PSTAT126 Project Step-2

Liuqian Bao, Stella Ramirez, Andrew Cheng

2023-11-12

Introduction

Our data source is from: Anna Montoya, DataCanary. (2016).House Prices - Advanced Regression Techniques. Kaggle.Obtained from https://kaggle.com/competitions/house-prices-advanced-regression-techniques (https://kaggle.com/competitions/house-prices^advanced-regression-techniques)

The population that we are inferring our results on are all residential houses in Ames, Iowa.

We are using the variables BsmtFinSF1 and SalePrice as our variables of interest. The BsmtFinSF1 variable is our predictor variable that we will use for hypothesis testing and plotting. The BsmtFinSF1 variable refers to the basement finished area square feet in the overall housing data. The SalePrice variable refers to the property's sale price in dollars, and it is our response, or dependent, variable that is affected by BsmtFinSF1.

We first fitted a simple linear model, and after exploring the data and checking model assumptions, we did a log-transformation on our response variable in order to fit the model better.

Hypothesis: Our hypothesis is based on the predictor variable, BsmtFinSF1, and the response variable, SalePrice. Our null hypothesis is that BsmtFinSF1 and SalePrice are not linearly correlated, $\beta_1 = 0$. Our alternative hypothesis is that BsmtFinSF1 and SalePrice are positively correlated, $\beta_1 > 0$. #### Simple Linear Model

