HW 3

Ashesh Shrestha

4/7/2021

ECO 395: Exercises 3

What causes what?

1. Cities with high crime rates might have deployed higher number of police personnel to control the crimes. In such case, we see that crime rates and number of police are positively correlated leading us to falsely conclude that increasing the number of police personnel increases the crime instead of reducing it. Therefore, we cannot just get data from a few different cities and run the regression of "Crime" on "Police".
2. The researchers sought to investigate what happens to crime rate when number of cops is increased because of the reasons unrelated to the crime rate. During high alert days, i.e. when the risk of terrorist attack is high, police increase their presence in Washington DC area. The researchers looked whether there was a significant decline in crime rate during high alert days or not . In this way, they were able to solve for endogeneity problem. As seen in table 2, during high alert days the expected decline in the number of crimes in Washington DC is 7 crimes per day, holding all else fixed. The result is statistically significant at 5 percent level. Controlling for Log of midday metro ridership, the expected decline in the number of crimes is 6 crimes per day, holding all else fixed, and the result is still significant at 5 percent level.
3. It is possible that the total number of tourists is lower during high alert days and therefore the potential victims of the crime are also reduced, which ultimately reduces the total number of crimes. Hence, to see if this was the reason behind the decline in the number of crimes during high alert days, the researchers controlled for the midday metro ridership which proxied for the number of tourists.
4. During the periods of high alert, average number of crimes decline by 2.621 crimes per day when there are crime incidents in the first police district area, holding all else fixed . The result is statistically significant at 1 percent level. The average number of crimes decline by .571 crimes per day when there are crime incidents in other districts during the high alert days, holding all else fixed . However, the result is not statistically significant.

Predictive model building: green certification

In this exercise, I have analyzed data set on green buildings and tried to build the best predictive model for prediction of revenue per square foot per calendar year. First of all, I started with data cleaning. I detected and deleted all the null values. Then, I created our variable of interest revenue per square foot per calendar year by multiplying rent and leasing rate.

I then started with a base model. As product of rent and leasing rate was the variable of interest, using interaction between these two as a feature variable did not make sense, hence I did not include them in the model. Furthermore, I removed 'CS_propertyID' which is building's unique identifier and contributed nothing to the model. I also deleted 'total_dd_07' due to the nature of its collinearity with the variables 'cd_total_07' and 'hd_total07'(total_dd_07 = cd_total_07 + hd_total07). The effect of cluster in rent is more or less reflected in 'City_Market_Rent', which is average rent per square-foot per calendar year in the building's local market. So, I have not included cluster in my model. In my model I have only considered 'green_rating' and not included LEED and EnergyStar separately. Hence, LEED and EnergyStar have also been removed.

To find the best predictive model, I have built 5 different models and compared their performance to come up with the best model.

Forward Selection Model

Forward selection model starts with a model having no variables and add all possible one-variable additions to it, including every interaction. The model with the lowest AIC which we get from forward selection process is:

revenue_persquarefoot ~ cluster_rent + size + class_a + class_b + amenities + cd_total_07 + green_rating + age + hd_total07 + Electricity_Costs + net + cluster_rent:size + amenities:green_rating + size:amenities + green_rating:age + size:Electricity_Costs + cluster_rent:hd_total07 + cd_total_07:hd_total07 + hd_total07:Electricity_Costs + size:class_a + size:class_b + size:age + class_a:age + class_a:cd_total_07 + size:cd_total_07 + cluster_rent:Electricity_Costs + cluster_rent:age + age:Electricity_Costs + cd_total_07:Electricity_Costs + class_a:Electricity_Costs + amenities:Electricity_Costs + cd_total_07:net + class_b:amenities + size:green_rating

The AIC for this model is 108159.9 and the number of variables is 34.

