MISSISSIPPI STATE UNIVERSITY

Data Analysis One

Emily Hubbard

September 12, 2025

HW 1

Executive Summary:

A. Introduction:
Linear Regression is a way to investigate whether or not two variables or more have a linear relationship. As one increases or decreases, does the other increase or decrease? Linear Regression is important because it allows us to make predictions about certain things if we are able to build a reliable model. This is practical in all life applications. Therefore, creating a best fit model from a dataset and checking its reliability is extremely important. Through various graphs and summary statistics, one can make inferences on the strength of a model. Once a model is constructed, you then have to check it against the linear assumptions. If it meets all of these assumptions, it can be extrapolated to other data and use for predictive itive analysis.

This data set includes 9 variables on 20,000 houses in the King County, Washington Area. The purpose of this report is to analyze the variables, sqft and price, creating a best fit model to be able to best predict their linear relationship. We will test various matches of transformations on both the explanatory and response variable to investigate which model gives us the most powerful predictive property. With each model, we will check whether or not it meets the underlying assumptions of linear regression: 1. Linearity, 2. Normality, 3. Constant Variance, 4. Independence. We check these assumptions to see whether or not a certain model could be extrapolated to other data. Towards the end of the report, we also explore a secondary variable to help supplement the accuracy of the model.

A linear model is based on the following formula:

$$Y = \alpha + \beta_1 X + \epsilon$$

B. Data Collection
Housing prices and log transformations. The dataset kingCountyHouses.csv contains data on over 20,000 houses sold in King County, Washington [Kaggle, 2018a]. The dataset includes the following variables:

- price = selling price of the house
- date = date house was sold, measured in days since January 1, 2014
- bedrooms = number of bedrooms
- bathrooms = number of bathrooms
- sqft = interior square footage
- floors = number of floors
- waterfront = 1 if the house has a view of the waterfront, 0 otherwise
- yr-built = year the house was built
- yr-renovated = 0 if the house was never renovated, the year the house was renovated if else.

B. Summary Information
The best fit model when only considering one covariate (single regression) is given when running a linear model on regular price and sqft, price being the response variable and sqft being the explanatory variable. This is the model that gives us the highest adjusted R^2.
The best fit model when considering two covariates (multiple regression) is given when running a linear model on log(sqft) and log(price) while also considering number of bedrooms. In this case, log(sqft) is the response variable, log(price) is the first covariate, and number of bedrooms is the second covariate. This model gives us the highest R^2 of any of the models tested in this report.
We can safely say that there is a linear relationship between price and square footage. However, there is still much work to be done to find the best fit model when it comes to including multiple covariates.

(a) Generate appropriate graphs and summary statistics detailing both price and sqft individually and then together. What do you notice?

Data Analysis:
Graphical Representations of Square Footage Data

(i.)

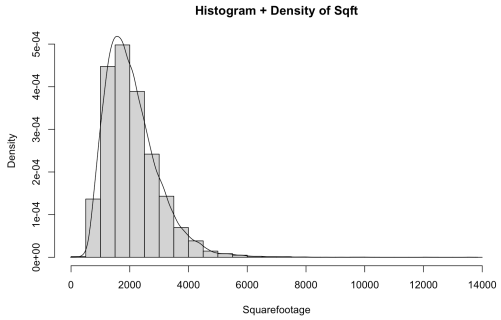

Figure 0.1: Histogram

Figure 0.2: Box-Plot

Summary Data:

Mean = 2080 ft^2

Median = 1910 ft^2

Difference = 170 ft^2

Explanation:
By looking at Figure 0.1, one observes a heavy tail to the right which skews this data. This can also be seen by the box-plot that more clearly expresses the many observations outside of the 3rd quartile. These are considered outliers. Simply by looking at these graphs, normality is brought into question. The summary data also supports this. Normality implies a similar mean and median, if not equal. This data has a 170 ft^2 difference. When this occurs, it is obvious that the outliers are heavily affecting the mean and pulling it to the right. These observations will be kept in mind throughout the rest of this report.

Graphical Representations of Price Data

(ii.)

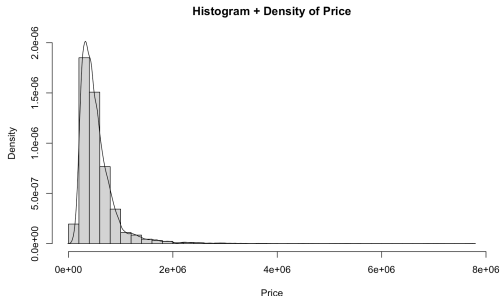

Figure 0.3: Histogram

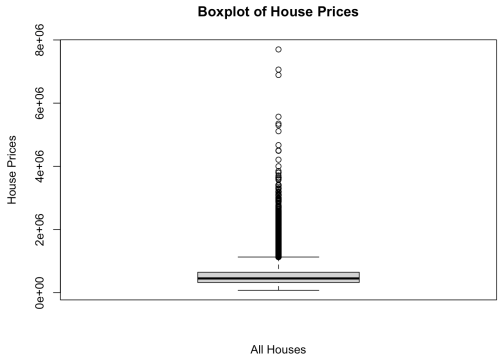

Figure 0.4: Box-Plot

Summary Data:

$$\text{Mean} = \$540,183$$
$$\text{Median} = \$450,000$$
$$\text{Difference} = \$90,183$$

Explanation:
Similar to the square footage data, the price data shows a heavy right skewness. This skew is even more extreme. This can be easily seen in the histogram, the box-plot, and the summary data. There is a difference of 90,183 dollars, meaning that the outliers are heavily effecting the typical price.

Graphical Representations of Price and Square Footage Data Together

(iii.)

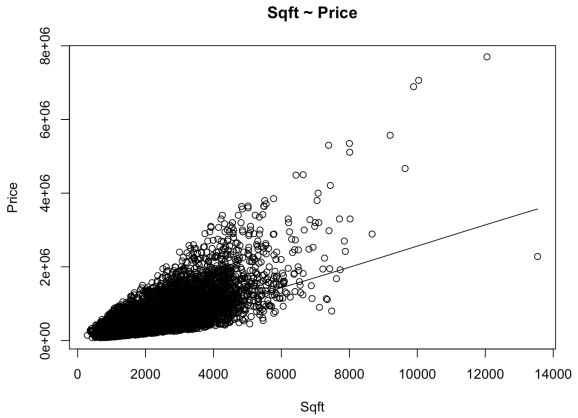

Figure 0.5: Scatter Plot

Explanation:
This scatter plot shows a linear relationship between the two variables. However, as you move along X, you see an increasing rise in Y. It is not perfectly liner. Some transformations may be able to help this.

