HW3 Key

Ryan O'Donnell

1. Quantitative variables used in the model were lot size, age, land value, living area, pct. college, bedrooms, bathrooms, and rooms.

Note: If a student manually processed the number of beds variable, it will be graded on a case-by-case basis.

```
houses <- read.csv("~/Spring TA job/houses.csv", row.names=1)
attach(houses)
model_all<- lm(Price~Lot.Size+Age+Land.Value+Living.Area+Pct.College+Bedrooms +Bathrooms+Rooms)
```

2.The null hypothesis is that none of the predictor/explanatory variables in the regression have a relationship with our response variable price. The alternative is that at least one of the variables has a relationship with our response variable.

$$H_0: B_1= B_2= B_3=\ldots=B_8=0.$$

$$H_A: \text{There exists at least one } B_i \text{ such that } B_i \neq 0.$$

3. The R^2 value is 0.6269, which is the variation in house prices that is explained by the

Model. This can be found in the summary of the model.

```
summary(model_all)
```

4. Land Value is most significant because it has the largest t-score, but the difference between it and living area, number of bathrooms, number of bedrooms, lot size and number of rooms is not really practically significant as they are all so far to the right on the t-distribution. Note, essentially the same information can be gleaned from the p-values, as the p value is directly related to the t-statistic. Again, this information can be found in the summary of the model.

```
summary(model_all)
```

5.

Unsurprisingly, all variables are significant in the model_signif. The intercept was not significant in the previous model, but it is here. Somewhat interestingly, that model did not have a drop in r^2.

The model insignif uses only pct college, and lists pct. college as a significant variable. This is somewhat surprising, as we removed this variable from model_all because it was insignificant. In a way, this new model is blaming the change in y on the wrong variable and inflating the coefficient on pct college. An inflated beta can lead to an inflated t statistic, which in turn can lead to a variable seeming statistically significant when it really shouldn't be. If other variables were available in the model, it would be clear that this variable is statistically insignificant.

Also, notice the insignificant model is not practically significant, as the r^2 is so low. The beta value is extremely large, but that alone does not indicate practical significance of the model. The model would be better than nothing, but you should strive to develop a better model when r^2 is this low.

```
model_signif<-lm(Price~Lot.Size+Age+Land.Value+Living.Area+Bedrooms+Bathrooms+Rooms)
model_insignif<-lm(Price~Pct.College)
summary(model_signif)
summary(model_insignif)
```

6.

Deciding between models and deciding to remove variables, can be a rather complicated topic. Any good, logical answer would be accepted for this question. Two possible answers are:

The t statistics for pct. College indicates insignificance, so we would want to remove it and take the smaller model.

We can focus on the adjusted r^2. The adjusted R^2 is a touch better for the smaller model, so we should use the smaller model. The pct. college variable is not helping us enough when it comes to predicting prices, which can be seen by the adjusted r^2 being bigger for the smaller model.

Note the below framework for model selection:

1) Model assumptions should be checked. If assumptions are violated for model A but not model B, you would usually want to go with model B.

2) If the f statistic indicates model A is insignificant, but model B is significant, we would prefer model B.

3) If the t statistics for some of the variables in model A indicate insignificance, then we want to throw those variables out.

4) Compare the models based on adjusted R^2 or some other statistic.

(Instructor Note: While multicollinearity or other assumption violations could appear in the dataset, any good, logical answer should be accepted for this question. Even an argument based on the r^2 not "increasing enough" would be fine for now)

