

zillow project Barrett Jones

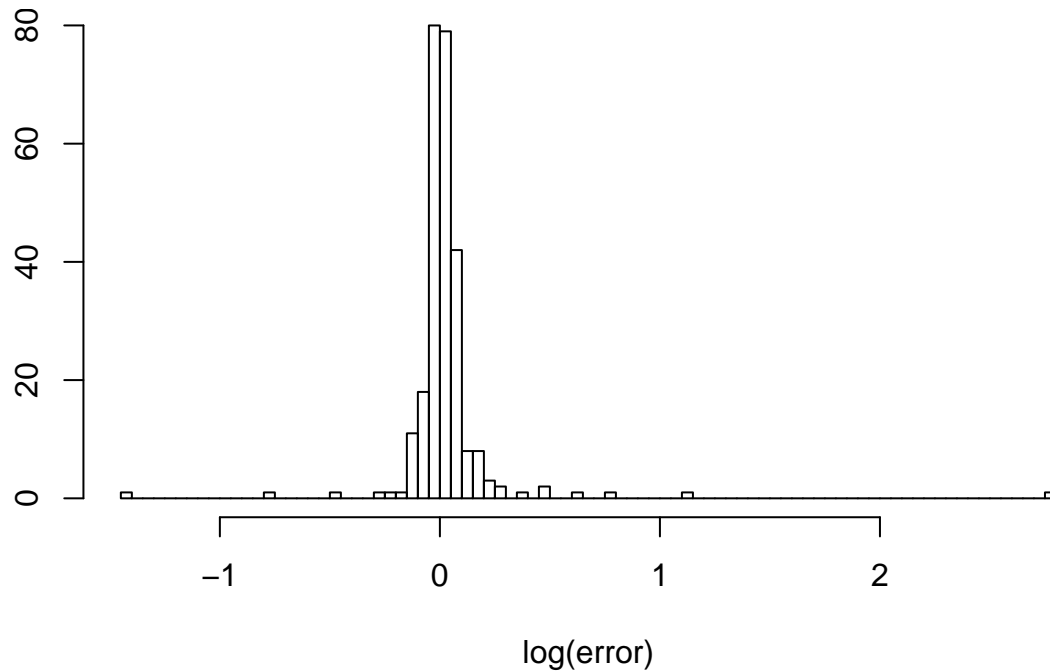
Barrett Jones

10/7/2017

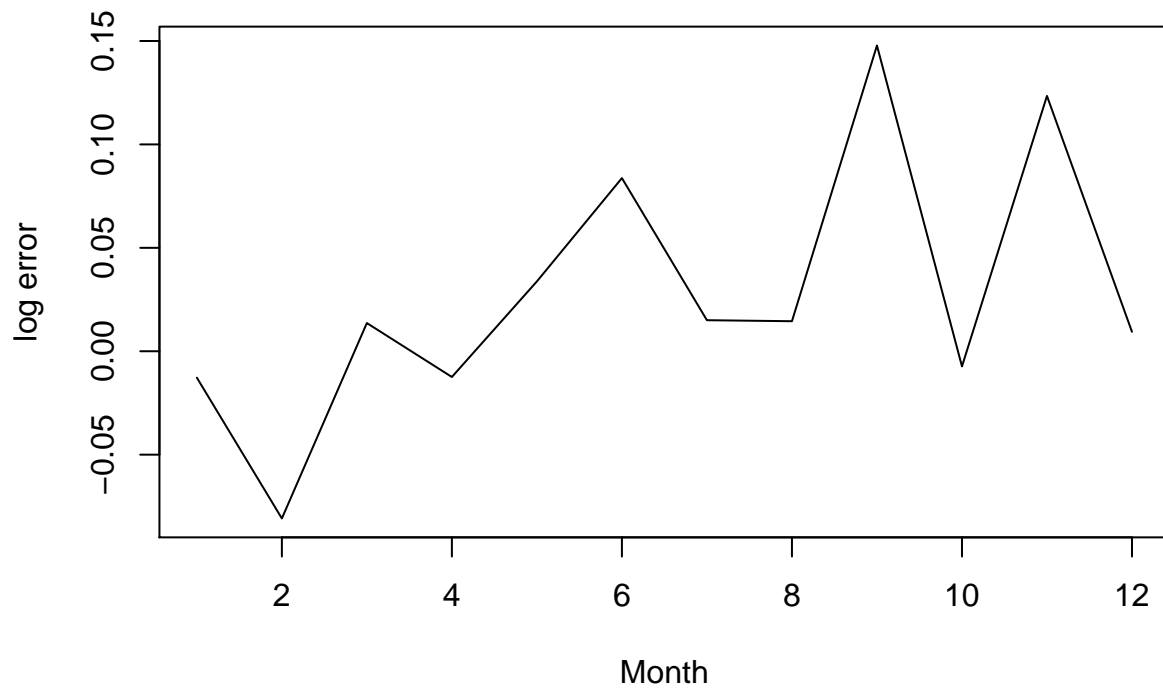
below we take a quick look at the data and some summary stats. I have reduced the data size by quite a bit (to 10,000 observations) so I can do some test modeling. I am trying to predict the log error in the zillow housing price model. You can see above that the log error follows a symmetric distribution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

