Experimental Statistics I

Isabelle Kallenberg

Andrea Rubio

Ozcan Ozturk

Zheran Wang

STATS 7000

**Housing Affordability in the United States**

**Abstract:**

The Housing Affordability Data System (HADS) is a set of files from the 1985 and later American Housing Service (AHS) and the 2002 and later Metro AHS. This data set contains information pertaining to the affordability of housing units and the housing cost burdens of households in regard to the area median incomes, poverty level incomes, and Fair Market Rents (FMRs). 300 observations were randomly selected for this research study. The seven variables used were the price of the house, the year in which it was constructed, the number of rooms in the house, the closeness of the house to the city, the condition of the house, the insurance costs, and the total income of the household. Using SAS analysis, the matrix plot showed that some of the variables had linear relationships. According to the QQ-plot and the residual plot, we determined that the model assumptions were satisfied. The multiple linear regression method was used to test the model, giving an R-square value of 25%. Backward elimination was then used to eliminate insignificant variables, giving an R-square of 24%. Interactions were selected based on the correlation table produced by SAS. The new model, a combination of the old model and the interacted variables, was then tested with multiple linear regression. The R-square value of the new model was 27%. After another backward elimination, an R-square value of 27% was produced. There exists a linear relationship between the house price and the number of rooms, condition of the house, insurance, and total income of the household. The model improved after the interaction of the variables.

**Introduction:**

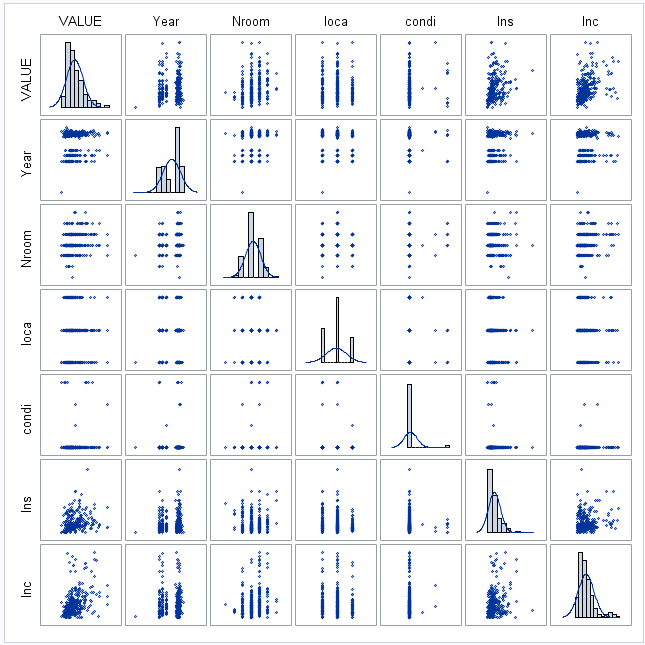
In September 2014, it was reported by the United States Census Bureau that the real median household income of Americans was $51,939 in the year 2013 compared to $51,759 in 2012. Statistically, this difference in median household income between 2012 and 2013 is insignificant. However, the real median household income was 8.0 percent lower in 2013 than it was in 2007. According to the same source, the United States population lives more often in urban areas rather than in rural. In 2000 and 2010, Americans living in urban areas accounted for approximately 80% of the population whereas those living in rural areas made up approximately 20% of the population. The data files were taken from the American Housing Service (AHS) public use files and the published income limits and Fair Market Rents (FMRs). The Housing Affordability Data System (HADS) is a set of files from the 1985 and later national AHS and the 2002 and later Metro AHS. The data set contains information pertaining to the affordability of housing units and the housing cost burdens of households in regard to the area median incomes, poverty level incomes, and FMRs. Of over 10,000 observations available in this data file, 300 were randomly selected for the purpose of this research project. Additionally, among approximately one hundred variables measured in HADS, only seven were used for this statistical analysis: the price of the house (renamed VALUE), the year in which the unit was constructed (YEAR), the number of rooms in the house (NROOM), the closeness of the house to the city (LOCA), the condition of the house (CONDI), the insurance costs or other home fees (INS), and the total income of the household (INC). For the purpose of this study, NROOM was measured from 0-4, with 0 indicating that the unit was a studio, and 4 indicating that there are 4 or more bedrooms in the unit. LOCA was a characteristic variable indicating whether the unit was located in the central city, the suburb, or in nonmetropolitan areas. These characteristics were renamed as 0, 1, and 2, respectively, in the SAS analysis. CONDI indicated whether the unit was in adequate condition, moderately inadequate, severely inadequate, or was vacant and contained no information. These characteristics were changed to 0, 1, 2, and 3 in the SAS analysis. The purpose of this statistical analysis was to analyze the relationship between the house prices with the year of construction, number of rooms, closeness to the city, condition of the houses, insurance, and total income of the households.

