Final_Assignment

Jacob Shore & Derek Huang

April 26, 2018

Introduction

In this study, we wish to analyze which factors best predict housing prices in King County (Seattle), WA. Our dataset describes over 21,000 homes sold between May 2014 and May 2015 in King County, so we anticipate the results are applicable to 2018 prices, as the economy has not shifted that much since 2015. The dataset consists of 19 variables: price, the number of bedrooms/bathrooms, the size of the house/parking lot/basement, the number of floors, whether waterfront or not, overall condition/grade, built year, renovation year, zip code and latitude/longitude coordinate. The data was taken from Kaggle, published by author "harlfoxem" under the title "House Sales in King County, WA." It is a well-known dataset on kaggle for performing regression analysis, but from what we can tell, our study is by far the most thorough.

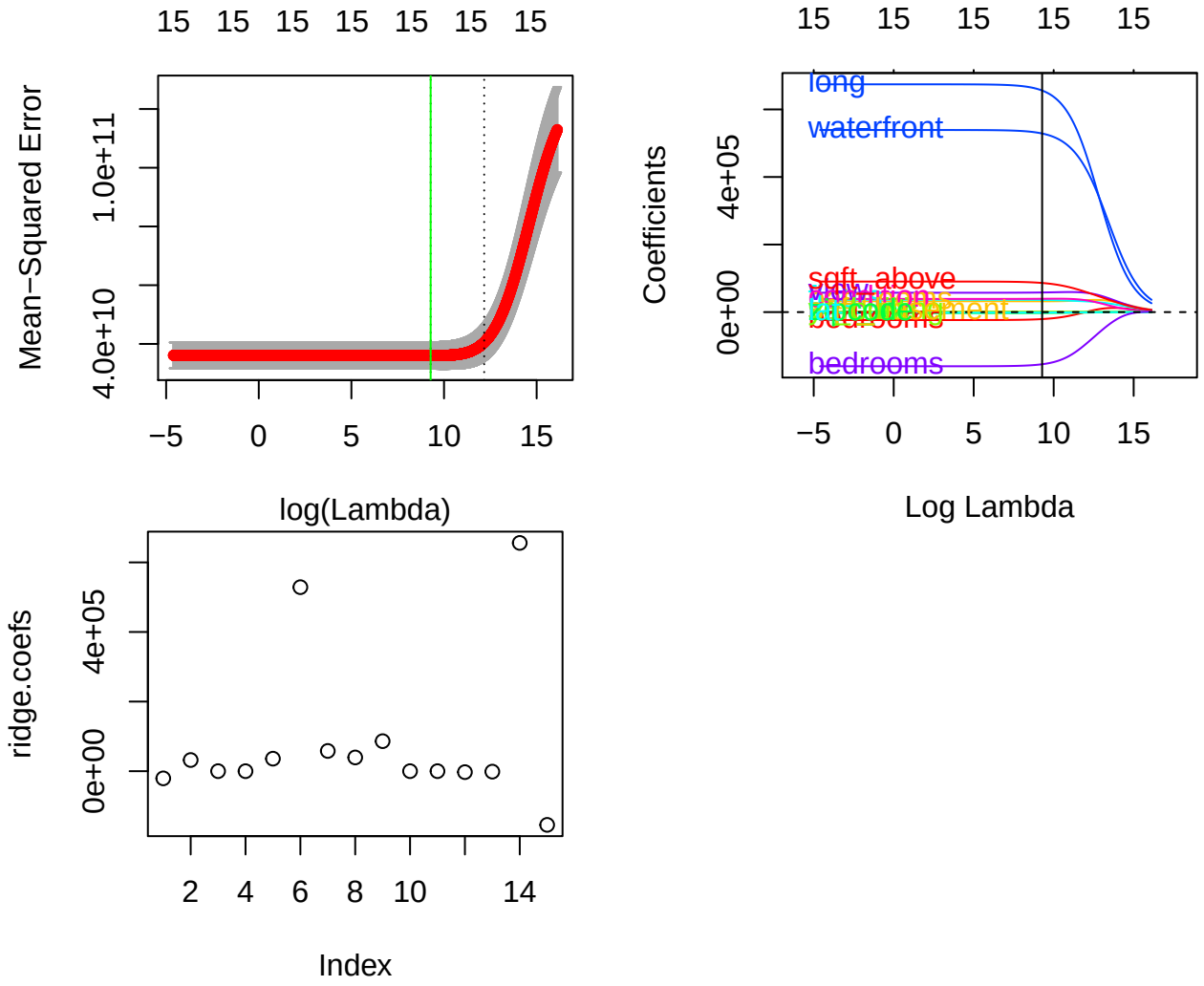
Our primary task is to examine the relationship between price (response variable) and the various explanatory variables to see which are the most predictive. Some of the variables that we anticipated to be important predictors were square footage of living space, square footage of property, condition, view, longitude, and latitude (we were right on some of these!), since these are the variables that intutively seem like they should most affect the price. We have randomly selected 1,000 of the 21,613 observations to be included in our model analysis, as this sample size is enough to generalize to the full population, makes plotting results cleaner, and eases computational load.

The variabes will be occasionally referred to as follows in R output (note that some of them are factor variables with multiple levels): sqft_living: the size of the house by square footage sqft_lot: the size of the property by square footage condition: appraiser's report on condition of the property, scale 1-5 view: appraiser's report on how nice the view is, scale 0-4 long: longitude of the house in degrees lat: latitude of the house in degrees bedrooms: number of bedrooms bathrooms: number of bathrooms waterfront: binary variable if the house is on the waterfront grade: appraiser's report on grade of the house, King Co. specific sqft_above: square footage of the attic (if applicable) sqft_basement: square footage of the basement (if applicable) yr_built: the year the house was constructed zipcode: the zipcode of the property

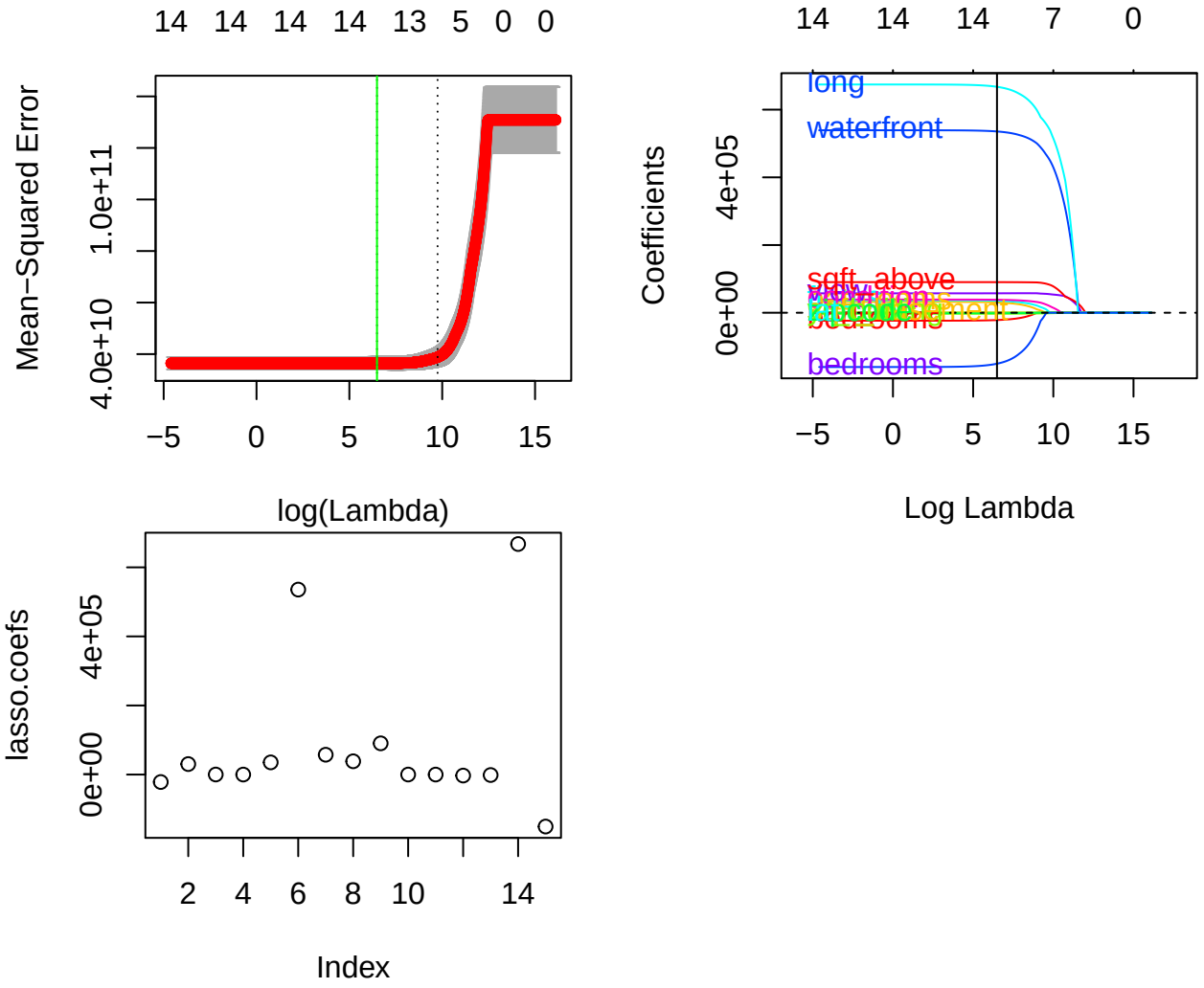
Let us now review what we have accomplished so far. In the first assingment, we described our data in greater depth and gave summary statistics on some of the interesting variables. In the second assingment, we focused on the SLR model regressing price on square footage of living space, which we had a hunch would be the most interesting and best candidate for a linear model. Our procedure, including the eventual transformations of the variables, can be found in those notes. Moving on to MLR, we first determined that multicollinearity between variables was not a concern (see pairs plot from assingment 3). We then fit multiple models using many different procedures, testing for interaction, forward/backward selection methods, and BIC selection criteria to name a few. The results of this analysis can be found in those notes, and will be brought up again as we move forward with this summary.

