PROJECT II ECON 494 F20

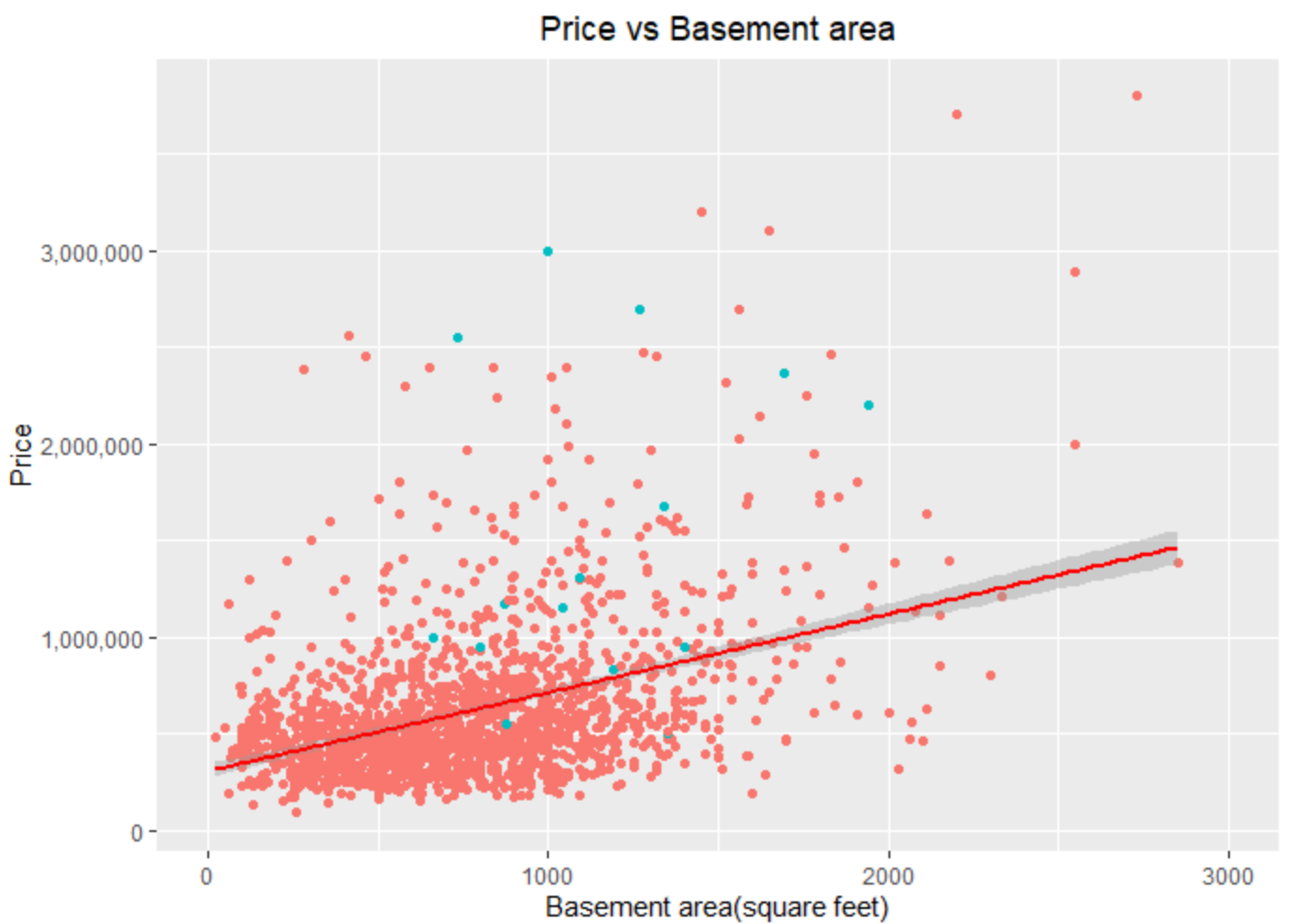
