Report
On

Sales Price of Homes:

Which Attributes Are Most Significant and Contributes the most to the Sale Price of a Home?

By
Thomas Heinz
Romnikh Ortega
Brigham Young University
20 April 2020

Table of Contents

Section 1: Executive Summary

This project aims to use statistical models in order to appraise homes in the Ames, Iowa region. The data set we used for this analysis consists of 517 observations of homes in Ames Iowa. The data fields for each observation includes the home’s sale Price (measured in dollars),a transformed longitude and latitude points, ground living area (measured in square feet), style of dwelling, remodel date, number of full baths, number of half baths, number of bedrooms above the ground, garage car capacity and if the home includes a central air system. We found that size of the ground living area, a two story house, the year remodeled, whether the house had central air, the number of bedrooms above ground, and the number of cars that could fit in the garage to be the most statistically significant attributes in determining the cost of a home.

Introduction and Problem Background

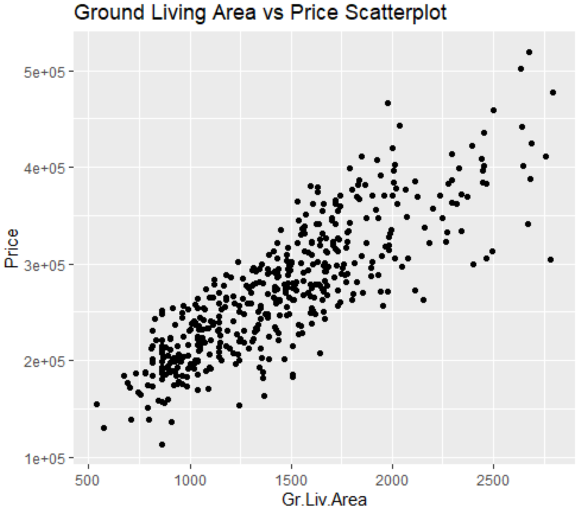
If someone decides to buy a new home and take out a mortgage, they will need to get an appraisal of the value of the home to be used as collateral against your loan. A home appraisal is an unbiased determination of the fair market value of the by a professionally trained third party appraiser.

Our analysis attempts to determine how well the given home characteristics explain sales price, which factors increase the sale price of the home, see if there is a correlation between the variability of sale price and the size of the home, and predict the sales price of homes given the characteristics given.

Data Description

The data set we used for this analysis consists of 517 observations of homes in Ames Iowa. The data fields for each observation includes the home’s sale Price (measured in dollars),a transformed longitude and latitude points, ground living area (measured in square feet), style of dwelling, remodel date, number of full baths, number of half baths, number of bedrooms above the ground, garage car capacity and if the home includes a central air system.

The boxplots below reveal relatively few outliers than what is expected from a home price data set. Also note that scatter plots below indicate that both the ground living area and remodel date covariates have a linear relationship with appraisal price. These observations indicate that a log-transform of the response variable would not be necessary for this analysis.



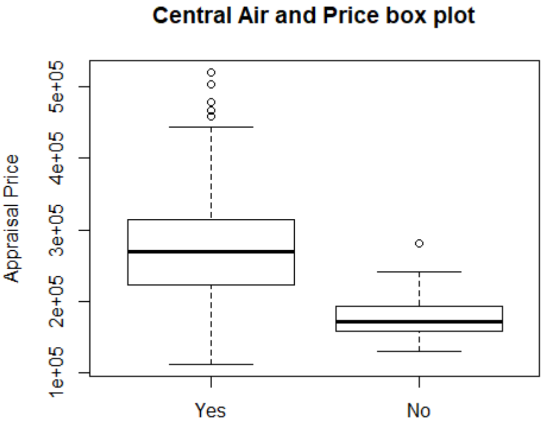
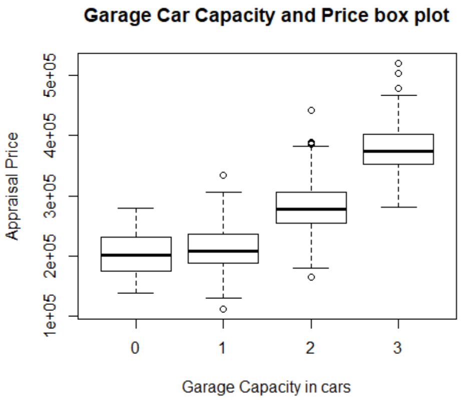


Figure 1: Note the linear relationship between price and ground living area.

