, it seems this is due to the ordinal nature of the data. Additionally, heteroscedasticity is evident in the “Scale-Location Plot” since the residuals are not equally spread along the ranges of the predictors. This suggests that the variance of the error term is different across observations, hence model uncertainty varies from observation to observation. Heteroscedasticity is also evident in the “Residuals vs Fitted Plot” since there are distinct patterns in the data. Also, the “Residuals vs Leverage Plot” shows that extreme outliers exist, which are located in the upper and lower right corners of the plot and may influence the regression results if they are included or excluded from the model. The regression diagnostics (refer to the [Multiple Linear Regression Model #1 Diagnostics](#mlr1d) in the appendix for more details) for this model suggest that multiple linear regression is not an appropriate statistical treatment. Additionally, the diagnostics show that the regression output is unreliable because the model does not fit the data. Another indicator of poor fit is the adjusted R squared, which suggests that only 0.3416% of the variation of the dependent variable is explained by its relationship to the independent variables. Therefore, the summary statistics of the model will not be analyzed further (refer to the [Multiple Linear Regression Model #1 Summary Statistics](#mlr1s) in the appendix for more details).

Due to the limitations and OLS violations of the multiple linear regression model, this is not an appropriate statistical method. The outcome variable is binary and independent variables are ordinal. Multiple linear regression ignores the ordinal measurement and treats the data as continuous. Therefore, a logistic model should be a more appropriate and suitable method.

### Logistic Regression Model #1

*Model #1:*

Logistic (“logit”) regression does not share the key assumptions of linear regression based on OLS, such as linearity, normality, homoscedasticity, and the measurement of each variable. In addition to these key OLS assumptions not holding true for the multiple linear regression model, the measurements of the variables are noteworthy. Logistic regression can handle a dependent variable that is dichotomous/binary and independent variables that are ordinal.

Logistic regression expresses the probability of a change in the direction of the median HPA for next year, based on the SHED respondents’ past home values and expectations of future home values in their neighborhood. Odds ratios will be calculated to determine the likelihood of the HPA direction change. The proportional odds assumption underlies ordinal logistic regression. That is, the relationships between the independent variables and logits are the same for all logits. Therefore, there are a set of parallel lines, thus one for each category of the outcome variable. For the purposes of this study, we assume that the proportional odds assumptions holds true.

When the SHED respondents reported in increase in their past home value (H1), the estimated odds of an upward change in the median HPA decreases. This is statistically significant at the 5% level. Additionally, when the SHED respondents expected home values in their neighborhood (H4) to increase, the estimated odds of an upward change in the median HPA also decreases. This is statistically significant at the 1% level. Since both variables are statistically significant, the null hypothesis is rejected so a relationship exists between residents’ opinions about the value of their homes & homes in their region versus the median HPA. Refer to the [Logistic Regression Model #1 Summary Statistics](#logit1s) in the appendix for details.

The odds ratio for H1 reveals that (holding all else constant) for each SHED respondent who reported an increase in their past home value, the odds of an increase in the median HPA will increase by a factor of 0.9516369, which means that the odds of an increase in the median HPA actually decreases. The odds ratio for H4 reveals that (holding all else constant) for each SHED respondent who expected neighborhood home values to increase, the odds of an increase in the median HPA will increase by a factor of 0.9242772, which means that the odds of an increase in the median HPA actually decreases. There is an indirect relationship between the SHED respondents past home values and the expected neighborhood home values versus the median HPA. Refer to the [Logistic Regression Model #1 Odds Ratios](#logit1odds) in the appendix for details. These results are contrary to the basis of this study, that is homeowners are able to accurately predict actual home prices based on past home values and expectations of future neighborhood home values.

The results of the logistic regression suggest that the SHED respondents are not successful forecasters of actual home values. An analysis of the estimated probabilities showed that when H1 and H4 are equal for every category, H0 is 0 for each category of H4, and H4 is 0 for every category of H1, the estimated odds and probability that there will be an increase in the median HPA always decreases. Refer to the [Logistic Regression Model #1 Predicted Probabilities Tables A – C](#logit1a) in the appendix for details.

## 3.3 Model #2

### Multiple Linear Regression Model #2

*Model #2:*

The same procedure was carried out for multiple linear regression model #2 as multiple linear regression model #1. The results of model #2 were similar in that all OLS assumptions were violated and a low adjusted R squared proved that the model does not fit the data. Given these preliminary results, further analysis of the summary statistics and coefficients would have been inefficient since the model is unreliable as multiple linear regression is not an appropriate statistical treatment of the binary and ordinal data. Refer to the Multiple Linear Regression Model #2 [Diagnostics](#mlr2d) & [Summary Statistics](#mlr2s) in the appendix for details.

### Logistic Regression Model #2

*Model #2:*

An ordinal logistic regression model was also run for model #2 since multiple linear regression was deemed as an inappropriate method.

