Part 4

Nick George

5/3/2018

Nick George and Spencer Louie

Introduction

This study is based on a data set of home sales in King County, Washington from 2014 to 2015 (May) and comes from the Center for Spatial Data Science. The relevant variables are price, the sale price of the home; sqft_living, the square footage of living space; sqft_lot, the square footage of lot space; the number of floors; whether the property is watefront; the number of bedrooms/bathrooms, and the condition of the house (based on King County's grading system). From this data we are hoping to infer more about the general population of home sales in the U.S. As such specific variables like condition, will be extrapalated as roughly how much the structure of the house matters rather than the specific value because it is based on King County's system. The goal of this research is to figure out which factors are important in deciding the sale price of a home.

We ran ridge regression and lasso models to compare coefficients with our MLR model. The two lists below show the RR and lasso coefficients, respectively.

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                                              1
## (Intercept)                       8.563047e+00
## floors                            8.729544e-02
## waterfront                        3.778780e-01
## view                              5.662611e-02
## condition                         3.592832e-02
## grade                             2.399240e-01
## sqft_basement                     5.237174e-05
## bedbath                          -2.655432e-02
## bedbathi                          2.365024e-02
## lsqft_living <- log(sqft_living)  4.118588e-01
## lsqft_lot <- log(sqft_lot)       -4.000150e-02
## yr_built.adj <- yr_built - 1899  -5.153840e-03
```

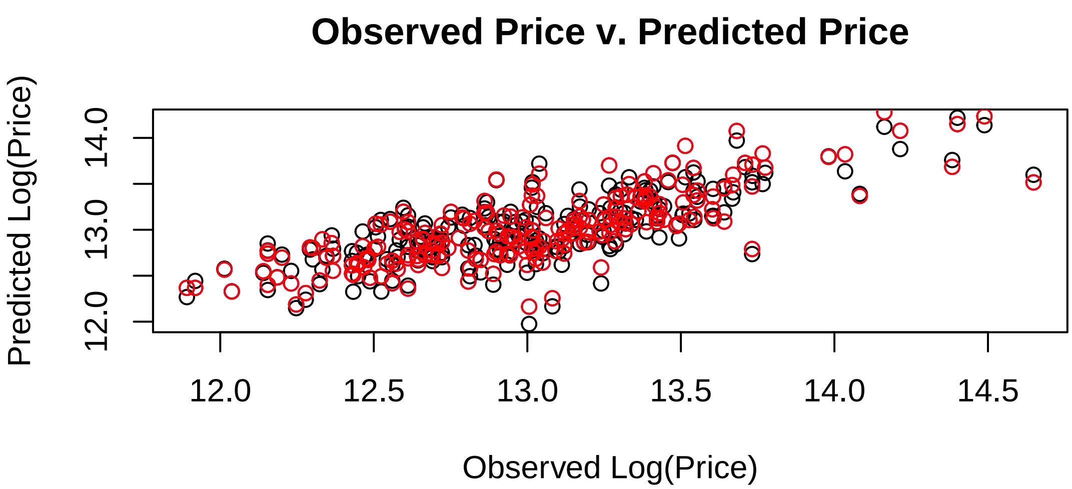
```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                                              1
## (Intercept)                       8.564988e+00
## floors                            8.723517e-02
## waterfront                        3.778592e-01
## view                              5.662709e-02
## condition                         3.590159e-02
## grade                             2.399753e-01
## sqft_basement                     5.236779e-05
## bedbath                          -2.559326e-02
## bedbathi                          2.278571e-02
## lsqft_living <- log(sqft_living)  4.115004e-01
## lsqft_lot <- log(sqft_lot)       -3.998905e-02
## yr_built.adj <- yr_built - 1899  -5.154885e-03
```

The ridge regression model suggests that we use all of the coefficients just as we would expect. The lasso model similarly suggests we use all of the variables which includes one more than we did in our own model.

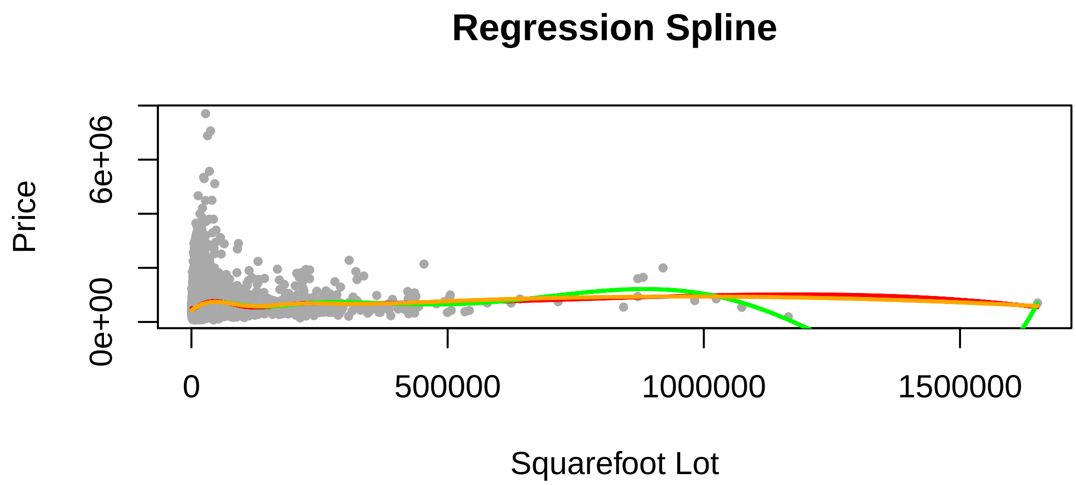
It suggests that we should include the amount of square footage in the basement, as well as the year the home was built. Note that the year built variable has been slightly adjusted so that 1900, the oldest home, is considered year 1 and then goes up normally from there. We then compare this with our full multiple linear regression model. Also because ridge regression and lasso both reccomend using all of the variables their coefficient estimates are nearly identical.