LASSO and RR

As a reminder, the final model from assignment 3 regressed the log of price on square footage of living space, latitude, view, condition, and interactions between square footage of living space and view, view and latitude, and square footage of living space and condition. This model will now be compared to the models from ridge regression and LASSO

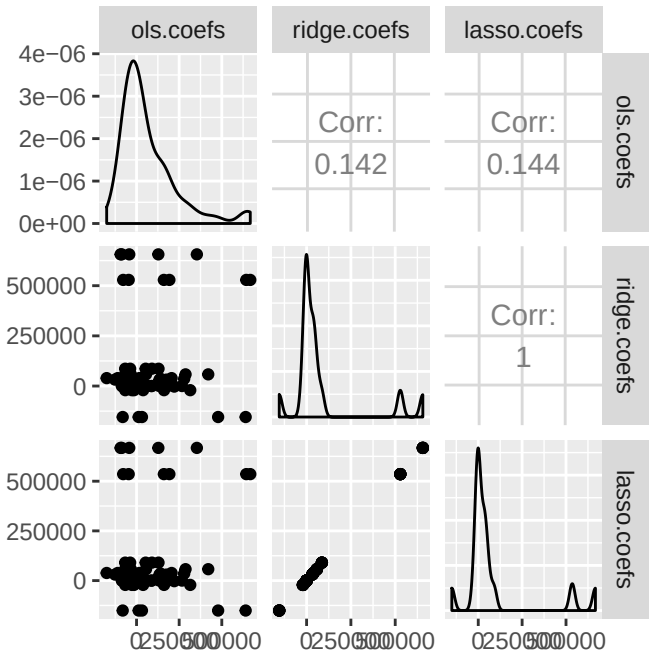


λ was cross validated and the value that minimized cv error was found to be 10701, as shown in the first plot above (in log scale). The second plot shows the shrinkage of the coefficients as λ increases. Finally, the third plot shows the coefficients of the 15 predictor variables under the value of λ which minimized cv error. These coefficients are radically different — the OLS model was very robust and involved a large number of variables, since there was interaction between continuous and factor variables. As such, a direct comparison of coefficients is difficult, although the positive and negative ones are preserved between the mdoels. We can say that ridge regression really elevated waterfront and longitude (at leasy didn't shrink these coefficients as much), and while latitude made it into the final OLS model, waterfront did not. We were also surprised that square footage of living space was shrunk so much in the ridge regression setting, as numerous OLS methods presented this variable as the most important! On to LASSO.

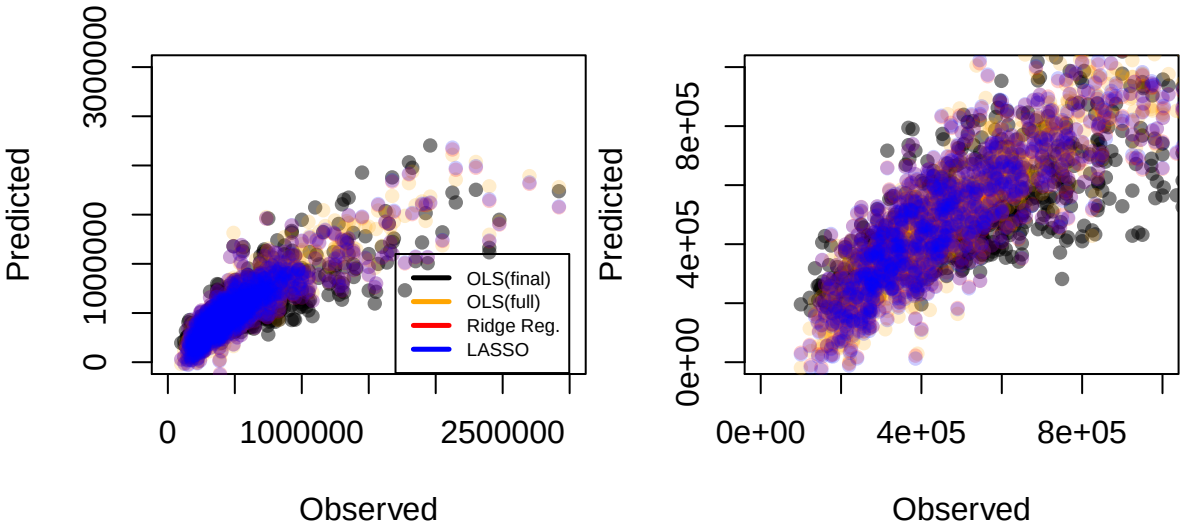


λ was cross validated and the value that minimized cv error was found to be 657, as shown in the first plot above (in log scale). The second plot shows the shrinkage of the coefficients as λ increases. Finally, the third plot shows the coefficients of the 15 predictor variables under the value of λ which minimized cv error. Just like in RR, LASSO selected very similar coefficients. Surprisingly, only one variable was shrunk to 0 (square footage of basement space), and the rest of the coefficients were very similar to RR. Thus, the same curiosities arise in comparison to the OLS coefficients.

Below, a pairs plot is presented for the three sets of coefficients. NOTE: a linear model was run on the full variable set in OLS, and this is NOT the same model as was compared above. Rather, the OLS model in the comparisons below contains no interaction terms, and is just the full explanatory variable set to match with RR/LASSO.