For multiple linear regression models #1 and #2, the adjusted R squared penalizes the models based on OLS methodology. For logistic regression models, the Akaike Information Criterion (AIC) measures the penalized likelihood. For OLS models, the model with the highest adjusted R squared is usually the best model, assuming that the OLS assumptions hold true. For logistic regression models, the AIC is used to compare models. The model with the lowest AIC is typically the “best” model. Comparatively, logistic regression model #1 is superior to model #2 based on the AIC. Logit model #1 has an AIC of 12,612 compared to logit model #2 with an AIC of 12,644. The smaller AIC for model #1 indicates that model #1 fits the data better than model #2.

Although logit model #1 is superior to logit model #2, it is worth noting that the independent variables do vary by model by magnitude but not direction. When the SHED respondents reported in increase in their past home value (H1), the estimated odds of an upward change in the mean HPA decreases. This is not statistically significant. Additionally, when the SHED respondents expected home values in their neighborhood (H4) to increase, the estimated odds of an upward change in the mean HPA decreases. This is statistically significant at the 1% level. Therefore, the null hypothesis cannot be rejected for both variables in logit model #2. Both variables were statistically significant in logit model #1 whereas only H4 is significant in logit model #2. Additionally, the odds of an upward change in the median HPA for logit model #1 decrease by less of a percentage in logit model #1 as opposed to logit model #2 for the mean HPA. This suggests that the SHED respondents are more likely to accurately predict the median HPA rather than the mean HPA. Since logit model #1 has a lower AIC and both predictors are statistically significant, this is the superior model. Since logistic model #1 is better than logit model #2, we will rely on the results of logit model #1 rather than performing deeper analysis on logit model #2. For completeness, refer to the appendix for the [Logistic Regression Model #2 Summary Statistics](#logit1s), [Logistic Regression Model #2 Odds Ratios](#logit2odds), and [Logistic Regression Model #2 Predicted Probabilities Tables A – C](#logit2a).

# 4 Conclusion

Multiple linear regression models #1 and #2 were unreliable because OLS is not a suitable method for these type of data given their binary and ordinal nature. Therefore, the logistic regression models #1 and #2 were built. Of the logit models, model #1 was superior based on the AIC test and comparison of the independent variables. Model #1 used median HPA values as the dependent variable whereas model #2 used the mean of the HPA. The logit model that used the median HPA may have been a stronger fit and more reflective of actual home prices since the median is not as sensitive to outliers as the mean. Therefore, logit model #1 is the most reliable model in this study.

The original null hypothesis was that no relationship exists between residents’ opinions about the value of their homes & homes in their region versus the median HPA. Since logit model #1 yielded statistically significant results and the assumption is that the logistic assumption of proportional odds held true, we can reject the null hypothesis, which suggests that a relationship exists between residents’ opinions about the value of their homes & homes in their region versus the median HPA. Although there is a relationship between the variables, the type of relationship was unexpected, since, regardless of the SHED respondents’ answers to the questions (H1 and H4), there were always decreasing odds of the median HPA increasing. This may be a result of the housing market recovery since the Great Recession. The somewhat turbulent real estate market may have contributed to this overall trend of downturns in the median HPA.

## 4.1 Limitations & Future Research

The initial goal of this study was to choose the best model and then break down that model by region, using training and testing data, to build regional confusion matrices to understand how well the SHED respondents predicted the HPA. However, these granular models by region were not possible due to severe imbalances of observations across regions. The [distribution of regional data](#_Distribution_of_Regional_1), as shown in the appendix, revealed that the data are mostly concentrated in the South Atlantic region, with very few cases, comparatively, in the New England, East-South Central, West-North Central, and Mountain regions. Although confusion matrices were built for some regions, there were not enough data to build a confusion matrix for every region. Therefore, a confusion matrix was built for all regions for logistic regression model #1. The matrix shows that the model has a prediction accuracy rate of 53.38%, which is largely driven by correct predictions of downturns in the HPA. Refer to the [Logistic Regression Model #1 Confusion Matrix](#logit1confu) in the appendix for details.

Another limitation of this study was that the SHED respondents were asked about changes to neighborhood home values (H4) rather than regional home values, which is how the data were aggregated. If the city of each SHED respondents’ residence was included in the SHED data, the HPA data could have been compared against the SHED data on the city level rather than the regional level, which may have yielded more accurate prediction accuracy. Furthermore, the availability of city level SHED data would have eliminated the necessity to convert the HPA to the median value for each region. This lack of granular data may have contributed to an under specification problem as the local data were omitted from the model since the SHED data were recorded at the state level and HPA at the city level. Therefore, this study may significantly improve by using such granular data.

The goal of these algorithms was to help home seekers attain summary statistics of the past (H1), current, and expected future home values of a particular neighborhood (H4). Such data would help home seekers in choosing a home in an area for which they completely understand that granular housing market (neighborhood). Ideal data for future applications of this study would be collected using the SHED respondents’ zip codes, which would be more easily compared to HPA data, which would be collected on the zip code level as well. Rebuilding the models proposed in this study on the zip code level would give homebuyers access to data about the neighborhood in which they are house hunting. This type of model would be profitable for real estate mobile apps and websites. Perspective home buyers would be drawn to real estate services that utilize this algorithm because it would give them access to past and expected home values, in addition to the data already available about current home values for each neighborhood.