Dr. Levkoff

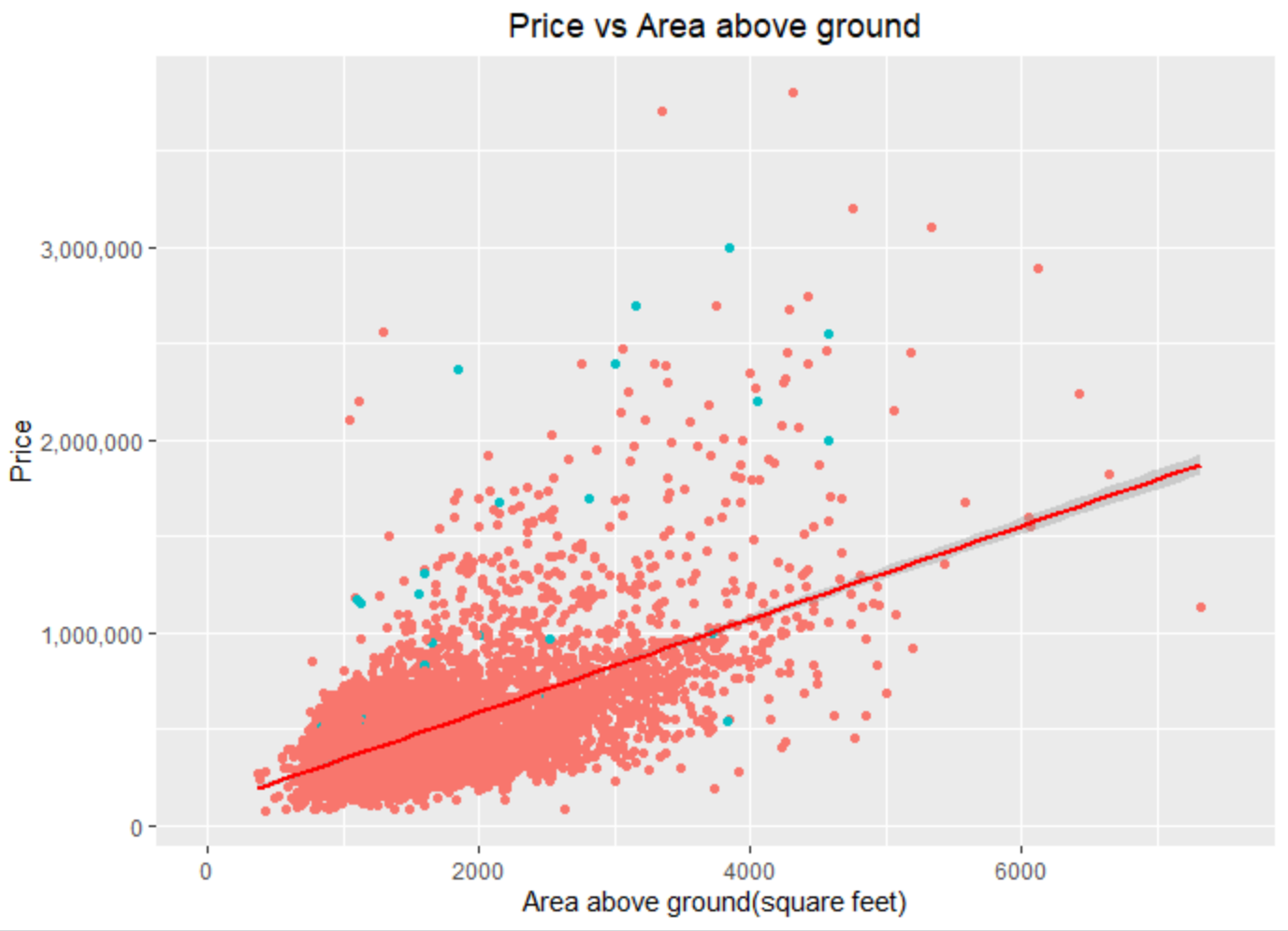