Figure 2: Note that homes with central air seem to be more expensive.



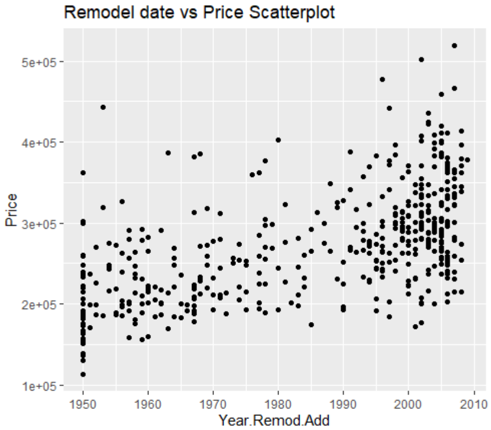
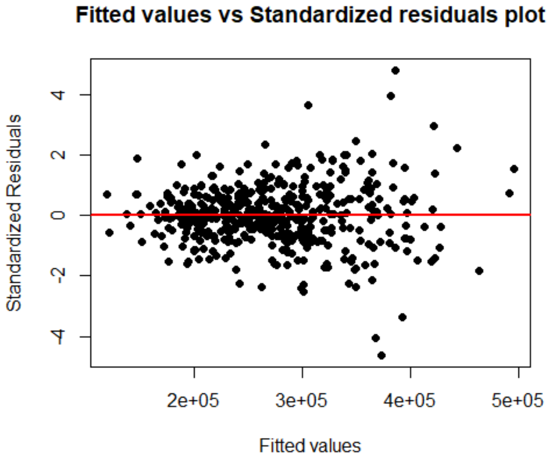


Figure 3: Bigger garages translates to higher appraisal price

Figure 4: Note the approximate linear relationship between Remodel date and appraisal price.

Issues with the Data and its Consequences

Based on the figures representing the variogram of the residuals and the fitted value versus standardized residuals plot below, it’s clear that the data has issues regarding heteroskedasticity and spatial correlation. If these issues are ignored, the estimates for the standard error will be inaccurate. This will lead to erroneous statistical inference conclusions since inaccurate standard errors will lead to unreliable confidence intervals, t-tests and prediction intervals.



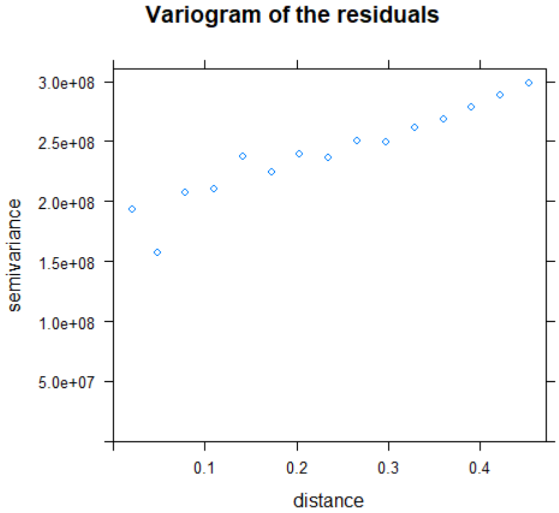


Figure 5: Note the "fan" shape of this plot. This indicates a Heteroskedastic problem.

Figure 6: Note that the variogram slants upwards. This implies the existence of spatial correlation in the data.

Section 2: Statistical Model

Model Definition

$y=X\boldsymbol{\beta}+\varepsilon$
$\mathcal{E}\sim\mathcal{N}(0,\sigma^2/R)$

Parameter Definition

Y - This represents the sale price of each home.
X - This represents the design matrix which includes linear terms for above ground living area, house style, remodel date, central air number of full bathrooms, number of half bathrooms, number of above bedrooms above ground and size of the garage in car capacity.
$\boldsymbol{\beta}$ - This is the coefficients of our model for the above mentioned terms.
R - This is the correlation matrix where the ijth element is given by (1-w) exp{-d_{ij}/ø} where d_{ij} is the distance between location i and location j and ø is the range parameter controlling how strong the correlation is between locations and w is the nugget parameter controlling the correlation at a single location.