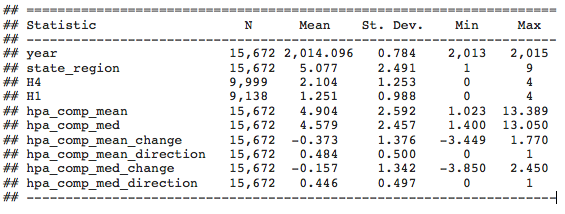
# Appendix

## Table 1

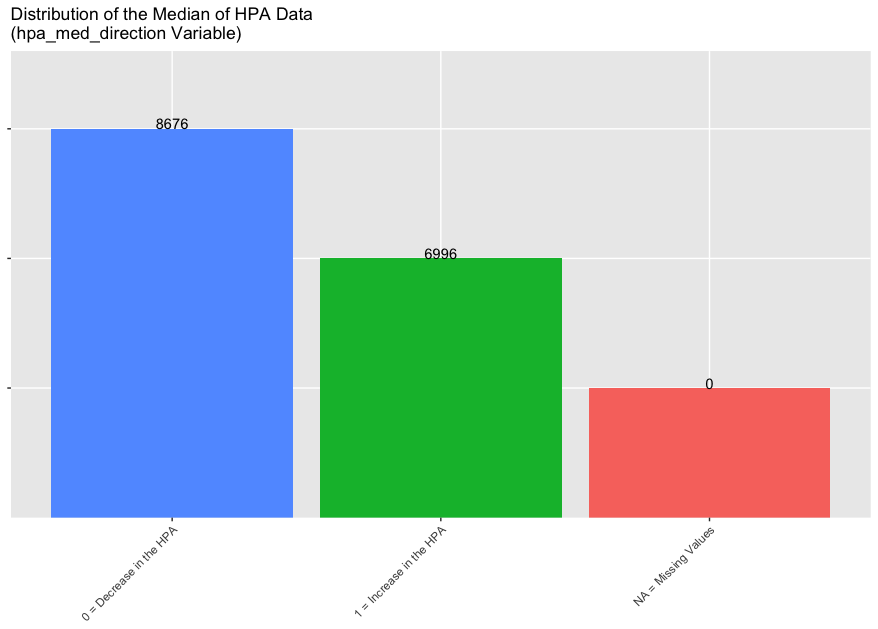
|  |  |  |
| --- | --- | --- |
| **File Name** | **Subfolder Name** | **Purpose** |
| EDA\_v1.R | Code | Exploratory Data Analysis (EDA) of the SHED. Refer to the R file for more details. |
| EDA\_v2.R and EDA\_v3.R | Code | Create a SHED dataset with the variables relevant to this project. Refer to the R file for more details. |
| EDA\_v4.Rmd | Code | Continue to modify the SHED dataset with the variables relevant to this project and clean the CoreLogic data. Refer to the R file for more details. |
| EDA\_v5.Rmd | Code | Create a master dataset. Refer to the R file for more details. |
| EDA\_v6.Rmd | Code | Final analysis of the data. Refer to the R file for more details. |
| EDA\_v6.html | Code | Knit version of “EDA\_v6.Rmd” above. |

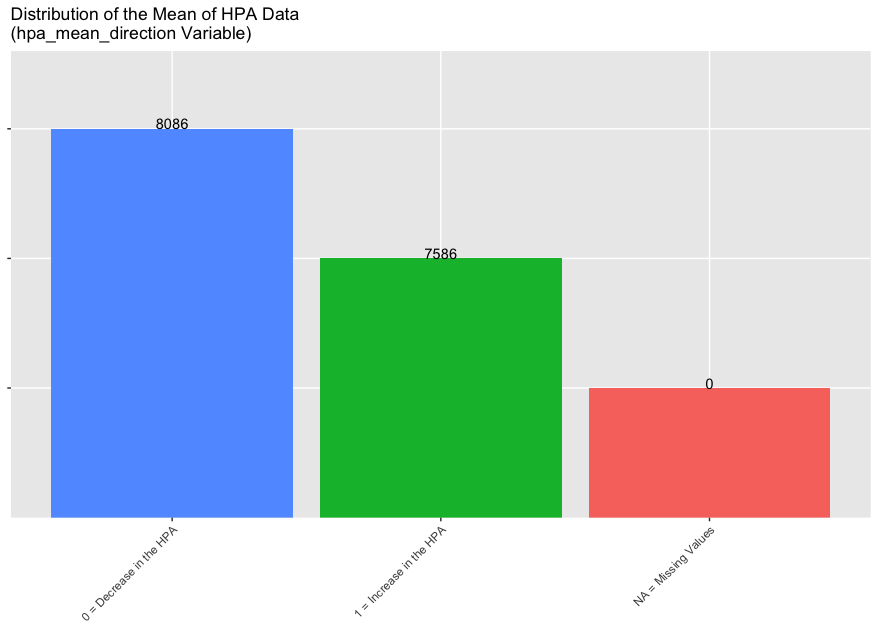
*\* NOTE: All input and output files referenced in the above R files are not saved in the submitted file package because the CoreLogic data were obtained through a paid corporate subscription and are not permitted to be shared outside of the corporation that maintains this subscription. Therefore, the input file of CoreLogic raw data as well as the output final master dataset file (which includes CoreLogic data) are not included in the file package provided. However, all code is included via R files for reproducibility purposes in the event that the reader’s goal is to reproduce the study using their own CoreLogic data. CoreLogic data used in this study were obtained on March 3, 2017.*