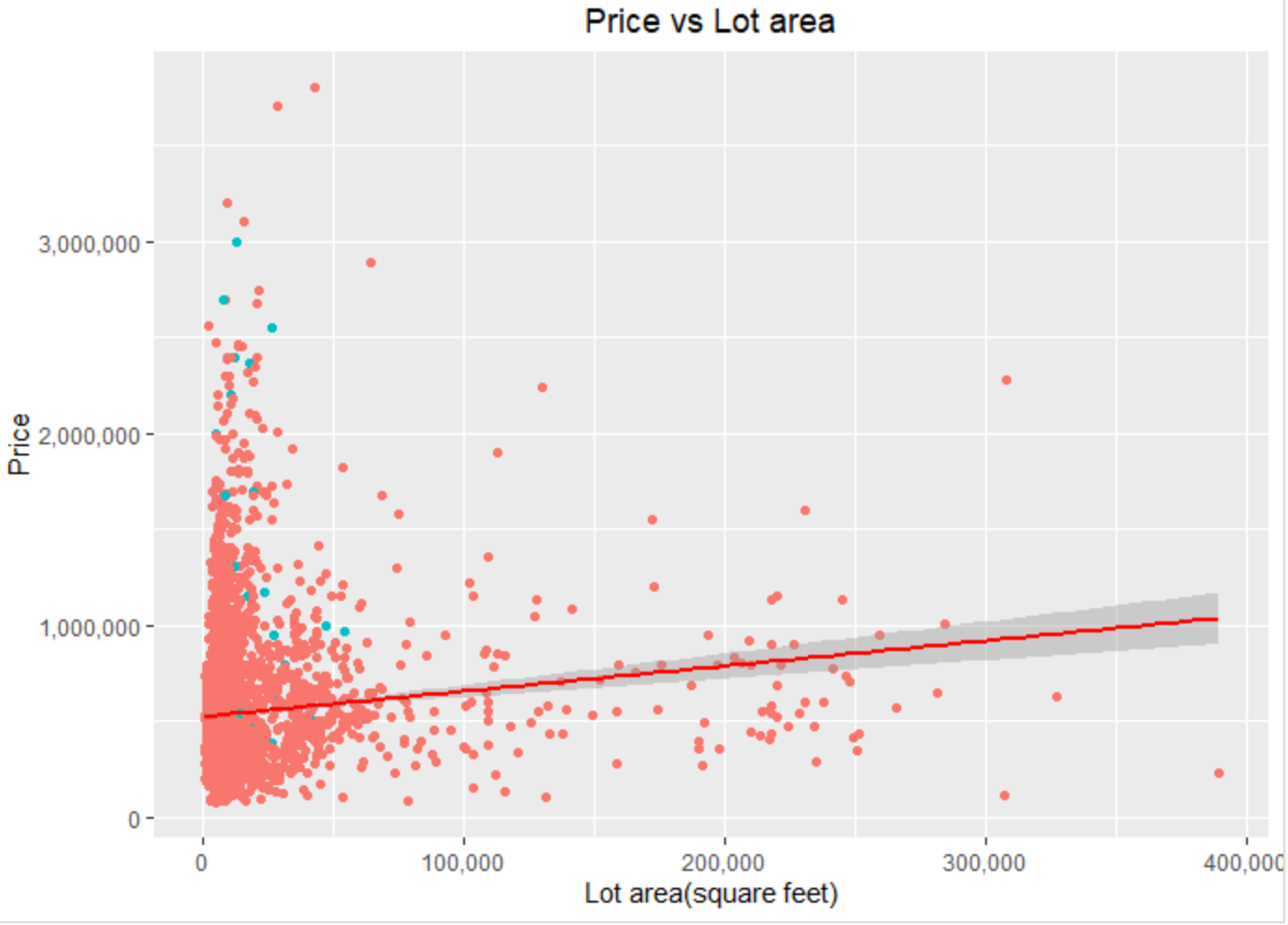
Xi Zhang(Simon)

