Izaak King

Professor Eleish

Data Science

November 4, 2025

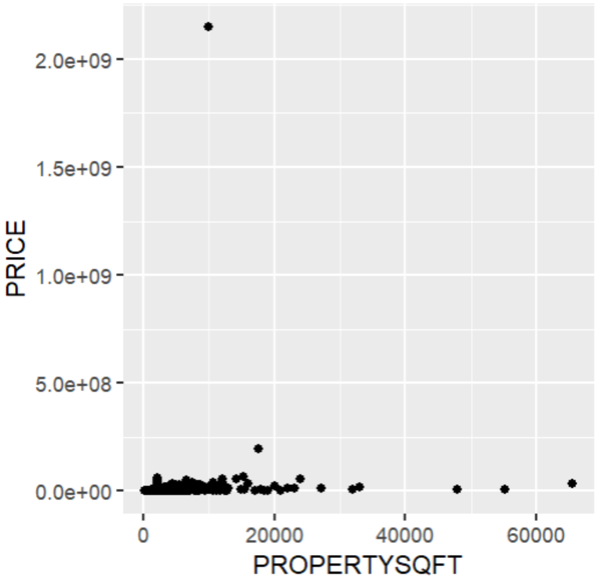
"Lab 5"

For this lab, I will use the NY House Dataset to train 3 regression models using different algorithms to predict price from square footage. Particularly, I will use a simple linear regression model, a linear SVM model, and a radial SVM model. I will then evaluate their performances using the MAE, MSE and RMSE metrics.

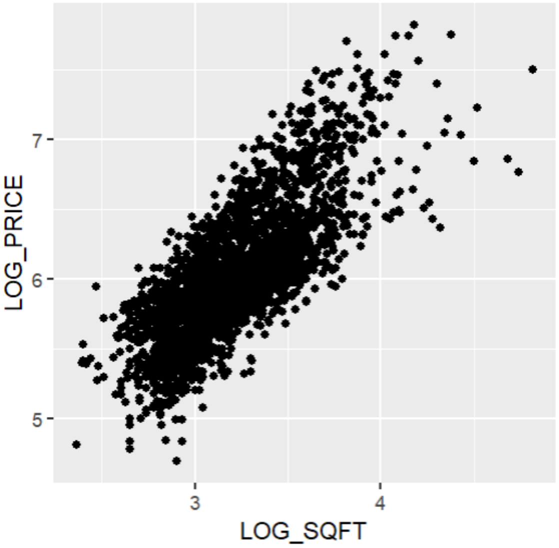
I. Dataset cleaning

Before beginning the analysis, I took time to clean the NY House Dataset. To do this, I visually inspected the dataset, and I made a plot of the data with respect to square footage and price. There were clear outliers for price, that I decided to remove. Further, I applied a log10 transformation to both price and square footage to bring the variables to a more comparable scale. Overall, this process makes the data much cleaner and efficient to analyze, as trends won't be swayed by outliers.

a. Scatter plot of PROPERTYSQFT and PRICE (before cleaning)

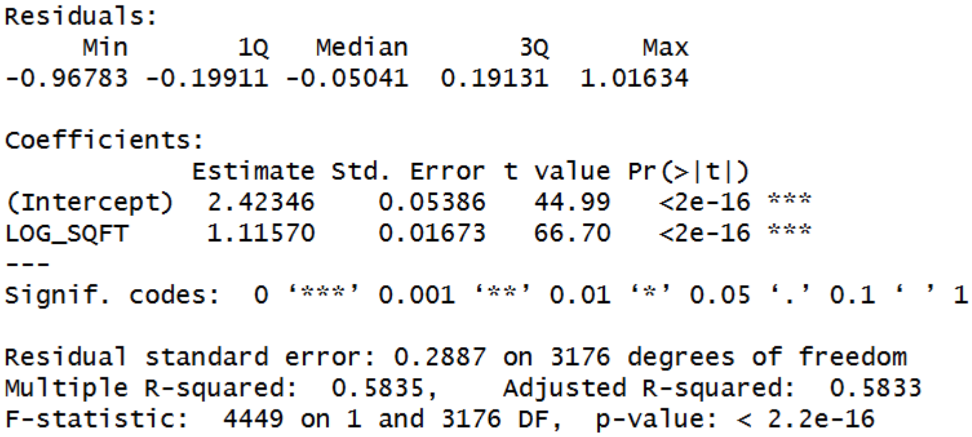


b. Scatter plot of LOG_SQFT and LOG_PRICE (after cleaning)