## Table 2

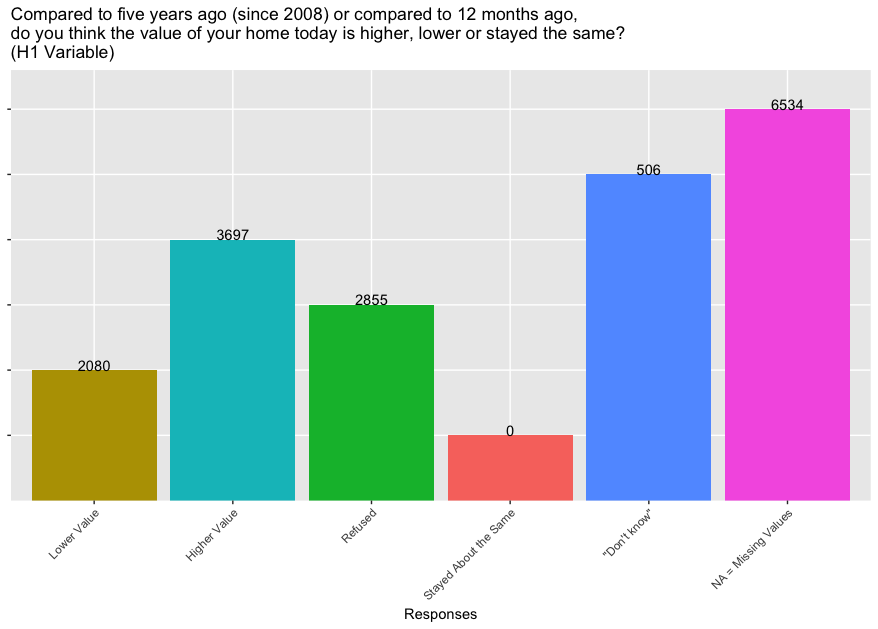


## Distribution of HPA Data

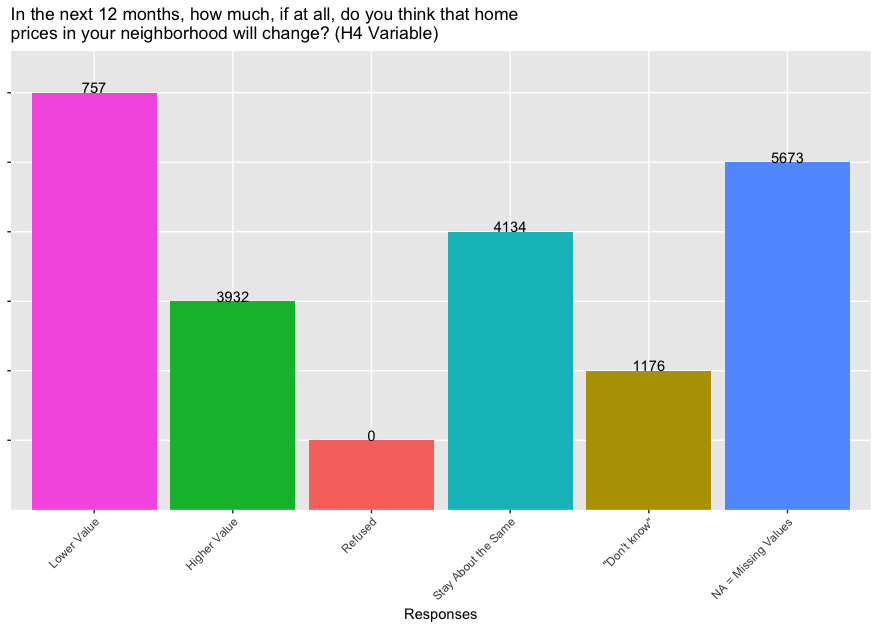




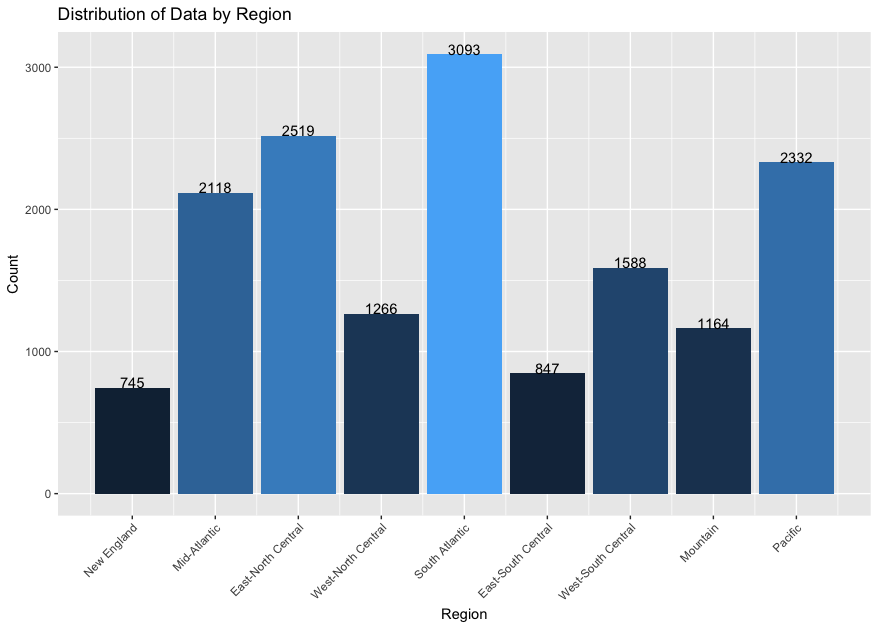
## Distribution of SHED Data (H1)