Backward Selection Model

Backward selection model starts with the full model that has all the variables including all of interactions, then improves its performance by deleting each variable. The model with the lowest AIC we get from backward selection process is:

revenue_persquarefoot ~ size + empl_gr + stories + age + renovated + class_a + class_b + green_rating + net + amenities + cd_total_07 + hd_total07 + Precipitation + Gas_Costs + Electricity_Costs + cluster_rent + size:empl_gr + size:stories + size:age + size:renovated + size:class_a + size:class_b + size:green_rating + size:cd_total_07 + size:hd_total07 + size:Electricity_Costs + size:cluster_rent + empl_gr:stories + empl_gr:renovated + empl_gr:class_a + empl_gr:class_b + empl_gr:Gas_Costs + stories:age + stories:renovated + stories:class_b + stories:cd_total_07 + stories:Precipitation + age:class_a + age:green_rating + age:cd_total_07 + age:hd_total07 + age:Electricity_Costs + age:cluster_rent + renovated:hd_total07 + renovated:Precipitation + renovated:cluster_rent + class_a:amenities + class_a:hd_total07 + class_a:Precipitation + class_a:Gas_Costs + class_a:Electricity_Costs + class_b:hd_total07 + class_b:Precipitation + class_b:Gas_Costs + class_b:Electricity_Costs + green_rating:amenities + amenities:Precipitation + amenities:Gas_Costs + amenities:Electricity_Costs + cd_total_07:Precipitation + cd_total_07:Gas_Costs + cd_total_07:Electricity_Costs + hd_total07:Precipitation + hd_total07:Gas_Costs + hd_total07:Electricity_Costs + Precipitation:cluster_rent + Gas_Costs:cluster_rent + Electricity_Costs:cluster_rent

The AIC for this model is 108044 and the number of variables is 68.

Stepwise selection Model Stepwise selection model starts with our base model lm(revenue_persquarefoot ~ . - Rent - leasing_rate - CS_PropertyID - cluster - LEED - Energystar - total_dd_07)' and we considered all possible one-variable additions or deletions including interactions. The model with the lowest AIC we get from stepwise selection model is:

revenue_persquarefoot ~ size + empl_gr + stories + age + renovated + class_a + class_b + green_rating + net + amenities + cd_total_07 + hd_total07 + Precipitation + Gas_Costs + Electricity_Costs +

cluster_rent + size:cluster_rent + stories:class_a + size:Precipitation + empl_gr:Electricity_Costs + green_rating:amenities + Precipitation:cluster_rent + hd_total07:Precipitation + amenities:Gas_Costs + amenities:Precipitation + stories:Gas_Costs + renovated:Precipitation + size:age + cd_total_07:Precipitation + stories:class_b + age:green_rating + class_a:Gas_Costs + class_a:Electricity_Costs + age:cluster_rent + age:Electricity_Costs + renovated:cluster_rent + Electricity_Costs:cluster_rent + cd_total_07:hd_total07 + age:class_a + renovated:hd_total07 + class_a:Precipitation + stories:renovated + size:renovated + size:Electricity_Costs + size:stories + size:hd_total07 + class_a:hd_total07 + empl_gr:renovated + age:hd_total07 + amenities:Electricity_Costs + class_a:amenities + renovated:Gas_Costs + size:green_rating

The AIC for this models is 108070.3 and the number of variables is 53.

The model derived though backward selection has lowest AIC if we compare AICs of all three models. Thus, on the basis of performance measured by AIC, we can conclude that the model we derived from backward selection is the best performing model among the three. Additionally, I have also performed k-fold cross validation to compare model performance of the above three models and also the base model.

The average of the root mean squared error calculated using 10 fold cross validation for our base model, forward selection model, backward selection model and step-wise selection model are 1037.316, 1006.796, 1006.068 and 1006.601, respectively. As the RMSE for backward selection model among the four models is lowest, this further hints that backward selection model is the best among the four.