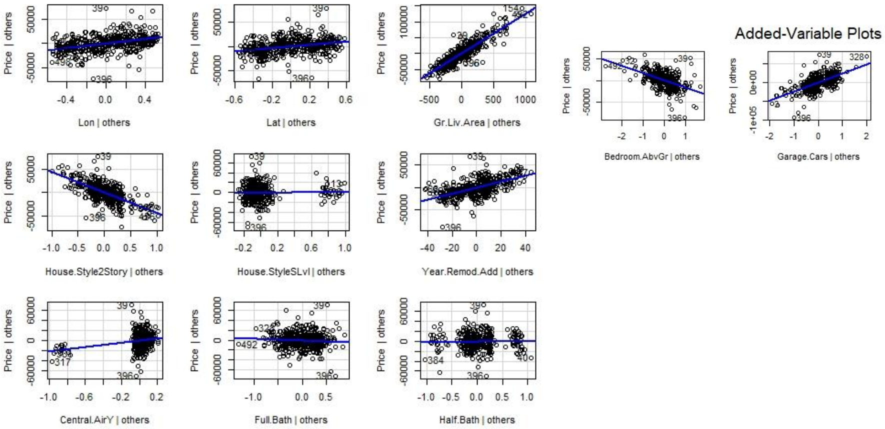
Model Assumptions

Our model assumes linearity in the slope parameters of the explanatory variables. We also assume independence between each sale price of each home, and independence for PM exposure within each child for different time periods. Finally, we assume equal variance and normality in the decorrelated residuals.

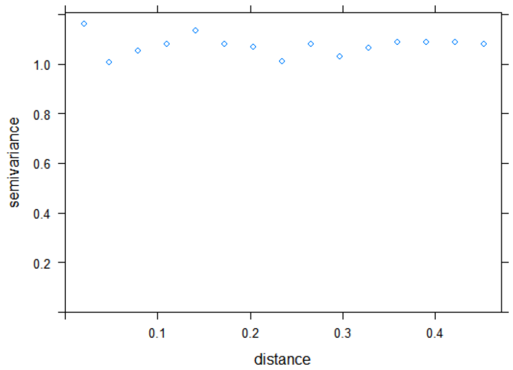
Section 3: Model Validation

The four assumptions for any linear model are the following:

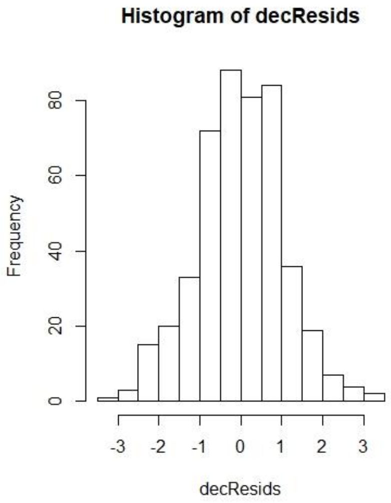
Linearity - The relationship between our explanatory variables and response variable. This assumption was checked using Added-Variable Plots. Though these plots, we can see that the effects of all of our quantitative explanatory variables are linear.



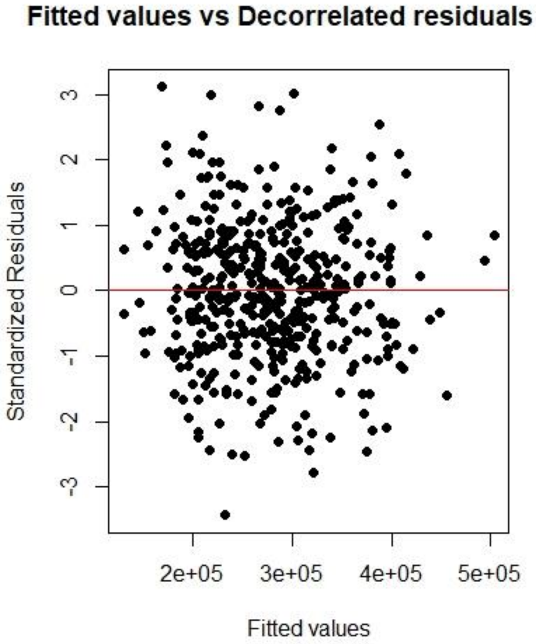
Independence - Note that the variogram plot of the standardized residuals below is roughly linear. This implies that there is no spatial correlation in the plot which means the independence assumption has been met.



Normality- The standard errors are normally distributed with a mean of zero. To show this we ran a K.S test and had a p-value of .9714 determined to not reject the null hypothesis. We have also included a histogram of the decorrelated residuals.