(b) Fit a simple linear regression model with price as the response variable and sqft as the explanatory variable (Model 1). Interpret the slope coefficient β_1. Are all conditions met for linear regression?

Regression Line Equation:

$$\text{Intercept}(\alpha) = -43866.0$$
$$\text{Coefficient}(\beta_1) = 280.8$$
$$\text{Y} = -43866.0 + 280.8X + \epsilon$$

Summary Statistics:

$$\text{Adjusted } \text{R}^2 = 0.4928$$
$$\text{Covariate P-value} = 2e^{-16}$$
$$\text{F-Test P-value} = 2.2e^{-16}$$

Interpretation:
As square footage increases by 1, the price of a house is predicted to increase by $280.80. The summary of this model shows a significant p-value for the square footage covariate. This means that square footage does give us useful information that informs the value

of price. The summary also shows an adjusted R^2 value of 0.4928. This tells us that the model is missing out on 50 percent of the information that could be gleaned from the data. So, although square footage has helpful information, this model needs work. The p-value associated with the F-test is extremely small, much smaller than 0.05. What this conveys to us is that the model is better at predicting response variable Y (price) when the covariate X (square footage) is considered and kept in the model.

Assumptions Check:

Linearity: By looking at the scatter plot (Figure 0.5), I would subjectively say that this data meets the assumption of linearity. Basically, this means that a linear regression model would give us good information on predicting the response variable because it acts linearly.

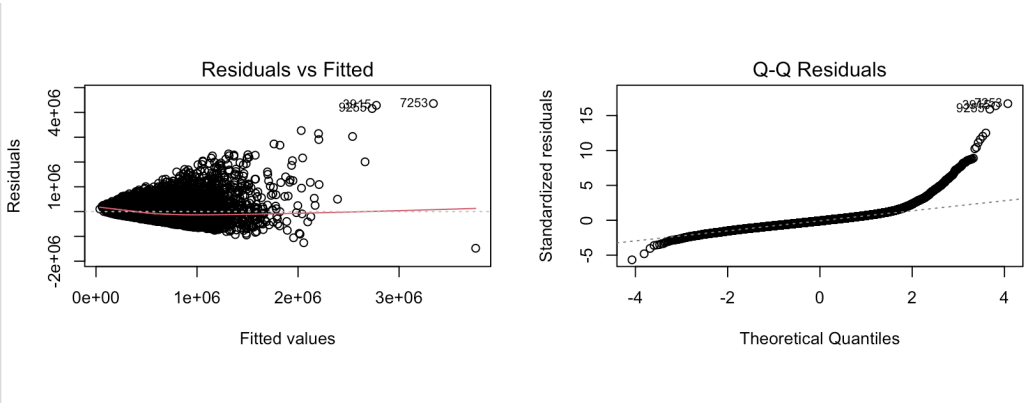

Figure 0.6: Visuals for Normality and Constant Variation Model 1

Normality: By using the lillie.test in R and observing the Q-Q plot, a conclusion can be made about the data's normality, namely that it is not normal. The P-value associated with the lillie.test is well below 5 percent meaning that we reject the null. In this case, it means that the residuals vary significantly from normality. It does not meet the normality assumption. This is also witnessed by observing th stark deviation from the line towards the rights in Figure 0.6's Q-Q plot.

Constant Variance: The Residuals vs Fitted plot in Figure 0.6 indicates that variance is not constant. Moving from the beginning towards the end, an increasing funnel shape is made by the residuals. If there was constant variance, there would be symmetrical deviation from the red line along the entirety of the plot. This is not true. Therefore, the assumption of constant variance is not met.

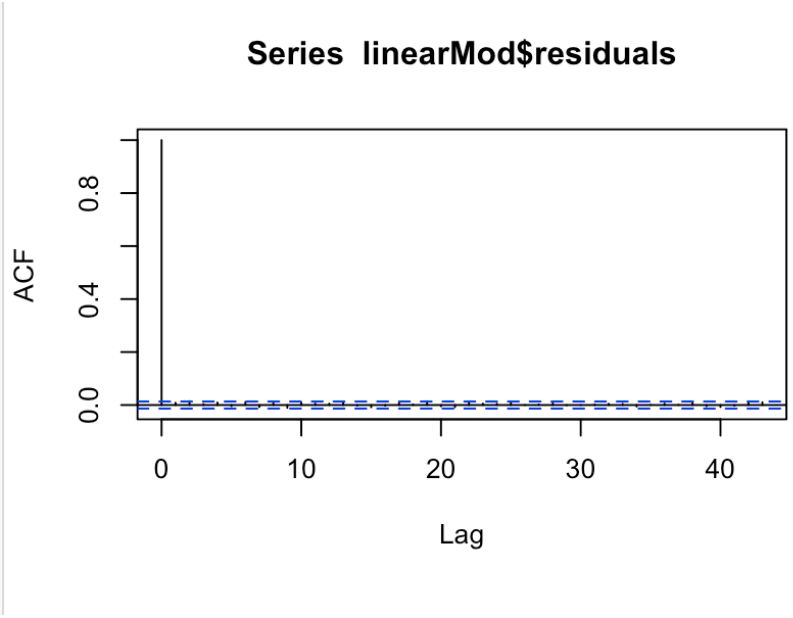

Figure 0.7: ACF Graph Model 1

Independence: The ACF Graph of Model 1 (Figure 0.7) indicates that the residuals are independent of each other. All residuals fall within the blue dotted lines, meaning they fall within the 95 percent confidence interval limits. This indicates zero correlation between the error points. This is important because if they we correlated (non-independent), then the model would be inflating the effect of some conclusions. This happens because the model is treating them as independent.

Assumptions Summary:

Linearity = TRUE
Normality = FALSE
Constant Variance = FALSE
Independence = TRUE

(c) Create a new variable, log(price), the natural log of price. Fit Model 2, where log(price) is now the response variable and sqft is still the explanatory variable. Write out the regression line equation.

