Regression on House Sale Price in King County, WA

Ethan Ancell, Kacy Craig, Christine Huffman, Samuel Rands

Introduction: In this project, we will be examining the relationship between the sale price of a home in King County, WA (including Seattle), and various information about the home, such as number of bedrooms, number of bathrooms, square feet of living room, total square feet, square feet of lot, whether it is a waterfront property or not, the year built, and the grade of the land. This data was found from the following link: https://www.kaggle.com/harlfoxem/housesalesprediction

Exploration: Running a preliminary OLS regression showed that this data set has a plot of potential in a linear model.

Root MSE	219033	R-Square	0.6442
Dependent Mean	540088	Adj R-Sq	0.6441
Coeff Var	40.55512		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	6840365	117498	58.22	<.0001	0
bedrooms	1	-41956	2039.73798	-20.57	<.0001	1.62124
bathrooms	1	55747	3344.92127	16.67	<.0001	2.98959
sqft_living	1	176.62089	3.28211	53.81	<.0001	4.09340
sqft_lot	1	-0.25121	0.03674	-6.84	<.0001	1.04313
waterfront	1	722210	17416	41.47	<.0001	1.02280
yr_built	1	-3879.49370	61.86850	-62.71	<.0001	1.48772
grade	1	130701	2118.33606	61.70	<.0001	2.79305

Figure 1: Preliminary output from ordinary least squares showing R-Square and parameter estimates for included predictors.

All of the variables that are included have a significant p-value, which tells us that they each are almost certainly useful in making a prediction in the response (sale price of home). The average variance inflation factor is a little bit above one, but it is probably not high enough to have a huge affect on the model. In the future, we may explore addressing the slight multicollinearity, but there are larger issues to take care of in influential points and outliers. Diagnostics for these are shown below.

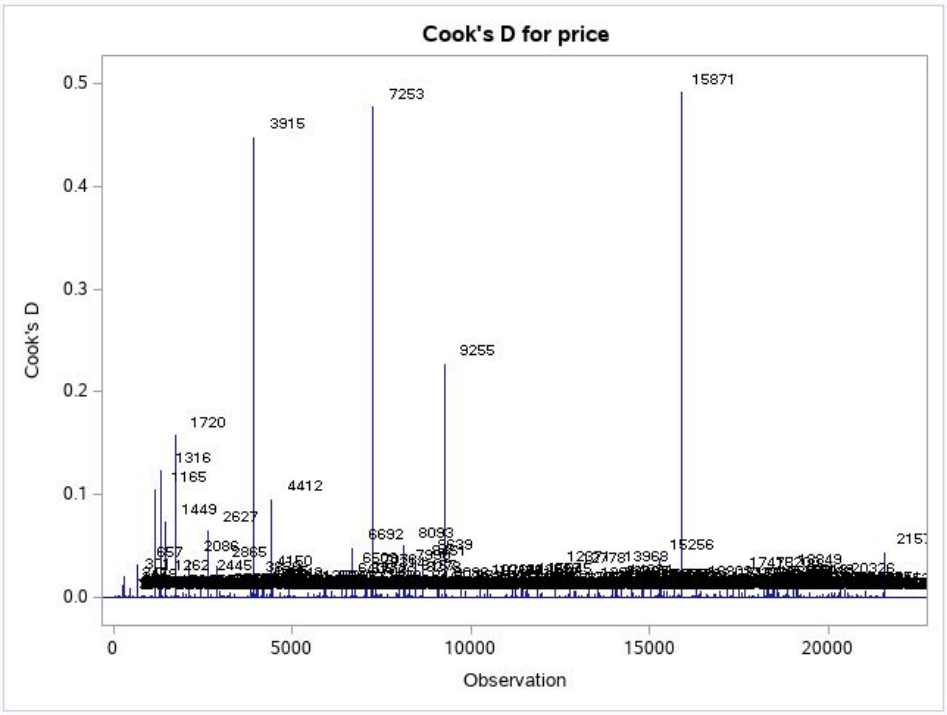

Figure 2: Cook's D diagnostic to identify influential points.

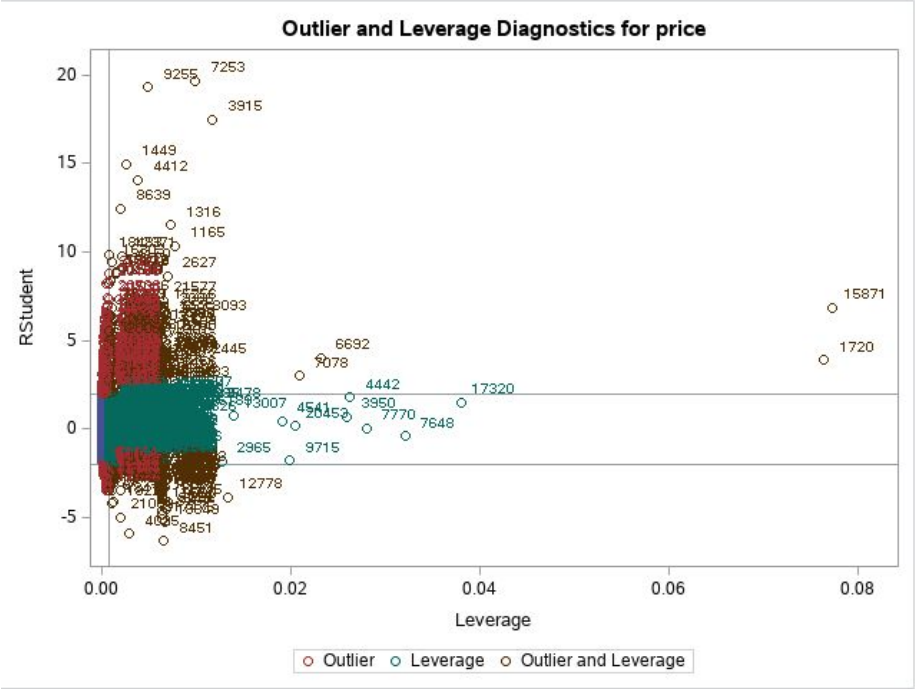

Figure 3: Diagnostic to show outliers and leverage from problematic points.

It appears that there are a few influential points in the data. King County is home to billionaires like Bill Gates and Jeff Bezos, so some neighborhoods are expected to have extremely high sale prices compared to a typical neighborhood. The plan to deal with these influential points is to use robust regression as to not give these points a lot of weight in determining the model coefficients.

Methods: Our proposed regression approach to deal with influential points is robust regression. There are some extremely expensive houses in King County that affect the model as a whole. Robust regression seems to be an effective measure in dealing with these points. Note that the model predicts sale price from bedrooms, bathrooms, square footage of living room, square footage of lot, if it's waterfront, total square footage, year built and grade.

Implications: Our model will be useful for predicting the sale price of houses in King County based on number of bedrooms, number of bathrooms, square footage of living room, total square feet, square feet of the lot, if it is waterfront property, the grade of the lot, and the year built.

Exploratory code used:

```
proc import
datafile='/folders/myfolders/EPG194/stat5100project/kc_house_data.csv'
replace
      out = house
      dbms = CSV
      ;
run;

proc reg data=house plots(maxpoints=220000 label)=(CooksD
RStudentByLeverage DFFITS DFBETAS);
      model price = bedrooms bathrooms sqft_living sqft_lot waterfront
yr_built grade / vif collin;
run;
```