```
##      (Intercept)           floors       waterfront             view
##            8.140           -0.015            0.380            0.087
##        condition            grade          bedbath log(sqft_living)
##            0.094            0.202           -0.019            0.482
##    log(sqft_lot)
##           -0.054
```

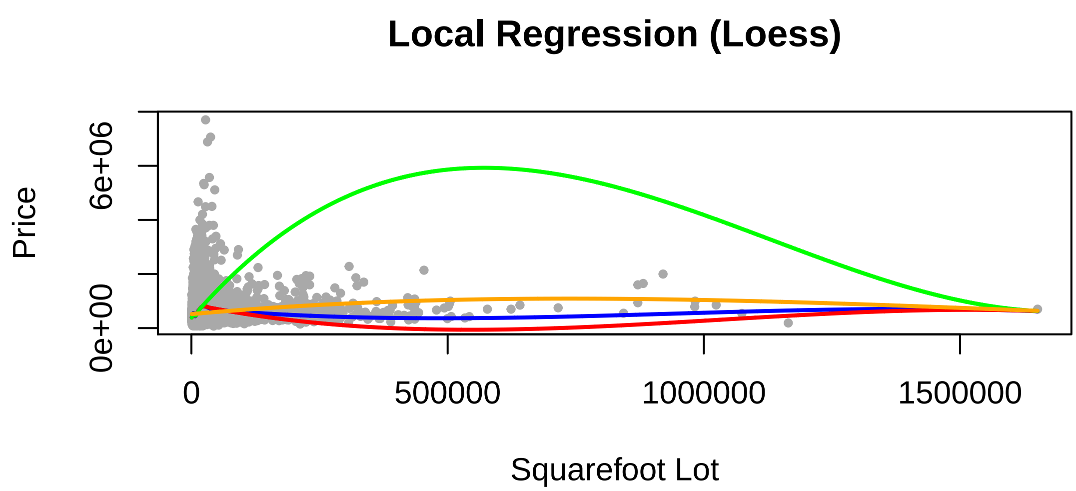
All of the coefficients for predictors in all three models are the same across the model in terms of direction, though not necessarily in magnitude (but not great changes in magnitude either). Except for floors, which went from having a negative coefficient in our multiple linear regression model to a postive one in the lasso and ridge regression models. We originally did not include the variables added in these models for logistical reasons. First the amount of squarefootage in the basement was problematic because it did not include a way to qualify when the basement was finished or not. Furthermore for the year the house was built, some of the homes have had renovations, but we have no way of qualifying the quality or scale of renovations. And due to those problems we did not originally include those predictors.



This graph shows the true values of the log of price on the x axis and the fitted values for each of the regression types on the y axis. Ideally we would a one-to-one relationship or a slope of 1.. In black are multiple linear regression values, in blue are the ridge regression values and in red are the lasso values. The data plotted comes from test data and the model was built using training data. We split our data into those two groups becasue we have so many observations. That is also why the plot has far fewer points than the overall data set. This separation allows us to better analyze how accurate our predictions are. The lasso and ridge regression models are very similar as you can see in the table with their coefficients, so in most places they are just overtop of each other. The black multiple linear regression points are also quite close to their blue/red counterparts implying that the models predict fairly similarly. Overall we have decided to stick with our originally multiple linear regression model because we are warry of the two logical reasons for excluding the two added variables in ridge regression and lasso and including them may cause some kind of unintended bias.



The graph above illustrates four different possible regression splines for square footage of lot against price. All four models contain knots at 100,000, 200,000 and 300,000 in an effort to better explain the early variability. The blue line contains those three knots at degree 3. The red line contains those three knots as well as another early knot at 10,000, at degree 3 in an effort to sort out the variability we see extremely early on. The orange line adds a knot at the end of the blue model, at 750,000 to look at the differences at the higher levels. Lastly, the red line contains the blue model's knots at degree 5. We see that all four models seem to be fairly similar, but each has a slight difference based on the added portion to it. For example the orange is line is far more straight in the middle to end with it's knot at 750,000 and the green line has more curvature due to it's increased degree.