**Methods and analysis:**

**Model 1**

The function for model 1 and its variables are as follows:

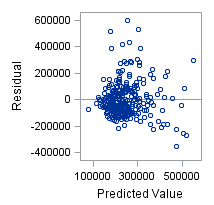
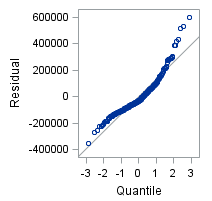
***Value***

****

**Figure 1.** Matrix Plot

Insurance and income of the household show some linear relationship but is unclear and some variables do not show a clear pattern. Insurance and income show a right-skew.

**Model 1- Results**



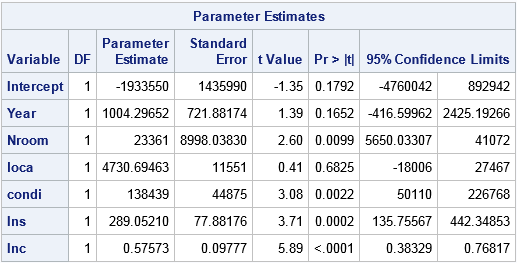
**Figure 2.** QQ plot (Left) and Residual Plot (Right) for Model 1.

**QQ-plot:** In the QQ-plot we can observe a normal distribution with most of the variables. Outliers are observed but don´t seem affecting the data.

**Residual plot:** In the residual plot, some values seem disperse, however, most of the data is together, no clear pattern is noticed.

Model assumptions are satisfied for this model.

**Regression- Model 1**





**Figure 2.** Regression and R square for model 1.

The fitted regression function for model 1:

**Figure 3.** Parameter Estimates and R-square table of model 1.

**Fitted linear regression function**

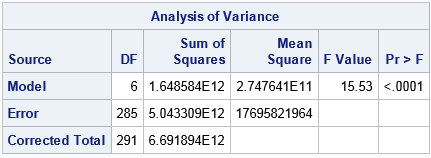
**Value**

Some variables in model 1 have p-values larger than 0.05 which may cause the need for eliminating some insignificant variables in this model. The R-square is 0.2464 which means that 24% of our data can be explained by this model.

**Linear Association- Model 1**

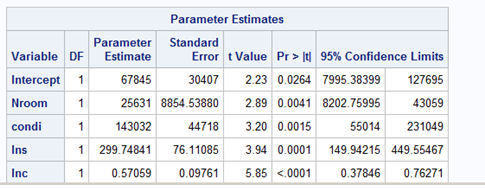
H0: There is no linear relationship between value and other 6 variables.

Ha: There exists linear relationship between value and other 6 variables.



**Figure 4.** ANOVA table for model 1.

The p-value for this linear association is less than 0.05, so we reject H0. There exists a linear association between the responses and the predictor.

**Model-1 After backward elimination**



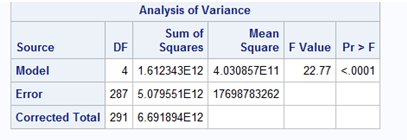
**Figure 5.** Model 1 after backward elimination.

**Fitted linear regression function**

Value =25631 ∗𝑿𝒏𝒓𝒐𝒐𝒎+ 25631 ∗𝑿𝒄𝒐𝒏𝒅𝒊+ 299.784 ∗𝑿𝒊𝒏𝒔+ 𝟎.5705∗𝑿𝒊𝒏𝒄

The R-square for model 1 is 0.2409, which means that 24% of our data can be explained by the model.