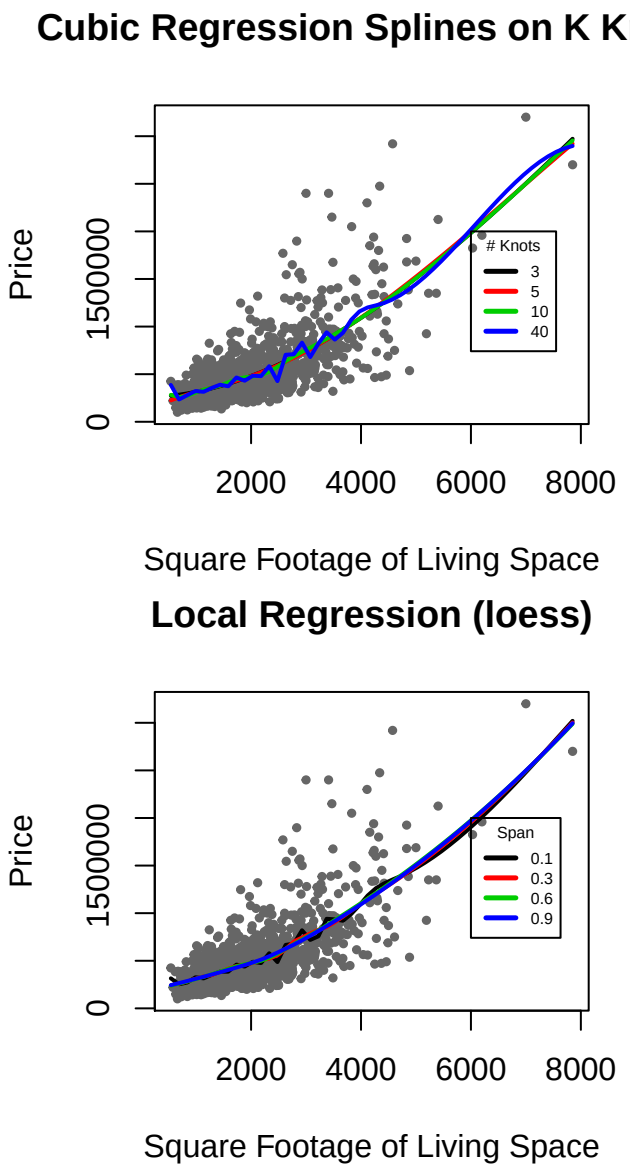


The pairs plot shows near identical selection of variables and coefficient values from RR and LASSO, while there is seemingly no correlation whatsoever with the OLS coefficients. These models are very, very different indeed.

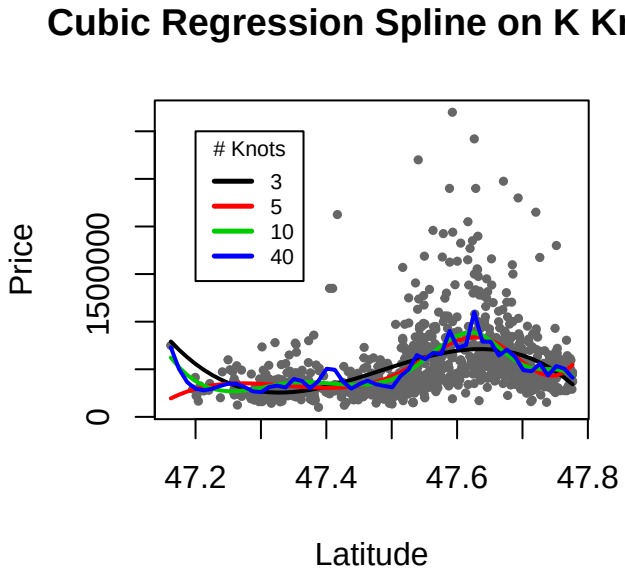


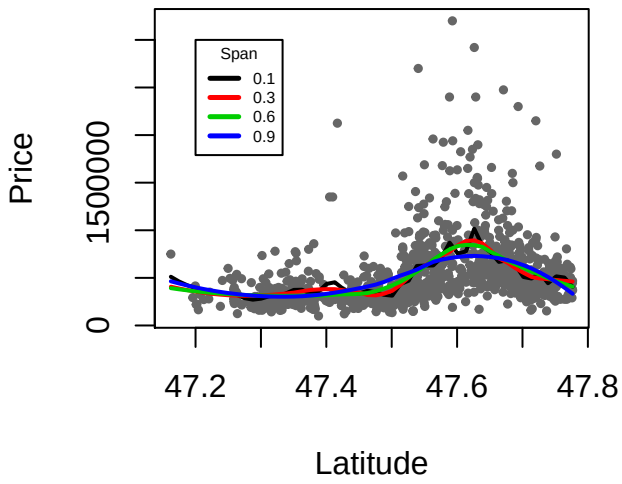
The above plots show the observed values of price on the x-axis, and the predicted values from all four models on the y-axis. Note that OLS(final) refers to the final model from last assignment, described above, while OLS(full) refers to the model on all expanatory variables with no interaction terms. Something to note — even though all the models are pretty different in makeup of coefficients, all models predict relatively well (and relatively the same). LASSO/RR predictions are nearly identical for the reasons described above. One thing that is particularly interesting is that while all predictions are pretty solid, all models besides the final model from have a slight bend in their slopes, predicting too low for small values and slightly too high for middle values (check the aspect ratio, R plots are not squares). This is particularly noticeable in the "heart" of the prediction zone between 0 and 1 million dollar homes. The final model is distinctively the most linear between observed and predicted — I guess we did alright in the last assignment!

Splines and LOESS



As we've discussed ad nauseam, regressing price on square footage of living space is already a pretty solid linear model without any additional changes. Thus, the smallest number of knots and the largest span seem to give the best estimated regression functions. We'd give a slight edge to the spline model, since prediction intervals are slightly easier to come by / to interpret, and it is slightly less computationally expensive. But these aren't all that interesting, since square footage of living space is already a really solid single variable predictor of price. Let's look at a variable which has less of a classic linear relationship with price but still has some interesting predictive properties — latitude. In previous assignments, latitude has always shown up as a significant predictor, but the relationship between latitude and price is not that straightforward.





Again, both styles provide some good fits in the middle ranges of their respective parameters to a relationship which is clearly non-linear (the downtown area has a spike in prices). We give a slight edge to LOESS in this case (on a span of .3) for very nice smoothness and easy interpretability. The behavior towards the fringes of the latitude range is particularly better in the LOESS model. Also, we cheated and cross-validated, and this model had the lowest cv error (we would have picked it anyways!). All in all, none of these models would suggest a strong relationship between latitude and price, of any form.

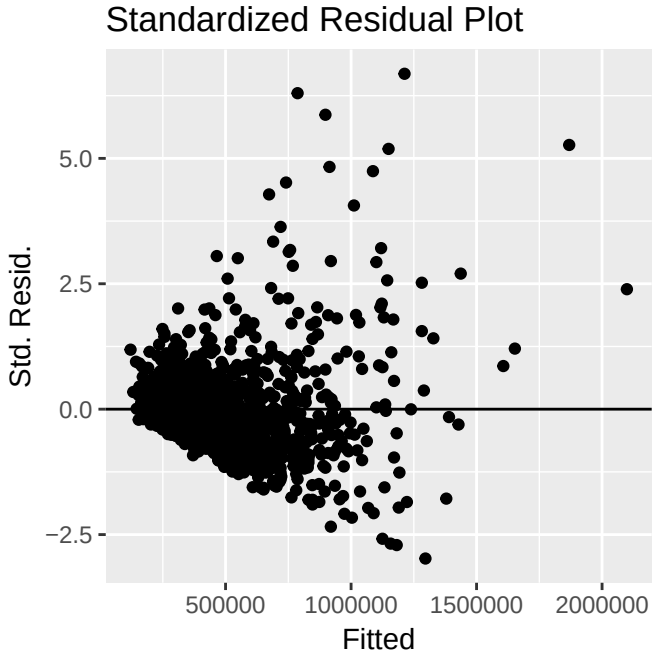
While none of the models presented above are "bad" fits per se, it is apparent that the work we did in MLR / OLS is better suited to this dataset. We have seen in homework how powerful these methods are in dealing with multicollinearity and non-linearity, but sadly (or happily?) these concerns do not burden us with this data. For the shrinkage models, sacrificing bias to decrease the variance when the variance is already very low just doesn't make sense. Furthermore, while the relationship between some of the other variable (besides sq. ft. living space) and price is interesting, none of them follow a shape that smoothing techniques really "apply" to. Thus, we are comfortable with our results from the last assignment, and would not alter that final model in any way shown above. Moving forward, something that we think could be very interesting is to look at demographic data, as this was not included in our dataset. Data about race, class, and age is out there, and in a rapidly changing city like Seattle could be a predictive category with interesting results.