The graph above shows another way to smooth our curve in a local regression method called Loess. The four lines predict on the same variables for the same values, but differ in the span for each curver. The blue curve has a span of .2, the red curve has a span of .5, the green curve has a span of .7, and the orange curve has a span of 1. The span effects the curve in that it dictates how many of the nearby points are used to predict the line, amongst those points the ones closest to it are given higher weights, and that is again dependent on the span. The smaller the span, the closer the points have to be to be considered in the local regression, and the lower the weight they are given. We see that he blue, orange, and red lines seem to fit the data fairly well, while the green line seems to be fairly off. This likely is because the points that were used for the local regression, particularly in the middle, caused a strange fit. The blue and red curve's use only the data right around each point to predict to it seems to fit the data well. The orange curve uses all of the data and runs the regression (though with weights) and also seems to fit alright.

Overall we would the blue Loess model to predict future values. We think it is important to have a low span so that you are only regressing using the points right around the point in question. This is especially true for this variable because the data does not have a clear functional form. That is also why we are not choosing to use regression splines because, again, there is not a clear functional form for the data, so doing a local regression seems more beneficial.

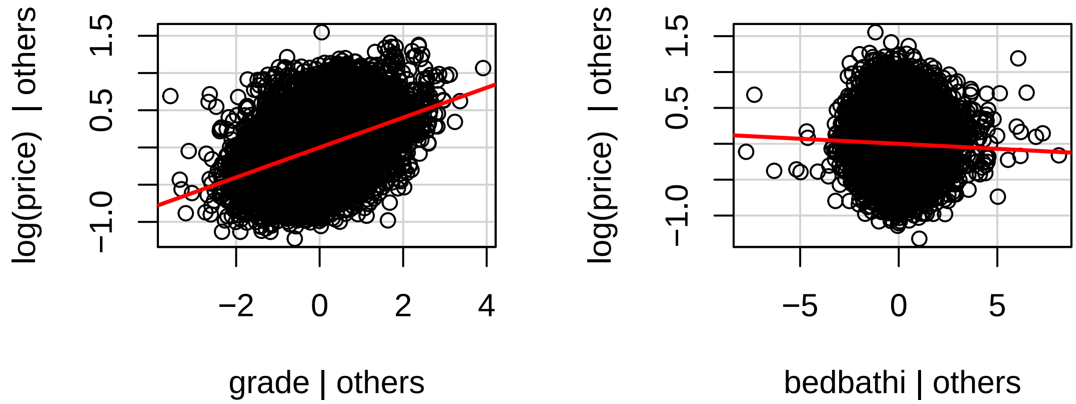
Conclusion

Ridge regression and lasso yielded the results we were expecting. With the number of observations far outweighing the number of predictors, it was expected that the RR and lasso to find all variables significant to the model. However, due to the way the variables are constructed it makes more sense to use the MLR that was determined from the last part of the project. Further, given the type and number of variables we have, RR or lasso are solving a problem that is not present in the data.

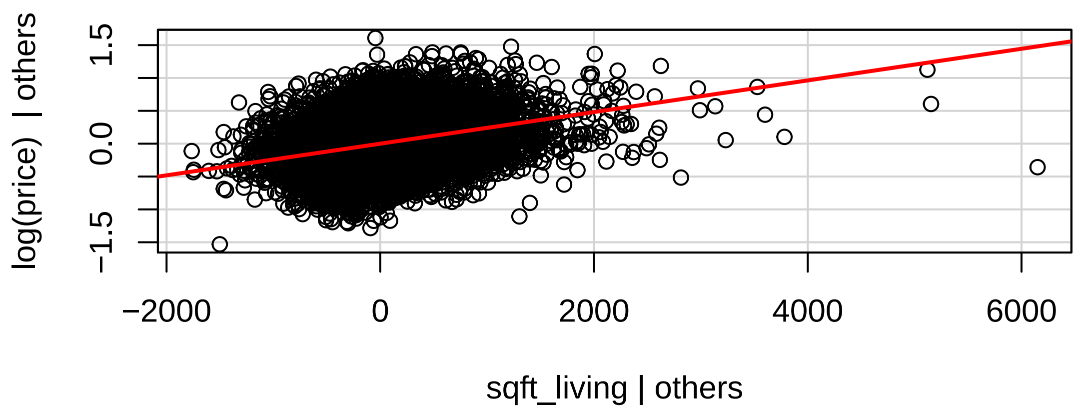
Loess and splines also yielded expected results. Choosing the appropriate number and location of knots was an important determination, but ultimately the spline curves looked quite similar. Much more important was choosing the span of the Loess, which made the smoothed curve vary widely. Overall, smooth regressions are much more appropriate for the kind of data we have compared with sparse models. Choosing an appropriate smooth regression function yielded curves that fit the data well. Loess seemed to be the optimal method because it did not attempt to fit an actual spline to it rather running individual local regressions along the way. We also picked the smallest span so that only points immediately surrounding the highlighted part would be considered in the local regression.