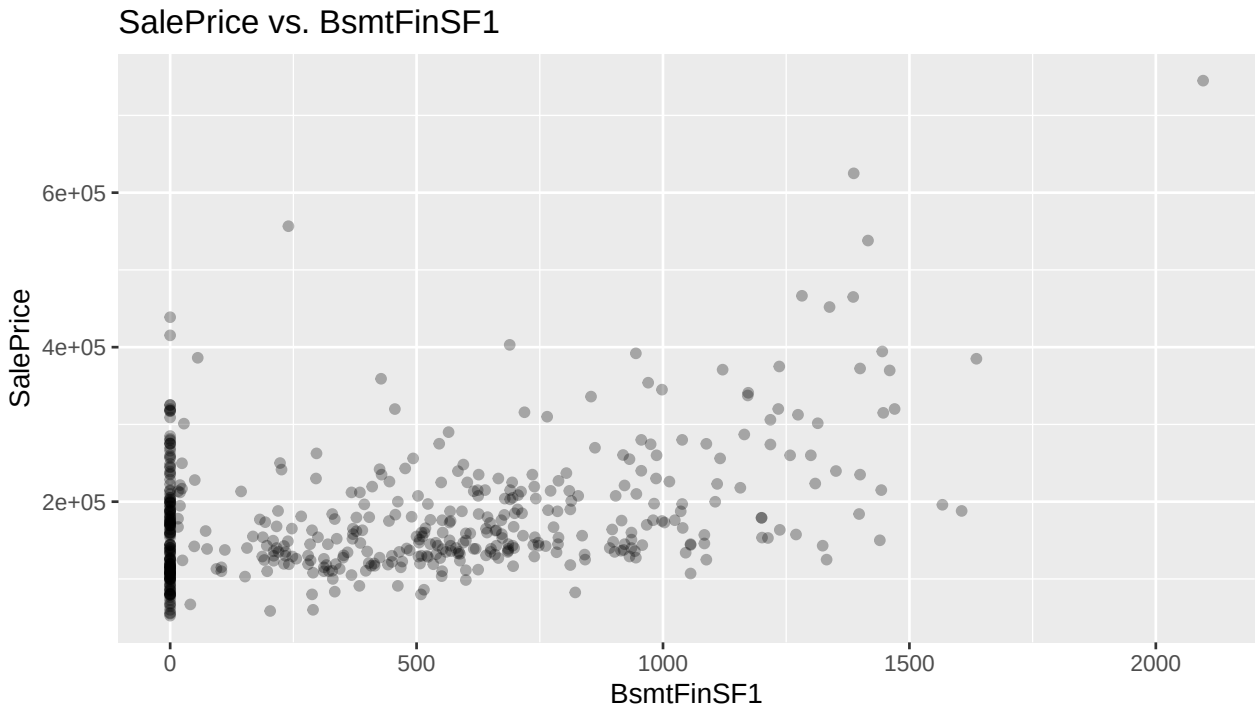


Figure 1: Simple Linear Model with BsmtFinSF1 and SalePrice

Assumption Checks for the Simple Linear Model

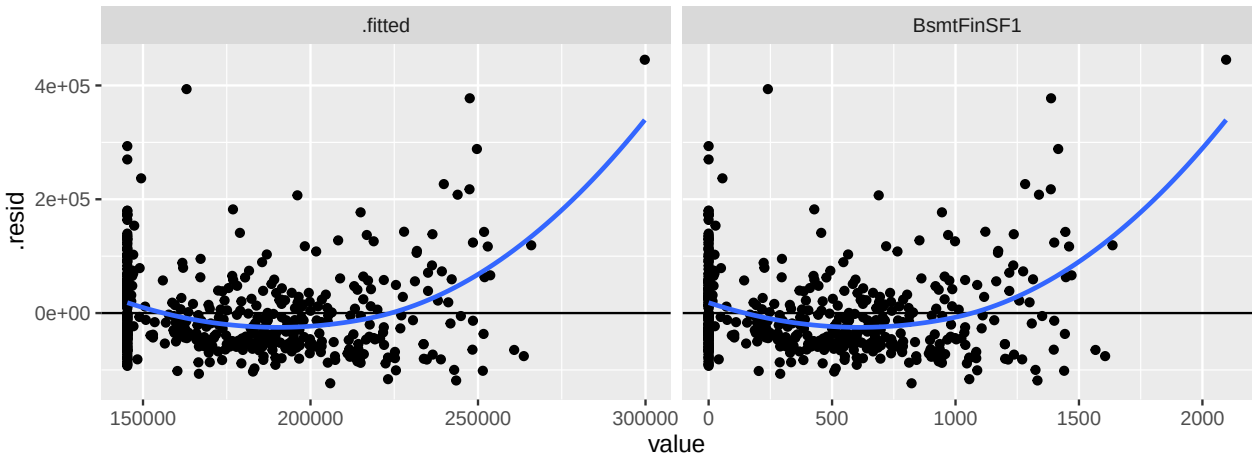


Figure 2: Residual vs. fitted and BsmtFinSF1

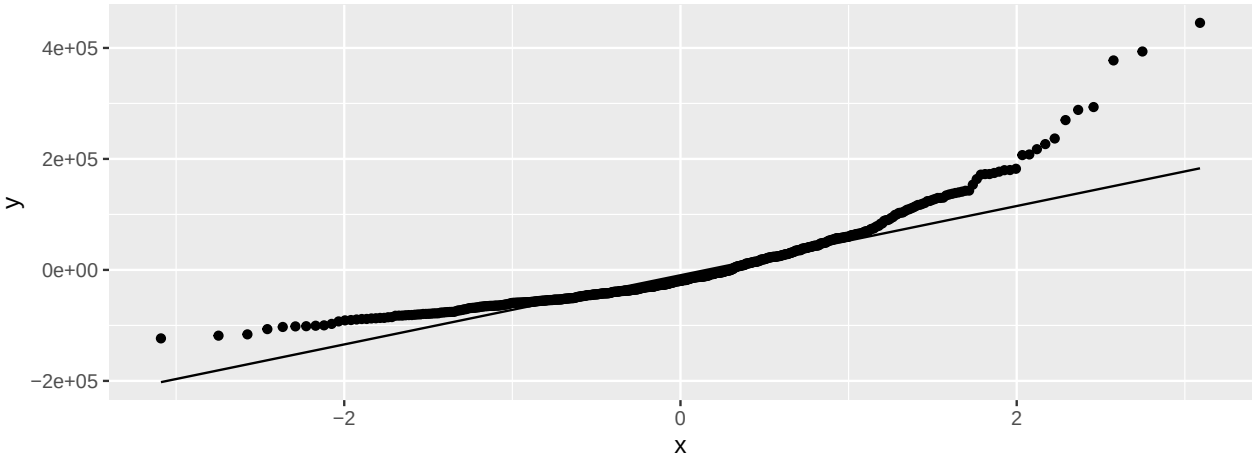


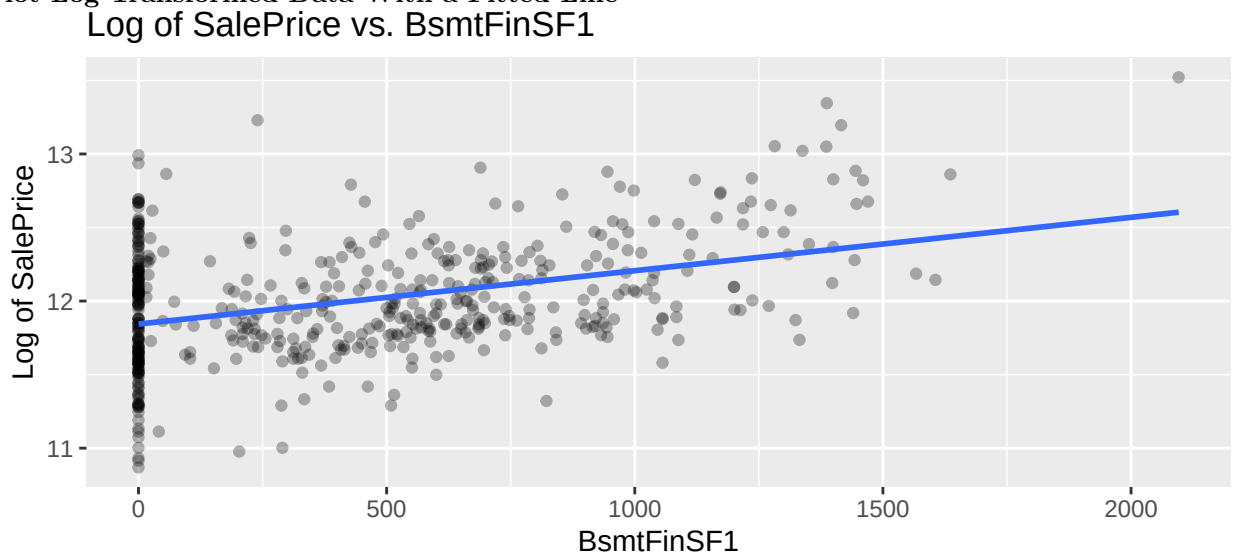
Figure 3: Linearity Check

Comments: The residual vs. fitted plot showed some slight pattern(i.e. greater variance at the two ends and smaller variance in the middle), and the qq plot also showed a slightly curved line, so we decided to try log transforming our response variable SalePrice to get a more constant variance.

Log Model

In the following section we attempted to do a log transformation on the response variable, SalePrice, and fit the explanatory variable, BsmtFinSF1, linearly with the log transformed response.