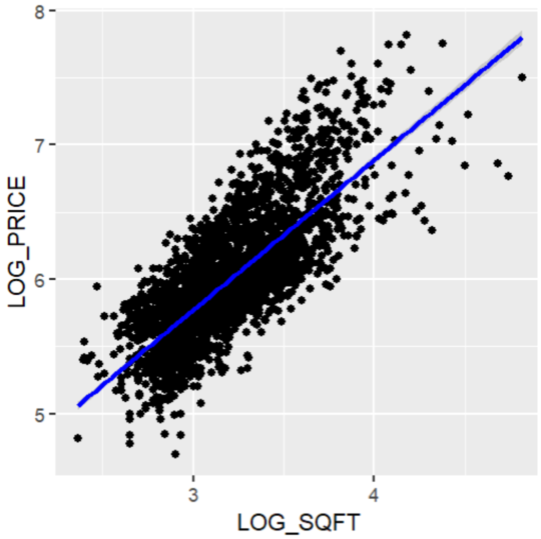
II. Linear Regression Model (LOG_SQFT predicting LOG_PRICE)

a. Summary

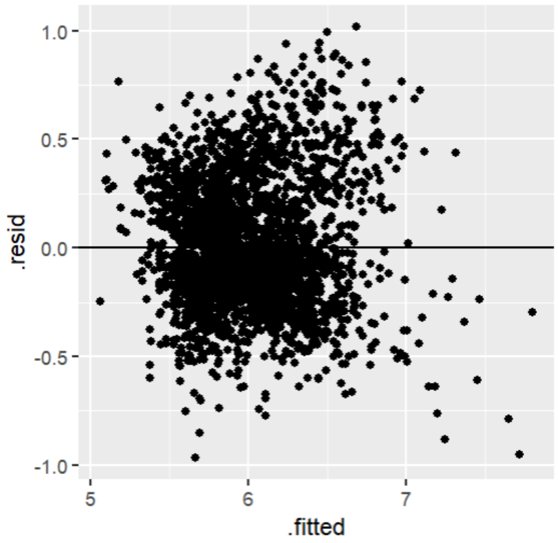


```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.96783 -0.19911 -0.05041  0.19131  1.01634 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.42346    0.05386   44.99   <2e-16 ***
LOG_SQFT     1.11570    0.01673   66.70   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2887 on 3176 degrees of freedom
Multiple R-squared:  0.5835,	Adjusted R-squared:  0.5833 
F-statistic:  4449 on 1 and 3176 DF,  p-value: < 2.2e-16
```