11/18/2020

**Introduction**

Since my data is the same as the last project, the data collection process would remain the same. The variable of interest would be housing prices since our dataset has many different variables on the features of the house related to price. This dataset is a typical cross-sectional dataset having quantitative variables: “date”, “housing price”, “living square footage”, “lot square footage”, “above ground square footage”, and “basement square footage”. Also, there are categorical variables such as “number of bedrooms”, “numbers of bathrooms”, “floors”, “view”, “condition”, and a dummy variable “waterfront”. Recall the results from Project I that all the independent variables have a positive linear relationship with respect to price.





Also, since from the scatter plot above we could observe that the trend line is not entirely catching the price change, I decided to add the quadratic form to variables “living area” and “basement area”. Additionally, to avoid perfect collinearity, I excluded the variable “area above ground” since living area = area above ground + basement area. But I will explicitly use a regression model replacing the variable “basement area” with the variable “area above ground” to test which one will explain and predict price better. I also excluded the lot area since it is including the living area and the scatter plot shows a very strange relationship. Since none of the scatter plots have shown any logarithm function shaped trend, we will not convert any of our variables into a log form. At last, I separated 70% of the data into a training set and 30% into testing data.

**Linear Regression Models**

As we have observed that every independent variable has a positive linear relationship with the dependent variable, the first linear regression model I proposed is including all of them as original form:

Model 1

In the next regression model, I have included the quadratic form of sqft\_living to check if it could explain my model better.

Model 2

In the third regression model, I added the quadratic form of sqft\_basement, which I wished it can explain or predict the price better.

Model 3

In model 4, I replaced sqft\_basement with sqft\_above.

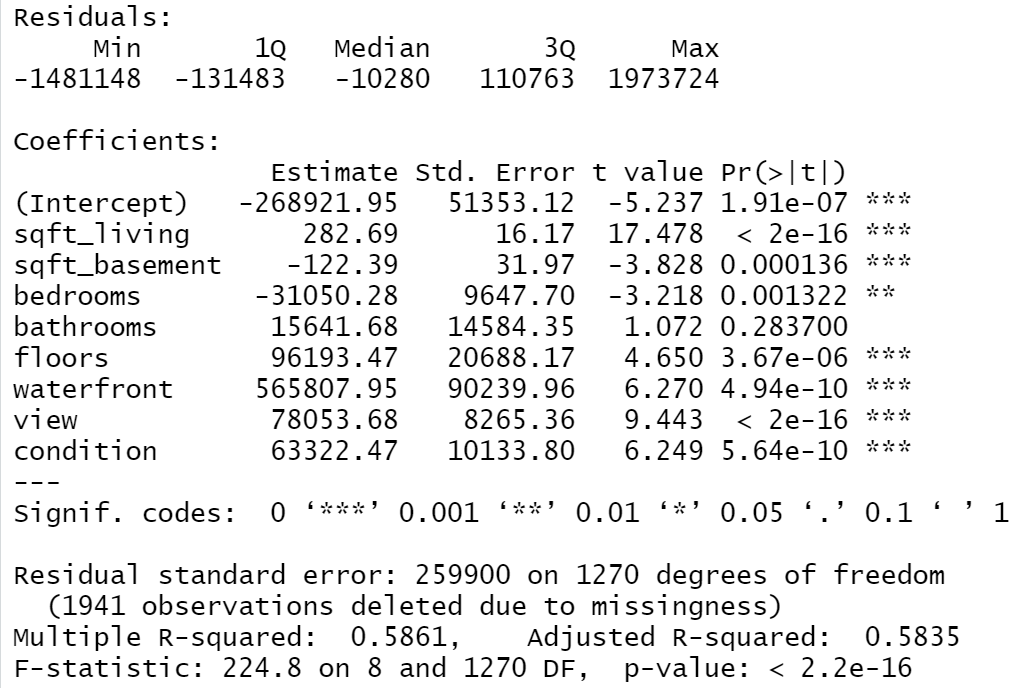
Model 4

In model 5, I added the quadratic form of sqft\_above to test if it would explain or predict the price better.

