Assignment 2

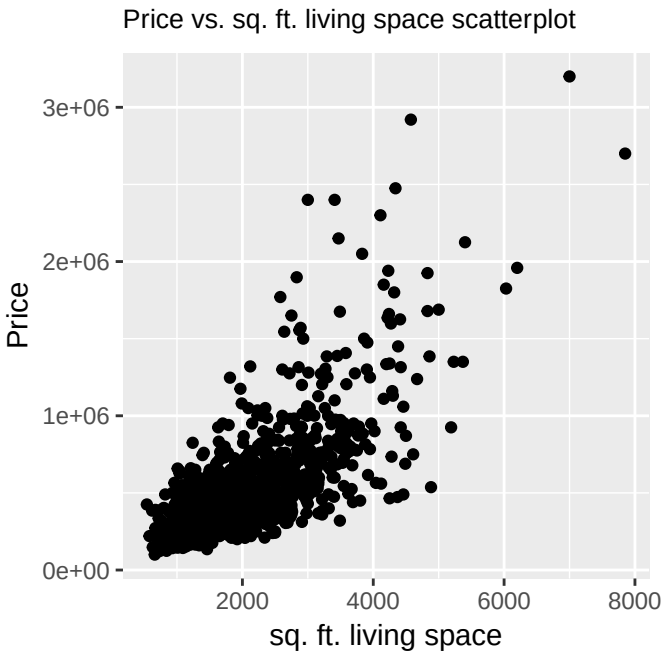
Jacob Shore & Derek Huang

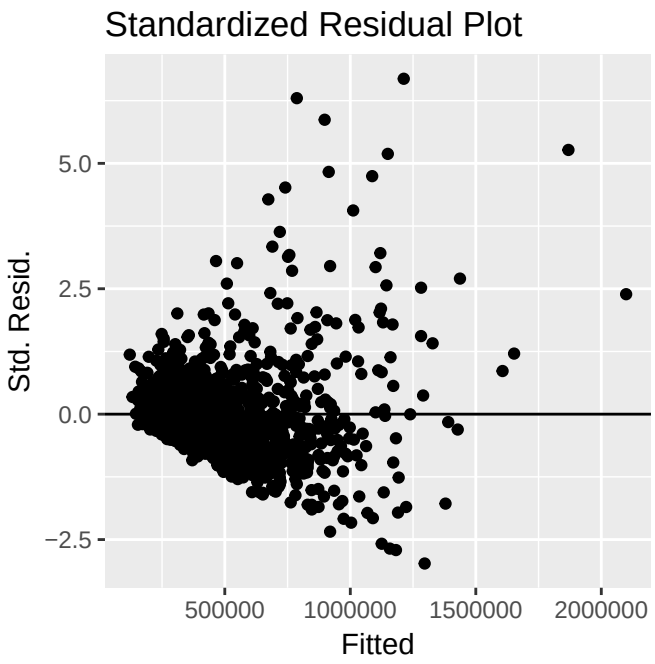
February 12, 2018

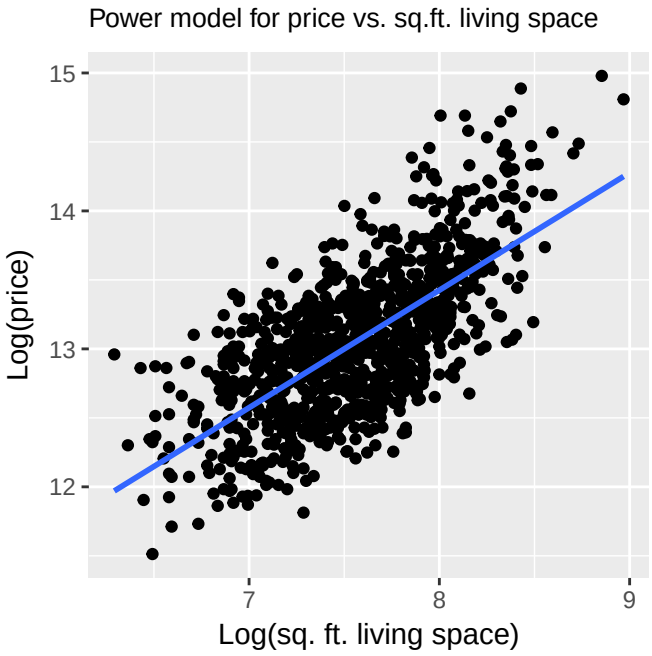
Introduction:

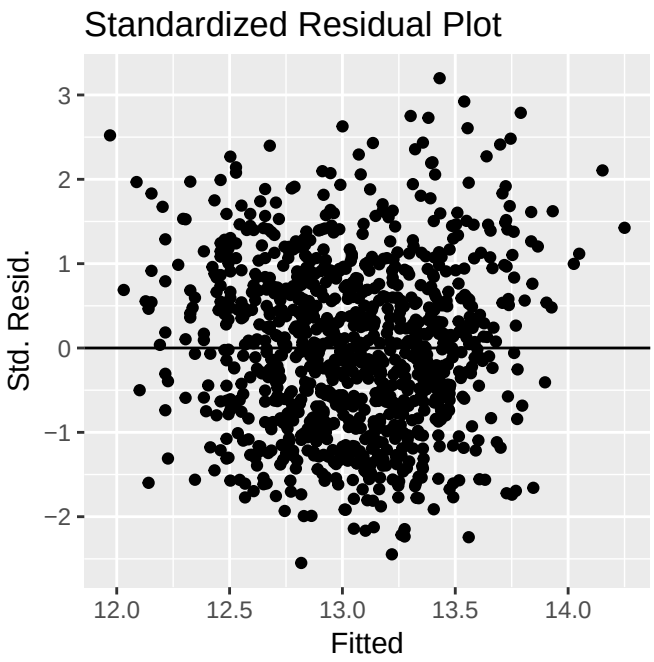
We are interested in examining the relationship between square footage of living space of a house and that corresponding lot's price for properties sold in 2015 in King County, WA (mostly Seattle). Specifically, we are curious if square footage of living space is a good predictor of price for these houses. Our original dataset consisted of 21,613 observations (property sales), but we have randomly selected 1,000 of these observations to be included in the following models, as this should not compromise the relationships between the variables and is less taxing on our computers.

Analysis









While the pre-transformed data is certainly linear in nature, it does not have constant variability. In fact, the data is severely heteroskedastic, as errors increase along with the predictor variable (sqft_living). We first tried a logarithmic transformation on the predictor, and while the variance was more stable, residual analysis made it clear that this transformation was introducting unintended curvilinearity into our model. We then tried a power model, tranforming both the predictor and response logarithmically. This model seems to address our concerns, as the data is still linear but now is shown to have constant variance as well.

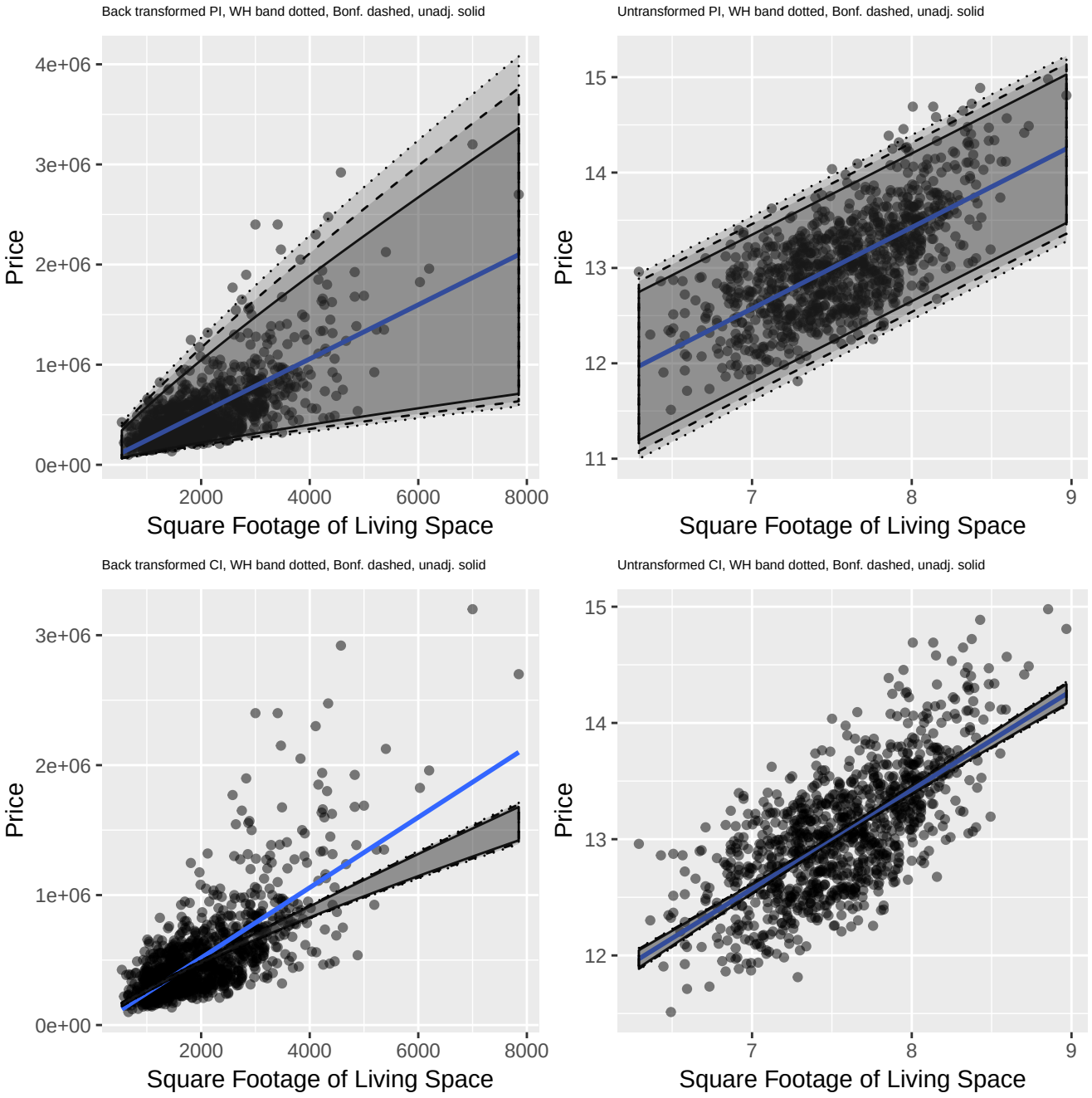
```
##
## Call:
## lm(formula = log(price) ~ log(sqft_living), data = housing2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0044 -0.2947 -0.0006  0.2707  1.2602
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        6.6118     0.2161    30.6   <2e-16 ***
## log(sqft_living)   0.8517     0.0286    29.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.394 on 998 degrees of freedom
## Multiple R-squared:  0.471,  Adjusted R-squared:  0.471
## F-statistic:  889 on 1 and 998 DF,  p-value: <2e-16
```