RMSE from k-folds cross validation

Baseline model : Mean RMSE

```
## [1] 1034.455
```

Forward selection model: Mean RMSE

```
## [1] 1002.047
```

Backward selection model: Mean RMSE

```
## [1] 996.7358
```

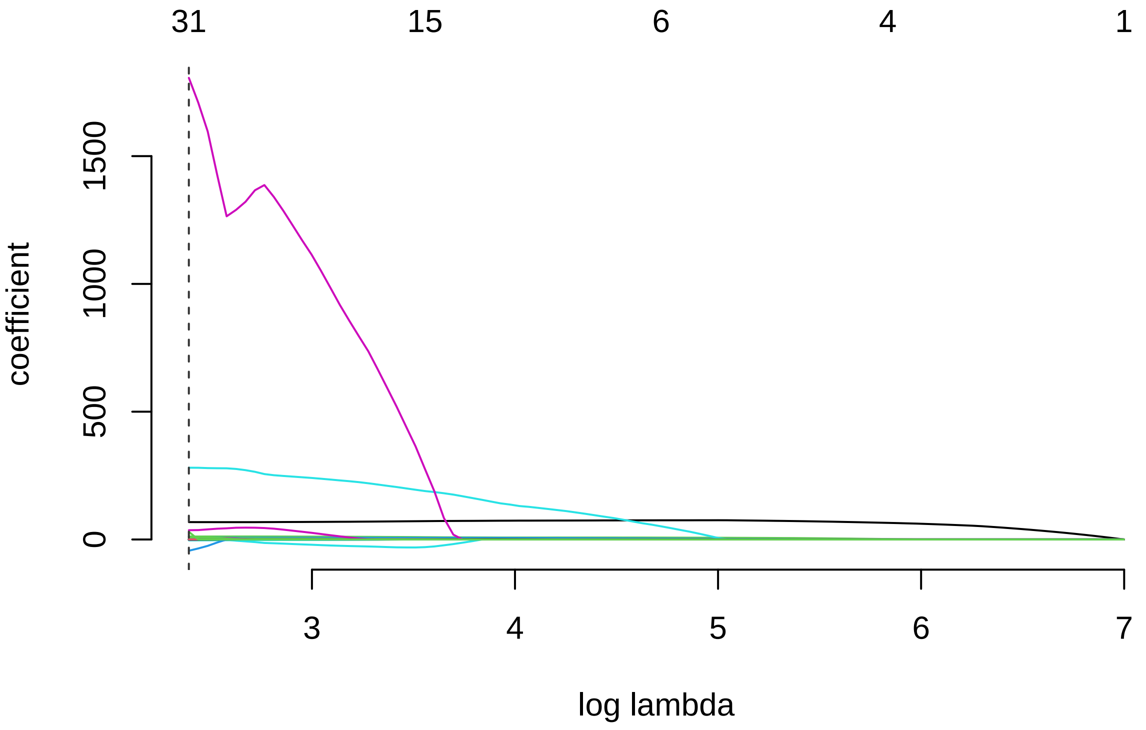
Stepwise selection model: Mean RMSE

```
## [1] 998.0623
```

Lasso Regression

I then fit lasso regression to see if lasso can beat the best model derived from backward selection. For lasso, I have used the full model including all the variables and all the two way interactions. The path plot which we get from running lasso regression is given below.

Figure 2.1 Path plot of lasso regression



```
##    seg100
## 2.393814
```

```
## [1] 31
```

The optimal value of lambda in a log scale is 2.39. The lowest AIC value is 108547.3 and the corresponding number of variables is 31 including the intercept.