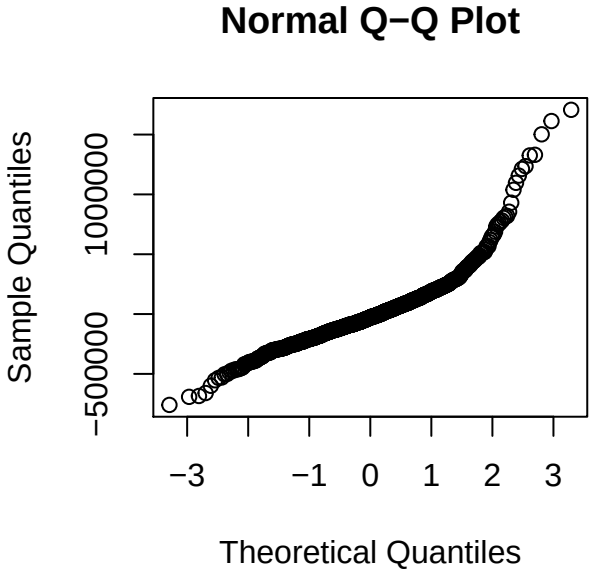
Normal Probability Plots

Perhaps the least talked-about technical assumption is normality of error terms. While slight deviations from normality usually do not pose issues, and the magic of the Central Limit Theorem suggests that even larger deviations from normality vanish under large sample sizes, normaly of error terms plays a major role in the theory underlying prediction intervals. In fact, the width of these intervals (specifically the multipliers) is a direct result of normal theory — the lack of normality means we are lost when searching for the correct multiplier (assuming the distribution of errors is unknown). Thus, even if the regression fit is solid, it may be worthwhile to check normality of errors — prediction intervals may not be as trustworthy as they seem. And in the realm of housing prices, prediction intervals are extremely important — it is common for someone to want to know what the price range of houses is for a given set of predictors. For instance, a family which needs a house with 3 bedrooms, at Z latitude, and with W square feet of living space needs to have a price range to adequately compare to similar houses available in another location.

The prevailing method for checking normality of errors is the normal probability plot, a special subtype of probability plot. In a normal probability plot (or qqplot), the residuals are plotted against their expected value under normality. That is, how likely a residual of that magnitude would be under the assumption that the residuals are normally distributed. The plot reads "quantiles," since this process implicitly compares the quantiles of the data (residuals) to the quantiles of a corresponding normal distribution (scaled by MSE). Thus, a plot that is linear suggests a normal distribution of errors, since each residual is nearly its expected value under the normality assumption. Furthermore, we can tell the exact manner in which the errors are distributed from this plot. For example, data which is skewed right would show up with a slope less than 1 for lower ordinal normal statistics (the x-axis) and greater than 1 for higher ordinal normal statistics. Data which has fat tails would show up at linear in the middle of the graph, but deviate to the bottom and top of the linear fit for low and high ordinal normal statistics, respectively.

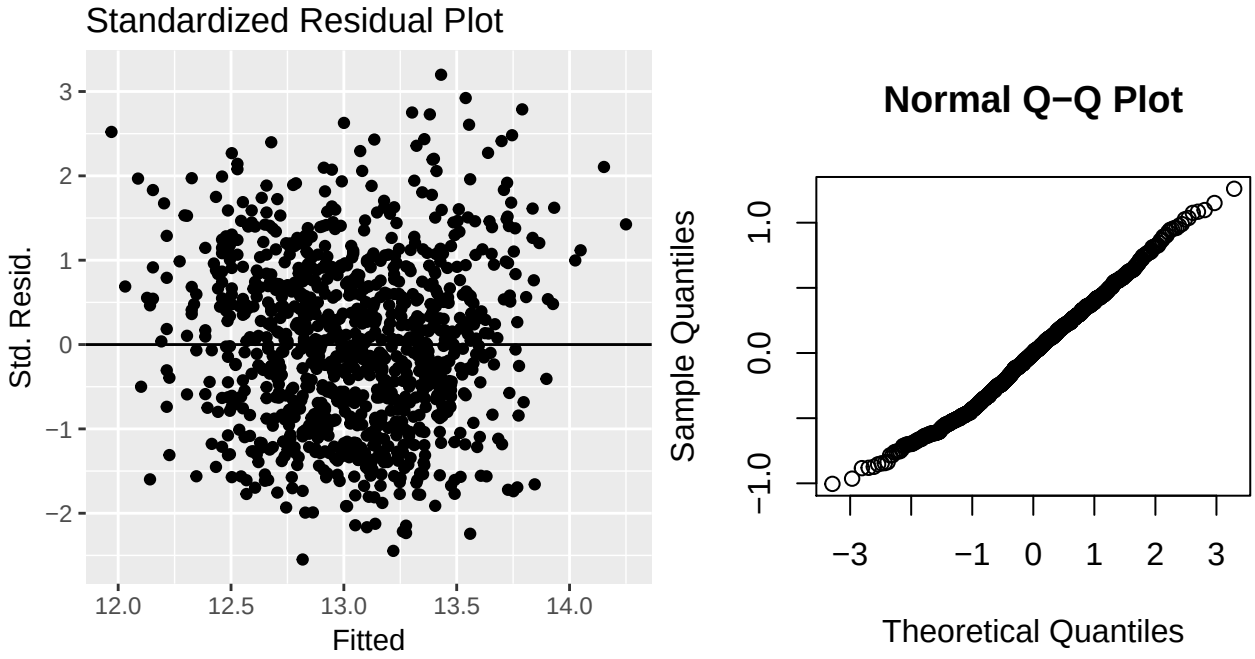
The function which maps residuals to their expected value under normality has been found to be $(k - .375/n + .25)\sqrt{MSE}$, where k is the kth smallest residual, and n is the total number of residuals. Since $\sqrt{MSE}$ is simply a scaling factor here, it may be omitted in the calculation without altering the shape of the plot. Let's move on to some examples for our data. As standardized residuals are simply another scaling factor, they to do not affect the nature of the plots, and we will use those as our input in the plots below.



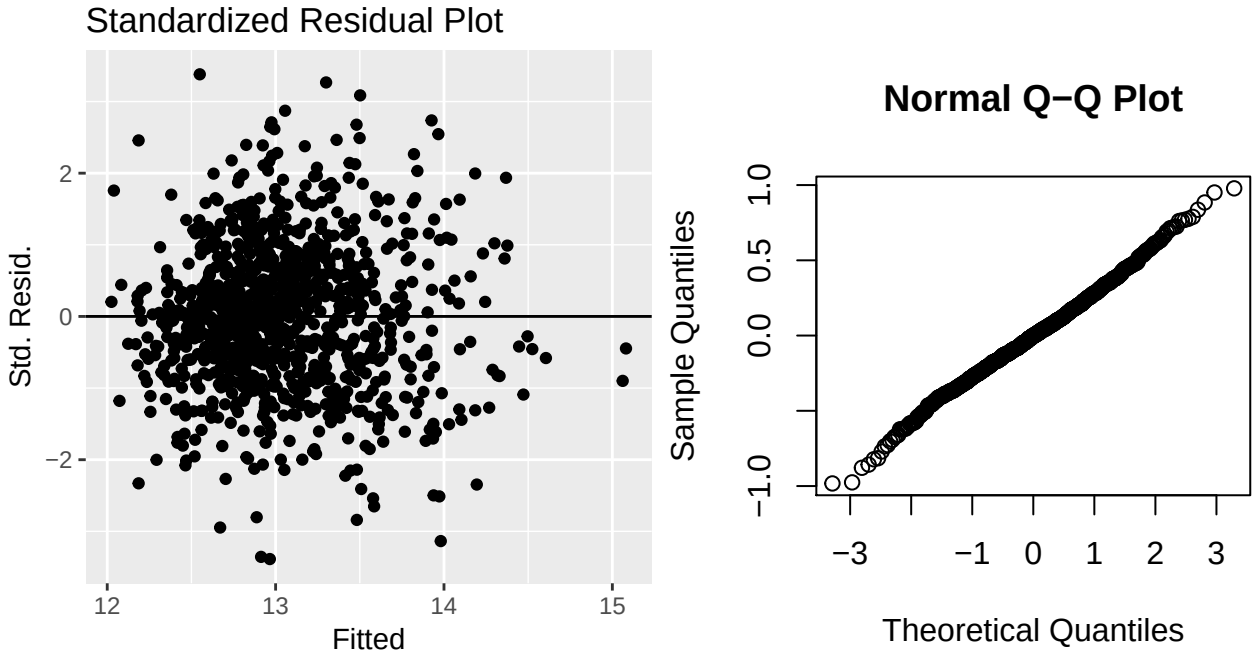


The above Q-Q plot is for the SLR model where price is predicted solely from square footage of living space. As shown by the non-linearity, normality of errors is not a great assumption to make here. The standardized residual plot shows many issues here on top of normality (it is severely heteroskedastic). In assignment 2, we

used a power model (log transformation on both x and y) to get around these issues. Let's see what the Q-Q plot looks like in that arena.