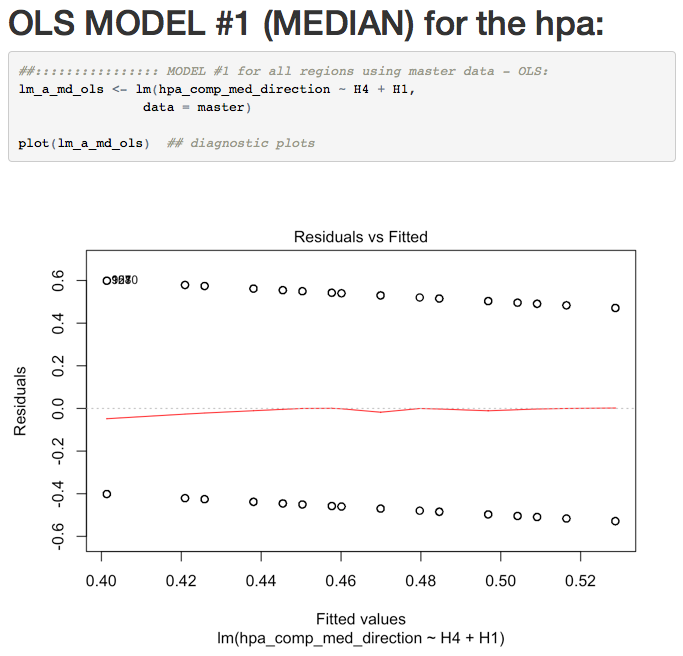
## Distribution of SHED Data (H4)