Lasso : Mean RMSE

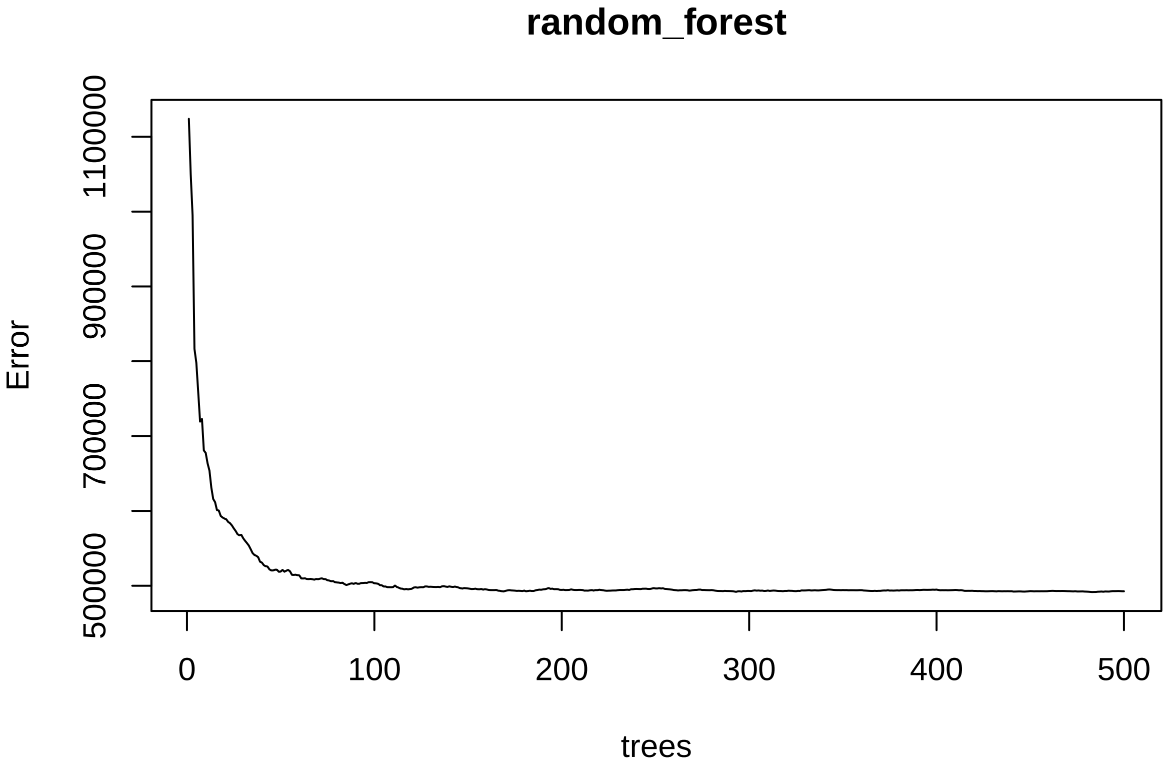
```
## [1] 1021.394
```

I conducted a k fold cross validation for lasso regression model. The RMSE in lasso regression is higher than any of the step-wise selection model, so we can conclude that step-wise selection model performs better than lasso regression in this case.

Random Forest

Lastly, I fit a random forest model using the base model. I have used 500 trees. As seen in the figure 2.3, using 500 trees is enough to reduce our errors.

Figure 2.3 Out of bag MSE as a function of number of trees



Random Forest : Mean RMSE

```
## [1] 705.7285
```

The RMSE from k-fold cross validation for Random Forest model is lower than RMSE of any model which we have used above. So, we can conclude that the model derived from Random Forest performs the best.

In order to find the average change in rental income per square foot per calendar year associated with green certification holding other features of the building constant, we used 'partial' function in 'pdp package'.

```
##   green_rating     yhat
## 1            0 2403.844
## 2            1 2474.367
```

The difference between the value of yhat when green_rating = 1 and the value of yhat when green_rating = 0, is the average change in rental income per square foot per calendar year associated with green certification, holding all other features of the buildings fixed.

Predictive model building: California housing

In this exercise, I have tried to build the best model for prediction of median market value of houses in the given census tract. I started with a baseline model which is a linear regression including all the variables without interactions. Then, I have built two additional models, namely Random Forest regression model and boosting model.