Plot Log Transformed Data With a Fitted Line



Assumption Checks for Log Model

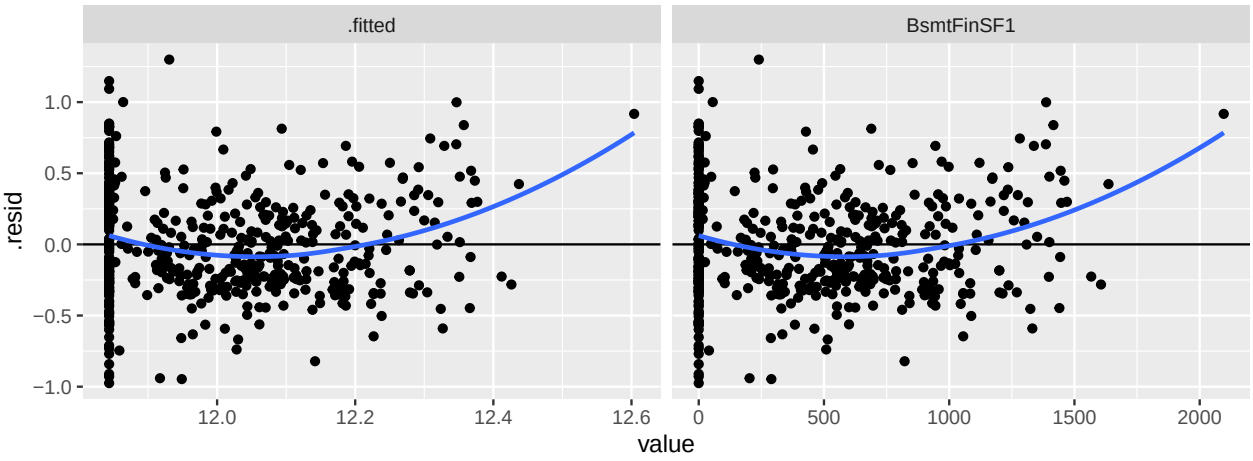


Figure 4: Residual vs. fitted and BsmtFinSF1, New Model

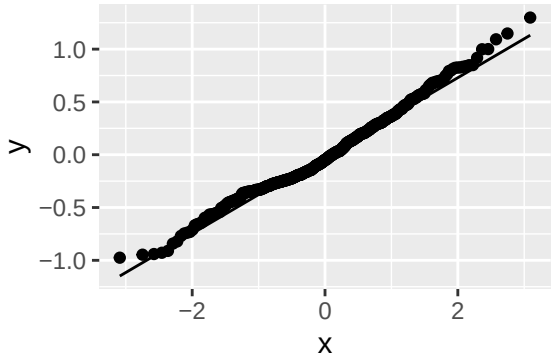


Figure 5: Linearity Check

As we can see from the residual vs. fitted plot and qq plot for the log model, the log transformation makes the variance of the residuals more constant throughout our data set, and the qq plot fits a straight line much

better after the log transformation. These tells us that the log transformed model meets the assumptions we made better, thus we decided to proceed with the log transformed model.

T-Test for β_1

We performed a t-test on the coefficient of BsmtFinSF1 in our linear model, β_1.

Hypothesis: Our hypothesis is based on the predictor variable, BsmtFinSF1, and the response variable, SalePrice. Our null hypothesis is that BsmtFinSF1 and SalePrice are not linearly correlated, $\beta_1 = 0$. Our alternative hypothesis is that BsmtFinSF1 and SalePrice are positively correlated, $\beta_1 > 0$.

p value:

```
##                 Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept) 1.184366e+01 2.395519e-02 494.408910 0.000000e+00
## BsmtFinSF1  3.626989e-04 3.868634e-05   9.375375 2.413948e-19
```

Conclusion: We are doing a one sided test, so the p-value will be divided by 2. As a result, we get a very small p-value, 1.205e-19, which allows us to reject our null hypothesis that $\beta_1 = 0$ and accept our alternative hypothesis that $\beta_1 > 0$. Our conclusion from the t-test is that BsmtFinSF1 is positively correlated with the log of the SalePrice.

Confidence Interval for β_1(coefficient of BsmtFinSF1)

The confidence interval we computed for β_1 is:

```
##                    2.5 %       97.5 %
## BsmtFinSF1 0.0002866904 0.0004387075
```

Interpretation: With 95% confidence, a 1 square foot increase in basement square feet is associated with an increase in average of the log of the sales prices between an estimated 0.0002866904 and 0.0004387075.

Confidence Interval for Mean and Individual Response

The following graph displays the mean of BsmtFinSF1(basement finished area), at which the following CI is calculated and the fitted value of response at the mean of BsmtFinSF1.