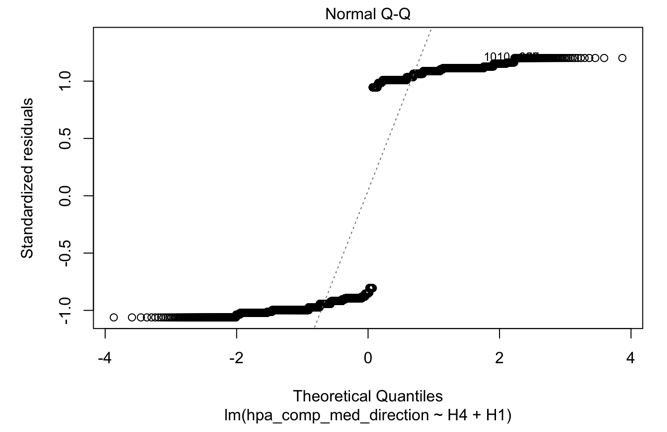


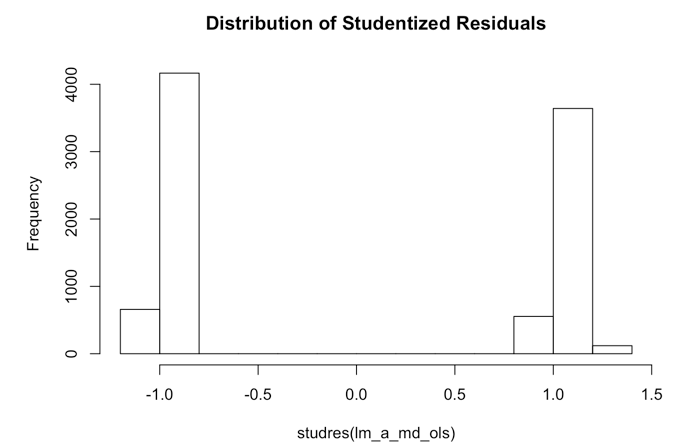
## Distribution of Regional Data

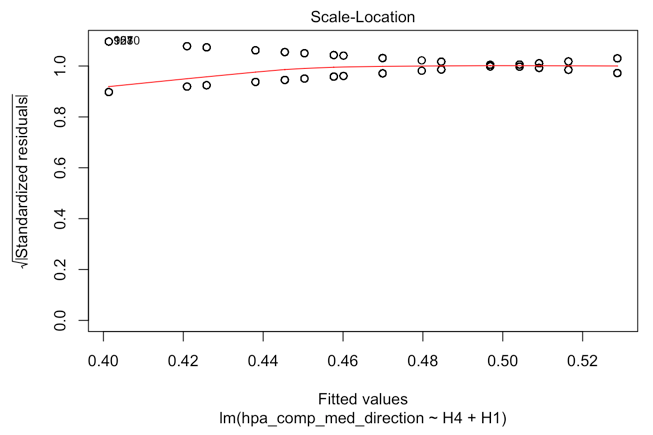


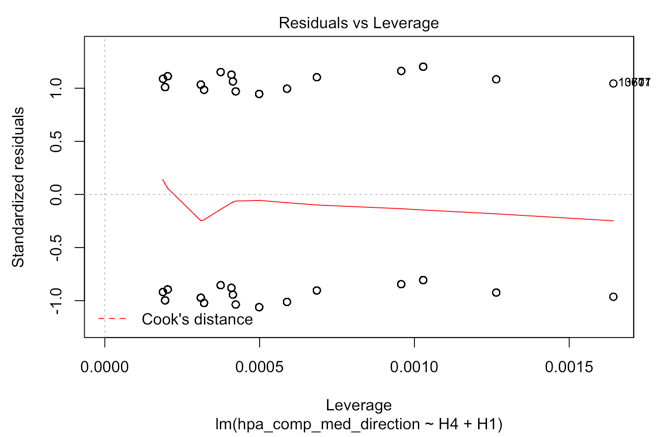
## Multiple Linear Regression Model #1 Diagnostics



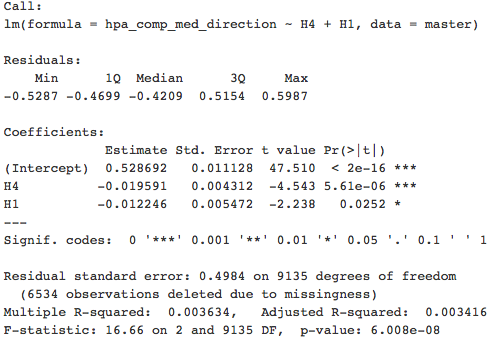




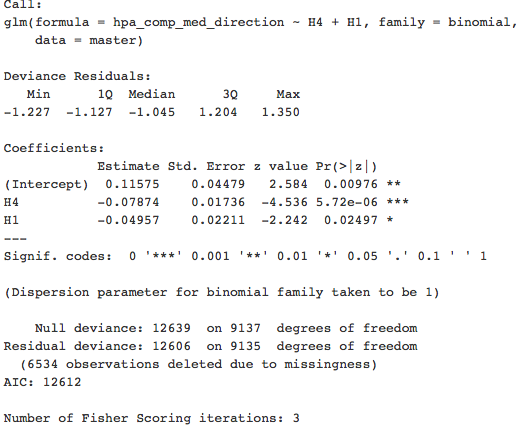




## Multiple Linear Regression Model #1 Summary Statistics



## Logistic Regression Model #1 Summary Statistics



## Logistic Regression Model #1 Odds Ratios



## Logistic Regression Model #1 Predicted Probabilities Table A

|  |  |
| --- | --- |
| H1 & H4 Response | Estimated Odds & Probability that the Median HPA Increased |
| 0 (Lower Value) | 0.5289064 |
| 1 (Higher Value) | 0.4968599 |
| 2 (Refused) | 0.4648393 |
| 3 (Stayed About the Same) | 0.4331059 |
| 4 (Don’t Know) | 0.4019119 |

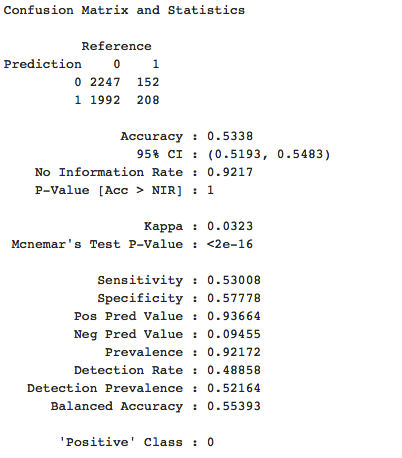
## Logistic Regression Model #1 Predicted Probabilities Table B

|  |  |  |
| --- | --- | --- |
| H1 Response | H4 Response | Estimated Odds & Probability that the Median HPA Increased |
| 0 (Lower Value) | 0 (Lower Value) | 0.5289064 |
| 0 (Lower Value) | 1 (Higher Value) | 0.5092518 |
| 0 (Lower Value) | 2 (Refused) | 0.4895685 |
| 0 (Lower Value) | 3 (Stayed About the Same) | 0.4699175 |
| 0 (Lower Value) | 4 (Don’t Know) | 0.4503594 |

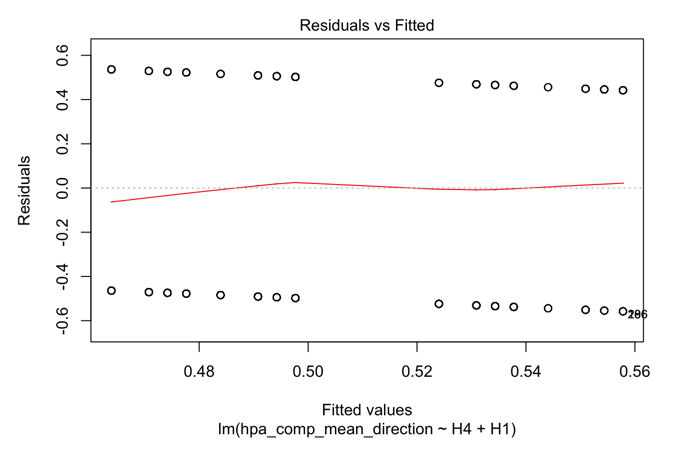
## Logistic Regression Model #1 Predicted Probabilities Table C

