HOUSING AFFORDABILITY IN THE SOUTHERN UNITED STATES

Okey Chidume, Zheran Wang

Auburn University, Department of Mathematics & Statistics,

Auburn University, AL 36849

Spring 2017 (Final Project) - April 18, 2017

# Table of Contents

[**Table of Contents**](#_1fob9te) **2**

[**Introduction**](#_mink1qier1mk) **3**

[Background](#_4xybildr9xcs) 3

[Purpose of the Study](#_5tuxezdi7ozb) 3

[**Research Questions**](#_6tjpktmakd2b) **4**

[**Methodology**](#_vrpr9ftc97ra) **5**

[Models](#_poucvpxavq3j) 5

[**Results**](#_km5e5q77z85r) **10**

[**Conclusion**](#_hn7mpnuillr8) **12**

[**Limitations**](#_l8g2bkp4jm8a) **13**

[**Reference**](#_hk075w11vkis) **13**

[**SAS Code**](#_gg1egbevgkdy) **13**

# Introduction

Housing is a fundamental necessity of life. Housing ownership is measured by whether or not a person or family earns enough income to qualify for a mortgage loan on a typical home; this mortgage is determined based on several factors, which include income and housing costs.

The U.S. Census Bureau reported in September 2014 that: U.S. real median household income was $51,939 in 2013 versus $51,759 in 2012, statistically unchanged. In 2013, real median household income was 8.0 percent lower than in 2007, the year before the latest recession.

The attempt to measure homeownership affordability for families and unrelated individuals (current owners and current renters) is the motivation for this project.

Data in this project reference 1985 – 2009 and the metropolitan files 2002 – 2009 of the American Housing Survey (AHS) national files. The represented population (population universe) is the civilian noninstitutionalized population living in the Southern United States.

## Background

This study will help avail investigators with the opportunity to use a consistent set of affordability measures. These findings may be adopted or adapted to alleviate housing cost burden for owner and renter households, which in turn, empowers communities.

## Purpose of the Study

The purpose of this study is to find out which factors are significant in determining the affordability of housing units, as well as the relationship/interaction between the aforementioned factors. We want to discover what effect, if any, the factors will the price of the house in southern US.

For this study, data from **The Housing Affordability Data System (HADS)** will be studied to figure out which model will be most suitable for our analysis. The data was randomly collected and includes reports of income, location, condition, age, race, sex, family type, number of rooms, year constructed, price, and insurance, among others.

We will only be using only the following in this analysis: Value, Number of Rooms, Insurance,and Location. These variables are coming from a survey, where **Value** is our **Response Variable**; due to the nature of the data, we decided to choose **Location** as our **Blocking Factor.**

| Value | the price of the house |
| --- | --- |
| Year | year in which the unit was constructed |
| Nroom | number of room in the house |
| Insurance | Insurance, condo, land rent, other mobile home fees |
| Loca | closeness to the city |

Table 1 : List of Variables

# Research Questions

1. Which factors are significant in determining affordability of housing units?
2. Which interactions among the factors are significant?

# 

# Methodology

After collecting the data of 215 observations from different southern states, we used **Factorial Design Analysis** with 3 effects and 1 blocking factor.

## Models

The models considered in this analysis include:

**Model 1- Random Effects mode**l with 3 effects (**Year, NRoom, Insurance**) and 1 blocking factor (**Location**).

**Model 1**: yijkl = µ + τi+ βj + γk + δl + εijkl

For i = 1, …,a, j = 1,…,b, k = 1,…,c, l = 1,…,abc.

Where

τi is the effect of ith level of year factor.

βj is the effect of jth level of number of room factor.

γk is the effect of kth level of insurance factor.

δl is the effects of blocking factor.

