University of California, Davis

Take Home Project 1

Question 1: Price & Square Feet

Ana Boeriu & Julia Tien
STA 108
Dr. Erin Melcon
October 27, 2017

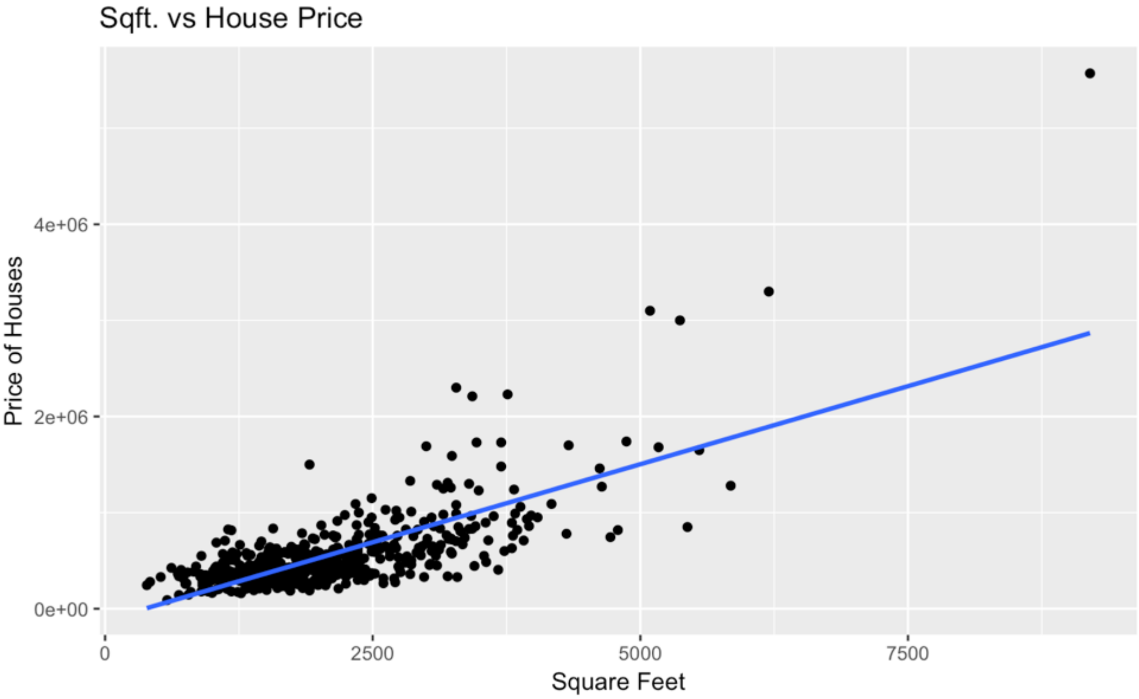
I. INTRODUCTION

The goal of this project is to build a model that can predict sales price of a house based only on its square footage. We will then use the model to calculate confidence interval and predictions that give us insight to what the true population parameter could be. House prices have been rapidly rising around the Bay Area and since we are both residents of the Bay, we were interested in this topic and wanted to use our statistical knowledge to investigate the relationship between house price and square footage. This may also be useful to real estate agents and for those who study and report on the housing market so they can make their own predictions. The approach we are taking is using simple linear regression because we believe that there is a significant linear relationship between square feet and the price of a house since we know larger houses cost more. We are regressing the response variable (Y) housing price against the explanatory variable (X) living square footage.