Added Variable Plots

One factor that we are particularly cognisant of in our research is the fit of our model. Without an approriate fit the inference of our predictors is fairly meaningless and inaccurate. It is also important because the shape of our model is not necesarilly obvious. So we have decided to look at a few things to help us ensure an appropriate model. First, we decided to look at added varaible plots. The premise behind these plots is to see how and if an additional predictor matters when you consider the other predictors already included in the model. We are more interested in the how part of the added variable plot than the if. The plot gives us the ability to again help evaluate fit. We can see if the fit is linear or something else. Fit is extremely important for our model and any model, because if the model does not accurately represent the data then the inferences from it will be innaccurate. Our model fit is particularly important because it is not obviously linear as is, and to address that we have already made several transformations to the data. The plots show the residuals of one variale given that the other variables are in teh model. Each point is $e_i(Y|X_2 + X_3 + X_4) = Y_i - \hat{Y}(X_2) - \hat{Y}(X_3) - \hat{Y}(X_4)$ on the Y axis. On the X axis it is, $e_i(X_1|X_2 + X_3 + X_4) = X_{i1} - \hat{X_{i1}}(X_2) - \hat{X_{i1}}(X_3) - \hat{X_{i1}}(X_4)$ In order to get these points we are doing a regression so we do require the normal assumptions or technical conditions for a regression, normality and constant variance for error terms, independence and linearity. With our dataset being so large we are able to satisfy the normality condition and we have checked for the previous assumption earlier.



Here we have a couple added variable plots for a couple variables of interest. Since none of them have a horizontal line we do see a relationship for all of them and therefore each one adds something to the plot. Bedbath seems to have a negative relationship while grade has a positive relationship. We also see the the lines themselves as well as the points seem to imply linear relationships, the lines are straight and the points are not mostly above or below the line at certain parts, rather they are even distributed at ever segement of the line.



Here we have a graph with all parts of the model included except for sqft_living, which we are looking at without a transformation. We do see that the line is straight, but the points are not as evenly distributed as we have seen the in past. Earlier and later on it seems more points are below and in the middle more are above. The curvature implies we should use a transformation, which we have done. These graphs are fairly hard to read since they include all of the data in the training set, so we have decided to include another method to more accurately and directly evaluate fit.

Lack of Fit

We decided to look at the Lack of Fit test. This is an F-Test with the simple goal of evaluating fit. One of the convenient aspects of this test is that it does not require any additional assumptions then the one we have used to run our regressions as it is simply looking at two regressions. In fact, it does not even require linearity as that is what it is meant to test for. This test uses a nested F-test, so it is comprised of both a full model and a reduced model. The full model looks at a regression where instead of assuming a linear relationship it counts each variable as a factor variable, so it measures the mean Y at each level of a variable. In other words what is the average value of Y when X1=1. The reduced model is the regression you wish to test, normally one that assumes linearity and so we would have our normal βs telling us the slope of a predictor instead of giving as a value at each level of it. This model only works if the variables you are testing are categorical in this sense. There must be multiple observations at each level. Below we have run this test on our variables that fit these requirements. Note that sqft_living and sqft_lot are not run as factor variables because they are continuous and there are not multiple observations at each level.

```
## Analysis of Variance Table
##
## Model 1: log(price) ~ log(sqft_living) + log(sqft_lot) + factor(floors) +
##     factor(waterfront) + factor(view) + factor(condition) + factor(grade) +
##     factor(as.integer(bedbath))
## Model 2: log(price) ~ log(sqft_living) + log(sqft_lot) + floors + waterfront +
##     view + log(condition) + grade + bedbathi
##   Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
## 1  21351 2269.7
## 2  21386 2424.4 -35   -154.72 41.583 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we find that we reject the null hypothesis, that the model fits, because our p-value is near 0, thus implying that the model does not fit. So instead of assuming a linear relationship with these stepped variables we will instead consider them as factor variables, finding the mean price at each given level, for example the mean price for a home with 5 beds and baths. The results for such a regression are shown below.