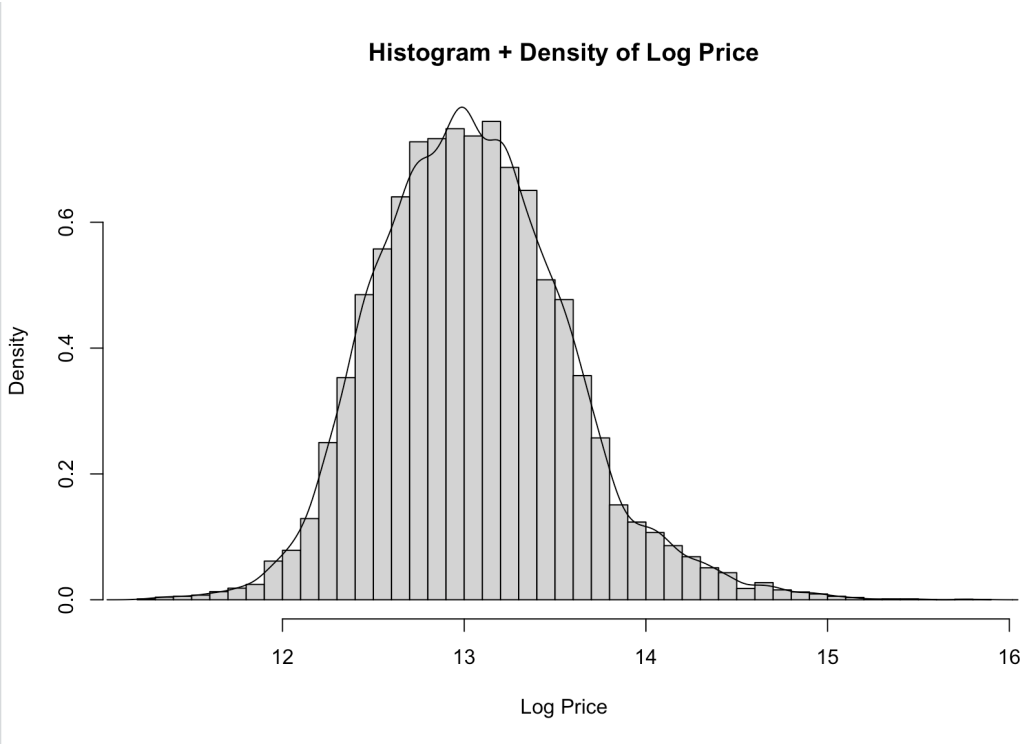

Figure 0.8: Histogram

Explanation:
This histogram of log(price)is starkly different from the histogram of price (Figure 0.3) at the beginning of the report. By taking the log of the data, the extreme outliers have been squeezed to the left. This creates a distribution that is much closer to normal. The histogram actually holds a bell-shape compared to the heavy skew of the other histogram.

Regression Line Equation:

$$\text{Intercept}(\alpha) = 1.222e^{01}$$
$$\text{Coefficient}(\beta_1) = 3.989e^{-04}$$
$$\text{Y} = 1.222e^{01} + 3.989e^{-04}X + \epsilon$$

Summary Statistics:

$$\text{Adjusted R}^2 = 0.4835$$
$$\text{Covariate P-value} = 2e^{-16}$$
$$\text{F-Test P-value} = 2.2e^{-16}$$

Interpretation:
The summary of this model shows a significant p-value for the square footage covariate similar to Model 1. This means that square footage does give us useful information that informs the value of price. The summary shows an adjusted R^2 value of 0.4835, a slight decrease compared to Model 1. This means we lost about another percent of information. Again, like the first model, the p-value associated with the F-test is extremely small, much smaller than 0.05. Therefore, the model is better at predicting response variable Y log(price) when the covariate X (square footage) is considered and kept in the model.

(d) How does log(price) change when sqft increases by 1?

As square footage increases by 1, the log(price) of a house is predicted to increase by 0.0004.

(e) Recall that log(a)-log(b)=log[a/b],and use this to derive how price changes as sqft increases by 1.

Explanation:
Using this logic from the notes: if y is log-transformed, one simply interprets e^{β} instead of just beta.

Therefore, when square footage increases by 1, the response variable Y, price, will increase by $e^{0.0004}$.

Derivation:

$$\log(\hat{\text{price}}) = \beta_0 + \beta_1 \text{sqft}$$
$$\log(\text{price})_{\text{new}} - \log(\text{price})_{\text{old}} = \beta_1(\text{sqft} + 1 - \text{sqft}) = \beta_1$$
$$\log(\frac{\text{price}_{\text{new}}}{\text{price}_{\text{old}}}) = \beta_1$$
$$\frac{\text{price}_{\text{new}}}{\text{price}_{\text{old}}} = e^{\beta_1}$$

(f) Are LLSR assumptions satisfied in Model 2? Why or why not?

Assumptions Check:

Linearity: Given that the F-test indicates that square footage is helpful in the model, I would subjectively say that this data meets the assumption of linearity. Because using the full model compared to the reduced does indicate a slight linear relationship. If it had no impact, you would simply drop the covariate.

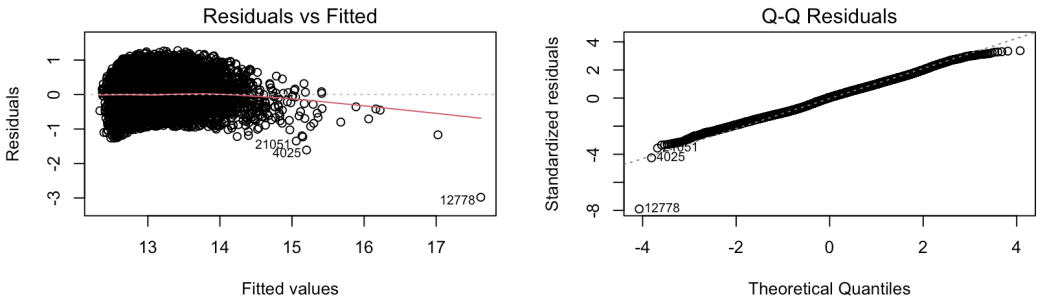

Figure 0.9: Visuals for Normality and Constant Variation Model 2

Normality: By using the lillie.test in R and observing the Q-Q plot, a conclusion can be made about the data's normality- it is not normal. The P-value associated with the lillie.test is well below 5 percent meaning that we reject the null. However, this data is much closer to normal when considering the Q-Q plot in Figure 0.9 alone. This plot shows only slight deviations at the endpoints whereas there was much stronger deviation in Figure 0.6's Q-Q plot. Regardless of it being only slight deviation when observing the Q-Q plot, it still indicates non-normal data. Therefore, it does not meet the normality assumption.

