zillow project Barrett Jones

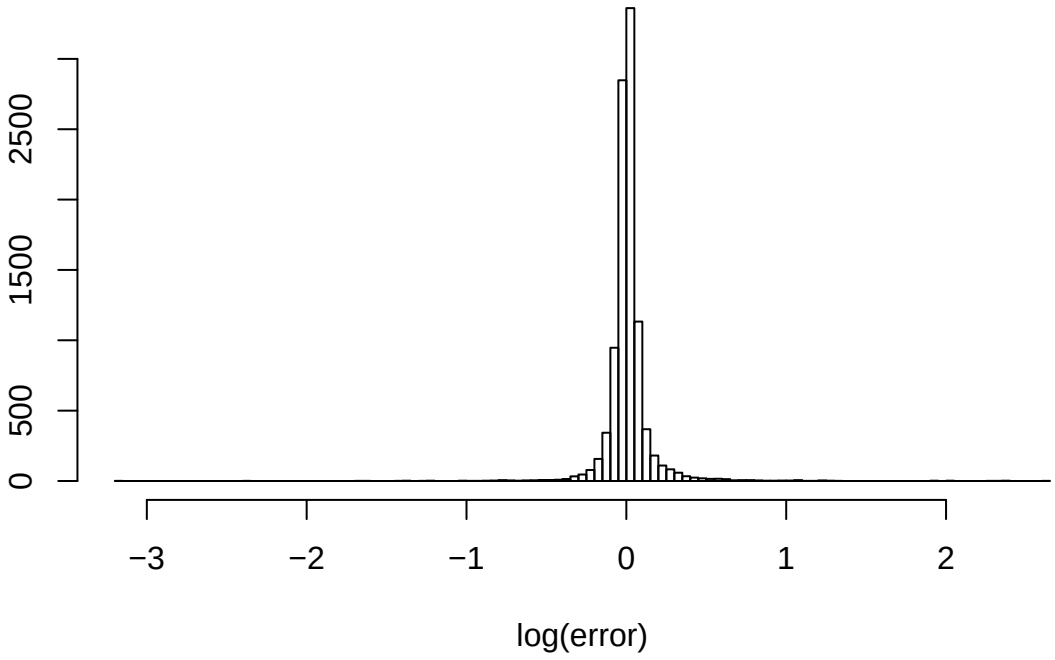
Barrett Jones

10/7/2017

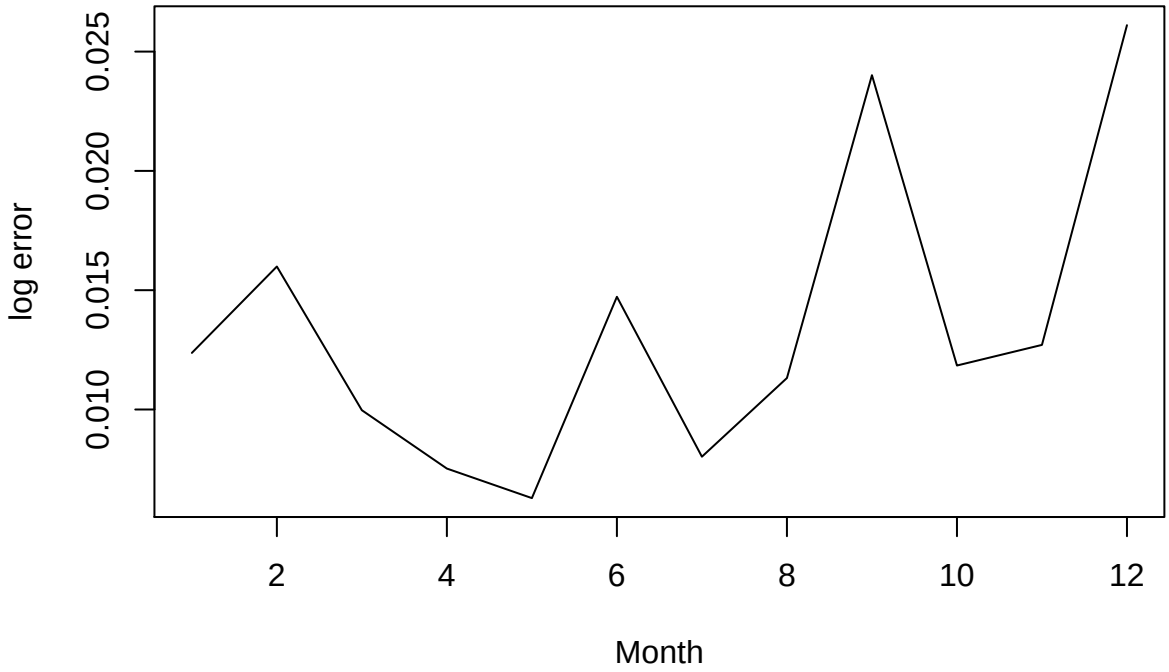
below we take a quick look at the data and some summary stats. I have reduced the data size by quite a bit (to 10,000 observations) so I can do some test modeling. I am trying to predict the log error in the zillow housing price model. You can see above that the log error follows a symmetric distribution with very long tails, and that log error seems to vary greatly month to month.