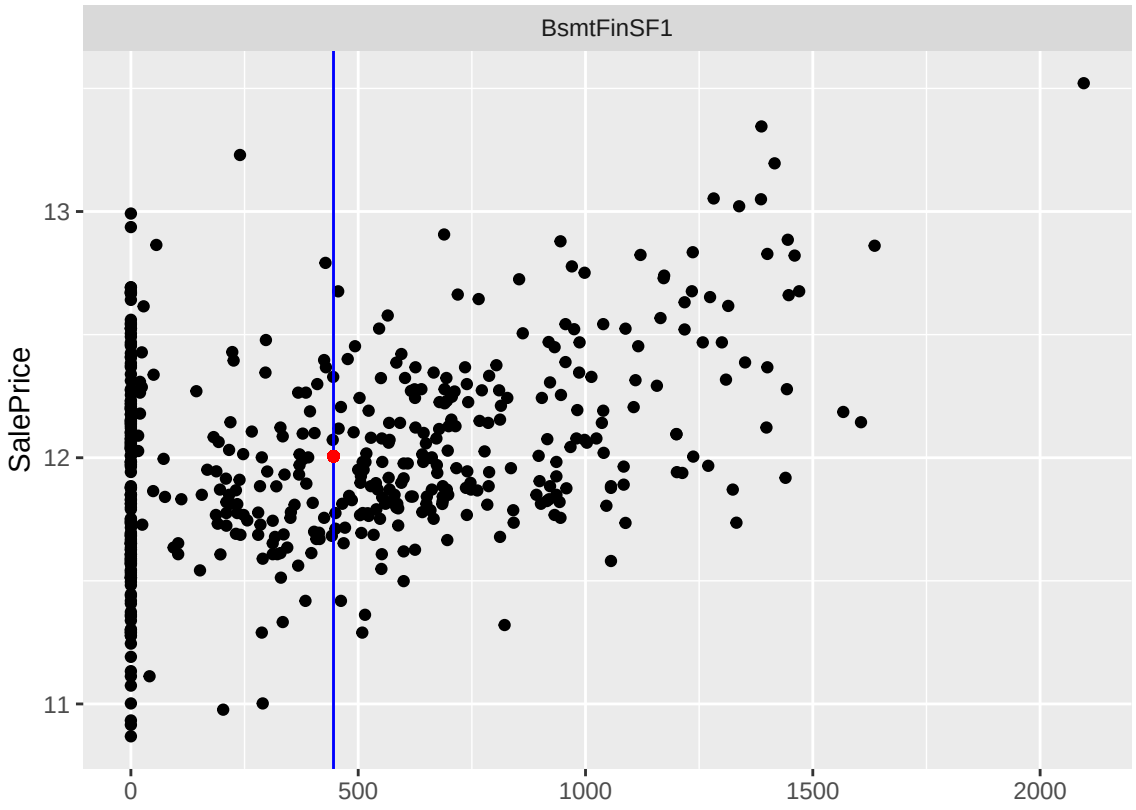


Figure 6: Visualization of Mean Line

Confidence interval:

```
##       fit      lwr      upr
## 1 12.0053 11.97263 12.03798
```

Interpretation: With 95% confidence, the mean of the log SalePrice for a house with basement area equal to the average in the data is estimated to be between 11.97263 and 12.03798.

Confidence interval for individual response at an interesting x value We used the point with x(basement finished area) value of 2096 as our point of interest, because this is the largest basement finished area in the 500 observations we used and the point is an outlier in our data.

In the following graph, the actual data point at x = 2096 is shown in blue and the fitted value of our model is shown in red.