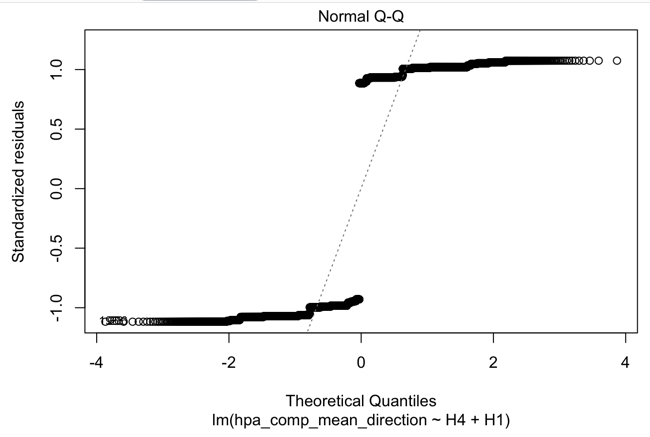
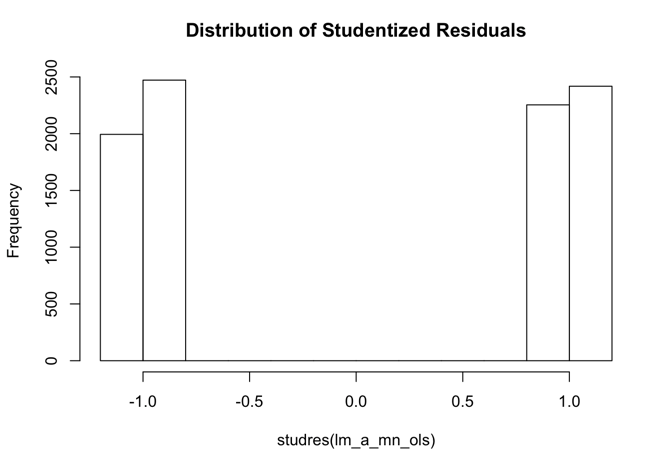
|  |  |  |
| --- | --- | --- |
| H1 Response | H4 Response | Estimated Odds & Probability that the Median HPA Increased |
| 0 (Lower Value) | 0 (Lower Value) | 0.5289064 |
| 1 (Higher Value) | 0 (Lower Value) | 0.5165397 |
| 2 (Refused) | 0 (Lower Value) | 0.5041527 |
| 3 (Stayed About the Same) | 0 (Lower Value) | 0.4917606 |
| 4 (Don’t Know) | 0 (Lower Value) | 0.4793787 |

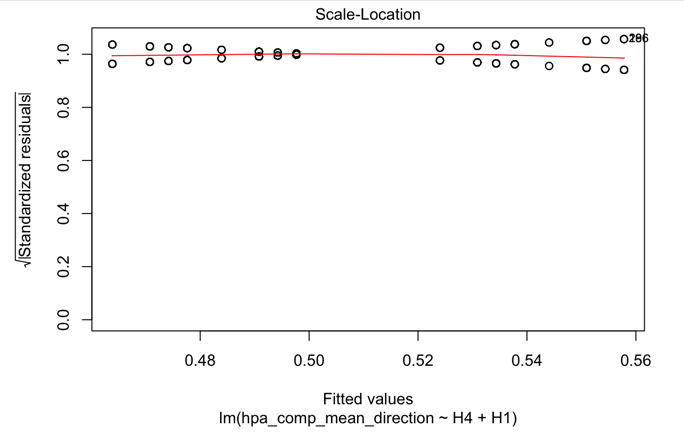
## Logistic Regression Model #1 Confusion Matrix