Our hypothesis for this model include:

**Year** effect: H0: στ2 = 0 ; Ha: στ2 ≠ 0,

**Nroom** effect: H0: σβ2 = 0 ; Ha: σβ2 ≠ 0,

**Insurance** effect: H0: σγ2 = 0 ; Ha: σγ2 ≠ 0,

**Blocking** effect: H0: σδ2 = 0 ; Ha: σδ2 ≠ 0.

**Model 2 - Random Effects model** with 3 effects with their interactions and 1 blocking factor.

Model 2: yijkl = µ + τi+ βj + γk + δl + (τβ)ij + (τγ)ik + (βγ)jk+ (τβγ)ijk +εijkl

Where

Τβ is the interaction effect for Year\*Nroom (denote as YR in SAS)

Τγ is the interaction effect for Year\*Insurance (denote as YI in SAS)

βγ is the interaction effect for Nroom\*Insurance(denote as RI in SAS)

τβγ is the interaction effect for factor Year\*Nroom\*Insurance(denote as YIR in SAS).

We fit both models by using PROC GLM in SAS.

First we decided to look at the data using PROC SGSCATTER in SAS to draw a scatter plot to look at the data.

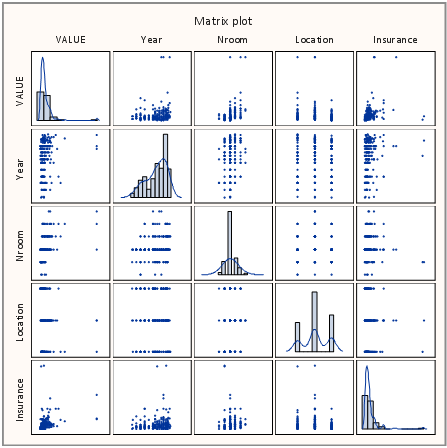


Figure 1: Matrix Plot

Upon looking at the data closely, we see that the **Value** and **Insurance** variables are skewed, but **Nroom** and location are showing normal distribution; therefore we decided to do a Log transformation of **Value** by using the statement log(value) in SAS.

We plotted the histogram of value after transformation to see if the variable is approximately normally distributed by using PROC SGPLOT in SAS.

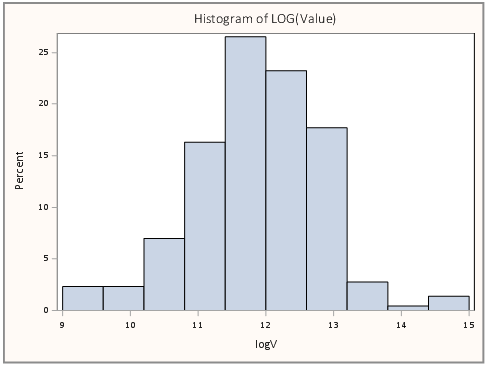


Figure 2: Histogram of Log(Value)

Then we tested the model assumptions which are the assumption of normality and the assumption of constant variance.

For test the assumption of normality, we using PROC UNIVARIATE to get the QQ-plot of the residual and the p-value of tests for normality.

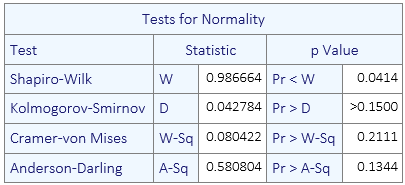


Table 2: Tests for Normality

The test for normality via the Kolmogorov-Smirnov shows that normality is not satisfied under the Kolmogorov-Smirnov test, we attribute this to the fact that the Kolmogorov-Smirnov has the most sensitivity out of the four.

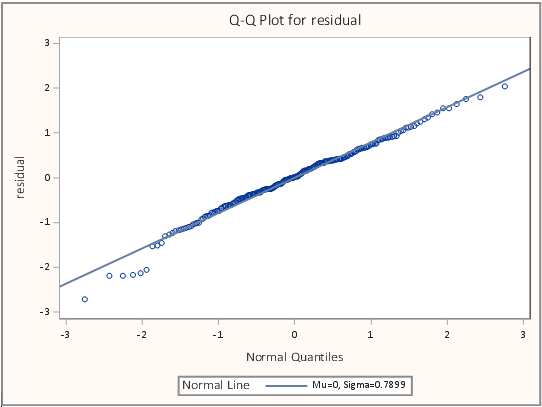


Figure 3: QQ-plot for Residual

Based on the results from the QQ-plot, the data looks about normal. And by looking at the table of tests for normality, the p-value for three out of four tests are showing significant. From both the results and tables, we can conclude that normality assumption satisfied.

For test the assumption of constant variance, we use PROC REG to plot the residual versus predicted value.