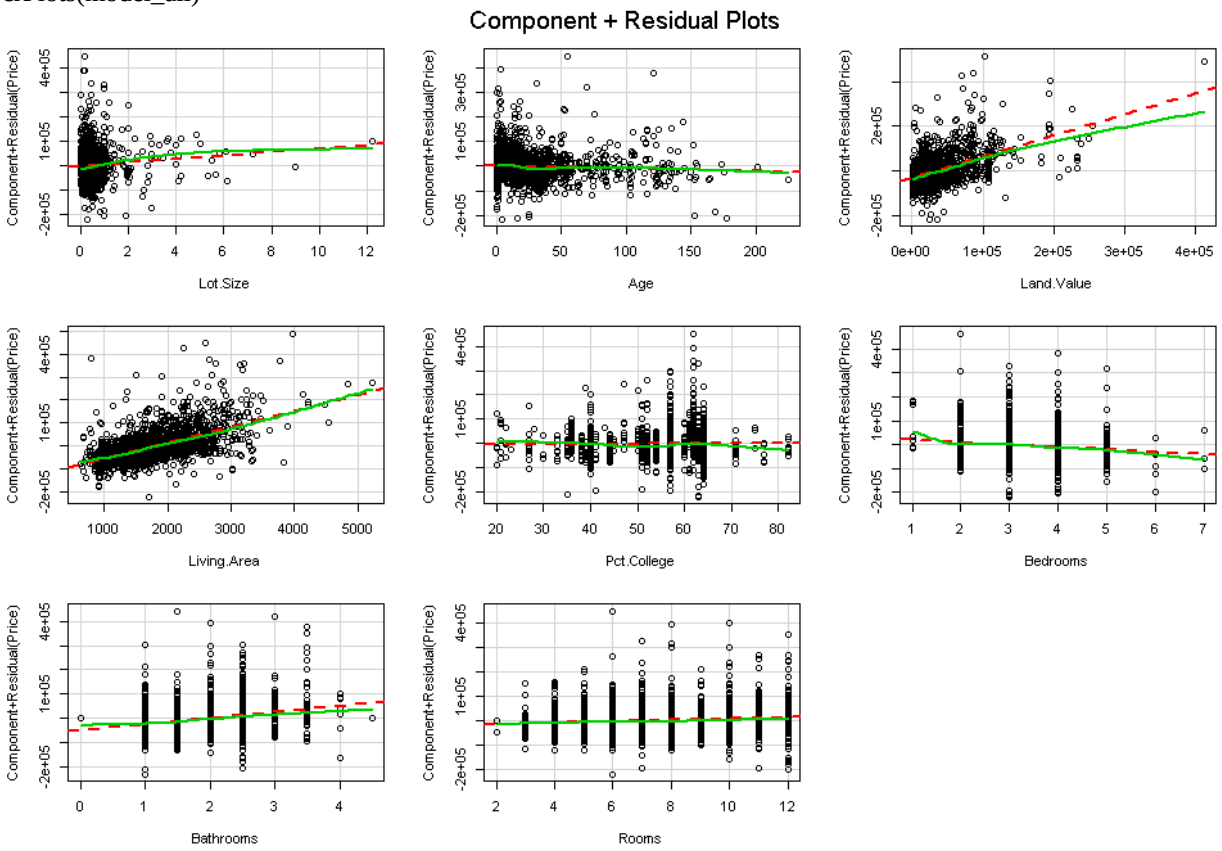
```
summary(model_signif)
summary(model_insignif)
```

7.

The rest of the variables are significant, so we can't throw out variables based on that. Again, there is more than one way to go about this. Again, any logical answer grounded in statistics would be accepted. Three possible example explanations are shown below.

If assumptions are violated for a particular variable, then that may indicate you should do something with the variable. For example, we could check the linearity assumptions. We will discuss outliers in more detail later. Below is one way to check for linearity. They all seem to look reasonable except for perhaps land value. See question 10 for more detailed information on assumption checking.

crPlots(model_all)



A more practicality focused method would be to examine the variables individually and present logical arguments for why or why not they may be important. For example, imagine a variable that is very expensive to gather data for. If this variable is statistically significant, but not very practically significant (adds very little to r^2), then we would perhaps throw it out of our model.

Alternatively, you could keep things simple and not throw out any of these variables. Your reasoning would be that they are all statistically significant.

As this class continues onward, we will expect you to check assumptions. For now, any good, logical arguments for a question like this would be given credit. \

8.

The intercept in this case is fairly meaningless, and is just a mathematical artifact to make the regression fit correctly. You would never have a house that is 0 square feet on a lot that is 0

square feet. Answers discussing the possibility that a house has some intrinsic value would also be accepted. Either answer is accepted as long as arguments are logical.

```
summary(model_signif)
summary(model_insignif)
```

9. It could be argued that the smaller model is a touch better because simplicity is nice. From a purely statistical perspective, the all variable model does have a higher adjusted R^2. Either answer is accepted as long as the argument is sufficiently grounded in logic. Either answer is accepted as long as arguments are logical.