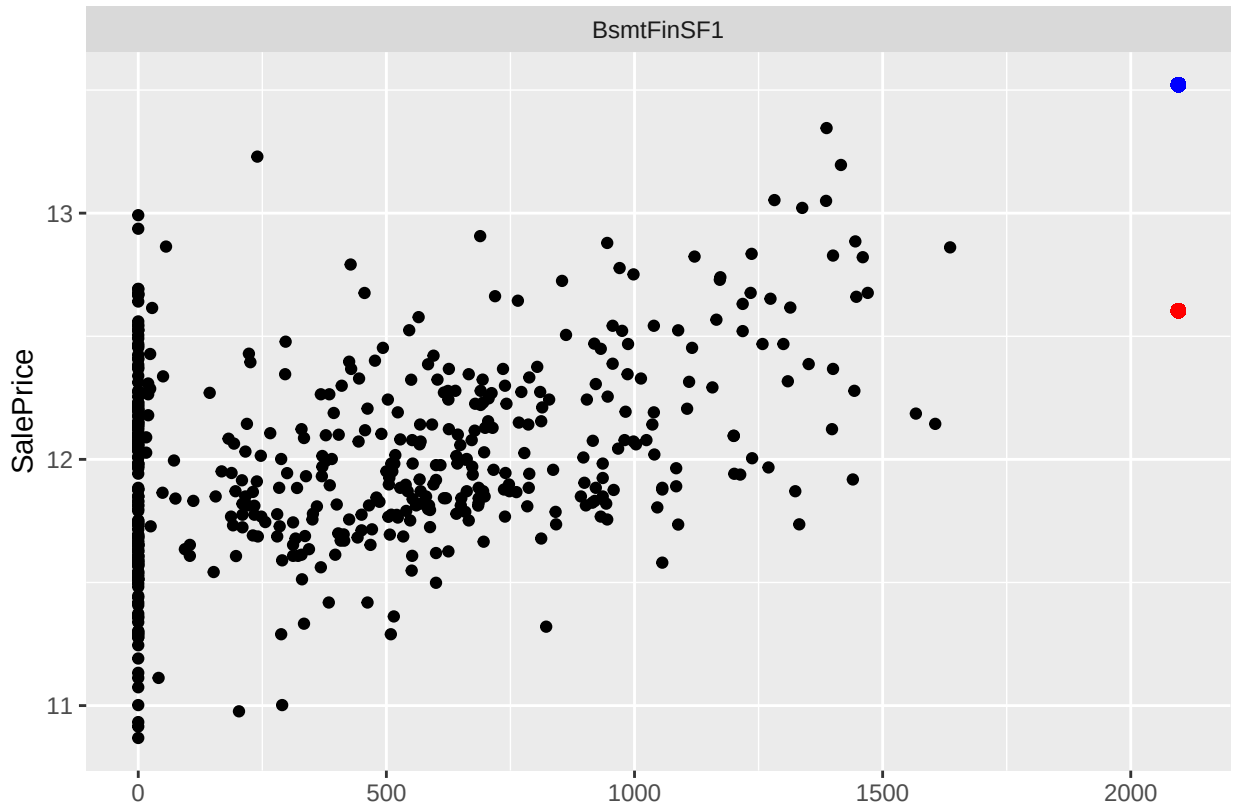


Figure 7: Visualization of the Prediction at an Interesting Point

Confidence interval:

```
##        fit      lwr      upr
## 1 12.60388 11.86181 13.34594
```

Interpretation: With 95% confidence, the predicted log of the SalePrice of a house with a finished basement area of 2096 square feet is estimated to be between 11.86181 and 13.34594.