II. SUMMARY

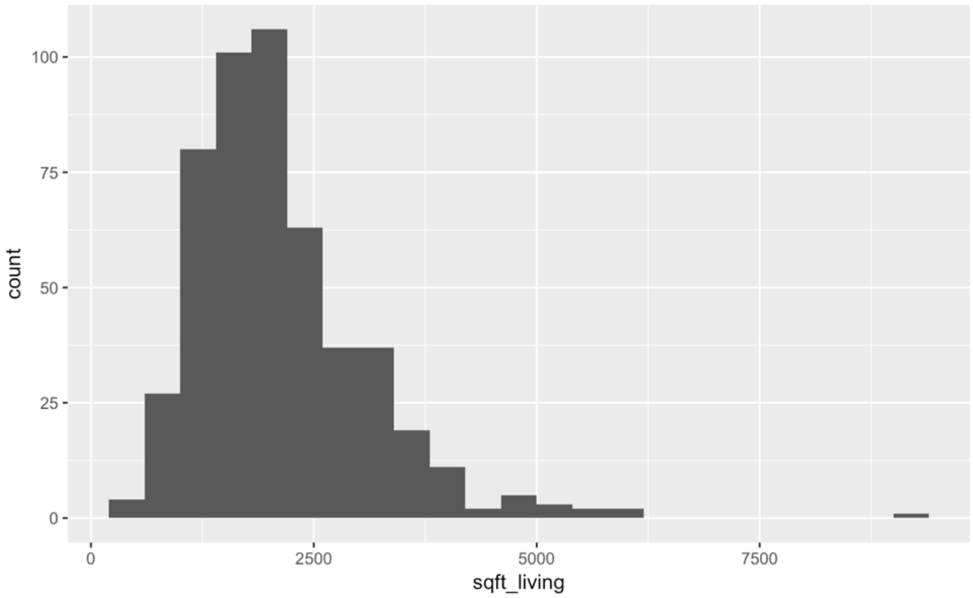
This section provides simple analyses to understand the data better. There is a positive relationship between price of houses and square feet overall.

A. Scatterplot:



As squared feet of a house increases, the price also increases.

B. Histogram:



Most of the homes in the data set have approximately 1,000-2,500 ft^2 and this data set is positively skewed so the mean is larger than the median.

C. Summary Statistics

The minimum house price sold in King County that year was $90,000 and the maximum was $5,570,000. The first quartile is at $333,722 which means that 25% of the house prices in this data set are less than $333,722. The middle point or median of the frequency distribution is $460,000 and the third quartile is at $646,250, which means that 75% of the house prices in this set is less than that amount. The average house price sold was $566,594 and the standard deviation was $434,558.60.

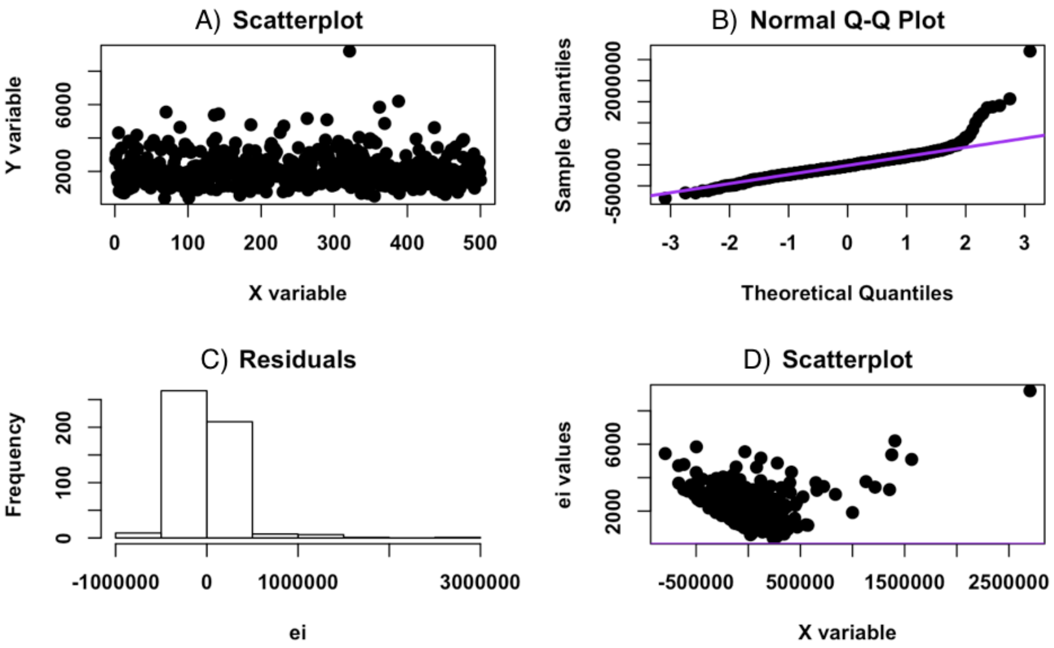
The minimum square footage in this data set was 390 ft^2 while the maximum was 9200 ft^2. The first quartile is at 1460 ft^2, the median is 1940 ft^2, and the third quartile is at 2562 ft^2. The average square footage of a house sold in the particular year in King County was 2115 ft^2 and the standard deviation was 976.9711 ft^2.