Much better! The power model fixed the issues with the variance and simultaneously fixed the issues with normality of errors. This is the model we'd move forward with in the SLR setting. Now let's check the final model we presented in the MLR setting, which regressed the log of price on square footage of living space, latitude, view, condition, and some interactions as described above.



Again, there is nothing to worry about here. The Q-Q plot shows slight bias towards fat tails, but that's why we use the t-distribution anyways! We should note that we weren't expecting a major deviation from normality in an of these scenarios — our sample size was large (1000) and the transformations made the linear model very appropriate.

Inverse Predictions

Inverse predictions refer to the situation when a new value appears on the response, and this value is used to infer what the value(s) of the predictors are using the existing regression of Y on X. In the housing market, situtations like this arise all the time. It is not uncommon for potential buyers to browse by price (Y) only at first, and having a tool which can build out the details of the house would be very valuable (obviously this information would typically exist, but maybe the buyer is extremely lazy).

Regression models are of course deterministic in nature, and so inverses are easy to come by. Note that the inverse prediction is NOT attained by regressing X on Y (called inverse regression), as this solves a different problem entirely (albeit with a similar result in most cases). The estimated regression function is assumed to be as always:

$\hat{Y} = b_0 + b_1 * X$

If we solve this equation for X given Y_new, we get the form

$\hat{X}_{new} = \frac{(Y_{new} - b_0)}{b_1}$

which is the MLE for X_{new} assuming that there is a linear relationship between X and Y (b_1 cannot be 0). Assuming normality of errors and constant variance as before grants the confidence interval:

$\hat{X}_{new} \pm t(1 - \frac{\alpha}{2}; n - 2)s\{predX\}$

where

$s\{predX\} = \frac{MSE}{b_1^2}[1 + \frac{1}{n} + (\frac{(\hat{X}_{new} - \bar{X})^2}{\sum(X_i - \bar{X})^2})$

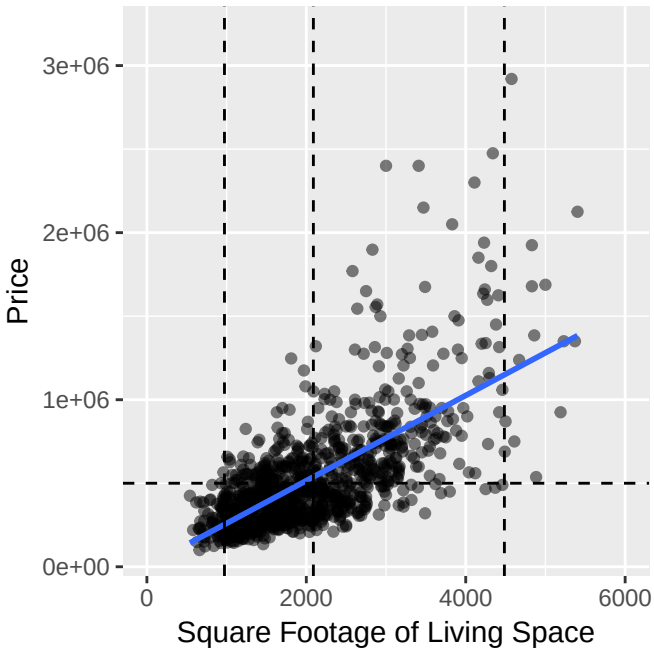
And that's it. If we can safely conclude that β_1 is NOT 0, then we are good to go with the inverse predictions. Note that the given interval above relies on normal theory — an exact confidence interval may be obtained through some tedious algebra. It will not be reproduced here, but may be found at https://journal.r-project.org/archive/2014/RJ-2014-009/RJ-2014-009.pdf in the documentation for the "investr" package, which we will employ to get inversion intervals.

Furthermore, in the case of multiple new observations, a Bonferroni or Scheffe procedure may be specified to adjust the critical values of the intervals accordingly. This will be illustrated below.

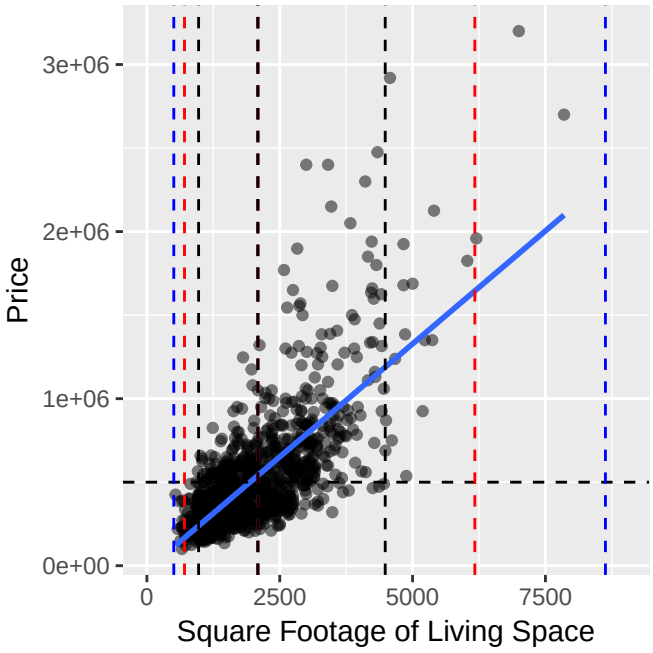
The model we will use is the same as above, where the log of price is regressed on the log of square footage of living space. The "investr" package was used to compute exact confidence intervals. This package automatically calculates the slope estimate, and performs a test of significance to make sure there is a linear relationship present. First, we will attempt to estimate what level of square footage of housing space led to a house listed for 500,000 dollars. As an example of the "investr" package, we leave the code in here. The calibrate function is used for inverse predictions.

```
library(investr)
housing_lm3 <- lm(log(price) ~ log(sqft_living), data=housing2)
res <- calibrate(housing_lm3, y0 = log(500000), interval = "inversion", level = 0.90)
cat("A 95% CI for the square footage of living space of a home listed for 500,000 dollars in
King County is (",exp(res$lower),",",exp(res$upper),") sq. ft.")
```

```
## A 95% CI for the square footage of living space of a home listed for 500,000 dollars in
## King County is ( 973.19 , 4484.9 ) sq. ft.
```



In the plot above, the shaded area represents the confidence interval around the fit of the regression function, while the dashed lines represent the new observation, the best guess for what value of square footage of living space corresponds to a 500,000 dollar house, and the inversion intervals for that guess at 90% confidence. Recall that the regression model was fit on the log of both predictor and response, so they have been back-transformed here.



The above plot shows simultaneous inference for 5 new houses that entered the market at 500,000. In red are the Scheffe bounds, and in blue are the Bonferroni bounds for the inversion intervals. Note that all intervals here are big — perhaps too big? We cannot determine if there is an error somewhere within the package or if the inversion intervals for simultaneous inference really are that large. Either way, nobody is going to be getting any information out of knowing that a house listed for 500,000 dollars is between 1000 and 45000 square feet.

Generalized Additive Models

GAMs provide a general framework for extending a standard linear model by allowing non-linear functions of each of the variables, while maintaining additivity. GAMs are extraordinarily flexible, as they can be applied with both quantitative and qualitative responses. The basic framework is as follows: the relationship between the predictor and each response is assumed to follow some pattern, which may or may not be linear. We estimate these relationships using some smooth fitting function, and then simply add up the estimates to get a prediction for some function of the response. Keeping track of the fitting functions is the hardest part, as the response variable must then be evaluated by the inverse of this amalgamation of functions.

An example of GAM would be, instead of having $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ...\beta_p x_{ip} + \epsilon_i$, we would write the model as $y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + ...f_p(x_{ip}) + \epsilon_i$. It shows the addictive characteristic because we calculate a separate f_j for each X_j, and then add together all of their contributions. Note that each f_j is smooth and non-parametric — its shape is fully determined by the data, much like kernel smoothers. In fact, splines are one of the most commonly used functions here. Since these functions are non-parametric, we don't need to worry about the confusion with high dimensional polynomials or other parametric methods which serve as a proxy for non-linear fits.