Constant Variance: The Residuals vs Fitted plot in Figure 0.9 indicates that variance is somewhat constant. I think it comes down to subjective judgement on whether or not this data meets constant variation. Until around 14.5, the variations are relatively symmetrical clay about the red dotted line. They start to move away after that. Overall, I would conclude there is not constant variance. This is mainly due to a few outliers towards the right endpoint.

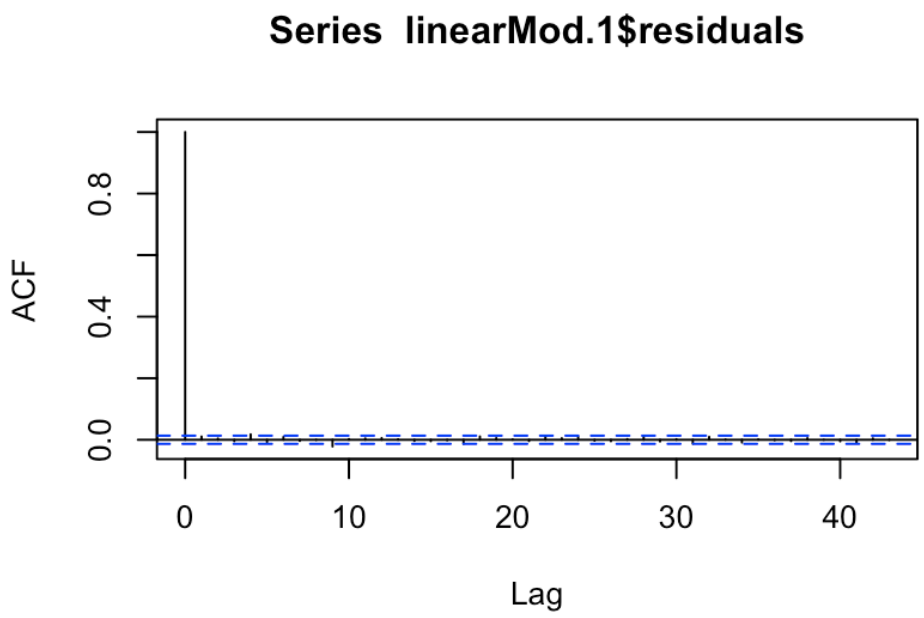

Figure 0.10: ACF Graph Model 2

Independence: The ACF Graph of Model 2 (Figure 0.10) indicates that the residuals are independent of each other. All residuals fall within the blue dotted lines, meaning they fall within the 95 percent confidence interval limits. This indicates zero correlation between the error points.

Assumptions Summary:

Linearity = TRUE
Normality = FALSE
Constant Variance = FALSE
Independence = TRUE

(g) Create a new variable, log(sqft), the natural log of sqft. Fit Model 3 where price and log(sqft) are the response and explanatory variables, respectively. Write out the regression line equation.

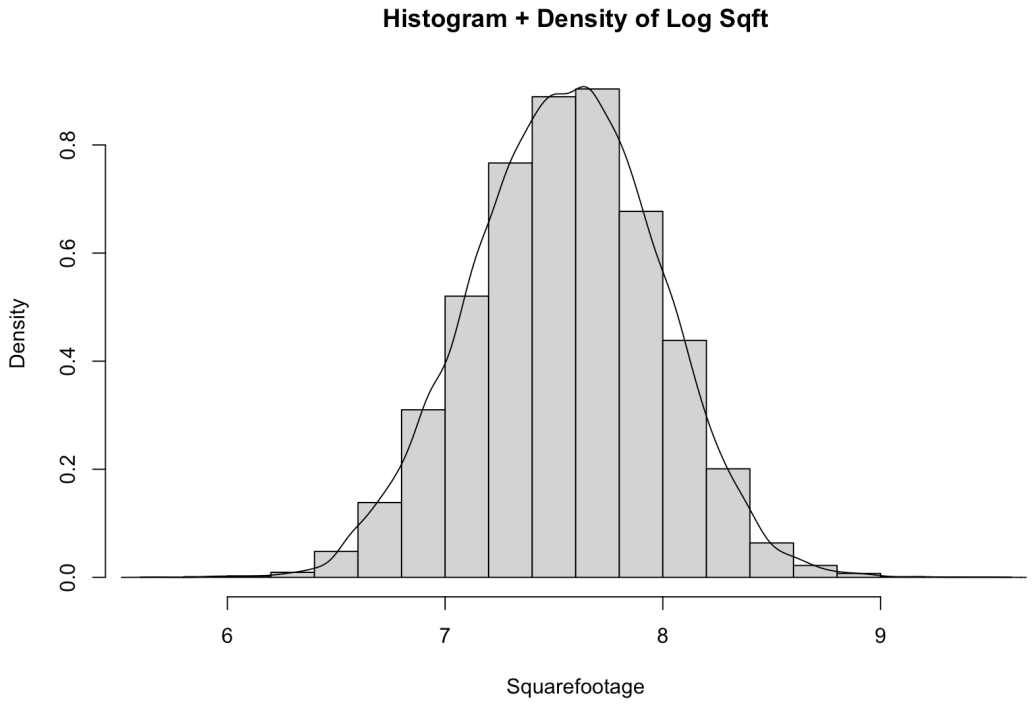

Figure 0.11: Histogram

Explanation:
This histogram of log(sqft) is starkly different from the histogram of sqft (Figure 0.1) at the beginning of the report. By taking the log of the data, the extreme outliers have been squeezed to the left as they were with the price earlier. This creates a distribution that looks much closer to normal.