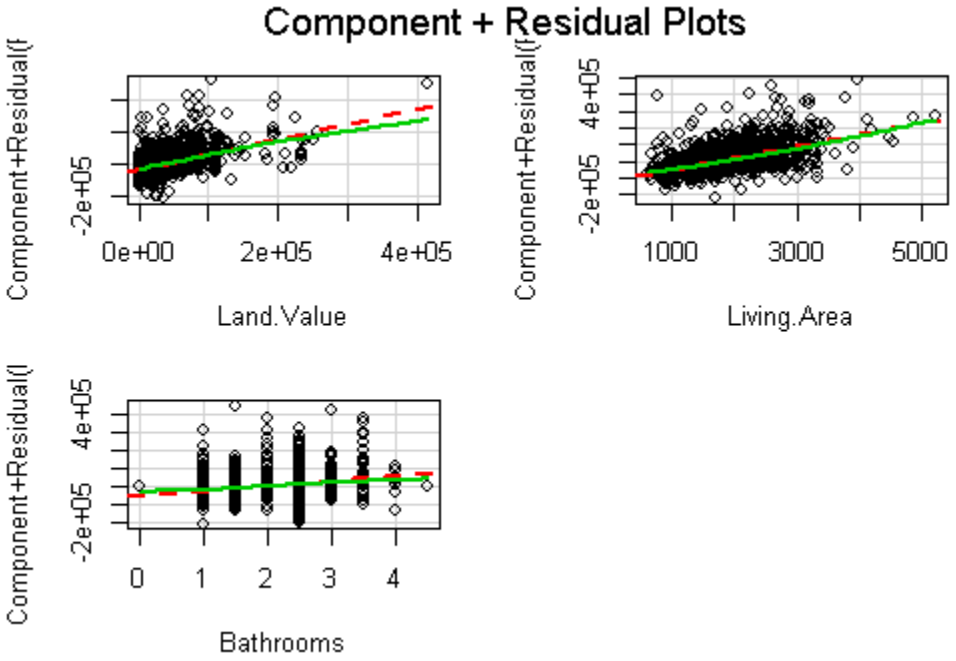
```
summary(model_signif)
summary(model_insignif)
```

10.

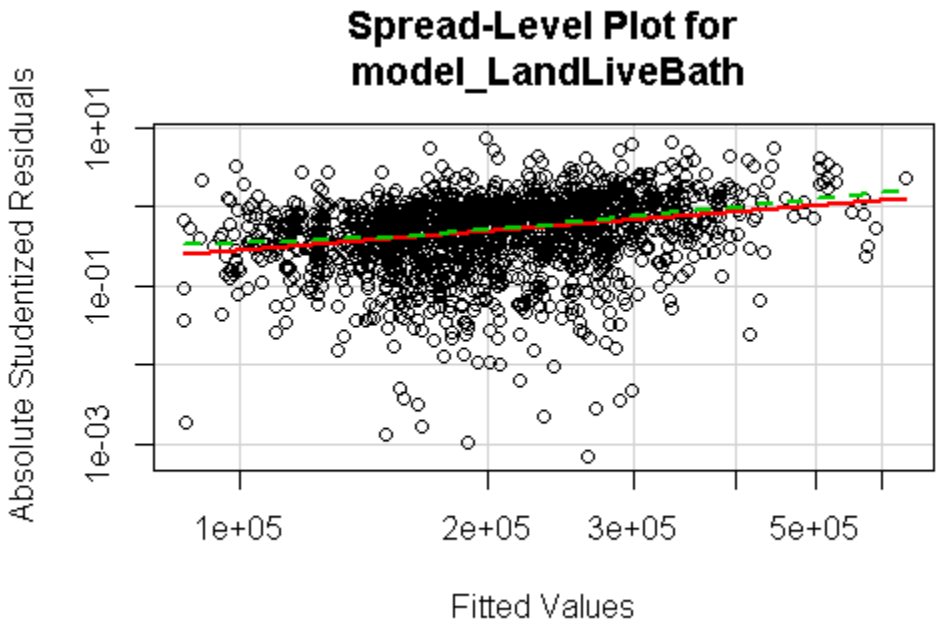
We have cross sectional data for just one year of home sales. The errors are probably independent just because the data is cross sectional. There are more formal ways to check this assumption that will be discussed later in the semester.

The below lines of code generate ways to check the other assumptions. Alternatively, you can use fitted vs residual plots, which are shown at the end of this question.