Hist of outcome

(Zillow Housing Price Model log(Error))



Trend Log Error by Month



I will try two different types of regression models in this data to see if I can get some good predictive power. First I will look at a multilevel model with random intercepts for zip code and city. After that I will take a look at a generalized additive model with some polynomial terms.

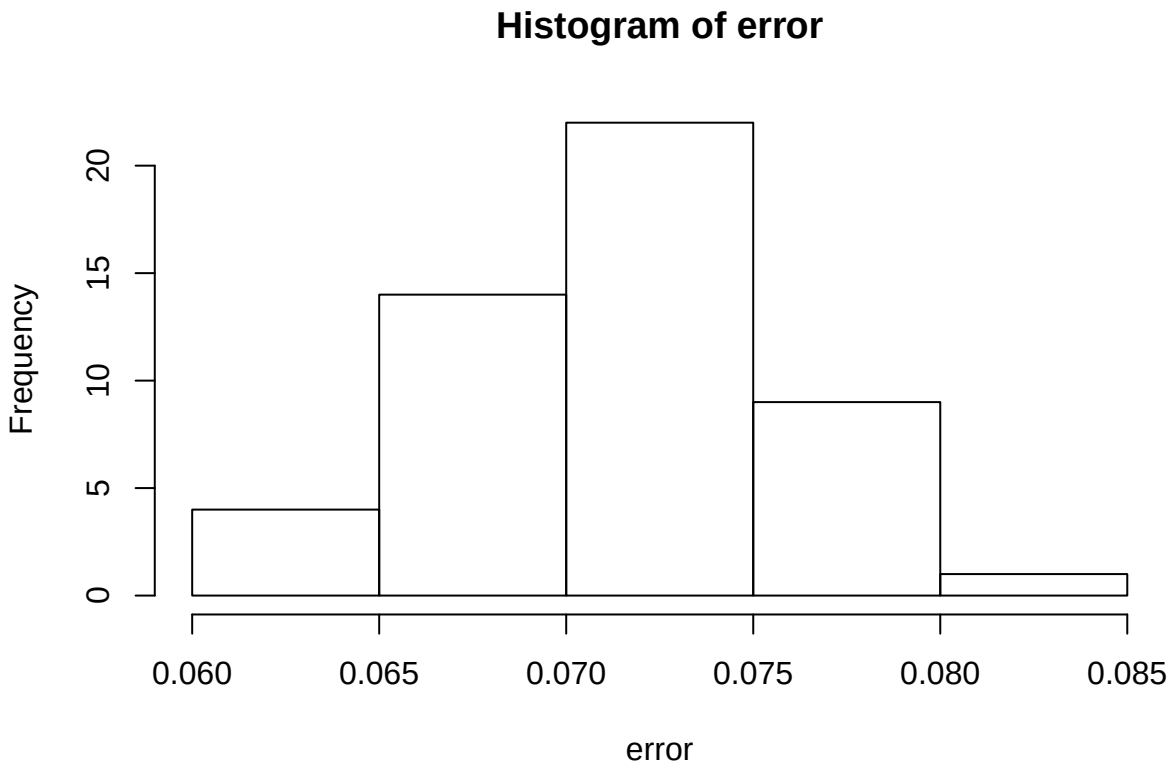
Model 1: Random Intercept Model

```
## lmer(formula = logerror ~ calculatedfinishedsquarefeet + month +
##     bathroomcnt + bedroomcnt + (1 | regionidzip) + (1 | regionidcity),
##     data = traindat)
```