Baseline linear regression model

medianHouseValue ~ longitude + latitude + housingMedianAge + population + households + totalRooms + totalBedrooms + medianIncome

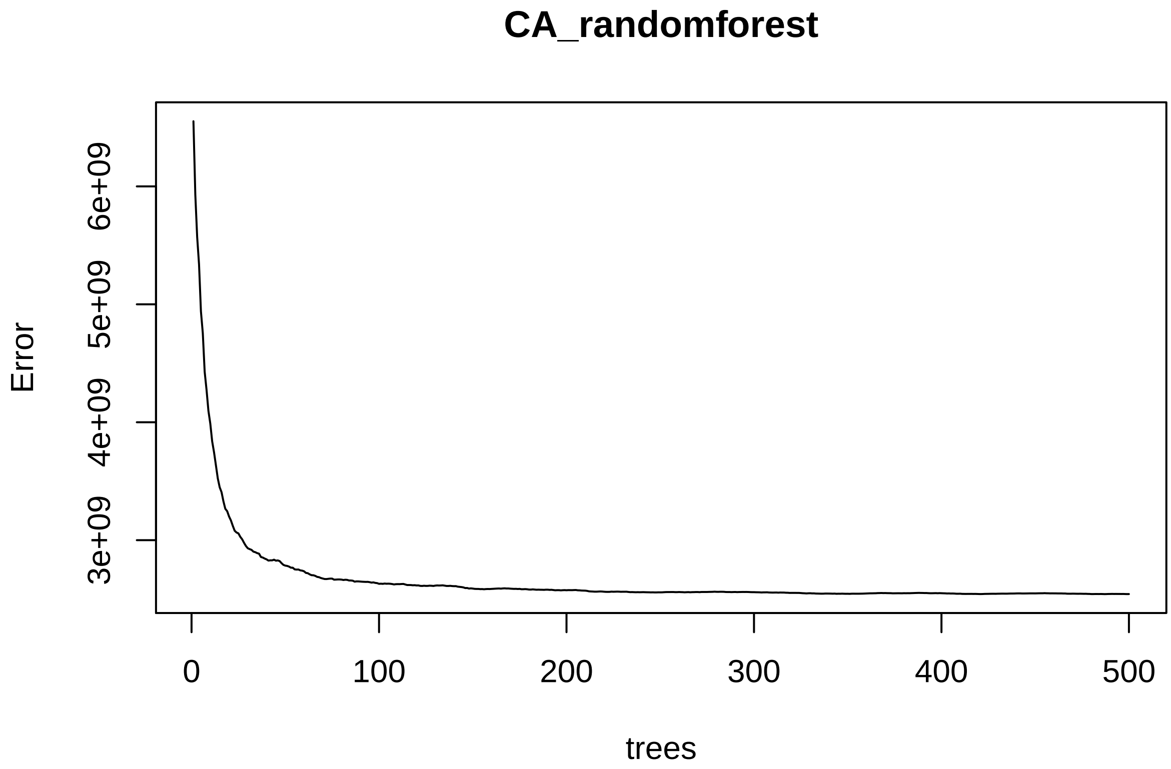
Baseline model : Mean RMSE

```
## [1] 69631.59
```

Random Forest model

I have fit a random forest model using the base model. As seen in the plot below which shows out-of-bag MSE as a function of the number of trees, using 500 trees is enough to reduce our errors. Hence, I have used 500 trees.

Figure 3.1 Out of bag MSE as a function of number of trees



Random Forest model : Mean RMSE

```
## [1] 50480.48
```

Boosting model

Lastly, I fit boosting model. As in the case of Random Forest model, we used the base model to begin with. The RMSE from the k-folds cross validation is which is a bit higher that that of random forest model.

Boosting model : Mean RMSE

```
## [1] 51649.15
```

As the k-folds cross validated RMSE from random forest model is the lowest, we use the random forest model for prediction.