III. DIAGNOSTICS

In this section, we are testing the validity of our data to ensure it is acceptable to use linear regression. Ideally our X and Y pairs are independent, and our errors have constant variance and are normally distributed.

A. Diagnostic Plots

These plots will be further analyzed in parts B and C.

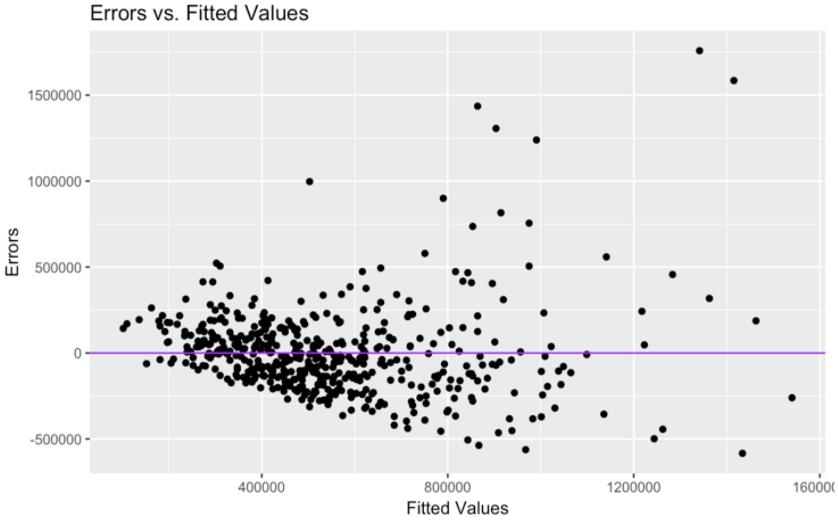


B. Testing for Constant Variance

We will begin by testing our data for homoscedasticity, which is constant variance between observations. This is important as all data points have equal weight in the regression and we do not want an outlier to skew the regression line as that would be an inaccurate representation of the actual relationship between X and Y.

1. Plot D: Plot of Errors Vs. Fitted Values

A closer look:



This plot of e_i versus x_i clearly violates homoscedasticity, since variance increases as the fitted values increase. This data has unequal variance so it is heteroscedastic.

2. Fligner-Killeen Test

This tests for normality as “X” is separated into larger and smaller parts, and then the variance of the errors are compared between both groups.

$Ho: Residuals\ have\ equal\ variance$ $Ha: Residuals\ do\ not\ have\ equal\ variance$

Fligner-Killeen:med	chi-squared = 26.109	df = 1	p-value = 3.227e-07

Since our p-value is relatively small, we reject the null, and there is sufficient evidence to support that our data does not have equal variance at any reasonable significance level (1%, 5%, 10%).

C. Testing for Normality

Normally distributed errors are important because that shows the data is random.

1. Plot B: QQ Plot

Because towards the 2nd and 3rd theoretical quantile the sample quantile increases away from the 45° line, this means the data is not approximately normal.

2. Plot C: Histogram of Errors

It is clear from this histogram that the errors in this data are not normally distributed and it is positively skewed.