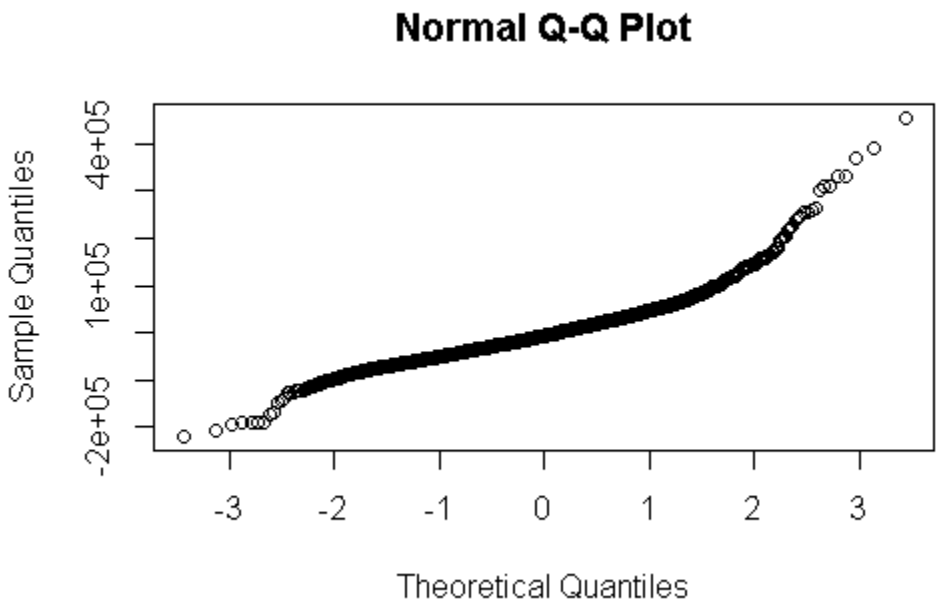
```
model_LandLiveBath<-lm(Price~Land.Value+Living.Area+Bathrooms)
crPlots(model_LandLiveBath) #This can be used to check linearity.
spreadLevelPlot(model_LandLiveBath) #This can be used to check heteroscedasticity.
qqnorm(model_LandLiveBath$residuals) #This can be used to check normality of residuals.
hist(model_LandLiveBath$residuals, xlab="Residuals", main="Histogram of Residuals") #This can also be used to check normality of residuals.
```



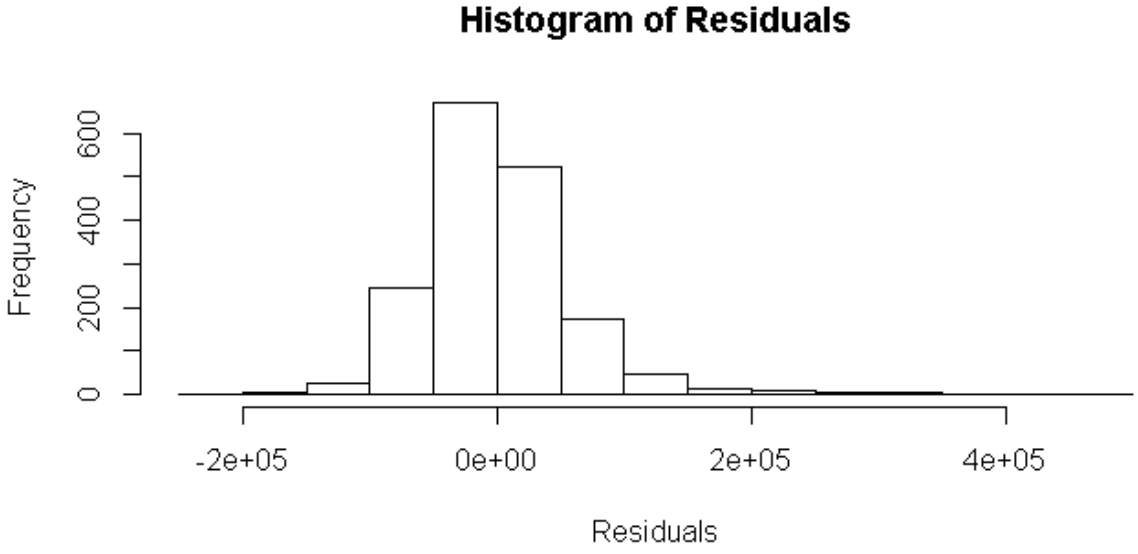
Above are plots that represent one way to check for linearity. These look okay, as the lines roughly match up.



Above is one way to check for heteroscedasticity (which is when the variance of the residuals is not constant). We want the lines to roughly match up, and for both of the lines to be roughly horizontal and pass through 0. The results here aren't great. It is possible we need to do a transform, or that some of our points are outliers (which we will discuss later).