Regression Line Equation:

$$\text{Intercept}(\alpha) = -345,3773$$
$$\text{Coefficient}(\beta_1) = 528,977$$
$$\text{Y} = -345,3773 + 528,977x + \epsilon$$

Summary Statistics:

$$\text{Adjusted } \text{R}^2 = 0.3741$$
$$\text{Covariate P-value} = 2e^{-16}$$
$$\text{F-Test P-value} = 2.2e^{-16}$$

Interpretation:
The summary of this model shows a significant p-value for the log(sqft) covariate similar to Model 1 and 2. The adjusted R^2 is 0.3741, a significant decrease of almost 10 percent compared to Model 1 and 2. This means we lost a lot more information that could have been provided with the data. The p-value associated with the F-test is extremely small, much smaller than 0.05. Therefore, the model is better at predicting response variable Y log(price) when the covariate X (square footage) is considered and kept in the model.

(h) How does predicted price change as log(sqft) increases by 1 in Model 3?

The predicted price change as log(sqft) changes by 1 would be an increase of 528,977 dollars. This is simply the normal way of interpreting the linear regression formula.

(i) How does predicted price change as sqft increases by 1 percent? As a hint, this is the same as multiplying sqft by 1.01.

As sqft increases by 1 percent, (log(1.01)), the price increases by 528,977 * log(1.01). This comes out to be approximately 5,263 dollars.

(j) Are LLSR assumptions satisfied in Model 3? Why or why not?

Assumptions Check:

Linearity: Given that the F-test indicates that square footage is helpful in the model, I would subjectively say that this data meets the assumption of linearity.

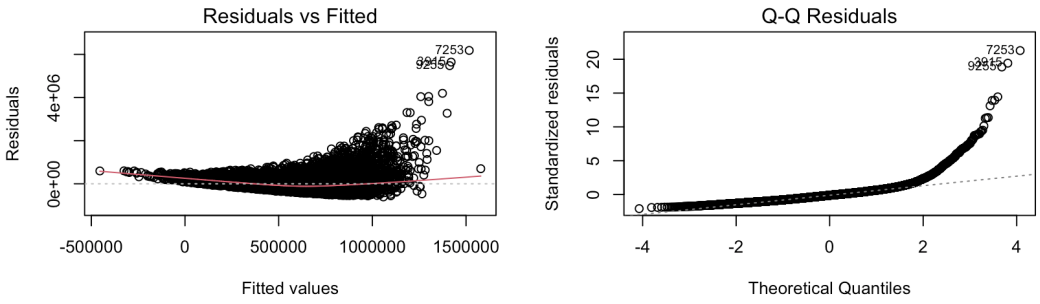

Figure 0.12: Visuals for Normality and Constant Variation Model 3

Normality: By using the lillie.test in R and observing the Q-Q plot, a conclusion can be made about the data's normality- it is not normal. The P-value associated with the lillie.test is well below 5 percent meaning that we reject the null. This Q-Q plot in Figure

0.12 shows a large deviation from normal after 2. Therefore, it does not meet the normality assumption.

Constant Variance: The Residuals vs Fitted plot in Figure 0.12 indicates that variance is not constant. Moving from the beginning towards the end, an increasing funnel shape is made by the residuals, similar to Model 1. If there was constant variance, there would be symmetrical deviation from the red line along the entirety of the plot. This is not true. Therefore, constant variance assumption is not met.

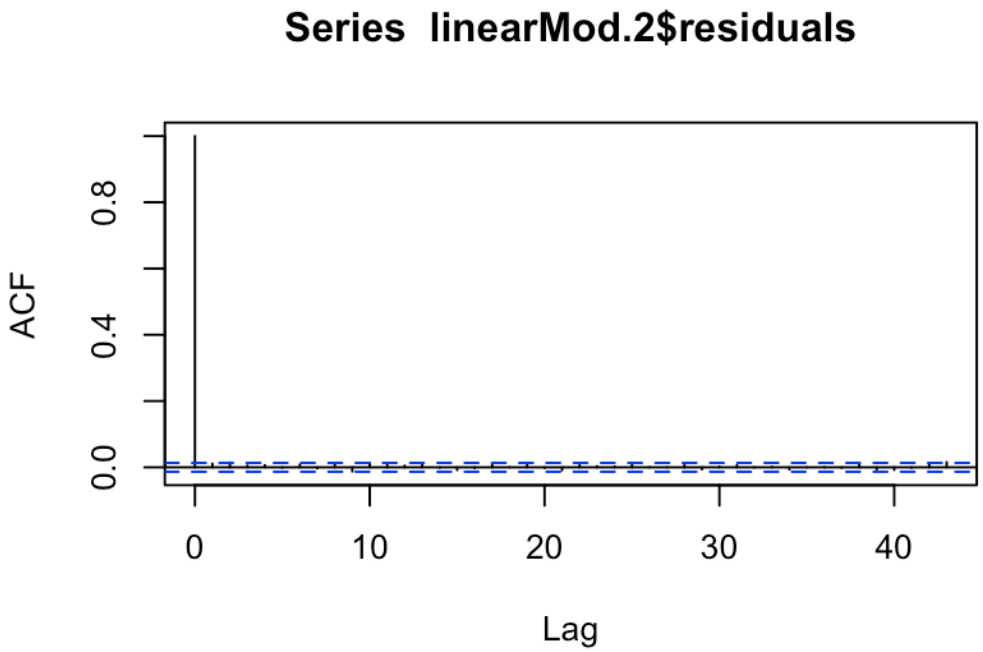

Figure 0.13: ACF Graph Model 3

Independence: The ACF Graph of Model 3 (Figure 0.13) indicates that the residuals are independent of each other just like Model 1 and 2. All residuals fall within the blue dotted lines, meaning they fall within the 95 percent confidence interval limits. This indicates zero correlation between the error points.