3. Shapiro-Wilk Normality Test

This is a modified correlation test on points from the QQ plot. A test statistic is formed from the correlation between theoretical and observed statistics. By failing to reject the null, there is evidence that errors are normal.

$Ho: Data\ is\ normal$ $Ha: Data\ is\ non-normal$

W = 0.82829	p-value < 2.2e-16

Since our p-value is relatively small, we reject the null, and there is sufficient evidence to support that our data is not normally distributed at any reasonable significance level (1%, 5%, 10%). However, because this particular dataset had a larger population of 500 observations, there is reason to believe that the test is overly conservative.

D. Overview

Although these diagnostics have shown that the errors in our data do not have constant and are not normally distributed, for the purposes of this report we will still be using simple linear regression. It is also important to note that we have removed five outliers (that were more than 4 standard deviations above the null value of 0), so the regression line can more accurately estimate price.

IV. ANALYSIS & INTERPRETATION

Before we use the model to predict anything, this section will help show that there is a significant relationship between X and Y. We calculate different values and draw conclusions to further analyze our data set.

A. Model Fit

$$\hat{Y} = 55119.02 + 2324X$$

When square footage increases by 1, house price increases by $232.4

B. Test Statistic and P-Value

Estimate	Std.Error	t-value	Pr(>|t|)
232.15	11.45	20.282	5.944613e-67

1. Test Statistic

The test statistics tells us the number of standard deviations b_1 is from the null value of zero. Thus, we can conclude the sample value of 20.282 is 20.282 standard deviations from the null hypothesis value of zero.

2. P-Value

The p-value helps us determine the significance of our results with respect to the null hypothesis. If in reality there was no linear relationship between square feet of a house and the price of a house, then the probability of observing our sample data or more extreme is 5.944613e-67.

We are testing for a linear relationship:

$Ho: \beta 1 = 0$ $Ha: \beta 1 \neq 0$

Since our p-value is relatively small, we reject the null, and there is sufficient evidence to support that there is a significant relationship at any alpha level (1%, 5%, 10%).

C. Confidence Interval

Confidence intervals give us plausible ranges of what the true value could be.

	2.5%	97.5%
Square Ft	209.6632	254.6421

We are 95% confident that when square footage of a house in King County increases by 1 square foot, price changes by between $209 and $254.6421 on average.

V. PREDICTION RESULTS

In this section we are using our regression model to predict various house prices and create ranges for plausible values the true house prices could be given an amount of square feet.

A. Average house price for houses with living square footage 2800

Estimated Y	Lower Bound	Upper Bound
705,146.4	679,449.7	730,843.2

The average house price for a house with 2800 ft^2 is estimated to be between $679,449.7and $730,843.2 on average, and the estimate is $705,146.4.

B. Price of a particular house with living square footage 3200

Estimated Y	Lower Bound	Upper Bound
798,007.5	353,908.2	1,242,107

The price of a house with 3200 ft^2 is estimated to be $798,007.5.

C. Average house price for houses with living square footage 8,000

Estimated Y	Lower Bound	Upper Bound
1,912,340	1,449,385	2,375,295

The average house price for a house with 8000 ft^2 is estimated to be between $1,449,385 and $2,375,295on average, and the estimate is $1,912,340.

VI. CONCLUSION

In this project we analyzed our data and used statistical measures and plots such as scatterplots and p-values to confirm a positive linear relationship between square feet of a house and the price it sells for. We performed numerous tests to diagnose the validity of our data by testing for constant variance and normality. Even though this data does not meet the assumptions for the linear regression model, for the purposes of project we still used it to find the estimated regression line that would characterize the relationship between square feet and house price. We then used this model to estimate various house prices and intervals the true house price could be in.