The R^2 value is 0.4711. This means that 47.11% of the variance in $log(price)$ is explained by a linear relationship with $log(sqft_living)$. The corresponding plot (found above) shows that there is seemingly a strong, positive relationship between $log(price)$ and $log(sqft_living)$, as we expected. In the residual plot, we see constant variability, so we are satisfied with the fit of this model.

We now run a hypothesis test on the value of β_1. Our null hypothesis is that $\beta_1 = 0$, and our alternative hypothesis is that $\beta_1 \neq 0$. As shown in the table above, the t-test statistic is 29.82, yielding a p-value of effectively 0. Thus, we may reject the null hypothesis at any reasonable level, and say with confidence that $\beta_1 = 0$ is false, and there is in fact some linear relationship between the logarithm of square footage of living

space and the logarithm of price for houses sold in King County in 2015.

We decided to look at predictions for properties with homes that were 2,000 square feet. This is considered a 'standard' size home, perhaps even a large one inside the city (where much of our data comes from). A 95% confidence interval for the mean price of lot with a 2000 sq. ft. house turns out to be ($470158.8, $493891.7) – pretty pricy! For an individual 2000 sq. ft. house, we are 95% confident that the cost will be between ($222152.5, $1045261). Again, not particularly reassuring for someone looking to buy a house in Seattle.



The above plots deal with simultaneous inference. It is important to adjust for multiple comparisons, as it is foolish to think that the intervals returned through standard methods at every point cover their true parameter with probability $(1 - \alpha)$. Thus, an adjustment is needed to widen the intervals, given by either Bonferroni's procedure or the Working-Hotelling procedure. The standard method leads to the lower bound on interval size of both procedures, but they introduce greater variance so as to provide greater accuracy

while not straying too far from this lower bound.

It is known that the Working-Hotelling procedure outperforms the Bonferroni procedure when larger subsets of the predictor are considered. Since we are examining the entire support of the predictor here, we prefer the WH procedure. The plots also clearly show that the WH band is tighter than the Bonferroni band. Both of these methods are preferred to the unadjusted, however. Finally, it should be noted that things get squirrely with the back-transformed CI, for all procedures, as it appears simultaneous inference breaks down.

Conclusion

Our analyses have shown that there is a positive, linear relationship between $log(sqft_living)$ and $log(price)$, meaning in general, the bigger the house, the more expensive the lot is. We don't doubt the linearity in the relationship between the untransformed variables, but the heteroskedastic nature makes fitting a linear model a rather suspicious endeavor, as the standard errors will be incorrect. While we suspect that house size is the primary driver of price, it will be interesting to examine how other variables (such as lot size, view, location) drive the price as well, and in the end, the R^2 score of .4711 is notable, but not as strong as we anticipated. It will be interesting to extend this to a multilinear model in the future.