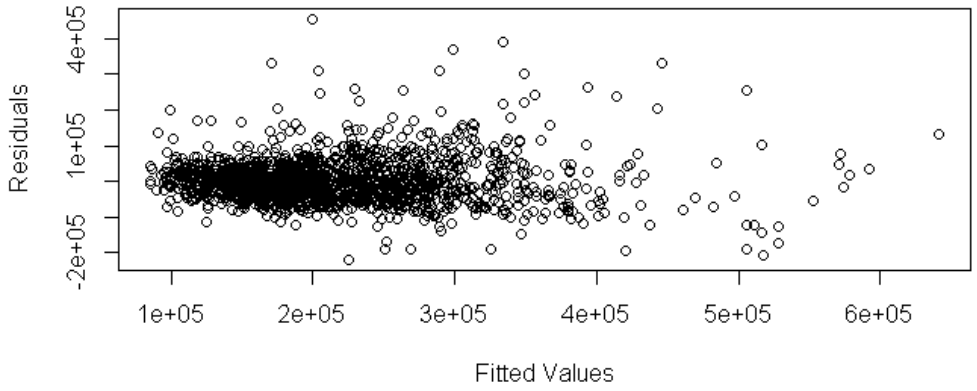
Above is one way to check for the normality of the residuals. For a QQ plot, we want the data to roughly conform to a 45 degree straight line. This also don't look too good, and could be caused by the possible heteroscedasticity from above.



The above is another way to check for normality of the residuals. We want this to roughly make a bell curve shape. This is skewed to the right some, so this also doesn't look great.

A more simplistic way to check for linearity and heteroscedasticity is to use fitted vs residual plots. If there is a pattern in the residuals, it implies linearity. If the residuals are spread differently as we move from left to right (such as a fan shape), it implies heteroscedasticity.

```
plot(model_LandLiveBath$residuals~model_LandLiveBath$fitted.values,ylab = "Residuals", xlab="Fitted Values")
```



11. It is easiest to just make a new data frame with all the points you want to predict. The below code makes a data frame, then runs the predict function that predicts a price based on the variable values present in the data frame.

```
newstuff<-data.frame(Land.Value=20000,Living.Area=1800,Bathrooms=1)
predict.lm(model_LandLiveBath,newdata=newstuff)
```

Answer=176693

12. Assume that the homeowner of the hypothetical house were to convert an unused room to a bathroom, making it a 2 bathroom, 1800 square foot house with a $20,000 land value. By how much would you expect the price to increase? (1 pt)

The price would increase by the coefficient on bathrooms, which is roughly 27100, based on the summary function in R.

This can be checked by making another data point that has two bathrooms instead of one. We already have a data frame called newstuff that holds the variable names, because of the previous question. So, we can just add more rows to this data frame using the rbind function (this stands for row bind). We can then run the prediction using just the second row, which is done by entering newstuff[2,]. This tells R to use only the second row, but use all the columns.

```
newstuff<-rbind(newstuff,c(20000,1800,2))
predict(model_LandLiveBath,newdata=newstuff[2,])
```

This yields a new prediction of 203796.2, so it increased by 27103.2.

13. It goes up by an amount of 10000*$Beta_{Land\_Value}$, which results in a new prediction of 186311.8. The beta value on Land Value was positive, so it is unsurprising that the home value increased. Again, these beta values can be seen by running the summary function in R, and can be checked by creating another prediction just like in #12.

```
newstuff<-rbind(newstuff,c(30000,1800,1))
predict(model_LandLiveBath,newdata = newstuff[3,])
```

14. The confidence interval is [20831.05,33375.36]. I can be 95% confident that the true effect of a change in the number of bathrooms lies in this interval. The 90% interval is [21840.3,32366]. The 95% interval is wider because it is more certain that the true effect is contained is wthin that interval. If a higher confidence level is desired, it will result in a wider interval.

```
confint(model_LandLiveBath,level = .95)
confint(model_LandLiveBath,level = .9)
```

15. The residual is -41837.2. The model overestimates the actual price of this house by 41837.2. This can be found manually with the following code:

```
temphouse<-houses[10,]
newstuff<-rbind(newstuff,c(temphouse$Land.Value,temphouse$Living.Area,temphouse$Bathrooms))
temp<-predict(model_LandLiveBath,newdata=newstuff[4,])
temphouse$Price-temp
```

Or automatically with the following code:

```
model_LandLiveBath$residuals[10]
```

16.

The land assessment is 27000. The predicted value is roughly 219900. This is much lower than Zillow's estimate of 273,000. This house could have an unusual combination of factors that make it an outlier for this data set. Alternatively, the data we have available to us may be old, biased, etc. Any logical answer is accepted.

```
realhouse<-c(27000,1935,2)
newstuff<-rbind(newstuff,realhouse)
predict(model_LandLiveBath,newdata=newstuff[5,])
```

17. The zestimate is [254000,293000]. Our prediction interval is roughly [100500,339300]. The Zillow model is likely much more sophisticated and has access to better data, which could lead to them having a much smaller, more accurate prediction interval. Additionally, Zillow is showing a 70% confidence interval. This information can be found if you look through their data explanations. Any logical answer is accepted.

```
predict.lm(model_LandLiveBath, newdata=newstuff[5,],interval = "prediction")
```