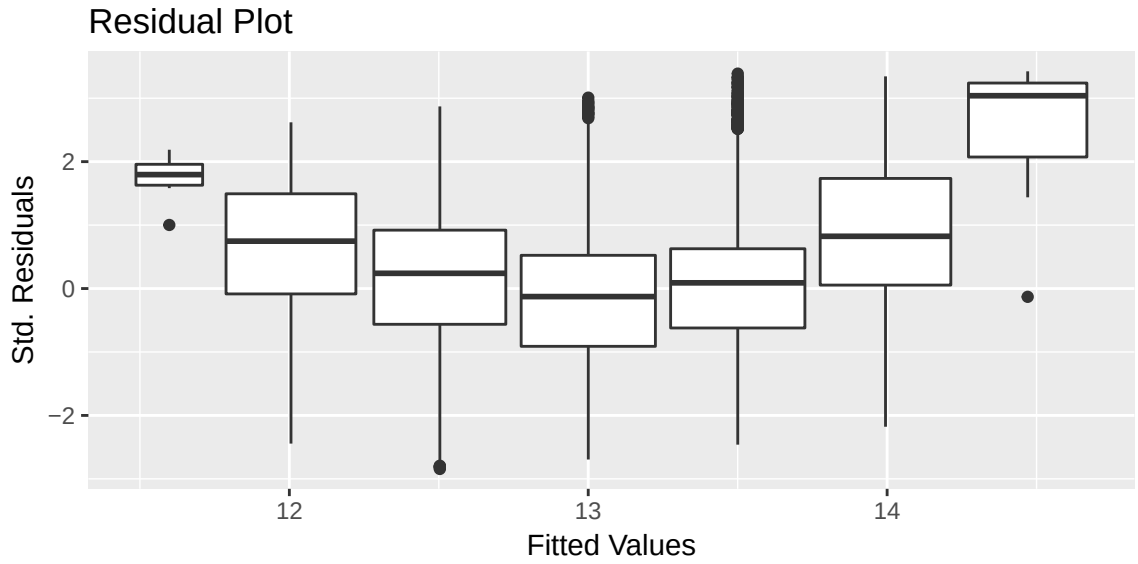
Math 158 Final Project: Part 2

Nick George and Spencer Louie

Introduction:

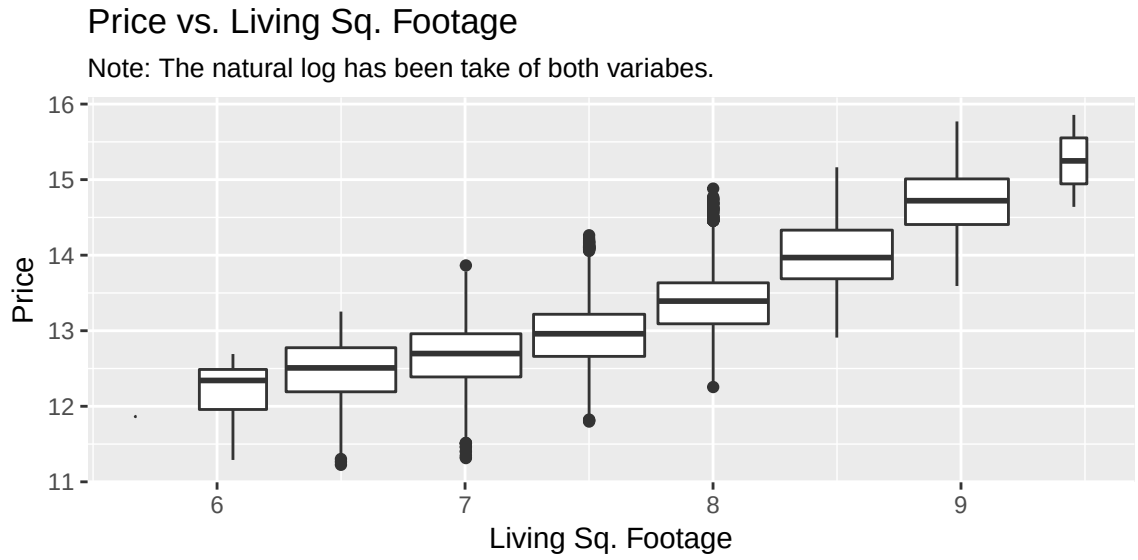
Here we'll be running a linear regression on housing prices, specifically from King County Washington. The explanatory variable will be the square footage of living area. While the response variable is the price of the house. We're interested in testing whether or not there is a positive relation between housing price and the square footage of living area. So $H_0 : \beta_1 \leq 0$ and $H_a : \beta_1 > 0$. Our null hypothesis is that there is a zero or negative relationship and our alternative hypothesis is that there is a positive relationship.