```
##                              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                    9.4875     0.3349  28.3258   0.0000
## log(sqft_living)               0.5316     0.0114  46.5243   0.0000
## log(sqft_lot)                 -0.0642     0.0030 -21.6534   0.0000
## factor(as.integer(floors))2  -0.0713     0.0062 -11.5648   0.0000
## factor(as.integer(floors))3   0.0134     0.0151   0.8864   0.3754
## factor(waterfront)1            0.3963     0.0323  12.2864   0.0000
## factor(view)1                  0.2188     0.0184  11.8935   0.0000
## factor(view)2                  0.1506     0.0111  13.5653   0.0000
## factor(view)3                  0.1938     0.0152  12.7310   0.0000
## factor(view)4                  0.2977     0.0236  12.6334   0.0000
## factor(condition)2            -0.0414     0.0695  -0.5961   0.5511
## factor(condition)3             0.0448     0.0651   0.6887   0.4910
## factor(condition)4             0.1198     0.0651   1.8395   0.0659
## factor(condition)5             0.2420     0.0655   3.6963   0.0002
## factor(grade)3                -0.9646     0.4252  -2.2686   0.0233
## factor(grade)4                -0.6598     0.3668  -1.7986   0.0721
## factor(grade)5                -0.6637     0.3621  -1.8331   0.0668
## factor(grade)6                -0.5102     0.3619  -1.4097   0.1586
## factor(grade)7                -0.3242     0.3620  -0.8957   0.3704
## factor(grade)8                -0.1245     0.3621  -0.3439   0.7309
## factor(grade)9                 0.1147     0.3623   0.3167   0.7515
```

```
## factor(grade)10                 0.3167     0.3625   0.8737   0.3823
## factor(grade)11                 0.4924     0.3628   1.3570   0.1748
## factor(grade)12                 0.7006     0.3638   1.9259   0.0541
## factor(grade)13                 1.0220     0.3749   2.7263   0.0064
## factor(bedbathi)1               0.8687     0.2215   3.9213   0.0001
## factor(bedbathi)2               0.4486     0.1371   3.2715   0.0011
## factor(bedbathi)3               0.4061     0.1350   3.0081   0.0026
## factor(bedbathi)4               0.2488     0.1347   1.8469   0.0648
## factor(bedbathi)5               0.2269     0.1347   1.6850   0.0920
## factor(bedbathi)6               0.1828     0.1346   1.3583   0.1744
## factor(bedbathi)7               0.2321     0.1348   1.7214   0.0852
## factor(bedbathi)8               0.2318     0.1348   1.7191   0.0856
## factor(bedbathi)9               0.3112     0.1364   2.2816   0.0225
## factor(bedbathi)10              0.2976     0.1372   2.1682   0.0302
## factor(bedbathi)11              0.2969     0.1458   2.0358   0.0418
## factor(bedbathi)12              0.4079     0.1566   2.6050   0.0092
## factor(bedbathi)13              0.1598     0.2127   0.7513   0.4525
## factor(bedbathi)14              0.5552     0.1854   2.9951   0.0027
## factor(bedbathi)15              0.0595     0.2694   0.2207   0.8253
## factor(bedbathi)16              0.0276     0.3556   0.0775   0.9382
```