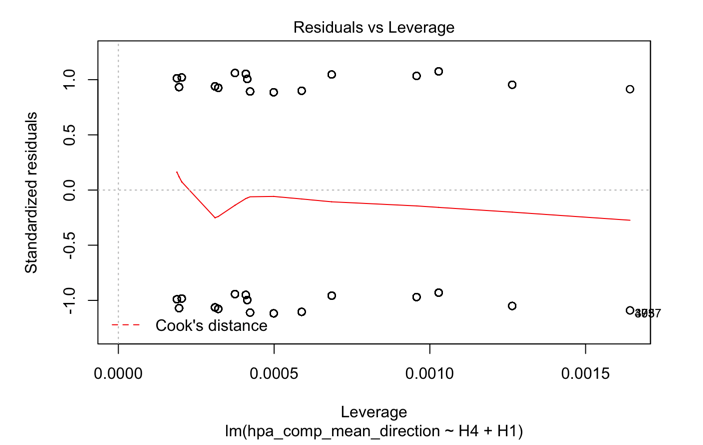


## Multiple Linear Regression Model #2 Diagnostics

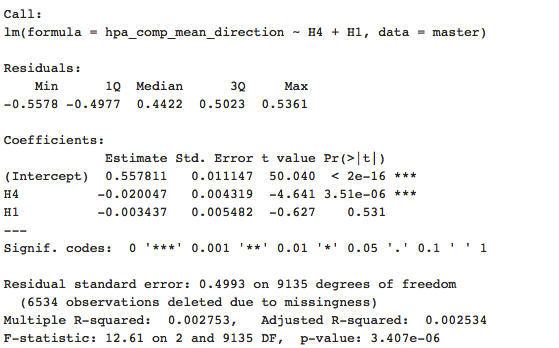




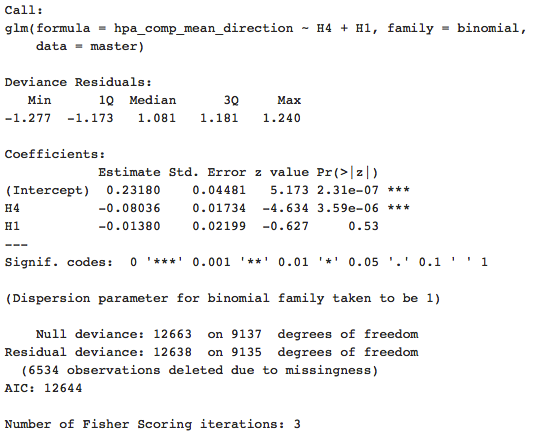




## Multiple Linear Regression Model #2 Summary Statistics



## Logistic Regression Model #2 Summary Statistics



## Logistic Regression Model #2 Odds Ratios

Macintosh HD:Users:StephanieLangeland:Downloads:Snip20170416_9.png

## Logistic Regression Model #2 Predicted Probabilities Table A

|  |  |
| --- | --- |
| H1 & H4 Response | Estimated Odds & Probability that the Median HPA Increased |
| 0 (Lower Value) | 0.557691 |
| 1 (Higher Value) | 0.5343564 |
| 2 (Refused) | 0.5108704 |
| 3 (Stayed About the Same) | 0.4873363 |
| 4 (Don’t Know) | 0.4638582 |

## Logistic Regression Model #2 Predicted Probabilities Table B

|  |  |  |
| --- | --- | --- |
| H1 Response | H4 Response | Estimated Odds & Probability that the Median HPA Increased |
| 0 (Lower Value) | 0 (Lower Value) | 0.557691 |
| 0 (Lower Value) | 1 (Higher Value) | 0.537788 |
| 0 (Lower Value) | 2 (Refused) | 0.5177639 |
| 0 (Lower Value) | 3 (Stayed About the Same) | 0.4976825 |
| 0 (Lower Value) | 4 (Don’t Know) | 0.4776086 |

## Logistic Regression Model #2 Predicted Probabilities Table C

|  |  |  |
| --- | --- | --- |
| H1 Response | H4 Response | Estimated Odds & Probability that the Median HPA Increased |
| 0 (Lower Value) | 0 (Lower Value) | 0.557691 |
| 1 (Higher Value) | 0 (Lower Value) | 0.5542847 |
| 2 (Refused) | 0 (Lower Value) | 0.5508732 |
| 3 (Stayed About the Same) | 0 (Lower Value) | 0.5474569 |
| 4 (Don’t Know) | 0 (Lower Value) | 0.5440362 |

# Works Cited

Allen, Susan. "The Role of Conditional Logic in AVM Cascade Creation." White Paper. 2009.

Board of Governors of the Federal Reserve System. 2013 Public SHED codebook. Codebook. Washington, DC: Consumer and Community Development Research Section of the Federal Reserve Board’s Division of Consumer and Community Affairs, 2014.

—. Codebook for 2014 Survey of Household & Economics Decisionmaking. Codebook. Washington, DC: Consumer and Community Development Research Section of the Federal Reserve Board’s Division of Consumer and Community Affairs, 2015.

—. Report on the Economic Well-Being of U.S. Households in 2013. Survey. Washington, D.C.: Consumer and Community Development Research Section of the Federal Reserve Board’s Division of Consumer and Community Affairs, 2014.

—. Report on the Economic Well-Being of U.S. Households in 2014. Survey. Washington, DC: Consumer and Community Development Research Section of the Federal Reserve Board’s Division of Consumer and Community Affairs, 2015.

—. Report on the Economic Well-Being of U.S. Households in 2015. Survey. Washington, DC: Consumer and Community Development Research Section of the Federal Reserve Board’s Division of Consumer and Community Affairs, 2016.

—. SHED 2015 public use codebook. Coebook. Washington, DC: Consumer and Community Development Research Section of the Federal Reserve Board’s Division of Consumer and Community Affairs, 2016.

CoreLogic. "Automated Valuation Solutions." Valuation. 2014.

—. Geo AVM Suite. 26 March 2017 <https://www.corelogic.com/about-us/events/asset\_upload\_file671\_12296.pdf>.

—. "MLS Data in AVMs: The Case For — and Against." White Paper. 2010.

U.S. Census Bureau. 25 March 2017 <https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us\_regdiv.pdf>.