R APPENDIX

II. Summary

A. Scatterplot

```
library(ggplot2)
ggplot(KingCounty,aes(x=sqft_living, y=price))+
  geom_point()+
  geom_smooth(method="lm", se=FALSE)+
  ggtitle("Sqft. vs House Price")+
  ylab("Price of Houses")+
  xlab("Square Feet")+
  theme_bw()
```

B. Histogram

```
ggplot(KingCounty, aes(x = sqft_living)) +
  geom_histogram(binwidth = 400)+
  theme_bw()
```

C. Summary Statistics

```
 summary(KingCounty)
sd(KingCounty$price)
sd(KingCounty$sqft_living)
```

III. Diagnostics

A. Multiple plots in a window

```
options(scipen = 7)
king_reg_line = lm(price~sqft_living, data=KingCounty)
par(mfrow=  c(2,2)) #Makes a 2 x 2 grid which plots will fill in to
qqnorm(king_reg_line$residuals)
qqline(king_reg_line$residuals)
plot(KingCounty$sqft_living, KingCounty$price, main="Scatterplot", pch = 19,font = 2,
     font.lab = 2, ylab = "Y variable",xlab = "X variable")
hist(king_reg_line$residuals,main = "Residuals", xlab = "ei",
     pch = 19,font = 2,font.lab = 2)
plot(king_reg_line$residuals, KingCounty$sqft_living, main = "Scatterplot",
     pch = 19,font = 2,font.lab = 2, ylab = "ei values",xlab = "X variable")
abline(h = 0, lwd = 2, col = "purple")
```

B. Testing for constant variance

Error vs fitted values

```
plot(king_reg_line$fitted.values, king_reg_line$residuals, xlab = "Fitted Values",
     ylab="Errors", main = "Errors vs Fitted Values")
abline(h = 0,col = "purple")
```

FK test

```
options(scipen = 1)
Group = rep("Lower",nrow(KingCounty)) #Creates a vector that repeats "Lower" n times
Group[KingCounty$price > median(KingCounty$price)] = "Upper" #Changing the appropriate values to "Upper"
Group = as.factor(Group) #Changes it to a factor, which R recognizes as a grouping variable.
cats$Group = Group
the.FKtest= fligner.test(KingCounty$ei, KingCounty$Group)
the.FKtest
```

Testing for Normality

Shapiro Wilks Test

```
options(scipen=1)
shapiro.test(KingCounty$ei)
```

Outliers

```
ei.s = king_reg_line$residuals/sqrt(sum(king_reg_line$residuals^2)/(nrow(KingCounty)-length(king_reg_line$coefficients)))
ri = rstandard(king_reg_line)
ti = rstudent(king_reg_line)
alpha = 0.01
n = nrow(KingCounty)
p = length(king_reg_line$coefficients)
cutoff = qt(1-alpha/(2*n), n -p )
cutoff_deleted = qt(1-alpha/(2*n), n -p -1 )
outliers = which(abs(ei.s)> cutoff | abs(ri) > cutoff | abs(ti) > cutoff_deleted)
outliers
newKingCounty = KingCounty[-outliers,]
```

IV. Analysis and Interpretation

A. Estimated regression line

```
no_outlers_model = lm(price~sqft_living, data=newKingCounty)
```

B. Hypothesis Testing

```
all.sum = summary(no_outlers_model)
HT = all.sum$coefficients
HT.b1 = HT[2,]
HT.b1
```

C. CI

```
confint(no_outlers_model, level=0.95)
```

VI. Prediction results

A. Predict the average house price for houses with living square footage 2800.

```
xs2 = data.frame(sqft_living = 2800)
pred.int = predict(no_outlers_model,xs2,interval = "confidence",se.fit = TRUE,level = 0.95)
```

```
CI = pred.int$fit
CI
```

B. Predict the price of a particular house with living square footage 3200.

```
xs3 = data.frame(sqft_living = 3200)
pred.int3 = predict(no_outlers_model,xs3,interval = "prediction",se.fit = TRUE,level = 0.95)
pred.int3$fit
```

C. Predict the average house price for houses with living square footage 8000.

```
xs4 = data.frame(sqft_living = 8000)
pred.int4 = predict(no_outlers_model,xs4,interval = "prediction",se.fit = TRUE,level = 0.95)
pred.int4$fit
```