Now we have significantly more coefficients as we have one for each factor level, or really we have one less than that as the intercept term contains all of the base factor levels, such as view=0 (not having a view) and grade=2 (the lowest grade in the data set). Some interesting notes are that grade is only significant for it's bottom three and final two (3,4,5 and 12 and 13), which implies that grades 6 through 11 all have the same effect on home price. So only when a home's grade is really high will it be important benefit in the price. Similarly condition is only significant for its two highest levels 4 and 5, so again only if the condition of a home is extraordinary will it be associated with a benefit in the sale price. when we look a the bedbath coefficients we can see them somewhat more like cutoffs. 6 is not necessarily significant but 7 is, so within a certain value for bedbath there is not necessarily an extra benefit, though having more is generally better as at certain points it becomes significant. View we find significant at every level and has positive coefficients so each factor higher in view is associated with an increase in price. This regresion provides a fairly good fit since we are only looking at specific factor levels, but having so many coefficients is not always helpful and we may instead want to look for some functional form for these variables so we have also decdided to look at Generalized Additive Models.

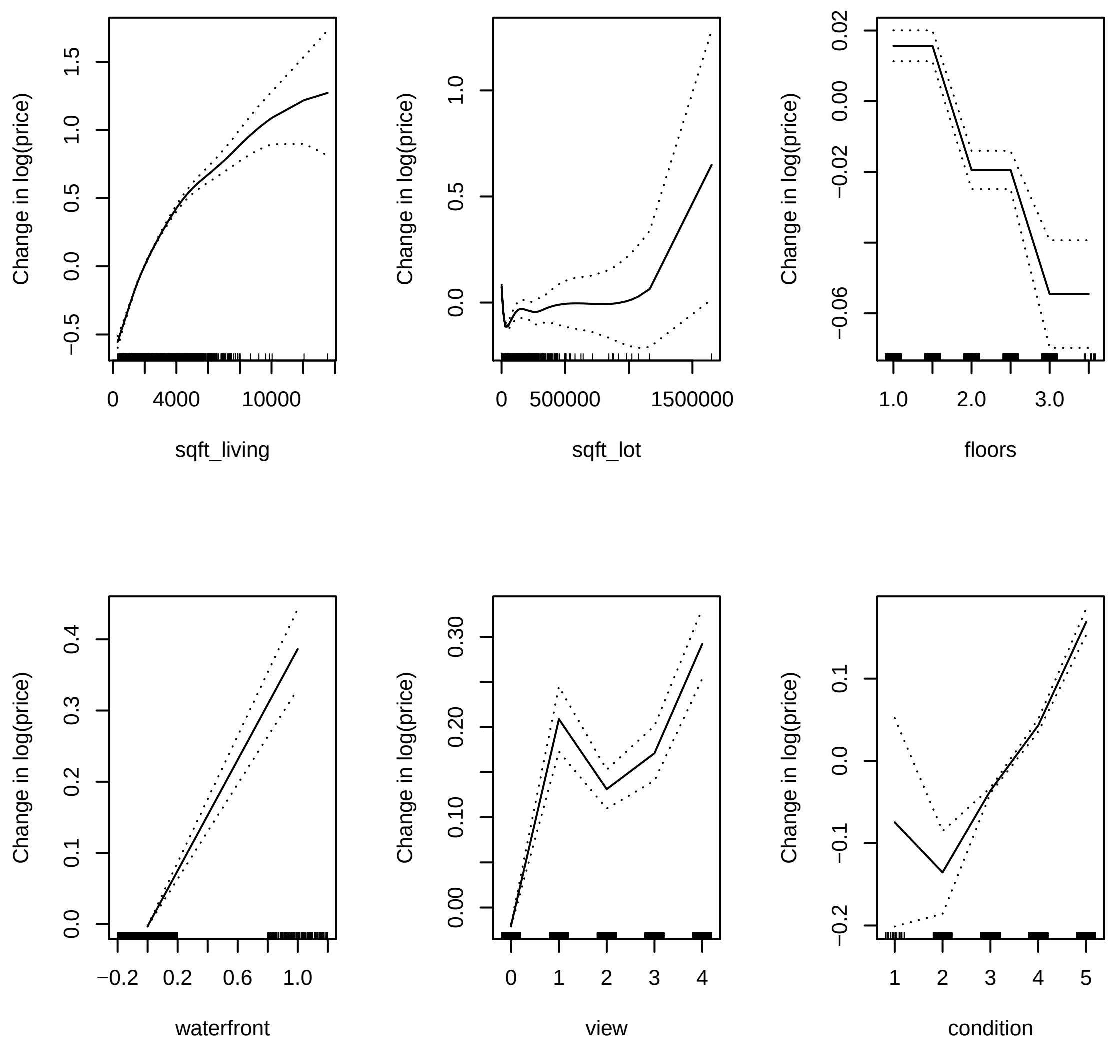
Generalized Additive Model

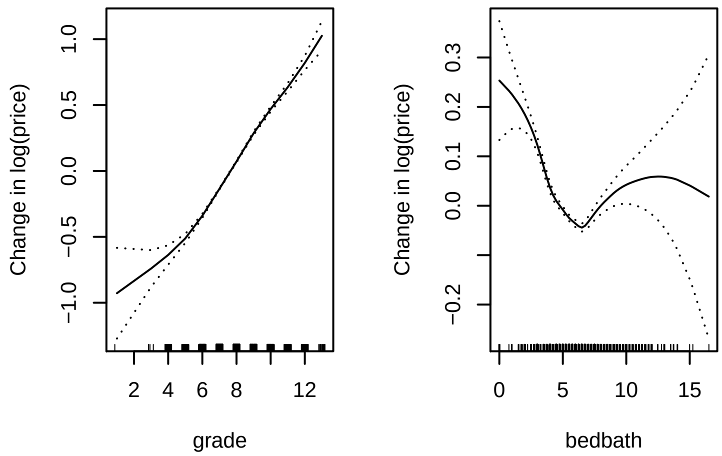
Generalized Additive Models help us get around problems of functional form. It is hard to tell what the form of each variable is specifically, but the GAM does not require us to pick a specific form (like linear) instead it uses unknown smooth functions. The GAM is of the form $g(E(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + f_3(x_3) + ... + .f_p(x_p)$ where g is some link function, in our case that will be a log function for price, β_0 is the intercept and f_i is the functional form for each predictor x_i. These functional forms can range from parametric to non-parametric. This allowance for smoother functions lets us deal with our data that is not clearly linear, but still allows us to define some functional form to it so that we can see an overall realationship. GAMs hold the same assumption as our normal linear models, except they relax two of them. First, and most obviously, is the need to use defined functional forms instead allowing us to using potentially non-parametric smoothers. Second, GAMs also have the flexibility to permit the use non-normal error distributions In our regression below we have used splines not unlike the ones we used above.