Hist of outcome

(Zillow Housing Price Model $\log(\text{Error})$)



Trend Log Error by Month



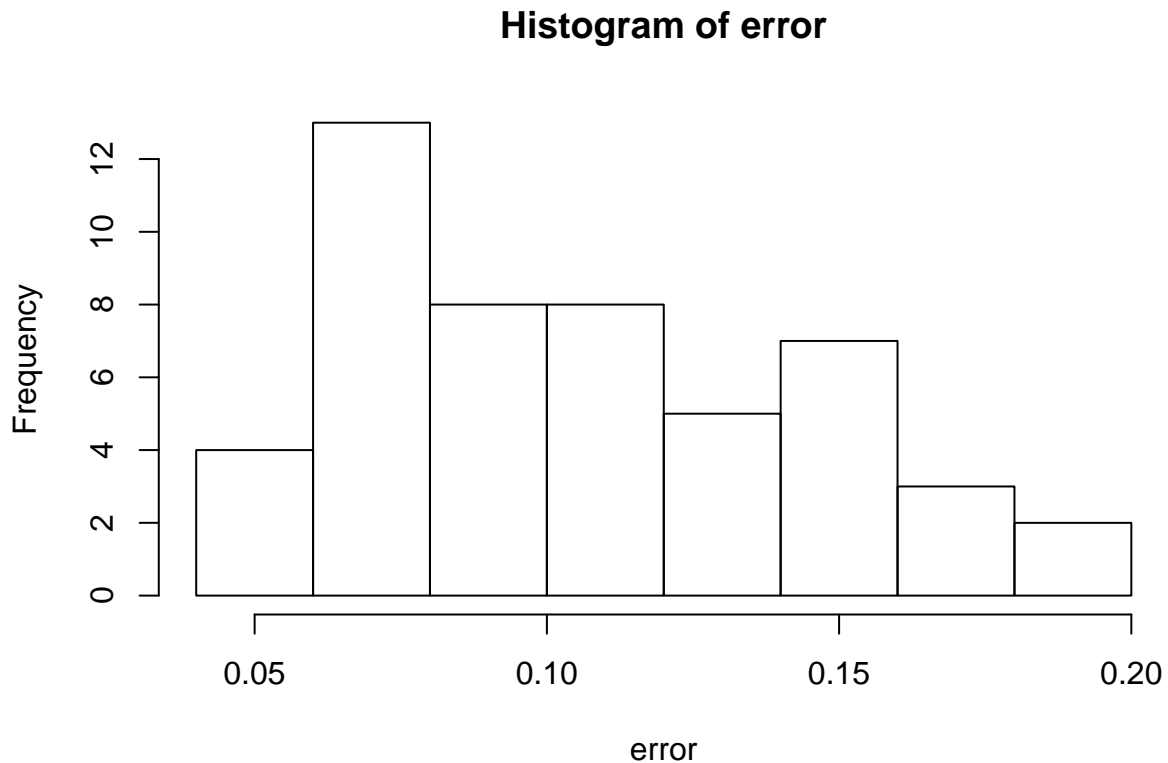
Model 1: Random Intercept Model

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

Summary of Errors

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.  
## 0.04161 0.07526 0.10010 0.10580 0.13590 0.19140
```

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 0.1001248, 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.