Figure 3.2 Mean house value across various latitudes and longitudes in California

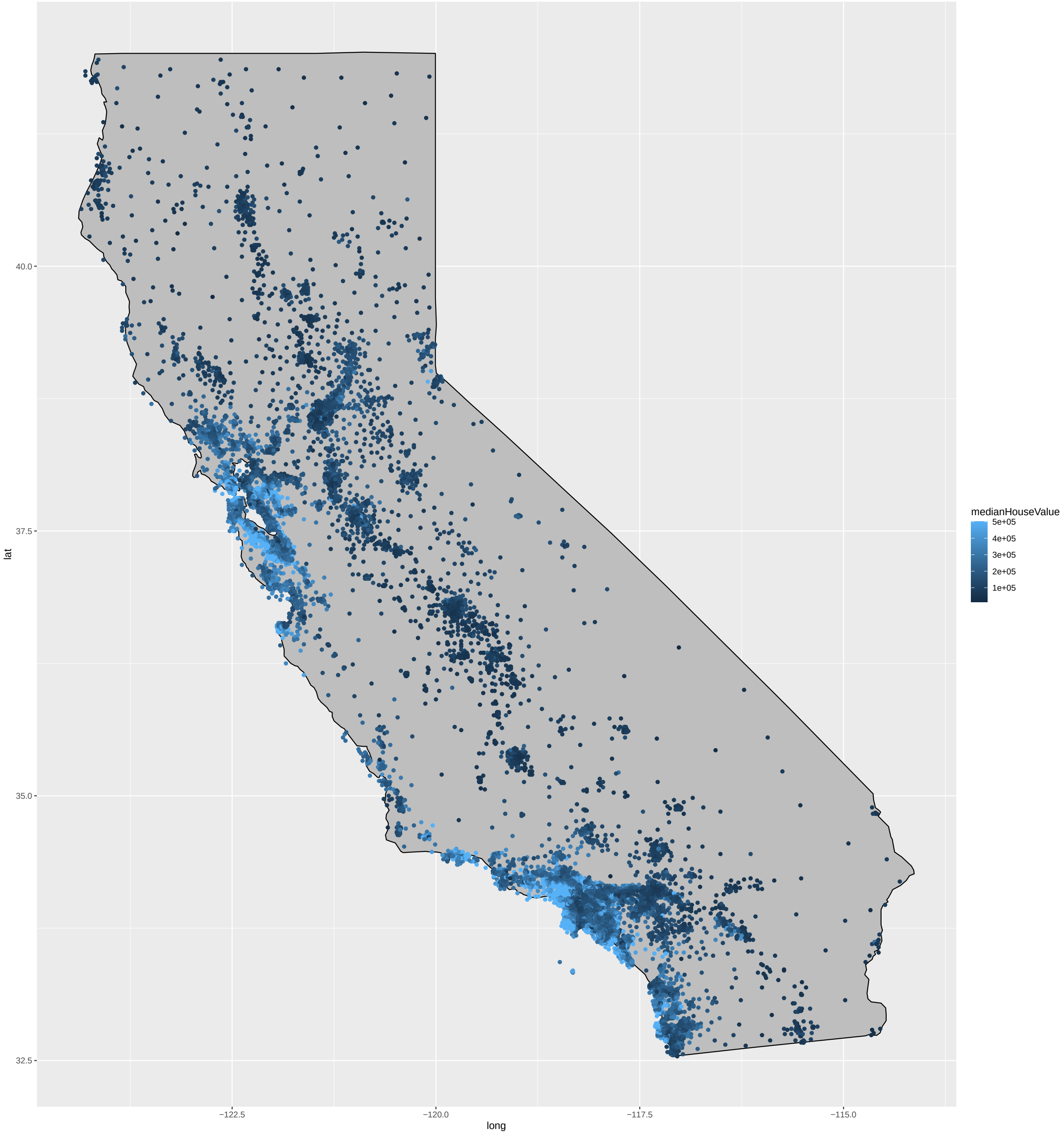


Figure 3.3 Prediction of Mean house value across various latitudes and longitudes in California using Random Forest