GAMs eventually have to combine the simultaneous smoothing they perform for each variable, and they do this by way of penalized likelihood maximization. We will not reproduce the mathematics here (see https://multithreaded.stitchfix.com/assets/files/gam.pdf for a walkthrough from a data scientist at stitchfix), but the basic idea is to find the set of smoothing functions which maximize the likelihood (defined in the normal way as the product of parameter pdfs given the data) subject to some constraint, which is governed by a variable λ. Increase λ, and we increase the penalty, thereby increasing the smoothness, and vice versa for decreasing this smoothing parameter. The penalty term may be specified ahead of time, or it may be chosen through a variety of cross-validation methods. Our package "mgcv" uses an approach known as "REML," or restricted maximum likelihood estimation, which essentially rewrites GAM as a parametric general linear model and updates λ from an initial guess through weighted least squares methods. Again, beyond the scope of this project, but interesting reading from the link above.

GAMs also possess the capacity to perform variable selection, using either forwards/backwards selection methods or shrinkage methods, both of which we've covered at various points in this project. However, as noted by the author of the above article, "throwing the kitchen sink at GAM may lead to weird results," and pre-screening of variables is advised.

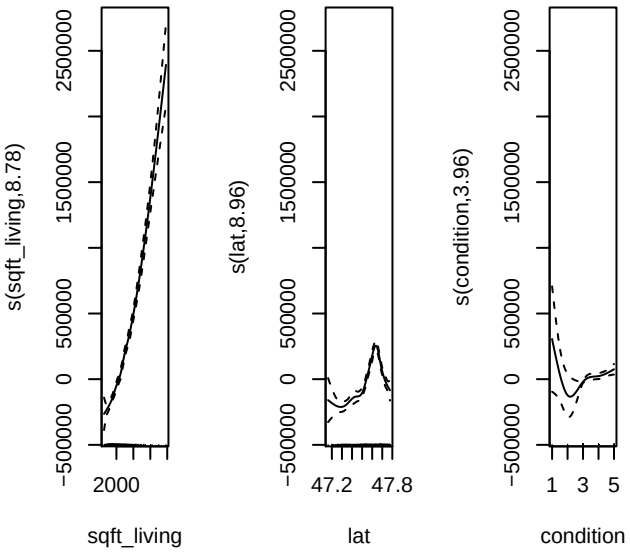
We picked GAMs because GAMs allow us to fit a non-linear f_j to each X_j, so that we can automatically model non-linear relationships between each predictor and the response that standard linear regression would miss. This is a major departure from standard linear regression, which assumes linearity across the board. Referring to housing prices in particular, we found in previous assingments that some significant predictors (such as square footage of living space) do have a lienar relationship with price, while other significant predictors (such as latitude) do not. See the above section on kernel smoothers for a direct comparison of these variables.

Like ridge regression or LASSO, much of the power in GAMs lies in their ability to do the work for us, as we do not need to assume what sort of functions fit each predictor variable and the response. These non-linear fits can potentially make more accurate predictions. For example, for latitude, we anticipate GAMs to yield better fit for us. Also, since the model is additive, we can still examine the effect of each X_j on Y individually while holding all of the other variables fixed. Hence if we are interested in inference, GAMs provide a useful representation, and are easily interpretable in non-technical terms (pricing increases linearly with square footage of living space, and increases initially with latitude before tailing off). It also makes the interpretation of factor variables much clearer, as the relationship between the different factors and the response can be visualized.

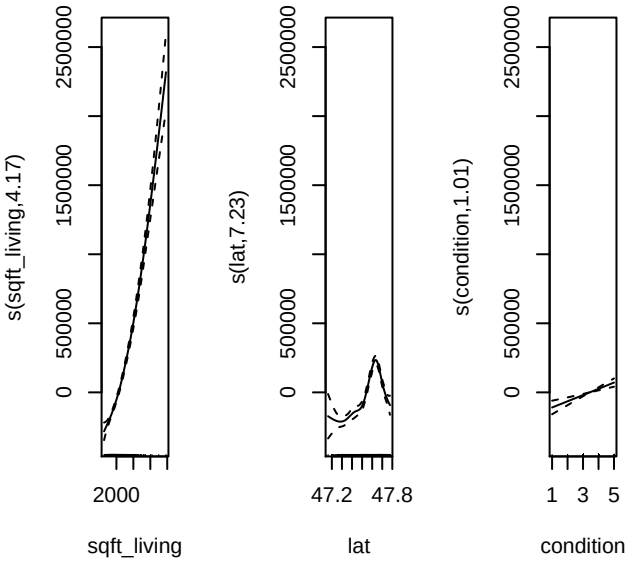
First, let's go through a quick example to see how GAM works. We will fit a GAM for regressing price on square footage of living space, latitude, and condition (which is a 5-factor variable). Note that we specify a lot here — we are fitting cubic regression splines on 10 knots for the first two variables, and on 5 knots (the

maximum) for the third variable, and we set λ to .5. The code is left in here as an example of the "mgcv" package, which we are using for our GAMs.

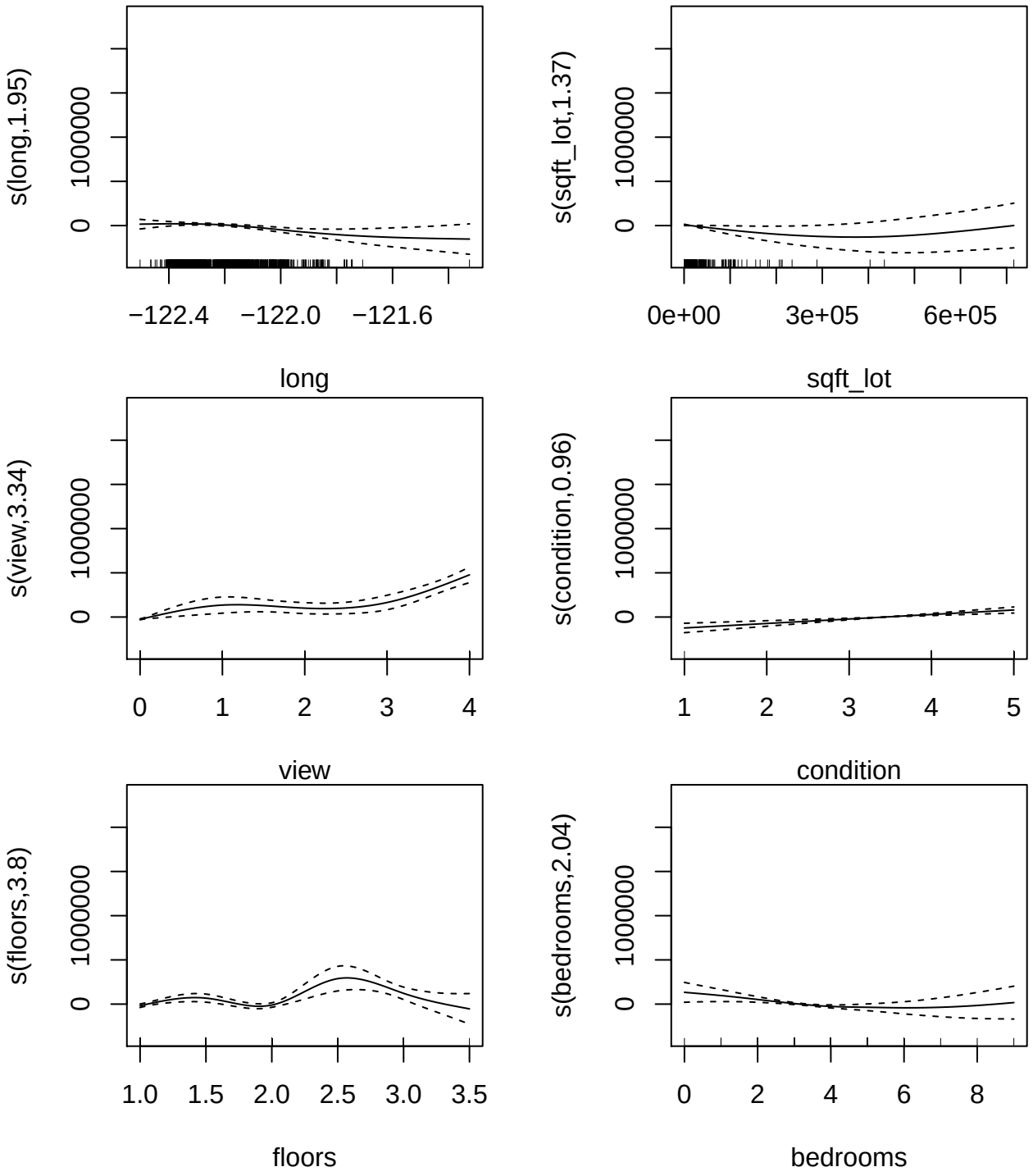
```
library(mgcv)
gam1= mgcv::gam(price ~ s(sqft_living,bs="cr", sp=.5)+s(lat,bs="cr",sp=.5)+s(condition, bs="cr",k=5,sp=
par(mfrow=c(1,3))
plot(gam1, se=TRUE,col="black")
```