Summary of Errors

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06290 0.06803 0.07165 0.07120 0.07419 0.08336
```

Distribution of Errors



So it looks like this model performs ok the median of the mean absolute error from the 50 cross validation cuts is 0.0716541, but I notice that the relationship between bedroom/bathroom count and logerror is not linear. I will try a gam to fit a polynomial regression model.

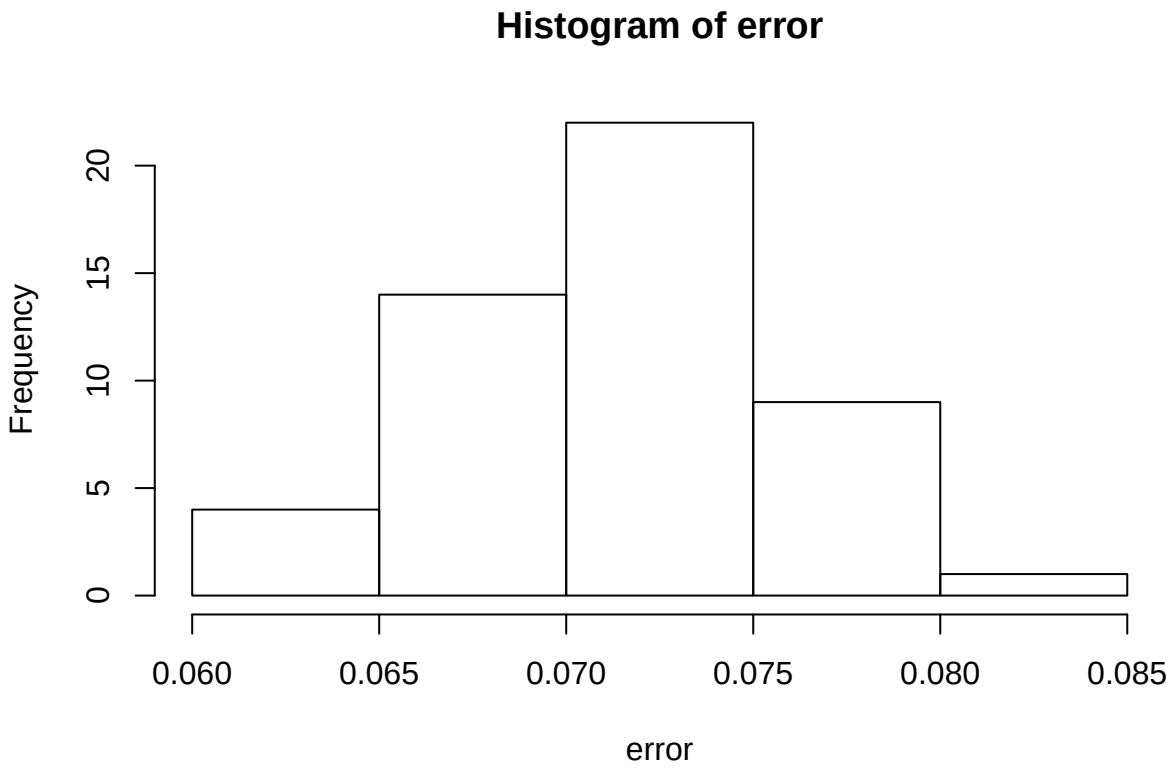
Model 2: Generalized Additive Model

```
## gam(formula = logerror ~ s(calculatedfinishedsquarefeet) + s(bathroomcnt) +
##     s(bedroomcnt) + as.factor(month), data = traindat)
```

Summary of Errors

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06302 0.06830 0.07168 0.07136 0.07436 0.08369
```

Distribution of Errors



looks like I got about the same performance from the gam. The median cross validate mean absolute error=0.0716812,is slightly different from the random intercept model, but not by much. Probably would be pretty similar results on the full data set. I will have to do some more testing, perhaps a hierarchial model with some polynomial terms with perform better.