**Linear Association after backward elimination Model 1**



**Figure 6.** ANOVA table for linear association in model 1 after backward elimination.

H0: There is no linear relationship between value and other 4 variables.

Ha: There exists linear relationship between value and other 4 variables.

The p-value for this linear association is less than 0.05, so we reject H0. There exists a linear association between the responses and the predictor.

**Model 2-Results**

Tested the correlation between the variables in SAS by using proc corr statement.

After the test of correlation, there were 5 more variables that needed to be created. The five new variables and their interactions are listed below.

Yearo = Year \* Nroom

Yearis = Year \* Ins

Roinc = Nroom \* Inc

Locins = Loca \* Ins

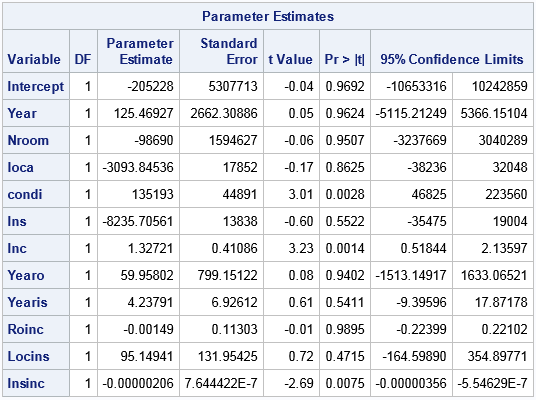
Insinc = Ins \* Inc

After the new variables were created, these variables combined with these 6 old variables formed a new model: model 2.

**Value =𝛽0+𝛽1 𝑦𝑒𝑎𝑟+𝛽2 𝑛𝑟𝑜𝑜𝑚+𝛽3 𝑙𝑜𝑐𝑎+𝛽4 𝑐𝑜𝑛𝑑𝑖+𝛽5 𝑖𝑛𝑠+𝛽6 𝑖𝑛𝑐+𝛽7 𝑦𝑒𝑎𝑟𝑜+𝛽8 𝑦𝑒𝑎𝑟𝑖𝑠+𝛽9 𝑟𝑜𝑖𝑛𝑐+𝛽10 𝑙𝑜𝑐𝑖𝑛𝑠+𝛽11 𝑖𝑛𝑠𝑖𝑛𝑐+𝜀.**

**Regression- Model 2**

Used multiple linear regression to test model 2. From the output of SAS the fitted regression function was determined.





**Figure 7.** Parameter Estimates and R-square table of model 2.

**Fitted linear regression function**

Value =−𝟐𝟎𝟓𝟐𝟐𝟖+𝟏𝟐𝟓.𝟒𝟕 ∗𝑿𝒚𝒆𝒂𝒓−𝟗𝟖𝟔𝟗𝟎 ∗𝑿𝒏𝒓𝒐𝒐𝒎−𝟑𝟎𝟗𝟑.𝟖 ∗𝑿𝒍𝒐𝒄𝒂+𝟏𝟑𝟓𝟏𝟗𝟑 ∗𝑿𝒄𝒐𝒏𝒅𝒊−𝟖𝟐𝟑𝟓.𝟕 ∗𝑿𝒊𝒏𝒔+𝟏.𝟑𝟐𝟕∗𝑿𝒊𝒏𝒄+𝟓𝟗.𝟗𝟓𝟖∗ 𝑿𝒚𝒆𝒂𝒓𝒐+𝟒.𝟐𝟑𝟖∗𝑿𝒚𝒆𝒂𝒓𝒊𝒔−𝟎.𝟎𝟎𝟏𝟒𝟗 ∗ 𝑿𝒓𝒐𝒊𝒏𝒄+𝟗𝟓.𝟏𝟒𝟗 ∗ 𝑿𝒍𝒐𝒄𝒊𝒏𝒄−𝟎.𝟎𝟎𝟎𝟎𝟎𝟐𝟎𝟔∗𝑿𝒊𝒏𝒔𝒊𝒏𝒄

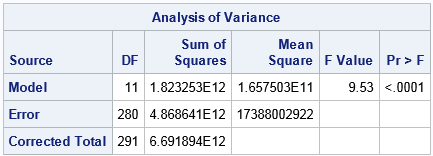
This time, 27.25% of the data could be explained by the model.