Equal Variance - Errors are equally distributed about a fitted value. To show this assumption is met we have included a plot of the decorrelated residuals against the fitted values of our model.



We also ran a cross validation 100 times, with 45 observations in each test. We found the mean of our width to be 50806.5 and our coverage to be .9651. This means that 96% of our model can be found within about $51,000 of the actual price.

We also had a mean bias of $325.14 and mean RPMSE of $13,159.7. Considering that a typical house in this data set cost $271,937, the model did a good job in predicting house prices since its prediction was only off target by about $13,000 on average.

Section 4: Analysis/Results

To answer how well the home characteristics in our dataset explain sale price, we fit a general least squares model with weights on the ground living area, year remodeled, the size of the garage and the numbers of bedrooms to account for heteroskedasticity. We also added a parameter to account for spatial correlation. The R^2 value of this model is .9312, which means that our model accounts for 93.12% of the variance of the sale price observations.

To find out which factors are most significant in determining the sale price of a home, we took a summary of our model and looked at the p-values. The size of the ground living area, a two story house, the year remodeled, whether the house had central air, the number of bedrooms above ground and the number of cars that could fit in the garage all had a p-value of zero to four decimal places. All of these attributes are deemed to be significant in determining the sales price. Below is a table of the estimates of these parameters and the 95% confidence intervals.

Attribute	Lower Est	Est	Upper Est	P-value
Gr. Living Area	116.9029	124.9000	132.8890	0.0000
2 Story Home	-46492.7160	-43115.3000	-39737.8987	0.0000
Year Remodeled	662.5105	714.8000	767.1474	0.0000
Central Air (Y)	17236.5952	21346.6000	25456.6689	0.0000
Garage Size	21133.5858	22903.1000	24672.6829	0.0000
Num Bed Above Ground	-17185.7434	-15413.8000	-13641.7583	0.0000

We interpret these intervals by the following:

We are 95% confident that for every 1 square foot increase in the size of a home, we expect the sales price of a home to go up between $116.90 and $132.89.

We are 95% confident that a two-story home will decrease the price of a home between $46,492.72 and $39,737.90

We are 95% confident that for every year that the home remodel occured closer to the current year, that the price of a home goes up between $662.51 and $767.15.

We are 95% confident that houses with central air increase the price of the home between $17,236.60 and $25,456.67.

We are 95% confident that as the size of a garage increases by 1 car, the value of a home increases between $21133.59 and $24,672.68.

We are 95% confident that as the number of bedrooms above ground increases by 1, the value of the house will decrease between $17,185.74 and $13,641.76.

To discover if the variability if the sale price of a home increases with the size of the home as given by living area measured by square footage, we looked at the theta coefficient of the variance function. This was estimated to be 0.00072 with a 95% confidence interval of (0.0007,0.0008). Since this is positive,the variability of a home's sale price goes up as the size of the house increases. This means that larger houses vary more in price based on other factors that are not included in our data set.

Lastly, below is a table of the first six predicted values based on the variables given in the data set .

Price	Gr.Liv.Area	House Style	Year Remodeled	Central Air	Num Full Baths	Num Half Baths	Num Beds Above Ground	Garage Car Size
227384.10	1144	1 Story	1960	Y	1	0	3	1
411242.60	2855	2 Story	2000	Y	2	1	4	3
210143.50	1114	1 Story	2004	Y	1	1	3	0
26540800	1576	S level	1961	Y	1	0	4	2
272788.80	1478	1 Story	1992	Y	2	0	3	2
332179.70	1483	1 Story	2001	Y	1	1	1	2

Section 5: Conclusions

In our analysis, we were able to fit a model that accounts for 93% of the variance in the sales price.

We also found that the following characteristics were deemed the most statistically significant for finding determining the price of a home:

- size of the ground living area, a two-story house,
- the year remodeled,
- whether the house had central air,
- the number of bedrooms above ground,
- the number of cars that could fit in the garage

We also concluded that the size of the home and the sales price have increasing variability.
In order to more accurately appraise homes, (particularly bigger homes) it would be beneficial to include extra amenities such as a pool or sports court, or distance to nearby school or parks.

