Step 3: P(e|s, t): Ending Station given starting station and time

Method1: Experimental Proportion

$$Y_{i,.,t}\text{the number of rides starting at station i at time t}$$

$$Y_{i,j,t}\text{the number of rides starting at station ending at station j at time t}$$

$$P_{i,j,t} = P(e_j|s_i, t) = \frac{Y_{i,j,t}}{Y_{i,.,t}}\text{ for each day}$$

$$\hat{P}_{i,j,t} = \frac{\sum_{d=1}^{n} P_{i,j,t}}{n}$$

There are 21(days) * 24(hours) * 319(starting stations) * 319(end stations) = `51,287,544 elements` and `51121047` of them (99%) are 0.

For this step, we get `mse = 0.0001493008`

Combine Step1 - 3

Now we know `P(s)`, `P(t|s)`, `P(e|t,s)`

$$P(e, t, s) = P(e|t, s) * P(t|s) * P(s)$$

1. experimental proportion(use total count in sample):
   $N_{i,t,j}$: number of bikes from i to j at t

   $$\hat{P}(e, t, s) = \frac{N_{ste}}{N_{st.}} * \frac{N_{st.}}{N_{s..}} * \frac{N_{s..}}{N_{...}} = \frac{N_{ste}}{N_{...}}$$

   Leave-1-out CV `MSE =4.517039e-11`

2. experimental proportion(use different total count):

   $$\hat{P}(e, t, s) = \hat{P}(e|t, s) * \hat{P}(t|s) * \hat{P}(s)$$

   $$\hat{P}(e|t, s) = E(\frac{N_{ste}}{N_{st.}})$$

   $$\hat{P}(t|s) = E(\frac{N_{st.}}{N_{s..}})$$

$$\hat{P}(s) = E(\frac{N_{s..}}{N_{...}})$$

Leave-1-out CV `MSE =8.826403e-12`
The average ride on each day is 8735.5.
The `sse of count` is 16394.68.

3. experimental proportion

$$\hat{P}(e, t, s) = \hat{P}(e|s, t) * \hat{P}(s|t) * \hat{P}(t)$$

Leave-1-out CV `MSE = 4.923364e-11`

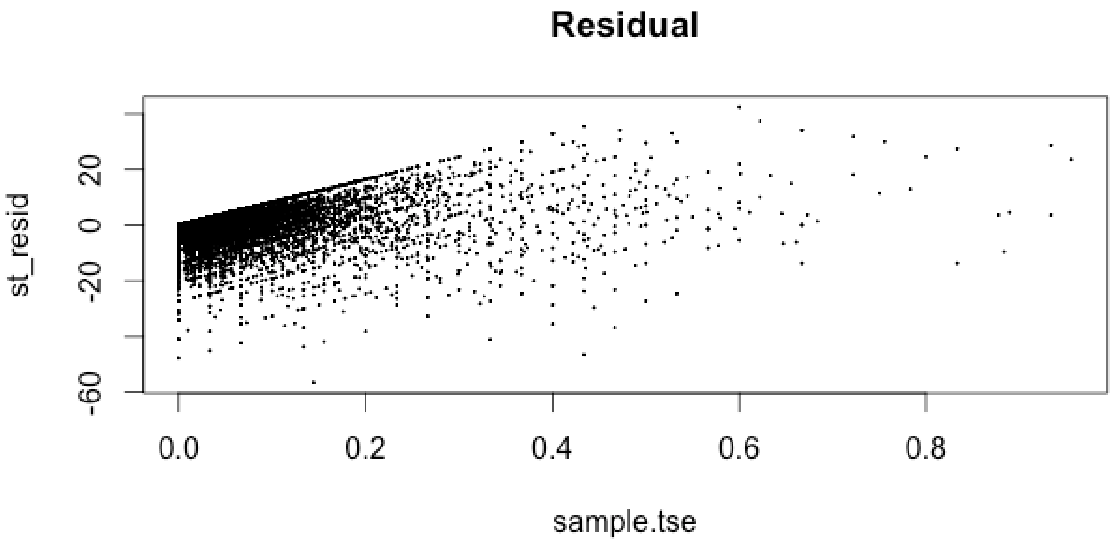
$$N_{ste} = \hat{P}(e|s, t) * \hat{P}(s|t) * \hat{P}(t) * N$$

$$N_{ste} = \hat{P}(e|s, t) * \hat{P}(s|t) * N_t$$

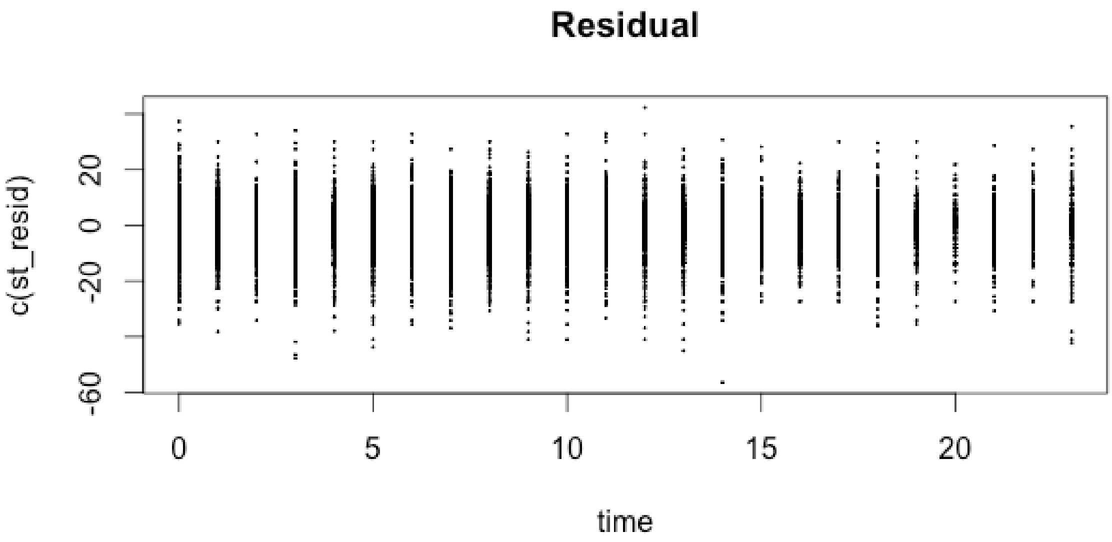
We did regression on N_t in Kaggle.