By taking log-log transformations of our data we were able to get a residual plot that does a decent job of satisfying some of our assumptions and is certainly better than before it was transformed. One of the key reasons we logged both sides was to get something closer to linearity. The residuals appear are somewhat equally distributed positively and negatively. The variance is not perfectly constant, but again significantly better than the pre-transformed data. This implies that our estimates will not be as efficient as they could be. If we were to take out some of the outliers, by potentially narrowing the focus of our study, we may be able to get a residuals that better fit the assumptions.

term	estimate	std.error	statistic	p.value
(Intercept)	6.729916	0.0470620	143.0011	0
log(sqft_living)	0.836771	0.0062233	134.4587	0

The p-value is roughly 0 and the t-stastic is quite large at $134/2 = 67$ so we are able to reject the null hypothesis and say that there is some positive relation between housing price and living square footage. This implies that a doubling in square footage would be associated with a $2^{.836} = 1.786$ multiplicative change in the median of price.




```
##        fit      lwr     upr
## 1 12.66269 11.90087 13.4245
```

This is a prediction interval for 1200 square feet of living space. So 95% of the log price values are between 11.9 and 13.42.

```
##        fit      lwr      upr
## 1 12.66269 12.65505 12.67033
```

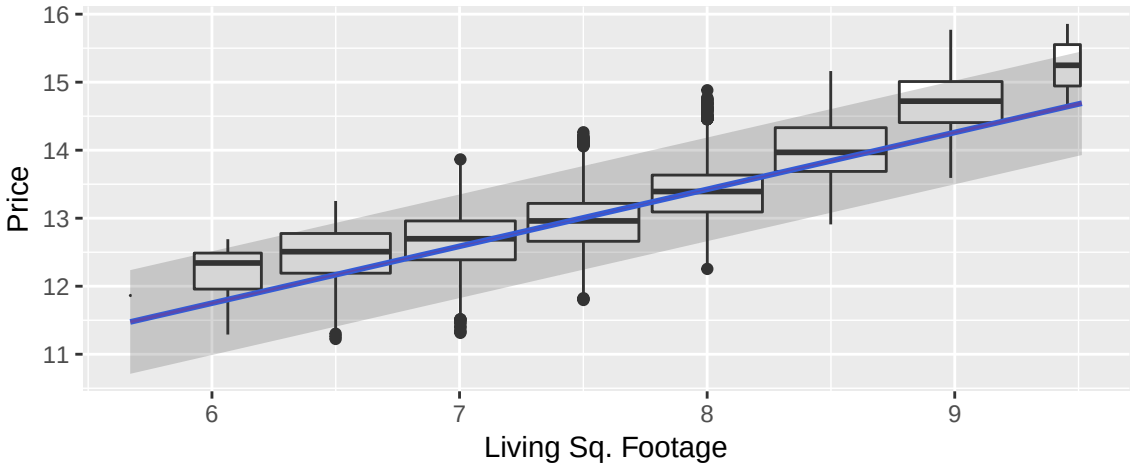
This is a confidence or mean interval for 1200 square feet of living space. We are 95% confident that the mean log price value for 1200 square feet of living space is between 12.65 and 12.67.

```
##
## Call:
## lm(formula = log(price) ~ log(sqft_living), data = kc_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.10511 -0.29300  0.01262  0.25701  1.33011
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.729916   0.047062   143.0   <2e-16 ***
## log(sqft_living) 0.836771   0.006223   134.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3886 on 21611 degrees of freedom
## Multiple R-squared:  0.4555, Adjusted R-squared:  0.4555
## F-statistic: 1.808e+04 on 1 and 21611 DF,  p-value: < 2.2e-16
```

The R squared of the model is .4555 which means that about 45% of the variance in log price is explained by log living square footage. Since we have only used one variable to attempt to explain price, our model has done quite a bit of work. Adding more variables should improve our R squared, allowing us to explain more of the variance in the response variable.