Model 5

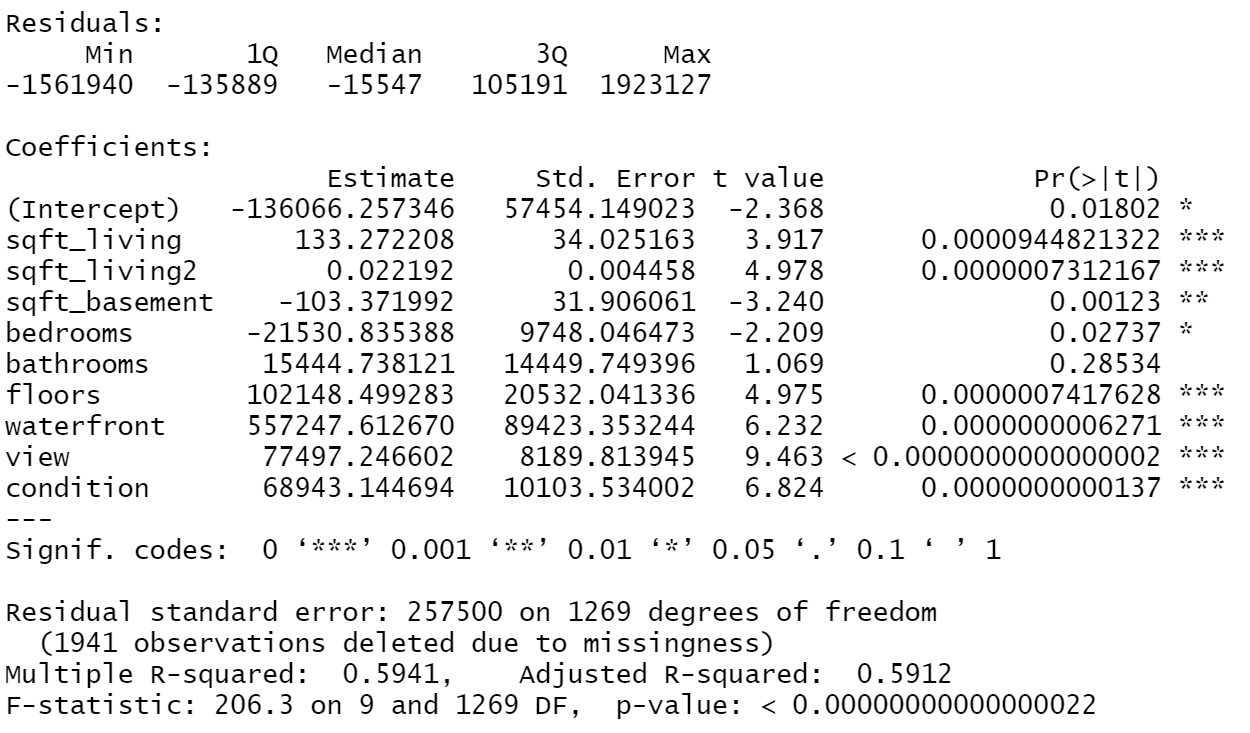
**Detailed results from Regression**

**Model 1:**



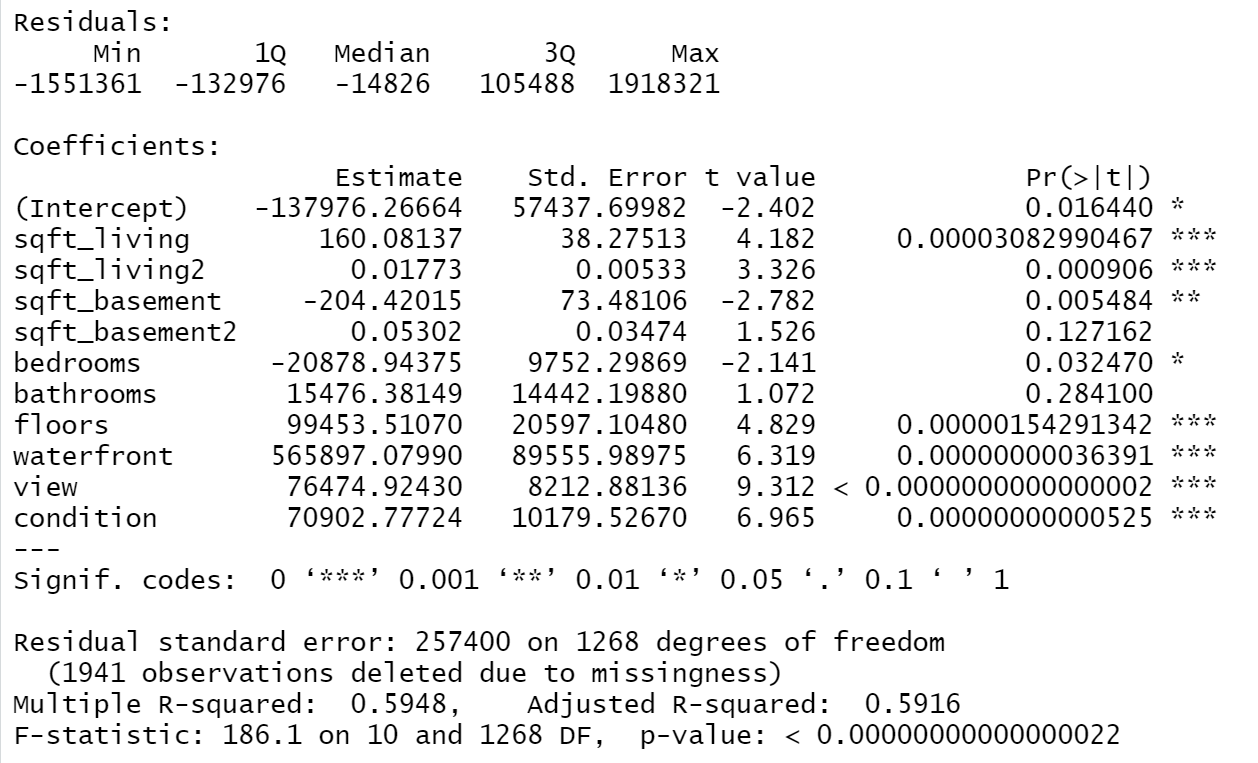
As the regression results from Model 1 show, this model explains 58.35% of the change in price. We can reject the null hypothesis that all of the regression [coefficients](https://www.statisticshowto.com/coefficient-definition/) are equal to zero from F-statistic. Also, all the variables were significant except for variable “bathrooms”. As expected, all the variables had a positive linear relationship related to the dependent variable. Surprisingly, two exceptions “sqft\_basement” and “bedrooms” had a negative coefficient whereas the area of basement and numbers of bedrooms increase, the price will go down. The reason is probably omitted variable bias since the house sold was different in years and locations. Also, it is possible that the old house tends to have more bedrooms and a bigger basement so the price will be lower. The actual equation would be:

**Model 2:**



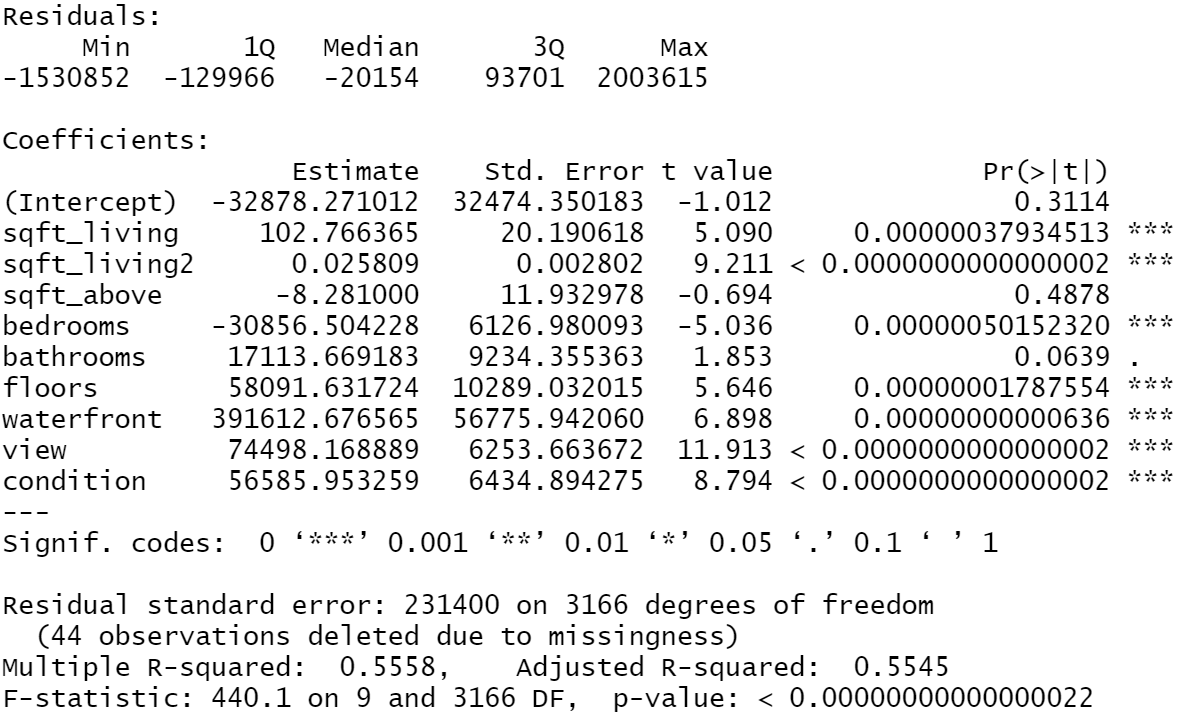
As the regression results from Model 2 shows, this model explains 59.12% of the change in price. We can reject the null hypothesis that all of the regression [coefficients](https://www.statisticshowto.com/coefficient-definition/) are equal to zero from F-statistic. Also, all the variables were significant except for variable “bathrooms”. As expected, all the variables had a positive linear relationship related to the dependent variable. Similar to the last model, two exceptions “sqft\_basement” and “bedrooms” had a negative coefficient whereas the area of basement and numbers of bedrooms increase, the price will go down. The reason is the same as above. The actual equation would be:

**Model 3**



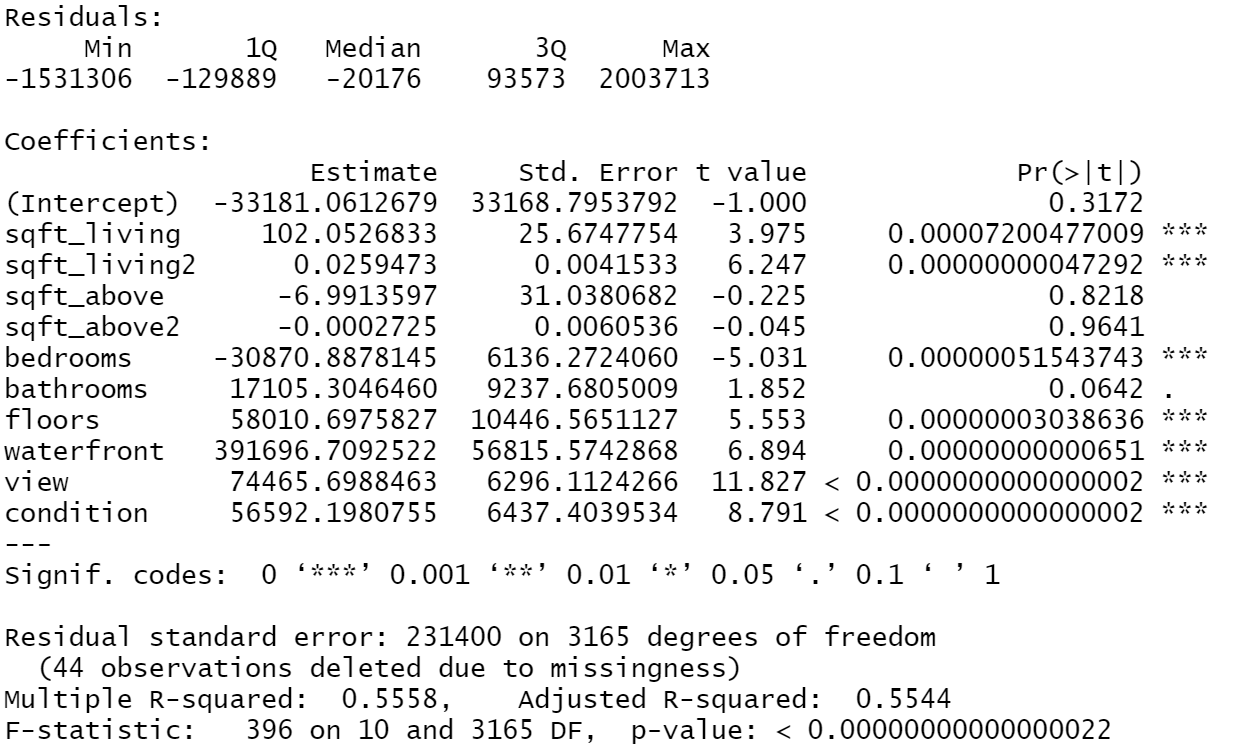
As the regression results from Model 3 shows, this model explains 59.16% of the change in price. We can reject the null hypothesis that all of the regression [coefficients](https://www.statisticshowto.com/coefficient-definition/) are equal to zero from F-statistic. Also, all the variables were significant except for variables “sqrt\_basement2” and “bathrooms”. As expected, all the variables had a positive linear relationship related to the dependent variable. Similar to the last model, two exceptions “sqft\_basement” and “bedrooms” had a negative coefficient whereas the area of basement and numbers of bedrooms increase, the price will go down. The reason is the same as above. The actual equation would be:

**Model 4**



As the regression results from Model 4 shows, this model explains 55.45% of the change in price. We can reject the null hypothesis that all of the regression [coefficients](https://www.statisticshowto.com/coefficient-definition/) are equal to zero from F-statistic. Also, all the variables were significant except for variables “sqrt\_above” and variable“bathrooms” was only significant at the 10% level. As expected, all the variables had a positive linear relationship related to the dependent variable. Similar to the last model, two exceptions “sqft\_above” and “bedrooms” had a negative coefficient whereas the area of basement and numbers of bedrooms increase, the price will go down. The reason is probably that old houses tend to have more bedrooms, which will result in a lower price. The actual equation would be:

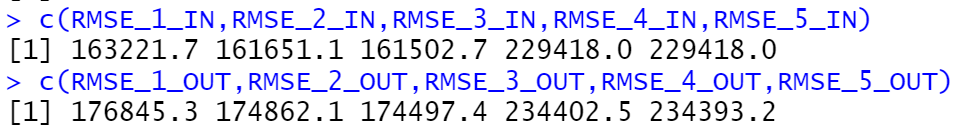
**Model 5**



As the regression results from Model 5 shows, this model explains 55.44% of the change in price. We can reject the null hypothesis that all of the regression [coefficients](https://www.statisticshowto.com/coefficient-definition/) are equal to zero from F-statistic. Also, all the variables were significant except for variables “sqft\_above”, “sqft\_above2”, and “bathrooms”. As expected, all the variables had a positive linear relationship related to the dependent variable. In this model, two exceptions “sqft\_above” and “bedrooms” had a negative coefficient whereas the area of basement and numbers of bedrooms increase, the price will go down. The reason is probably omitted variable bias since the house sold was different in years and locations. The actual equation would be:

We could observe from the adjusted R^2 that M3>M2>M1>M4>M5. Thus, from the training data, Model 3 is the best model in terms of explaining the change in the dependent variable.

**Predictions**



First, I chose to use root mean square error throughout the entire prediction process since I wanted the unit of the error to remain the same with my data. As we can see in the results above, the upper vector includes the in-sample-error, which shows that Model 3 had the smallest in sample error. This is consistent with our adjusted R^2 results from the last section. For the second vector, it included the out of sample error. We can observe that M3<M2<M1<M5<M4. Since we want the smallest error to have a better prediction, Model 3 would be the best predicting model. Surprisingly we have the same model for best testing data and training data. One explanation is that this model is the best model that we tested for both explaining and predicting the price. But it could be also a coincidence since we could only explain 59% of the change in price and it could be different if we add more variables.

**Conclusion**

In conclusion, since model 3 have the best performance in both explaining and predicting the housing price, it would be my best analysis model to use. But even though it is the best out of four models, it could only explain 59% of the change in price. Since there are variables “year” and “city”, I am supposing that the model would be better if I could add some fixed effect.