```
##
## Call: gam(formula = log(price) ~ s(sqft_living, df = 6) + s(sqft_lot,
##     df = 6) + as.integer(floors) + waterfront + s(view, df = 6) +
##     s(condition, df = 6) + s(grade, df = 6) + s(bedbath, df = 6),
```

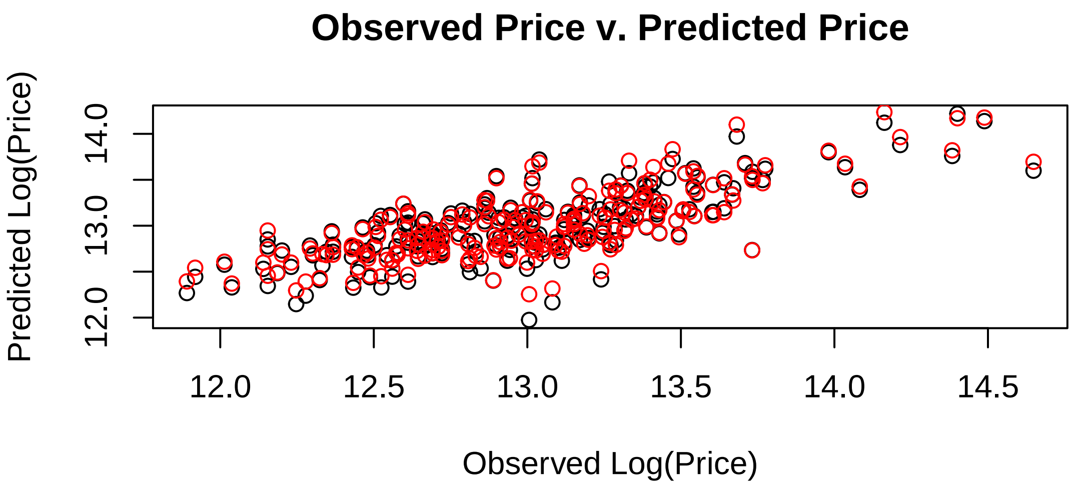
```
##     data = kc_data.train)
## Deviance Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2627 -0.2289  0.0128  0.2213  1.4074
##
## (Dispersion Parameter for gaussian family taken to be 0.1082)
##
##     Null Deviance: 5942.388 on 21394 degrees of freedom
## Residual Deviance: 2310.518 on 21360 degrees of freedom
## AIC: 13169.84
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##                           Df  Sum Sq Mean Sq    F value    Pr(>F)
## s(sqft_living, df = 6)     1 3025.08 3025.08 27965.9025 < 2.2e-16 ***
## s(sqft_lot, df = 6)        1    2.39    2.39    22.1103  2.59e-06 ***
## as.integer(floors)         1    1.02    1.02     9.4029  0.002169 **
## waterfront                 1   76.28   76.28   705.1820 < 2.2e-16 ***
## s(view, df = 6)            1   98.23   98.23   908.0762 < 2.2e-16 ***
## s(condition, df = 6)       1   46.99   46.99   434.4340 < 2.2e-16 ***
## s(grade, df = 6)           1  446.26  446.26  4125.4962 < 2.2e-16 ***
## s(bedbath, df = 6)         1    8.54    8.54    78.9600 < 2.2e-16 ***
## Residuals              21360 2310.52    0.11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                        Npar Df  Npar F     Pr(F)
## (Intercept)
## s(sqft_living, df = 6)       5  62.829 < 2.2e-16 ***
## s(sqft_lot, df = 6)          5 120.365 < 2.2e-16 ***
## as.integer(floors)
## waterfront
## s(view, df = 6)              3  25.522 2.220e-16 ***
## s(condition, df = 6)         3   6.617 0.0001831 ***
## s(grade, df = 6)             5   5.759 2.561e-05 ***
## s(bedbath, df = 6)           5  57.214 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The summary of the model above illustrates one of the major trade offs you make using GAMs, which is a loss of interpretation. We no longer have clear coefficients that illustrate the effect of each variable. Instead we have a series of F-tests showing that each individual predictor, in its form, is importantant to the model as they are all statistically significant at a .1% level. However, it is somewhat hard to tell what that effect is. Here we have used splines with six degrees of freedom for every varaiable, except floors and waterfront. That is because floors has only three levels 1, 2, and 3 floors, so a spline could not be created with such few levels and waterfront only has values of 0 or 1 so it does not make sense to use a spline as it simply turns "on" or "off". In order to better understand the effects of each predictor we have graphed the functions from the regression below.


Change in log(price)
sqft_living
sqft_lot
floors
waterfront
view
condition
0
4000
10000
500000
1500000
1.0
2.0
3.0
−0.2
0.2
0.6
1.0
0
1
2
3
4
5
−0.5
0.0
0.5
1.0
1.5
0.02
0.00
−0.02
−0.06
0.1
0.2
0.3
0.4
0.00
0.10
0.20
0.30
0.1
0.0
−0.1
−0.2



Now we can get at least some interpretation of the effect of each variable. We don't have specific coefficients, but we do have graphs. We can see that increasing the square footage of living space is always associated with a positive bump in price, but begins to taper off after awhile. The square footage of the lot on the other hand seems to start off slowly and then increase greatly in his positive association with price. Floors, is a simply linear function, for each level since we could not use a spline, but we do find that it is a negative effect. Waterfront is similarly linear, and is associated with higher prices. View is strange, it always associated with higher prices, but dips down at the second and third level, suggesting that once you have a view the price does not benefit much more until it becomes the best possible view, but still it is a somewhat strange result. Prices increase significantly as condition increases, crosing out of a negative effect at level 4. Grade has a very similar relationship with price. Finally bedbath starts off being associated with high prices, but that association, while still positive, tapers of dramatically. It is important to remember that these effects hold the other variables constant, which can help explain some of the results we have seen. If the square footage of the lot is held constant it makes sense that having a bigger and bigger home on it might have diminishing returns. While having the same sized home on a bigger and bigger lot could have greater returns. For bedbath, if we have the same sized home and keep adding more and more bedrooms and bathrooms to it, then eventually the home will be overpopulated with them so the negaive function makes a bit more sense.



The plot above attempts to evaluate whether this new Generalized Additive Model is indeed any better than the model we got from multiple linear regression. This plot is of the same form as the on we used to evaluate the ridge regression and lasso models. Here the original multiple linear regression's predictions are plotted in lack and the GAM predictions are in red. The predicions are again made on test data, while the model is created on training data. We are looking for a 1-1 relaionship, the predicted values are the same as the true values so we want the points to clustered around a line with the slope of 1. It appears that the red points are slightly more clusted and centered the black ones, suggesting that the GAM is better at predicting. To ensure that is the case, below we have also found the AIC for the multiple linear regression and generalized additive models. They are 14038.84 and 13137.35 respectively. Therefore, the generalized additive model did indeed do a better job of predicting.