Assumptions Summary:

Linearity = TRUE
Normality = FALSE
Constant Variance = FALSE
Independence = TRUE

(k) Fit Model 4, with log(sqft) and log(price) as the response and explanatory variables, respectively. Write out the regression line equation

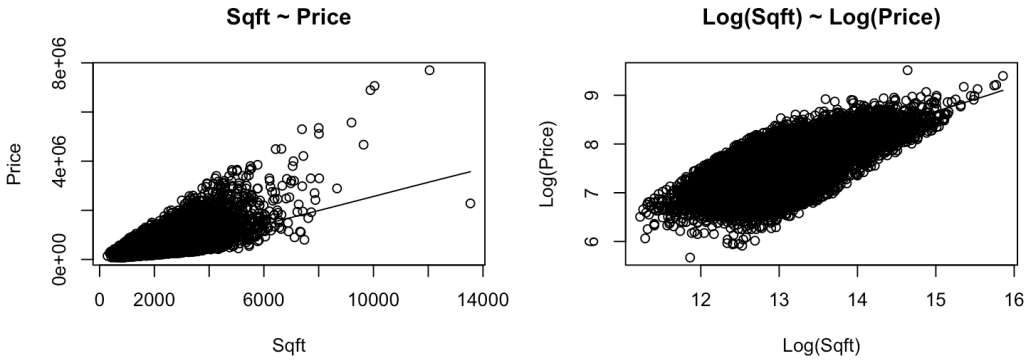

Figure 0.14: Comparison of Sqft and Price and Log(sqft) and Log(price) Scatter Plots

Explanation:
Before running the model, it is helpful to take a look at the linear relationship between Log(sqft) and Log(price) in comparison to our first scatter plot (regular price and sqft). When both variables undergo log transformations, it seems that the linear relationship gets stronger. That is what I get from the visuals. From end to end, there are point along the best fit model line. The same cannot be said of the first scatter plot. This gives me hope that maybe this model will give us the greatest information and predictive power of the response variable.

Regression Line Equation:

$$\text{Intercept}(\alpha) = 0.4493$$
$$\text{Coefficient}(\beta_1) = 0.5442$$
$$Y = 0.4493 + 0.5442x + \epsilon$$

Summary Statistics:

$$\text{Adjusted } R^2 = 0.4555$$
$$\text{Covariate P-value} = 2e^{-16}$$
$$\text{F-Test P-value} = 2.2e^{-16}$$

Interpretation:
The summary of this model shows a significant p-value for the log(price) covariate similar to Model 1 and 2. The adjusted R^2 is 0.4555, which is the 3rd worst of the 4 models we

have put together. This was not what I expected. This log transformation did not maximize the information gleaned from the data. However, the p-value associated with the F-test is extremely small, much smaller than 0.05, so we still use the full model instead of the reduced one.

(l) In Model 4, what is the effect on price corresponding to a 1 percent increase in sqft?

This is a log-log transformation model, meaning the coefficient becomes an elasticity. This means that, as the covariate increase by 1 percent, the response variable increases by the percentage of the covariate coefficient. This means that as price (the explanatory variable), increases by 1 percent, the sqft (response variable) increases by 0.5442 percent.

(m) Are LLSR assumptions satisfied in Model 4? Why or why not?

Assumptions Check:

Linearity: Given that the F-test indicates that square footage is helpful in the model, I would subjectively say that this data meets the assumption of linearity.

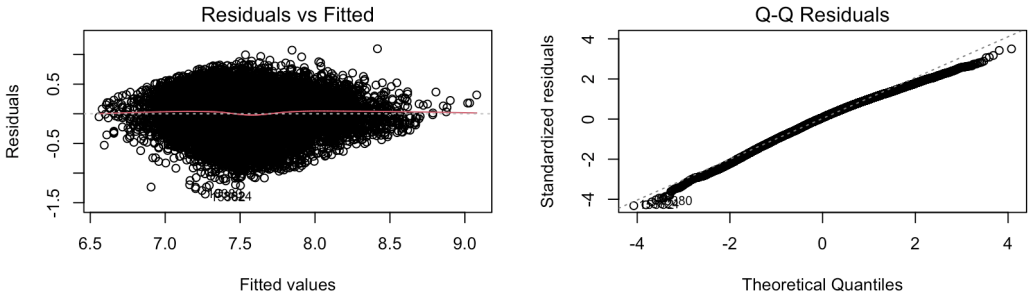

Figure 0.15: Visuals for Normality and Constant Variation Model 4

Normality: By using the lillie.test in R and observing the Q-Q plot, a conclusion can be made about the data's normality- it is still not normal. The P-value associated with the lillie.test is well below 5 percent meaning that we reject the null. This Q-Q plot in Figure 0.15 shows only a slight deviation towards the endpoints, but regardless, the result is the same, non-normality.

Constant Variance: The Residuals vs Fitted plot in Figure 0.15 indicates that variance is not constant. There seems to be a thicker concentration of deviation towards the middle of the plot compared to the endpoints. This indicates that the constant variance assumption is not met.

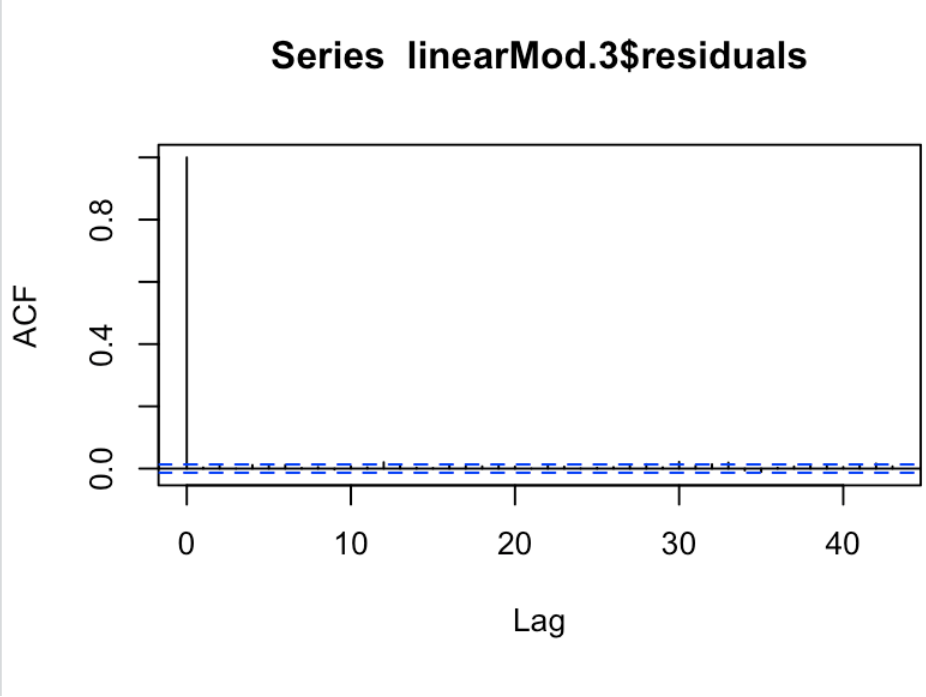

Figure 0.16: ACF Graph Model 4

Independence: The ACF Graph of Model 4 (Figure 0.16) indicates that the residuals are independent of each other just like Model 1 and 2 and 3. All residuals fall within the blue dotted lines, meaning they fall within the 95 percent confidence interval limits. This indicates zero correlation between the error points.