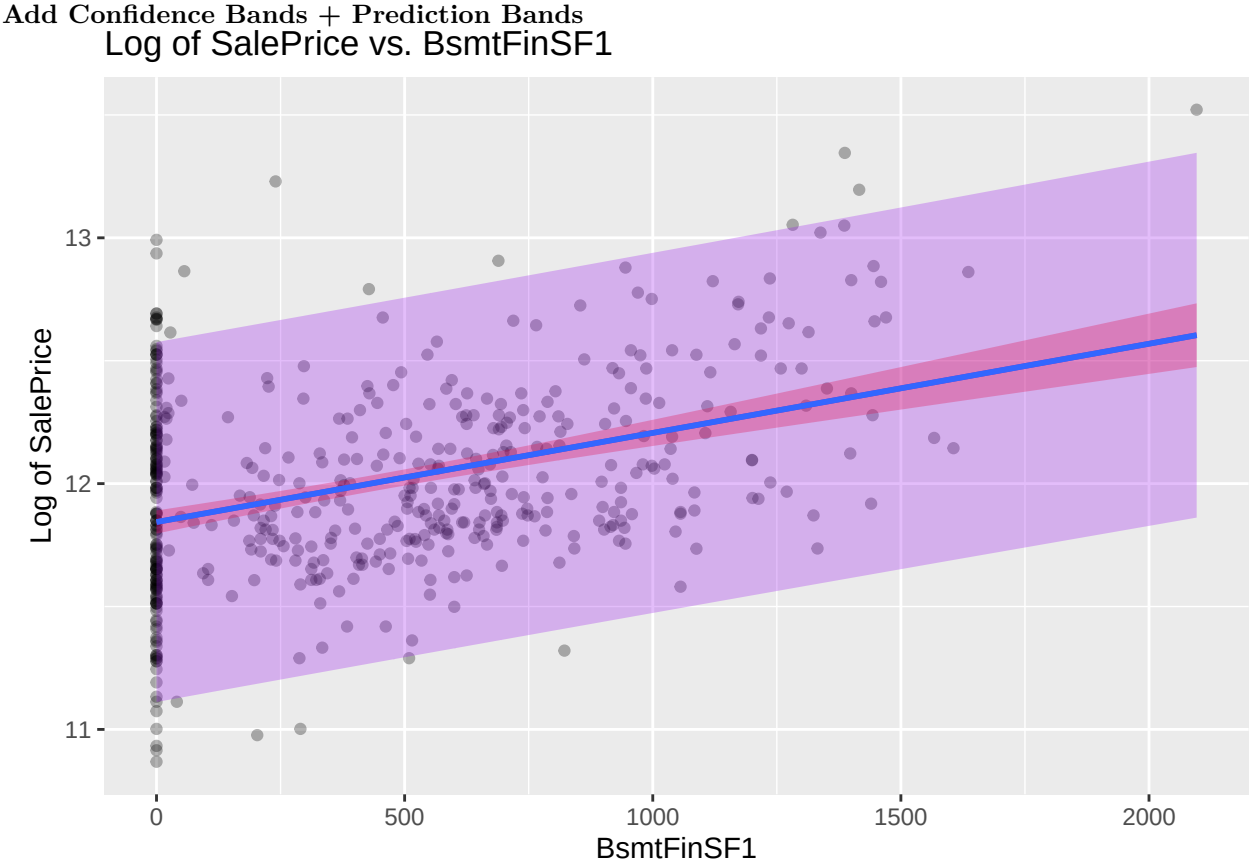
Interesting feature: The actual value of the log of the sale price of the house with a finished basement area of 2096 square feet is outside(greater than) the prediction interval we obtained using our model. This is interesting because it tells us it is very challenging to predict sale prices of houses that are extremely large(judged by its basement area), potentially some luxury properties, using a simple linear model as the one we built.

R^2

```
## [1] 0.1483154
```

Our model has an Adjusted R-squared value of 0.1483. This tells us that 14.82% of the variance in SalePrice is explained by our simple linear model with the log transformation with the explanatory variable basement finished area(BsmtFinSF1). This means our model explains only a relatively small amount of the variance, but it is reasonable considering we are only using one variable.

Add Confidence Bands + Prediction Bands



As you can see from this graph, the prediction band is much wider than the confidence band. This is accurate because while the CI depends on the estimates, the prediction interval at a specific value is influenced by both the estimates and the variation in the response. It is also interesting to note that the values moving to the right of the graph have a wider interval, implying higher uncertainty. Following our analysis, this conclusion supports our suggestion that it may be difficult to use existing data to predict the Sale Price of very large properties.

Conclusion

After checking the normality of our original model, we decided to take the log of our response variable to make the variance more constant. Then we performed several tests to check the significance of our variable, and see what it may predict or allow us to infer. We did a hypothesis test for the coefficient, constructed confidence and prediction intervals for the mean and individual response at an interesting level of x, assessed the fit of our model by looking at the R^2 and residual graphs, and graphed the prediction and confidence interval bands.

Interesting feature:

The conclusion we drew from the t-test for β_1 lines up with our expectation because we would expect the sale price of a house to be higher if it has a larger finished basement area. The t-test gives a very small p-value, which tells us that the correlation is statistically significant.

Our finding through the prediction interval at an x level that is the maximum(2096 square feet) of our finished basement area data is also interesting. The actual value of the log of the sale price of the house with a finished basement area of 2096 square feet is outside(greater than) the prediction interval we obtained using our model. It tells us it is very challenging to predict sale prices of houses that are extremely large (judged by its basement area), potentially some luxury properties, using a simple linear model as the one we built. Other questions that we would like to explore within the data is whether the response variable is influenced by other variables.