Code

```
###############################
##Reads in Required Libraries##
################################
library(multcomp)
library(ggplot2)
library(lmtest)
library(gstat)
library(car)
library(MASS)
library(nlme)
library(tidyverse)
source("DrHeaton'sStResGLS.R")
source("PredictGlsFunction.R")
##################
#Reads in Data####
##################
housingData <- read.csv(file = "C:/Users/heinz/Desktop/School/Stat
469/Homework/HousingPrices.csv", header = TRUE,
            dec = ".")

#Turns Categorical Variables into Factors
housingData$House.Style <- as.factor(housingData$House.Style)
housingData$Central.Air <- as.factor(housingData$Central.Air)

#Separated data into one containing Na and one with no NA
housingNaData <- housingData[rowSums(is.na(housingData)) > 0,]
noNaData <- na.omit(housingData)

#######
#EDA###
#######
#Central Air Boxplot
unique(noNaData$Central.Air)
YCentral <- noNaData[noNaData$Central.Air== "Y",]
NCentral <- noNaData[noNaData$Central.Air == "N",]
airNames <- c("Yes","No")
boxplot(YCentral$Price,NCentral$Price,ylab = "Appraisal Price",
      main = "Central Air and Price box plot", names = airNames)

#Scatterplot of Above ground lliving area and Price
ggplot(data= noNaData,mapping=aes(x= Gr.Liv.Area, y= Price)) + geom_point()
summary(noNaData$Price)
```

```
##Garage Car Capacity and Price
#Note that homes with grages that ca fit 4 cars were excluded since there was only one home
that
#fit this category.
unique(noNaData$Garage.Cars)
garage0 <- noNaData[noNaData$Garage.Cars== 0,]
garage1 <- noNaData[noNaData$Garage.Cars== 1,]
garage2 <- noNaData[noNaData$Garage.Cars== 2,]
garage3 <- noNaData[noNaData$Garage.Cars== 3,]
garage4 <- noNaData[noNaData$Garage.Cars== 4,]
carNames <- c("0","1","2","3")
boxplot(garage0$Price,garage1$Price,garage2$Price,garage3$Price,ylab = "Appraisal Price",
        main = "Garage Car Capacity and Price box plot", names = carNames)
#########################
#Fits appropriate model#
#########################
#Fits MLR model
housinglm <- lm(formula = Price ~ ., data = noNaData)
summary(housinglm)

###############################
##Checking For equal Variance##
###############################
#Using the BP test, I obtained a p-value of less than 2.2e-16. This means that we reject the Null
hypothesis
#which means that the assumption of equal variance is not met.
yPred <- fitted(housinglm)
plot(yPred,stdres(housinglm), main= "Fitted values vs Standardized residuals plot",
     xlab = "Fitted values", ylab = "Standardized Residuals", pch = 19)
abline(h = 0, col = 'red', lwd = 2)
bptest(housinglm)

#############################
#Checking for correlation####
#############################
#Variogram:  Variogram is non-linear. This suggests that there is spatial Correlation
myVariogram = variogram(object = Price ~ Gr.Liv.Area + House.Style + Year.Remod.Add +
               Central.Air + Full.Bath + Half.Bath + Bedroom.AbvGr + Garage.Cars,
               locations = ~Lon+Lat, data = noNaData)
plot(myVariogram,main = "Variogram of the residuals")

##############################
##Model Fitting###############
##############################
```

```
#housingExp = gls(model = Price ~ Gr.Liv.Area + House.Style + Year.Remod.Add +
#                Central.Air + Full.Bath + Half.Bath + Bedroom.AbvGr + Garage.Cars ,
#                data = noNaData,weights = varExp(form = ~ Gr.Liv.Area),
#                correlation = corExp(c(1.6e+15,1e-10),form = ~ Lon + Lat, nugget = TRUE)
,method = "ML")
#Gr.Liv.Area + Year.Remod.Add + Full.Bath + Half.Bath + Bedroom.AbvGr + Garage.Cars,
housingExp <- gls(model = Price ~ . -Lon - Lat,
              data = noNaData,weights = varExp(form = ~ Gr.Liv.Area + Year.Remod.Add +
Garage.Cars + Bedroom.AbvGr)
              ,correlation = corExp(form = ~ Lon + Lat, nugget = TRUE) ,method = "ML")
housingSph <- gls(model = Price ~ . -Lon -Lat ,
               data = noNaData,weights = varExp(form = ~ Gr.Liv.Area + Year.Remod.Add +