Assumptions Summary:

Linearity = TRUE
Normality = FALSE
Constant Variance = FALSE
Independence = TRUE

(n) Find another explanatory variable which can be added to Model 4 to create a model with a higher adjusted R^2 value. Interpret the coefficient of this added variable.

Explanation:
I chose number of bedrooms to be my second covariate because it would make sense that the number of bedrooms would have a predictive effect on price. That is what I wanted to test.

Figure 0.17: Histogram

Summary Data:

Mean = 3.371 bedrooms
Median = 3 bedrooms

Explanation:
There looks to be a normal distribution of number of bedrooms. Now we will add it to our model and see if it increases the predictive power of our model.

Regression Line Equation:

Intercept(α) = 1.3645
Coefficient(β_1) = 0.4220
Coefficient(β_2) = 0.2014
$Y = 1.3645 + 0.4220x_1 + 0.2014x_2 + \epsilon$

Summary Statistics:

Adjusted $R^2 = 0.627$

Covariate 1 P-value = $2e^{-16}$

Covariate 2 P-value = $2e^{-16}$

F-Test P-value = $2.2e^{-16}$

Interpretation:
The summary of this model shows a significant p-value for the log(price) covariate. It also shows a significant p-value for the second added covariate (number of bedrooms). This indicates that both price and number of bedrooms have an impact on square footage. The adjusted R^2 increase to 0.627, which is the best of the 4 models we have put together. The p-value associated with the F-test is extremely small, much smaller than 0.05, so we still use the full model instead of the reduced one.

Of all the models, this one is the most predictive. It utilizes the largest amount of information from the data to predict the response variable.