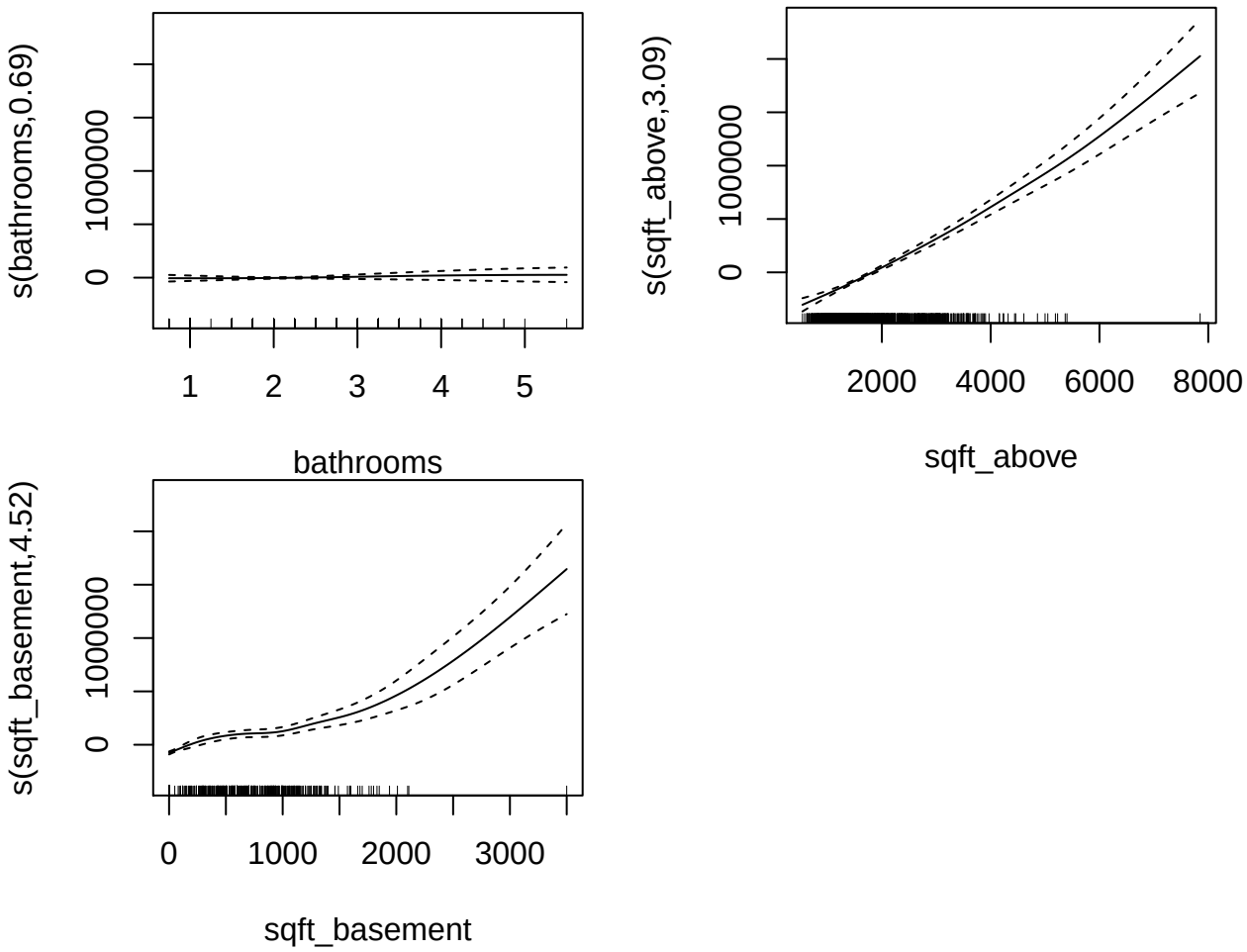


These plots tell us that square footage of living space is relatively linear in price (we already knew this from SLR), that price is higher when latitude is around 47.6 (downtown Seattle, confirms our LOESS model), and that condition starts off high, then decreases before rising gradually again. Moreover, we see the degree to which these variables influence price — the higher ranges of square footage of living really affect price, while there is not much difference across all conditions relative. So instead of having to fit many, many SLR models, or plot each LOESS/spline model seperately, this GAM (1 line of code!) got us the same "verbal" information. Although the model isn't as precise (a lot of estimation involved in non-parametric methods), if all we need is the quick and dirty, this is certainly the way to go. You'll notice that the model for condition doesn't make sense here — lower condition should correspond to lower prices, not higher. Perhaps there is something wonky with the specified smoothing parameter of .5? Let's see what happens when we let "mgcv" pick these values.

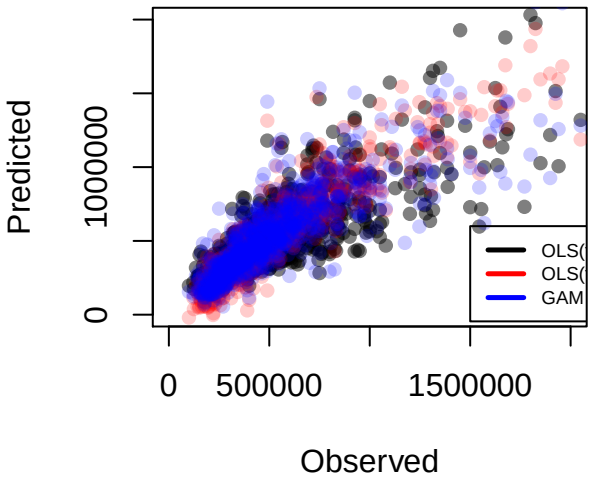


Now it's much more smoothed, and the result lines up with what we've seen from prior assingments. Let's add more variables in and see what happens. Note that model selection never actually worked for us — it just returned all inputted variables no matter what we tried. Also note that the binary variable "waterfront" we could not get to work, as it is below the minimum number of knots required. There is a way around this, but not in conjunction with continuous variables from what we can tell.





GAMs give us the chance to examine some of these "forgotten" variables, and the effect they may have on price. Some interesting caveats here — having 2.5 floors is pricier than having 3 floors, more bedrooms is not necessarily more expensive (but standard error is large), and having an outstanding view is way better than just having a mediocre one. However, nothing (except to a degree, view) predicts anywhere near the strength of square footage of living space (recall that view was deemed significant in previous models as well, though). Also note that square footage of attic and basement space are actually really solid linear predictors of price — we removed both of these variables at the beginning of all other analyses! However, both of these variables are very highly correlated with square footage of living space, which we've covered as a linear predictor of price ad nauseam (r = .87 and .62) respectively. Finally, let's take a look at how GAMs predict vs. some of the other models we've discussed before



Here, GAM predictions are compared to the final OLS model described many times above, and the full OLS model consisting of all explanatory variables. The GAM model is on the same variables as the final OLS model (square footage of living space, latitude, view, condition), just without interaction. Note that once again, the final OLS model is still the most preferred and grants the best predictions, although GAM really isn't bad (except at small price values, where it's pretty bad). Given all the additional information and graphical displays that GAM brings to the table, it's a solid thing to consider.

GAMs allow for as little or as much control as we want, and for that reason, they are very useful, especially when dealing with lots of data or non-linear relationships. We wouldn't present this model to the client as our final build, but beginning the process by doing GAM to explore how the various expplanatory variables interact with price in the regression setting is invaluable. We could easily anticipate preferring GAM to OLS if our data was not so dominated by the linear relationship between square footage of living space and price.