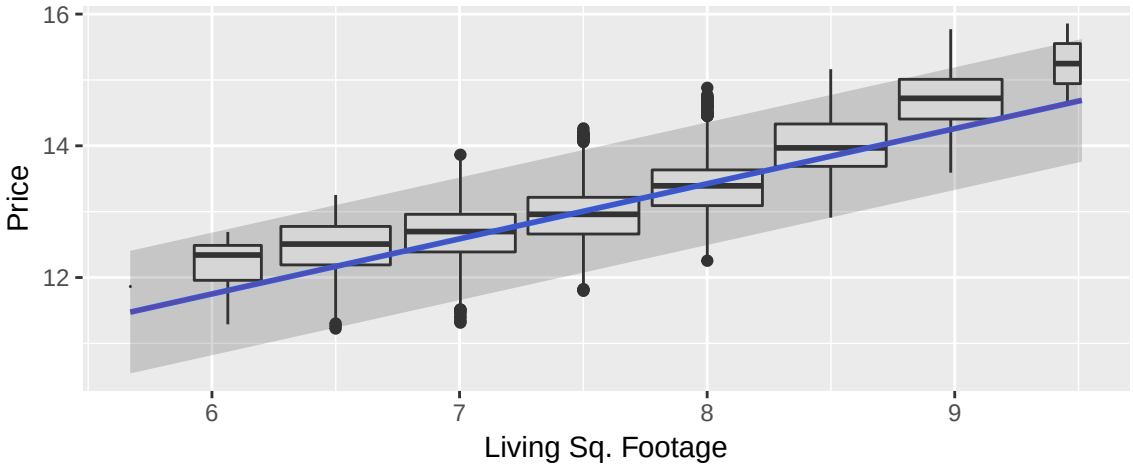
Price vs. Living Sq. Footage (Non−Adj. Bands)

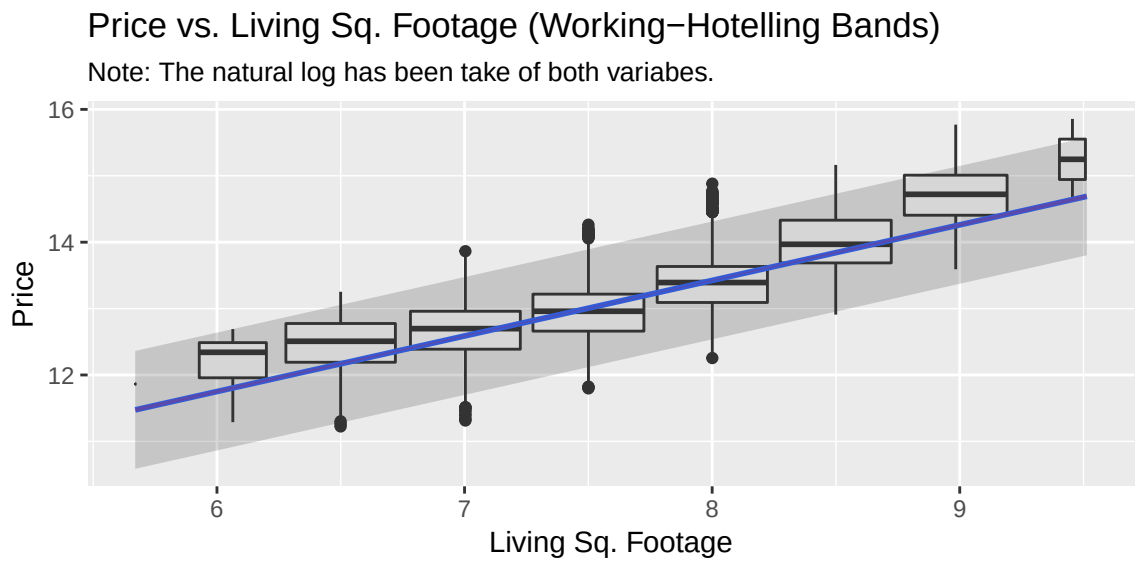
Note: The natural log has been take of both variabes.



Price vs. Living Sq. Footage (Bonferonni Bands)

Note: The natural log has been take of both variabes.





Note that in the three graphs above, the gray band is the prediction interval. There is also a band for the confidence interval included, however it is so thin that it is pratically impossible to see. As evident from the size of the p-value and the confidence interval for living sq. footage = 1200 shown above the confidence interval would indeed be quite small.

Adjusting for multiple comparisons is important because confidence intervals aren't a guarantee. The true value we are interested in is not necessarily in the interval we select. By doing multiple different versions, here we have three, we give ourselves a better opportunity to capture the true value. Furthermore when you are looking at a large number differences you run into the probelm of multiple comparisons. When observations differ by a number of factors then our discovery may be seem stronger than it should. Therefore we adjust so that our intervals are more realistic, this is why the bands for both adjusted versions are larger. Since we are only comparing on one dimension the non-adjusted version would probably work out well for us, however if we had to choose between the two adjusted, we would choose the Work-Hotelling / Scheffe versions as it is a tighter band than the other.

Conclusion: Overall our model does support our theory that there is a positive relationship between price and living square footage. However, we did have to transform our variables in order to reach the necessary assumptions for our model to work. There were two things that paritcularly suprised us. One was just how large our R squared was with only one variable. We were able to explain almost half of the variance in log price with log living square footage. We are surprised by that because it doesn't account for quality, only partially for size of the overall lot, and only partially for rooms and floors as well as a number of unquantifiables. The other factor that surprised was the implied magnitude of the effect. A doubling in square footage leads to a less than doubling of median price. In some respects we expected a doubling in square footage to lead to an even higher multiplicative increase in median price.