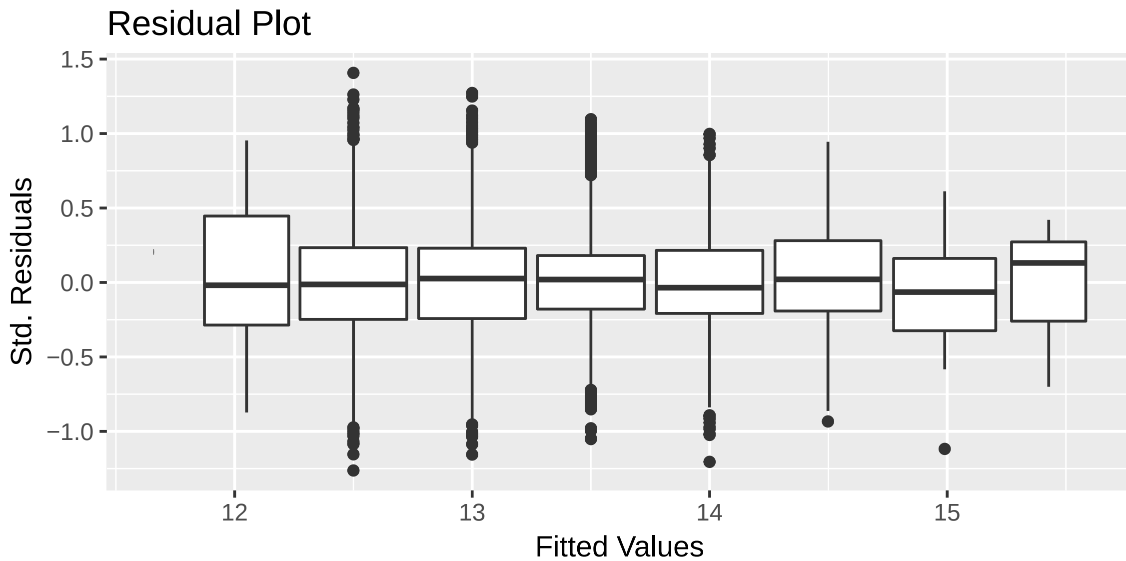
Summary

One of the most interesting findings from our analysis is how non-linear the relationships between predictors and price is. In fact practicially none of our predictors have a linear relationship with price. We have gone through a number of steps to attempt to navigate this reality, through data transformations, added-variable plots, lack of fit tests, and finally ending with the generalized additive model. That generalized additive model, that we detailed through above, is the model that we would recommend to observe the relationships between the highlighted predictors and housing prices. Specifically that model is

$$log(price) = \beta_0 + f_1(sqftliving) + f_2(sqftlot) + \beta_3 floors + \beta_4 waterfront + f_5(view) + f_6(condition) + f_7(grade) + f_8(bedbath)$$

As we have seen from the graphs above we are able to glean some level of the relationship between each predictor and price. Some of the more interesting results being that increasing the size of a home is only so beneficial and at some point the effect of the larger home on the same lot begins to taper off. We also found that condition and grade have deterimental effect at lower levels and positive effects only at extremely high levels. We also found that an increase in the number of floors is associated with a decrease in price, suggesting that people might prefer single floored homes rather than having multiple floors, assuming the overall square footage is the same. Furthermore, packing a home with bedrooms and bathrooms does not have a consistently positive effect, suggesting that perhaps people value having larger common rooms than more sleeping quarters.

The generalized additive model is the best model for us to use because it allows us to easily deal with non-linear relationships, while still being able to glean information. Unlike the factor level model, we are still able to see some overall relationship, rather than having coefficeints for each individual level, though that had some better interpretability. Below is the residual plot for the generalized additive model.



The Y axis has the residuals and the X axis has the fitted values. The boxplots are used to making the residuals easier to read. Each boxplot is of the same width. As you can see the residuals definitely seem to be centered around zero. Furthermore there seems to be a fairly even spread above and below 0. Lastly the variance does appear to be fairly constant throughout the plot. Since the residual plot seems fairly satisfactory and becase of the reasons stated above we decided to go with a generalized additive model.

Finally, we believe our model can be extroplated to all home sales in Washington state, and too a somewhat lesser degree, but still siginficantly, all home sales in the United States. While the housing markets certainly differ between states, we do not believe there is an particular reason our results in Washington could not be extended to say Florida. The model would probably not predict quite as well due to potentially different clientelle or tastes in that market, but the overall predictions and relationships would likely hold true. For example, though, people who live in King's County, Washington, may be older and thus having a multi-floored home could prove difficult while somewhere with a younger population may not have quite the same relationship. That is a pure hypothetical intended to reveal potential issues in transferring the model beyond King County, however we still believe that our results speak to some trends in the housing market as a whole.