b. Scatter plot with fitted line



c. Residual plot

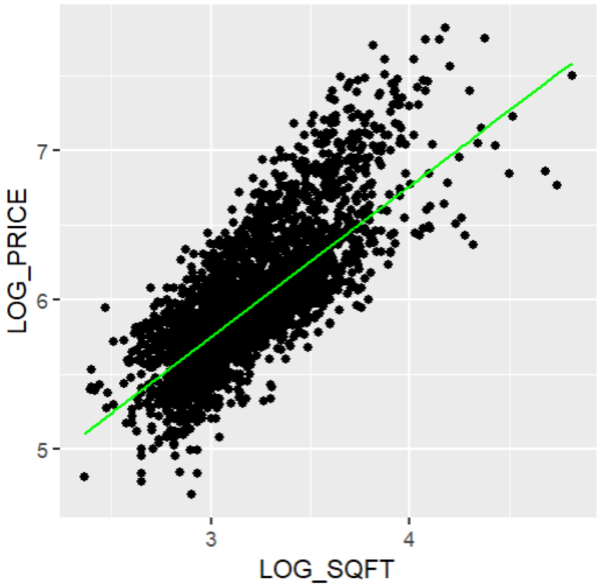


III. SVM - Linear Kernel (LOG_SQFT predicting LOG_PRICE)

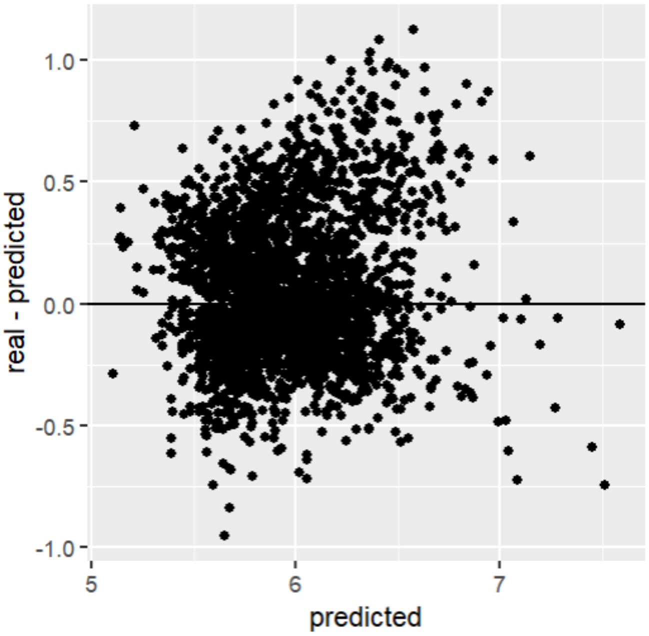
a. Summary

```
svm(formula = LOG_PRICE ~ LOG_SQFT, data = dataset, kernel = "linear")


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  linear
       cost:  1
      gamma:  1
    epsilon:  0.1


Number of Support Vectors:  2785
```

b. Scatter plot with fitted line

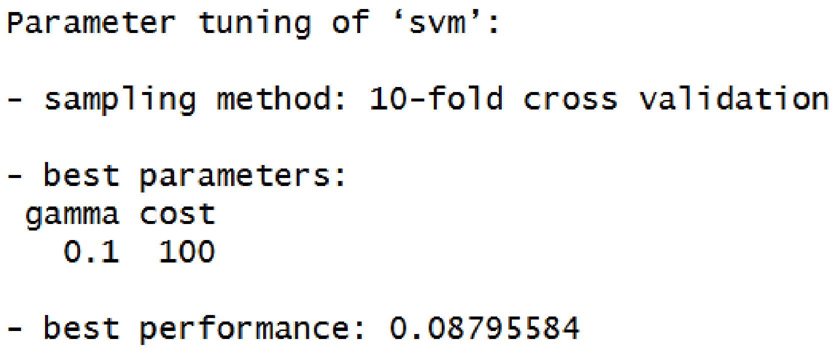


c. Residual Plot



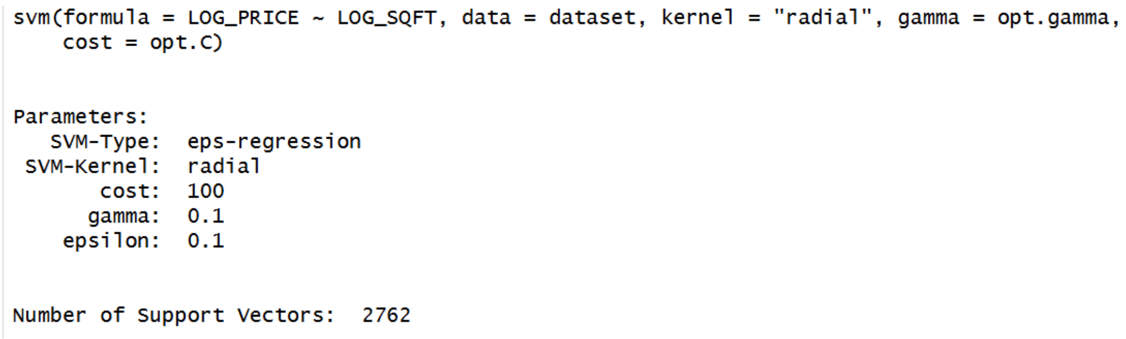
IV. SVM - Radial Kernel (LOG_SQFT predicting LOG_PRICE)

a. Tuning parameters (Random subset of 500)