**Linear Association- Model 2**

Then tested the linear association of the model.

H0: There is no linear relationship between value and other 11 variables.

Ha: There exists linear relationship between value and other 11 variables.



**Figure 8**. ANOVA table for model 2.

The p-value of the test from SAS output showed less than 0.05, therefore, H0 was rejected, claiming that Ha was true. There existed a linear association with the model and the data.

**Model-1 After backward elimination**

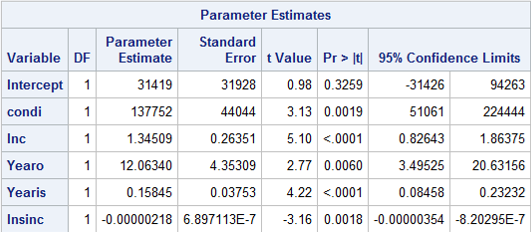
Still, within the model, there were some variables that were insignificant. Those variables were removed by testing significant.

SAS did the tests systematically by using backward elimination.

SAS suggested to eliminate 6 variables that were insignificant, the variables removed are listed below.

Roinc, Year, Loca, Nroom, Locins, and Ins.

After removing the variables listed above, we used multiple linear regression to test the model again.





**Figure 9.** Model 2 after backward elimination.

**Fitted linear regression function**:

**Value =31419+137752 \*Xcondi+1.345\*Xinc+12.063\*Xyearo-0.158 \* Xyearis-0.00000218\*Xinsinc**

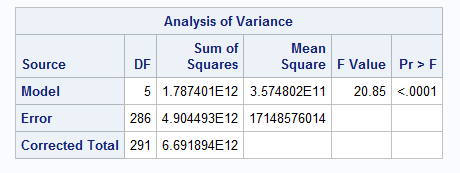
The R-square value showed that after removed the insignificant variables, there were 26.71% of the data can be explained by the new model. It was slightly decrease due to less variables.

**Linear Association after backward elimination Model 2**

Then tested the linear association of the model.

H0: There is no linear relationship between value and other 5 variables.

Ha: There exists linear relationship between value and other 5 variables.



**Figure 10.** ANOVA table for linear association in model 2 after backward elimination.

Since the p-value was less than 0.05 from SAS output, we rejected the H0, which means that there existed a linear association between the model and the data.

Overall, with the intersection of the variables, the R-square increased to 27.25%. Adjusted R-square value increased to 0.2439 after removing the insignificant variables. R-square value reduced to 26.71% due to the decreased number of variables, but the adjusted R-square value increased to 0.2543, which means the model improved compared to model 1.

**Conclusion**

In this study, we analyzed the factors that affected house values in the United States using The Housing Affordability Data System (HADS) from American Housing Survey: housing affordability data system for the period covering the years from 2002 to 2009.

For this purpose, we employed a linear regression model to test if there exists a linear association between the house value and the variables such as number of rooms, the year of construction, location of the house, insurance and income of the household. We used a backward elimination method to revise our model by omitting the variables that were not significant. Results showed that the number of rooms, condition of house, insurance and income had significant effects on house values. Results also showed that about 24% of the variation in the data was explained by these variables.

Moreover, we used an additional model to develop the existing model. For this purpose, we added interaction terms in our model. We chose interaction terms by checking the correlation matrix. Our results suggested that the interactions between year-insurance, year-number of room and insurance-income should be added to the model. Interestingly, the year variable itself was not significant, which was surprising because we would expect the year of construction to affect house prices. However, when we interacted the year variable with other variables, we got significant effects. After adding the interaction terms, our model slightly improved to the extent where about 27 percent of variation was explained by this model.

Reference:

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

https://www.census.gov/