APPENDIX

R code:

```
    #### HW 1 ####
data <- read.csv("kingCountyHouses.csv",header = TRUE)
data

#### a ####
#Examining sqft individually:
a <- data$sqft
summary(a)
hist(a,xlab = "Squarefootage", main = "Histogram + Density of Sqft",ylim = c(0, 0.0005),
lines(density(a))
#Similar to the price distribution, there are a few outliers of homes with a lot more sq
#The median is a better descriptor of the average home.
boxplot(a, ylab = "House Sqft", xlab = "All Houses", main = "Boxplot of House Sqft")

#Examining price individually:
b <- data$price
summary(b)
hist(b, xlab = "Price", ylim = c(0, 0.0000020), main = "Histogram + Density of Price", p
lines(density(b))
#It is heavily skewed to the right because of a few extremely expensive homes.
#The median is the best method of measuring central tendency.
boxplot(b, ylab = "House Prices", xlab = "All Houses", main = "Boxplot of House Prices")
#Similar to the Histogram and Density curve, this box plot shows many extreme outliers.

##Examining both sqft and price in relation to the other:
scatter.smooth(a,b, ylab = "Price", xlab = "Sqft", main = "Sqft ~ Price")
#From looking at these two data sets, it seems that a linear model might work very well.
#Especially if the data set's extreme values were trimmed a bit.

#### b MODEL 1 ####
?lm
linearMod <- lm(b~a)
linearMod
#Intercept = -43866.0 Slope Coefficient = 280.8
#Formula: Y = -43866.0 + 280.8X + E
#With every 1 unit increase in sqft, the price is raised by 280.8
summary(linearMod)
#The P-value is well below the comparison of alpha at 5%.
#This tells us that sqft has a high correlation with price.
#Also the R^2 is giving a slight indication that the model could give us some valuable i
```

```
#The slope coefficient, B, is saying with every 1 unit increase in squarefootage,
#the price is raised $280.80.


##Checking for normality of residuals (which will apply to the normality of the response
install.packages("nortest")
library(nortest)

linearMod$residuals
lillie.test(linearMod$residuals)
#P-value is well below 5% meaning that we reject the null.
#In this case, it means that the residuals vary significantly from normality.
#It does not mmeet the normality assumption.
#AKA, as of right now, this model can not be used on other data.

##Checking for constant variation
mean(linearMod$residuals)
par(mfrow=c(2,2))
plot(linearMod)
#The Residuals vs Fitted plot shows an increasing variation. This means that the assumpt
#constant variation is not met.
#The Q-Q plot matches the result of the lillie.test- not normal.
#The plot indicates heavy-tails/outliers. This means it is not normal.

##Checking for normality of residuals and independence
#The lillie.test told us that the residuals are not normal.

acf(linearMod$residuals)
#The acf function tells us that the residuals are independent.



##Checking for continuous response variable
class(data$price)
#The outcome of this function is "numeric" which most likely indicates a continuous vari

## Overall, the assumptions of normality, independence of residuals, and constant variat
# are unmet, met, and unmet, respectively.

#### c MODEL 2 ####
c <-log(data$price)
summary(c)
hist(c, xlab = "Log Price", main = "Histogram + Density of Log Price", probability = TRU
lines(density(c))
```

```
#compared to the histogram of the regular price, this one follows a much more bell curve
#implying a normality that we didn't see in the regular price.

linearMod.1 <- lm(c~a)
linearMod.1
#Y= 1.222e+01 + 3.989e-04x + E
summary(linearMod.1)
#Similarly to regaular price and sqft, the coefficient adds helpful information to the r
#The information lost in the model actually increased. Instead of capturing 49.28% of in
#it captured a percent less information, 48.35%

#### d ####
#When sqft increases by 1, the logprice increases by 0.0004.

#### e ####
#Using this logic from the notes: if y is log-transformed, one simply interprets
#e^beta instead of just beta.

#Therefore, when square footage increases by 1, the response variable Y, price,
#will increase by e^0.0004

#Without a log transformation, the price data is heavily skewed to the right. This means
#that a few extreme values shift the model dramatically. After using the log transformat
#the extreme values are shrunkated and this allows for a more normal distribution. This
#visualized in the histogram and desnity plot of price versus logprice.

#### f ####
##Checking for normality of residuals (which will apply to the normality of the response
install.packages("nortest")
library(nortest)

linearMod.1$residuals
lillie.test(linearMod.1$residuals)
#P-value is well below 5% meaning that we reject the null.
#In this case, it means that the residuals vary significantly from normality.
#The same can be applied to the normality of the response variable. Therefore, Y is not
#AKA, as of right now, this model can not be used on other data.

##Checking for constant variation
mean(linearMod.1$residuals)
par(mfrow=c(2,2))
plot(linearMod.1)
#The Residuals vs Fitted plot shows a fairly constant variation.The bulk of the residual
```

```
#Once it nears the end, the residuals begin to show a more dynamic variation, but I woul
#for constant variation is mostly met. This is different compared to our earlier graph u
#The Q-Q plot matches the result of the lillie.test- not normal. The bulk of the residua
#however, the plot indicates heavy-tails/outliers. This means it is not normal.


##Checking for normality of residuals and independence
#The lillie.test told us that the residuals are not normal.

acf(linearMod.1$residuals)
#The acf function tells us that the residuals are independent.


##Checking for continuous response variable

## Overall, the assumptions of normality, independence of residuals, and constant variat
# are unmet, met, and met, respectively.



#### g MODEL 3 ####
d<- log(data$sqft)
summary(d)
hist(d,xlab = "Squarefootage", main = "Histogram + Density of Log Sqft", probability = T
lines(density(d))
#Similar to the log price distribution, taking the natural log gives us a more bell-curv
#This indicates a more normal behavior.
boxplot(d, ylab = "House Sqft", xlab = "All Houses", main = "Boxplot of House Log Sqft")
#Converting the sqft to log helps shrink the effect of the outliers as can be seen when
#histograms and boxplots of the reg vs log sqft.

linearMod.2 <- lm(b~d)
linearMod.2
#Y= -3453773 + 528977x + E
summary(linearMod.2)
#Similarly to regular price and sqft, the coefficient adds helpful information to the re
#The information lost in the model actually increased. Instead of capturing 49.28% of in
#it lost a little over 10% more information, 37.41%.

#### h ####
#The predicted price change as logsqft changes by 1 would be an increase of $528,977. Th
#the normal way of interpreting the linear regression formula.
```

```
#### i ####
#As sqft increases by 1% (log(1.01)), the price increases by 528,977 * log(1.01). This c
#out to be approximately $5,263.

#### j ####
##Checking for normality of residuals (which will apply to the normality of the response
install.packages("nortest")
library(nortest)

linearMod.2$residuals
lillie.test(linearMod.2$residuals)
#P-value is well below 5% meaning that we reject the null.
#In this case, it means that the residuals vary significantly from normality.
#The same can be applied to the normality of the response variable. Therefore, Y is not
#AKA, as of right now, this model can not be used on other data.

##Checking for constant variation
mean(linearMod.2$residuals)
par(mfrow=c(2,2))
plot(linearMod.2)
#The Residuals vs Fitted plot shows us that the residuals have an increasing variation.
#The Q-Q plot matches the result of the lillie.test- not normal. The bulk of the residua
#however, after 2 there is heavy deviation from normal.


##Checking for normality of residuals and independence
#The lillie.test told us that the residuals are not normal.

acf(linearMod.2$residuals)
#The acf function, similar to the other two models, tells us that the residuals are inde


##Checking for continuous response variable

## Overall, the assumptions of normality, independence of residuals, and constant variat
# are unmet, met, and unmet, respectively.


#### k MODEL 4 ####
##Examining both log(sqft) and log(price) in relation to the other:
scatter.smooth(c,d, ylab = "Log(Price)", xlab = "Log(Sqft)", main = "Log(Sqft) ~ Log(Pri

linearMod.3 <- lm(d~c)
linearMod.3
```

```
summary(linearMod.3)
#Formula: Y = 0.4493 + 0.5442X + E
#This means that for every 1 unit increase in logprice,
#the logsqft increase by 0.5442.

#### l ####
#This is a log-log transformation model, meaning the coefficient becomes an elasticity.
#This means that, as the covariate increase by 1%, the response variable increases by th
#percentage of the covariate coefficient.

#This means that as price (the explanatory variable), increases by 1%, the sqft (respons
#increases by 0.5442%.

#Since l is asking us about a 1% increase in the response variable, we must invert our s
#AKA 1/0.5442 = 1.837. Now, a 1% increase in sqft will create a 1.837% increase in price

#### m ####
install.packages("nortest")
library(nortest)

linearMod.3$residuals
lillie.test(linearMod.3$residuals)
#P-value is well below 5% meaning that we reject the null.
#In this case, it means that the residuals vary significantly from normality.
#The same can be applied to the normality of the response variable. Therefore, Y is not
#AKA, as of right now, this model can not be used on other data.

##Checking for constant variation
mean(linearMod.3$residuals)
par(mfrow=c(2,2))
plot(linearMod.3)
#The Residuals vs Fitted plot shows an inconsistent variation meaning
#constant variation assumption is not met.
#The Q-Q plot matches the result of the lillie.test- not normal. The bulk of the residua
#however, the plot slowly deviates from normal at the extreme values. This means it is n


##Checking for normality of residuals and independence
#The lillie.test told us that the residuals are not normal.

acf(linearMod.3$residuals)
#The acf function tells us that the residuals are independent.
```

```
##Checking for continuous response variable

## Overall, the assumptions of normality, independence of residuals, and constant variat
# are unmet, met, and unmet, respectively.

#### n MODEL 4.adj ####
e <- data$bedrooms
summary(e)
hist(e,xlab = "Bedrooms", main = "Histogram + Density of Bedrooms", xlim = c(0,10), prob

linearMod.3.2 <- lm(d~c+e)
linearMod.3.2
summary(linearMod.3.2)
#The adjusted R^2 of this model is 0.627. This is a stark jump from the adjusted R^2 of
#By adding in a second covariate, number od bedrooms, we have significantly increased th
#that we have gathered from the data.
```