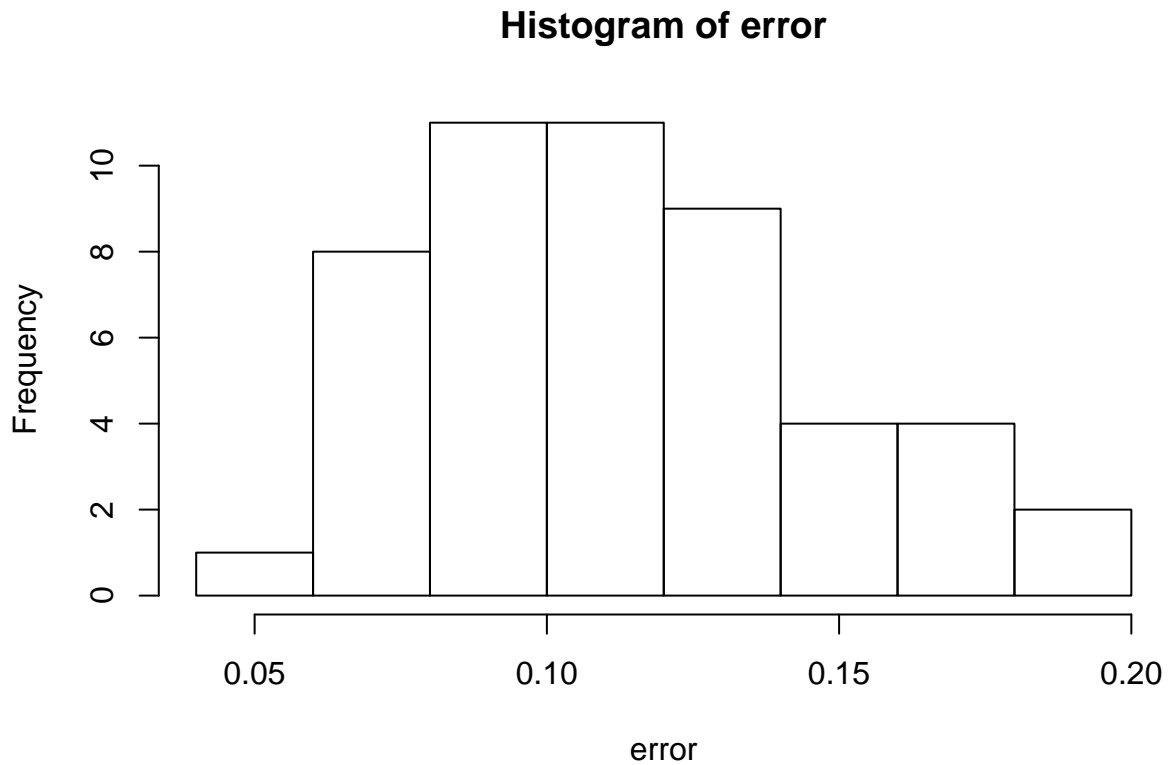
Model 2: Generalized Additive Model

```
## 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.  
## 0.05708 0.09181 0.10940 0.11390 0.13620 0.19120
```

Distribution of Errors



looks like I got about the same performance from the gam. The median cross validate mean absolute error=0.1093742, 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 hierarchical model with some polynomial terms will perform better.