Analysis of Residuals

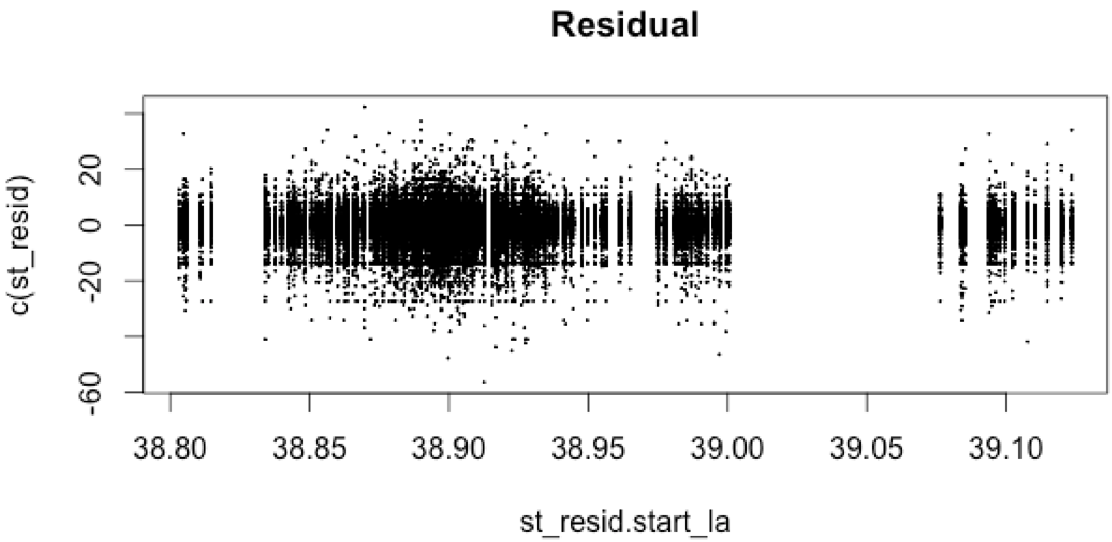
1. residual vs true value

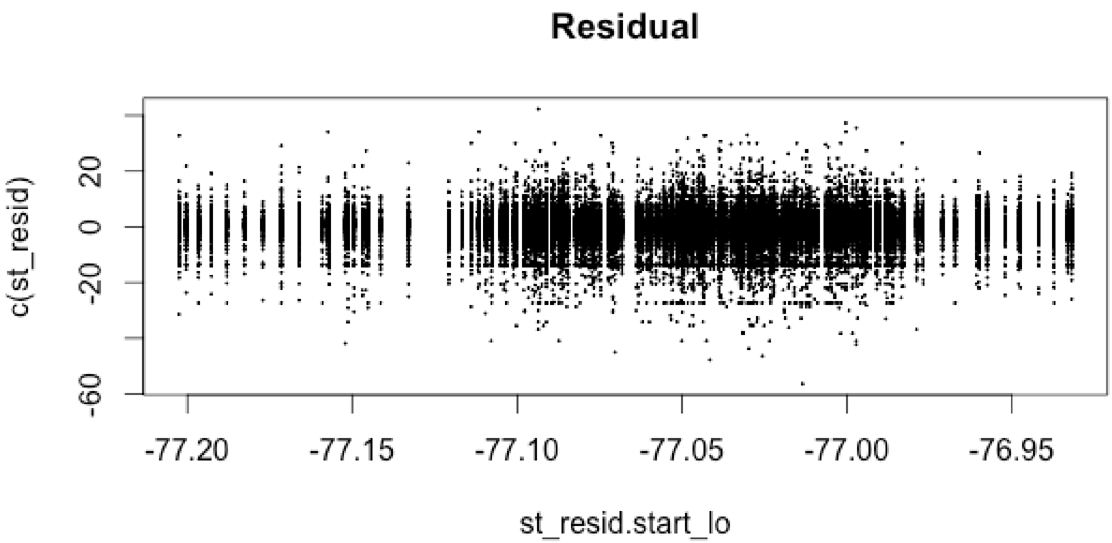


2. resid vs time



3. resid vs starting station





Feature: Hot spot

I pick most popular 3 subway stations in Washington D.C.

```
c("Galary Palace",38.898740, -77.021459)
c("Union Station",38.896993, -77.006422)
c("Capitol Hill",38.8897, -77.0111)
```

We cannot use more stations as hot spots because we can determine the location given the distances between it and 3 other locations.

Rebuild the dataset in sparse form

~~ For each `count != 0`, we create a row in the new data base with the following schema:c ~~