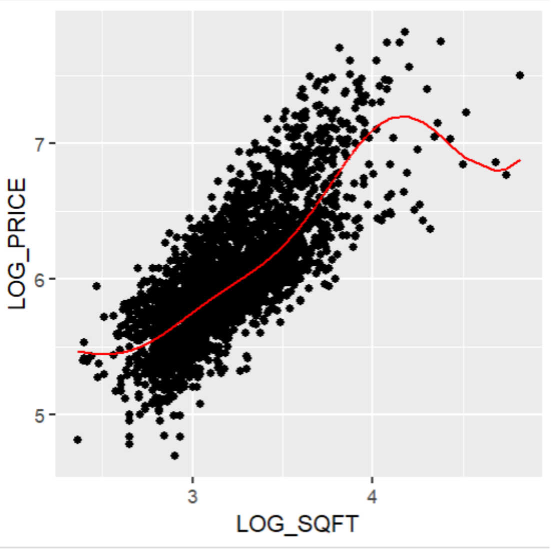
```
Parameter tuning of ‘svm’:

- sampling method: 10-fold cross validation

- best parameters:
 gamma cost
   0.1  100

- best performance: 0.08795584
```

b. Summary

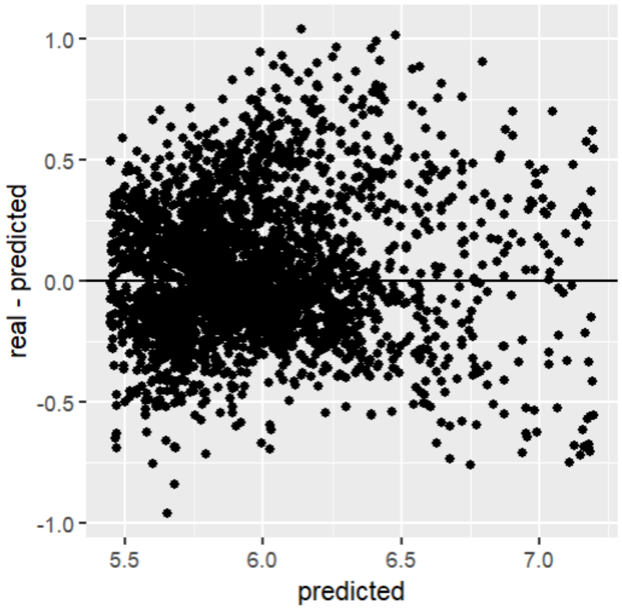


```
svm(formula = LOG_PRICE ~ LOG_SQFT, data = dataset, kernel = "radial", gamma = opt.gamma,
    cost = opt.C)


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  100
      gamma:  0.1
    epsilon:  0.1


Number of Support Vectors:  2762
```

c. Scatter plot with fitted line



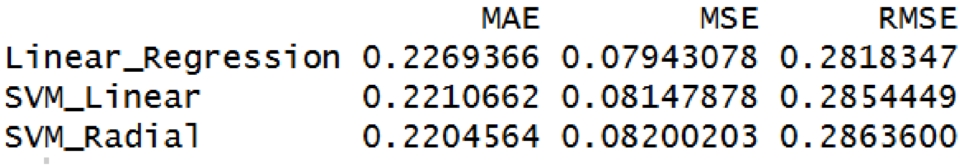
d. Residual Plot



V. Error Analysis

To analyze the errors of these three created models, we will compute the MAE, MSE, and RMSE of our linear regression model, our SVM with the linear kernel and our SVM model with the radial kernel. We first randomly divide the dataset into two, with 75% of the data being used for training, and the other 25% used for testing, which will be used to compute our metrics. During calculations, I chose output the metrics as one coherent matrix, featuring metrics for all three models.

a. Error Metric matrix



```
                        MAE        MSE      RMSE
Linear_Regression 0.2269366 0.07943078 0.2818347
SVM_Linear        0.2210662 0.08147878 0.2854449
SVM_Radial        0.2204564 0.08200203 0.2863600
```

Overall, based on these metrics all three models performed relatively similar. MAE was slightly higher for the linear regression model, whereas MSE and RMSE were slightly higher for the SVM models. Overall, these results reflect that the relationship between log square footage and log price is roughly linear, as linear regression models perform comparably to the more flexible SVM models. Thus, a linear regression model is likely sufficient for this dataset, as using SVMs do not provide a substantial improvement in predictive accuracy, as indicated by the error metrics.