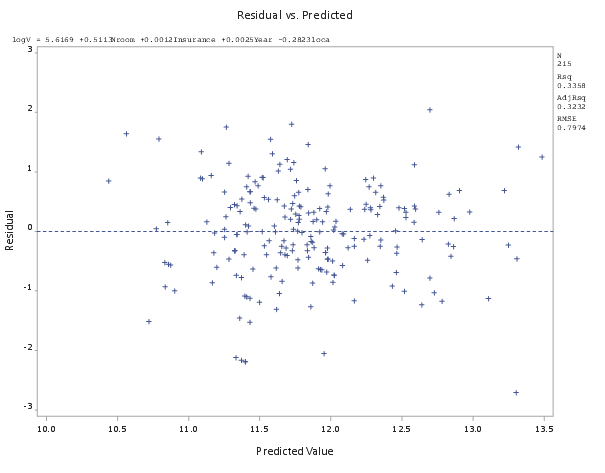


Figure 4: Assumption of Constant Variance

Based on the plot of the residual versus predicted value, we can see there is no clear pattern. Therefore the **constant variance assumption is satisfied**.

Normally, we would do the Levene's test to test the constant variance, but in our case, we cannot divide the response variable into different groups. Therefore, we only use the plot to show the assumption for constant variance.

# Results

**Model 1:** log(yijkl) = µ + τi+ βj + γk + δl + εijkl

The result of SAS output showing below.

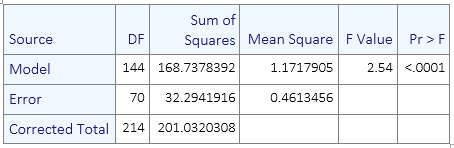


Table 3: Overall ANOVA Table For Model 1

For the overall ANOVA table, we are testing the hypothesis if all the effects are equal to zero. The **p-value shows significant, which we reject the null hypothesis**. We conclude there at least one effect not equal to zero.

The following table shows the **test of hypothesis that each individual effect is equal to zero**. The output shows below.

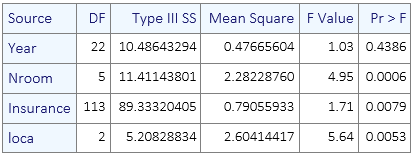


Table 4: Type III ANOVA Table For Model 1

The **p-value of the output shows number of room, insurance and location are significant.**

We fit the variables that are significant as our reduced model for model 1.

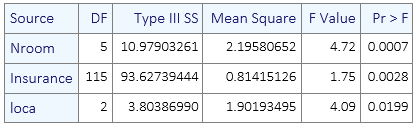


Table 5: Type III ANOVA Table For Reduced Model 1

The output shows all the variables of reduced model 1 are significant. Which means our new model is log(yjkl) = µ + βj + γk + δl + εjkl

**Model 2:**

log(yijkl) = µ + τi+ βj + γk + δl + (τβ)ij + (τγ)ik + (βγ)jk+ (τβγ)ijk +εijkl

The result of SAS output showing below.

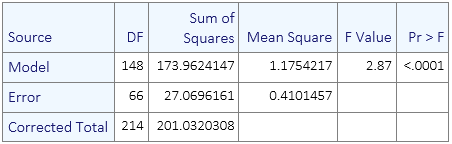


Table 6: Overall ANOVA Table For Model 2

Again, the overall ANOVA table shows significant on the p-value, which conclude that there at least one effect not equal to zero.

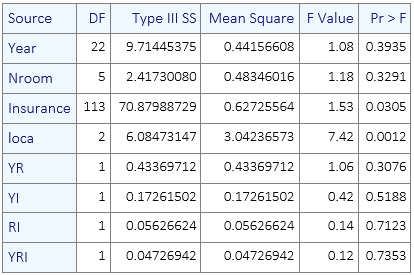


Table 7: Type III ANOVA Table For Model 2

For the table above, we conclude that with all the interactions, only **insurance** and **location** are **significant**.

We fit the variables that are significant as our reduced model for model 2.