Garage.Cars + Bedroom.AbvGr),
               correlation = corSpher(form = ~Lon+Lat, nugget=TRUE) ,  method = "ML")
housingGauss <- gls(model = Price ~ -Lon -Lat,
               data = noNaData,weights = varExp(form = ~ Gr.Liv.Area + Year.Remod.Add +
Garage.Cars + Bedroom.AbvGr),
               correlation = corGaus(form = ~Lon+Lat, nugget=TRUE) ,  method = "ML")

#glsControl(maxIter = 500, msMaxIter=2000, returnObject = TRUE)

############################
##Picking the best Model###
############################
# Exponential is the best model
AIC(housingExp)
AIC(housingSph)
AIC(housingGauss)

#############################
#Model Validation###########
#############################
#Linearity Assumption
avPlots(housinglm)

#Independece Assumption
decResids <- stdres.gls(housingExp)
residDF <- data.frame(Lon = noNaData$Lon,Lat = noNaData$Lat, decorrResid = decResids)
residVariogram <- variogram(object = decorrResid ~1,locations = ~Lon+Lat, data = residDF)
plot(residVariogram)

#Normality Assumption
hist(decResids)
ks.test(decResids,"pnorm")
```

```
#Equal Variance
decResids <- stdres.gls(housingExp)
fittedVals <- fitted(housingSph)
plot(fittedVals,decResids, main= "Fitted values vs Decorrelated residuals plot",
     xlab = "Fitted values", ylab = "Standardized Residuals", pch=19)
abline(h = 0, col = 'red')
##########################
##Cross-validation########
#########################
#Cross validation
nCV <- 50
nTest <- 45  #Number of observations in a test set
rpmse <- rep(x = NA, times = nCV)
bias <- rep(x = NA, times = nCV)
wid <- rep(x = NA, times = nCV)
cvg <- rep(x = NA, times = nCV)
pb <- txtProgressBar(min = 0, max = nCV, style = 3)
for(i in 1:nCV) {
  #Selects tests observations
  testObs <- sample(x = 1:nrow(noNaData), size = nTest)
  #Split into training and test data sets
  testData <- noNaData[testObs,]
  trainData <- noNaData[-testObs,]
  #Fits a linear model using the training data
  glsTrain <- gls(model = Price ~ . -Lon - Lat,
                data = trainData,weights = varExp(form = ~ Gr.Liv.Area),
                correlation = corExp(form = ~ Lon + Lat, nugget = TRUE) ,method = "ML")
  #Generates prediction for the test set
  yPredgls <- predictgls(glsobj = glsTrain, newdframe = testData, level = 0.95)
  #calculates bias
  bias[i] <- mean(yPredgls[,'Prediction'] - testData[['Price']])
  #Calculates RPMSE
  rpmse[i] <- (testData[['Price']] - yPredgls[,'Prediction'])^2 %>% mean() %>% sqrt()
  #Calculates coverage
  cvg[i] <- ((testData[['Price']] > yPredgls[,'lwr']) & (testData[['Price']] < yPredgls[,'upr'])) %>%
mean()
  #Calculates Width
  wid[i] <- (yPredgls[,'upr'] - yPredgls[,'lwr']) %>% mean()
  ## Update the progress bar
  setTxtProgressBar(pb, i)
}
close(pb)
#Diagnostic results
```

```
#Bias of prediction intervals:
mean(bias)
#RPMSE of prediction intervals:
mean(rpmse)
#Coverage of prediction intervals:
hist(cvg)
mean(cvg)
#Width of prediction intervals: gls mdodel had a narrower confidence interval at 1.92 as
opposed to
#the 5.39 for the lm model
mean(wid)

#########################
#Research Questions 1#####
#########################
#Rsquared
SSE <- (fitted(housingExp)-noNaData$Price)^2 %>% sum()
SST <- (noNaData$Price - mean(noNaData$Price))^2 %>% sum()
1 - (SSE/SST)
#Pseudo-Rsquared
cor(fitted(housingExp),noNaData$Price)^2
#########################
#Research Questions 2#####
#########################
summary(housingExp)
confint(housingExp)
#########################
#Research Questions 3#####
#########################
intervals(housingExp, level = 0.95)[3]
coef(housingExp$modelStruct, unconstrained = FALSE)
#########################
#Research Questions 4#####
#########################
housingNaData <- housingData[rowSums(is.na(housingData)) > 0,]
yPredicted <- predictgls(glsobj = housingExp, newdframe = housingNaData,level = 0.95)
housingNaData$Price <- yPredicted$Prediction
head(housingNaData)
```