Summary

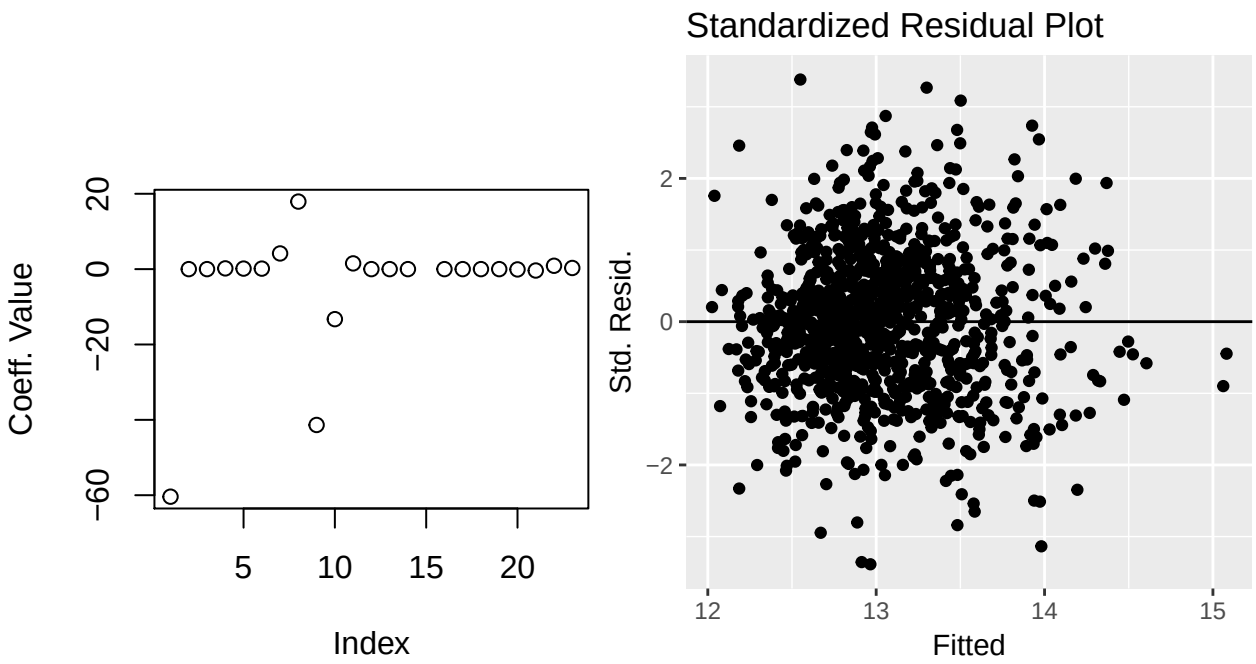
In this study, we examined which factors best predict housing prices in King County (Seattle), WA. At the beginning of the experiment, we were confident in being able to fit a linear model well, as it is well-known that housing prices are affected by some of the variables we had access to (such as square footage of living space, number of bedrooms, etc.). One aspect of the pricing dynamics that we hoped to uncover was whether objective (square footage, rooms, location) or subjective (appraisal scores on condition, view) carried more weight in the model. As it turns out, square footage of living space is unsurprisingly the strongest predictor, as it had a near perfect linear fit with the log of price, it was the most significant predictor in the MLR setting, and it was consistently the first variable in the model in forward / backward selection methods. Some of the appraisal scores were also very significant predictors — specifically the view rating and condition rating. Surprisingly, latitude ended up being a very significant predictor across all of the model types we analyzed, while longitude was not. Perhaps even more surprisingly, predictors which we thought might be important, such as square footage of property and appraisal ratings on grade and objective measures such as number of bathrooms and bedrooms were not as relevant as the variables listed previously. That is not to say they didn't have predictive power — they did, but the other variables were so noteworthy that adding in these additional variables was not worth the marginal benefit at the risk of overfitting.

As stated above, shrinkage and smoothing models did not provide any new insight in terms of model building, and confirmed our hunch that the final MLR model we constructed in assignment 3 was the best for our situation. Our data was staunchly linear between square footage of living space and price, but with significant interaction effects. As the GAM model showed, many of the other variables were not linear with price in the slightest. The major advantage of the MLR model ended up being these interactions effects, or so it seems.

To recap, the final OLS model regressed the log of price on square footage of living space, latitude, view,

condition, and interactions between square footage of living space and view, view and latitude, and square footage of living space and condition. This model was selected as follows: After some initial digging to find the most significant predictors (t-statistics), an F-test was run to see if these variables (square footage of living space, view, latitude, condition) had significant interaction terms. The resulting p-value of ~0 implied that interaction was significant. We decided to keep the interaction terms with the largest marginal t-statistics, even if they were single-factor interaction (e.g. we kept the interaction of view and latitude even though only three levels of view significantly interacted with latitude). R^2 ajusted for this model was .711, meaning that approximately 71.1% of the variance in the log of price can be explained by a linear relationship with the above listed explanatory variables. The value of the coefficients can be found below, as well as the residual plot. Pardon the clutter, but our model is fairly complex with the factor variables. As shown in the coefficient plot below, many of them are close to 0, which explains why LASSO did not shrink many of them (it physically couldn't!).

```
##                         (Intercept)                         sqft_living
##                         -6.0393e+01                          4.5746e-04
##              as.factor(condition)2              as.factor(condition)3
##                          -8.2381e-05                          1.9760e-01
##              as.factor(condition)4              as.factor(condition)5
##                           1.4109e-01                          1.3265e-01
##                   as.factor(view)1                   as.factor(view)2
##                           4.1612e+00                          1.7930e+01
##                   as.factor(view)3                   as.factor(view)4
##                           -4.1359e+01                         -1.3275e+01
##                                   lat sqft_living:as.factor(condition)2
##                           1.5228e+00                         -1.9434e-04
## sqft_living:as.factor(condition)3 sqft_living:as.factor(condition)4
##                          -9.2841e-05                         -3.9093e-05
## sqft_living:as.factor(condition)5      sqft_living:as.factor(view)1
##                                    NA                         -1.1012e-04
##      sqft_living:as.factor(view)2      sqft_living:as.factor(view)3
##                          -7.3276e-05                         -1.9605e-04
##      sqft_living:as.factor(view)4               as.factor(view)1:lat
##                          -1.6797e-04                         -7.6615e-02
##               as.factor(view)2:lat               as.factor(view)3:lat
##                          -3.6920e-01                          8.8976e-01
##               as.factor(view)4:lat
##                           3.0217e-01
```



The following variables were significant at the .05 level, with their corresponding p-values.

```
##                                                summary.full.model..coef.summary.full.model..coef...4.....0.05..
## (Intercept)                                                                                          7.4902e-65
## sqft_living                                                                                          4.0009e-35
## as.factor(view)3                                                                                     2.6777e-02
## lat                                                                                                  4.5627e-88
## sqft_living:as.factor(condition)3                                                                    1.1922e-02
## sqft_living:as.factor(view)1                                                                         4.7441e-02
## sqft_living:as.factor(view)3                                                                         1.5607e-03
## sqft_living:as.factor(view)4                                                                         6.3819e-04
## as.factor(view)3:lat                                                                                 2.3882e-02
```

As for why this model in particular worked well, it is clearly due to the interaction effects, which played a major role in the final predictions. Housing prices are affected by a great many factors beyond simply square footage of living space, and these variables interact in interesting ways. For instance, our model shows that houses with fantastic views (4 on the scale) go for far, far more than houses with mediocre view (2 or 3 on the scale) when latitude is taken into account. This naturally makes sense, as views in downtown Seattle will not be relatively scored the same by appraiser's as views outside of the city center. Square footage of housing also interacts with things like condition — a house in poor condition might still be listed for a lot of money, since the condition is easier to "fix" than square footage is.

We caution against generalizing our results beyond King County housing. While many of these factors no doubt predict housing across the country (or world), others may be Seattle specific. Latitude and longitude are two obvious ones, but things like the interaction between view and square footage may not hold true in other settings. Even though the data was collected three years ago, and housing prices change pretty rapidly, we do feel comfortable generalizing these results to the current Seattle market. It was stated somewhere above, but we really would like to examine demographic changes in Seattle and how these affect the housing prices (race, age, religion), as Seattle is changing very quickly on that front. We'd also like to do more spatial modeling with this dataset, perhaps out of the regression arena, and utilize the zipcode variable in conjunction with demographic data to perform spatial statistical analyses.

Overall, this project was very interesting for us. While it sort of felt like a pre-prepared dataset (and this dataset is in fact commonly used for introductory regression analyses), it was fun to apply what we've done in class to real-world data. Jacob is moving to Seattle very shortly for work as well, so it's nice to get an

idea about how prices in Seattle are figured. We were also glad to find out that our best fit was not the simplest solution (just using square footage of living space), but that the interaction effects were significant and relevant.