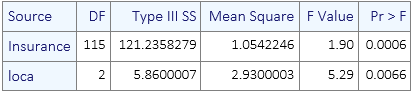


Table 8: Type III ANOVA Table For Reduced Model 2

The output shows all the variables of reduced model 1 are significant. Which means our new model is log(ykl) = µ + γk + δl + εkl

# Conclusion

After looking at the two reduced models

Reduced model 1: log(yjkl) = µ + βj + γk + δl + εjkl

Reduced model 2: log(ykl) = µ + γk + δl + εkl

We decide to use reduced model 1 as our model of choice. The is so because the reduced model 1 has more variables than the reduced model 2.

Moreover, due to the high insignificant interaction terms in the model 2, it will probably not be a good fit and make the **Nroom** variable end up being insignificant.

So, to answer the main research questions:

1. Which factors are significant in determining affordability of housing units?

**ANSWER: Number of Rooms, Insurance, Location**

1. Which interactions among the factors are significant?

**ANSWER: No interactions are significant**

# Limitations

Some limitations include:

* The data has more factors than the 4 we considered for analysis.
* Sample size is small, with an increase in sample size, the **Year** might end up being significant.

# Reference

*American Housing Survey: Housing Affordability Data System | HUD USER. N.p., n.d. Web. 18 Apr. 2017.*

# SAS Code

ods rtf file="I:\project\7030\SASOUTPUT.rtf" style=phil;

proc import out=house

datafile='I:\project\7010\house1.csv'

dbms=csv replace;

run;

/\*numbering the data\*/

data h;

set house;

logV=log(VALUE);

if location = 'Central City' then loca=1;

if location = 'Suburb' then loca=2;

if location = 'Nonmetro' then loca=3;

keep VALUE logV Nroom Insurance Year loca ;

label loca='Location';

if value='1' then delete;

run;

/\*matrix plot\*/

proc sgscatter data=h;

matrix value year nroom loca Insurance/

diagonal=(histogram kernel);

title'Matrix plot';

run;

/\*test of normaly for value\*/

proc glm noprint data=h;

model value=Nroom Insurance Year loca;

output out =diag1 R=residual P=pred;

run; quit;

Proc univariate data=diag1 normal;

var residual;

title'QQ-plot before transformation';

qqplot residual/normal(l=1 mu=est sigma=est);

ods select TestsForNormality QQPlot;

run; quit;

/\*histo for VALUE\*/

proc sgplot data =h;

title'Histogram of Value';

histogram value;

run;

/\*test of normal for logvalue\*/

proc glm noprint data=h;

model logV=Nroom Insurance Year loca;

output out =diag2 R=residual P=pred;

run; quit;

Proc univariate data=diag2 normal;

var residual;

title'QQ-plot after transformation';

qqplot residual/normal(l=1 mu=est sigma=est);

ods select TestsForNormality QQPlot;

run; quit;

/\*histo for LOGVALUE\*/

proc sgplot data=h;

histogram logV;

title'Histogram of LOG(Value)';

run;

/\*graph of constance varance\*/

proc reg data= h;

model logV=Nroom Insurance Year loca;

plot r.\*p.;

output out=diag3 r=residual p=pred;

title 'Residual vs. Predicted';

run; quit;

/\*regular model\*/

ods trace on;

proc glm data=h;

class loca nroom insurance year;

model logV= Year Nroom Insurance loca ;

random nroom insurance year;

run; quit;

/\*fit the regular again\*/

proc glm data=h;

class loca nroom insurance;

model logV= Nroom Insurance loca ;

random nroom insurance;

run; quit;

/\*create interaction\*/

data h1;

set h;

YR=year\*Nroom;

YI=year\*insurance;

RI=nroom\*insurance;

YRI=year\*nroom\*insurance;

run;

/\*model with interaction\*/

proc glm data=h1;

class loca nroom insurance year;

model logV=Year Nroom Insurance loca YR YI RI YRI ;

random nroom insurance year;

run; quit;

/\*fit the interaction model again\*/

proc glm data=h1;

class loca nroom insurance year;

model logV=Year Nroom Insurance loca RI;

random nroom insurance year;

run; quit;

/\*only insurance & location are significant\*/

/\*remove year\*/

proc glm data=h1;

class loca insurance;

model logV=Insurance loca;

random insurance ;

run; quit;

ods rtf close;