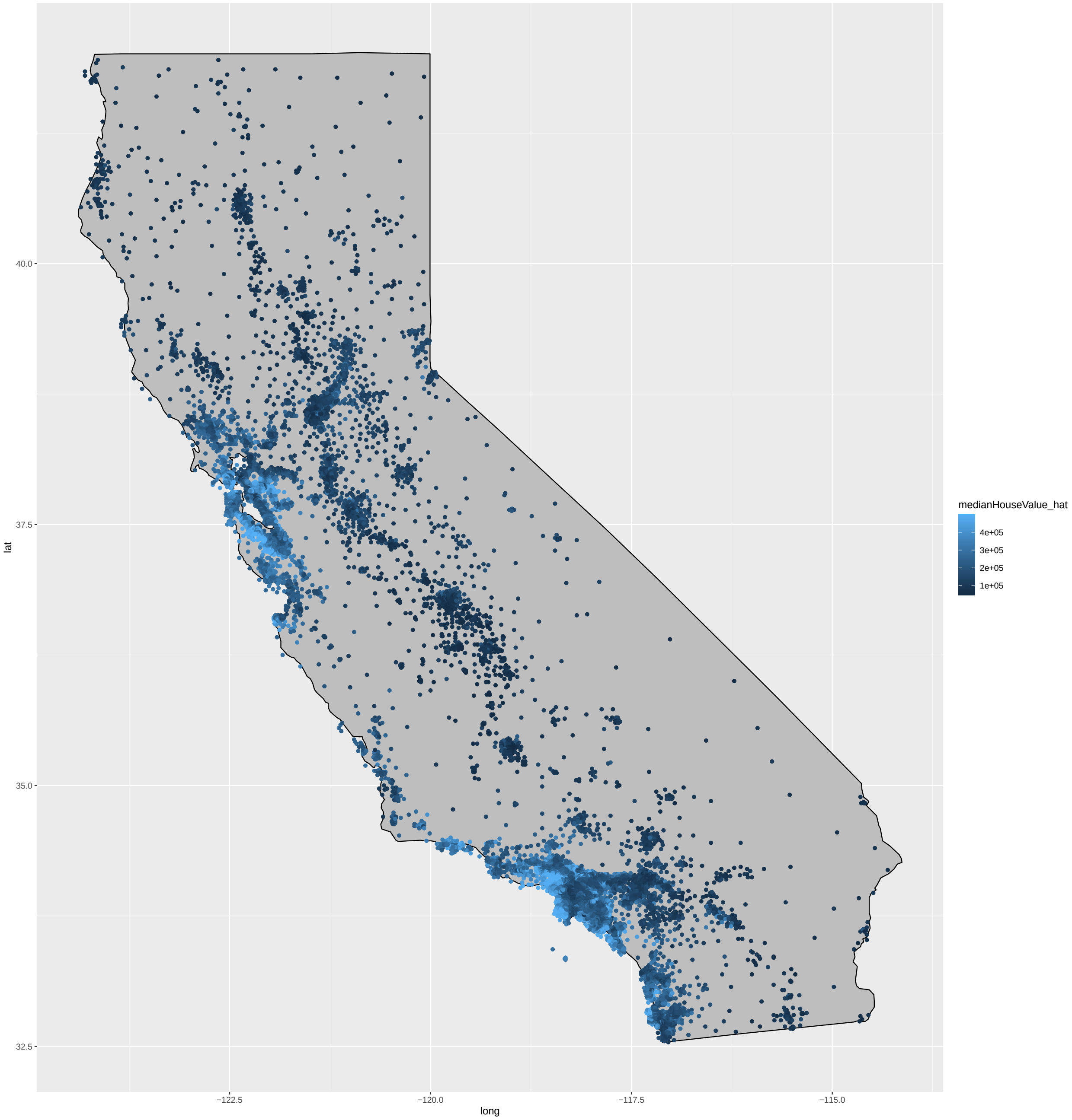
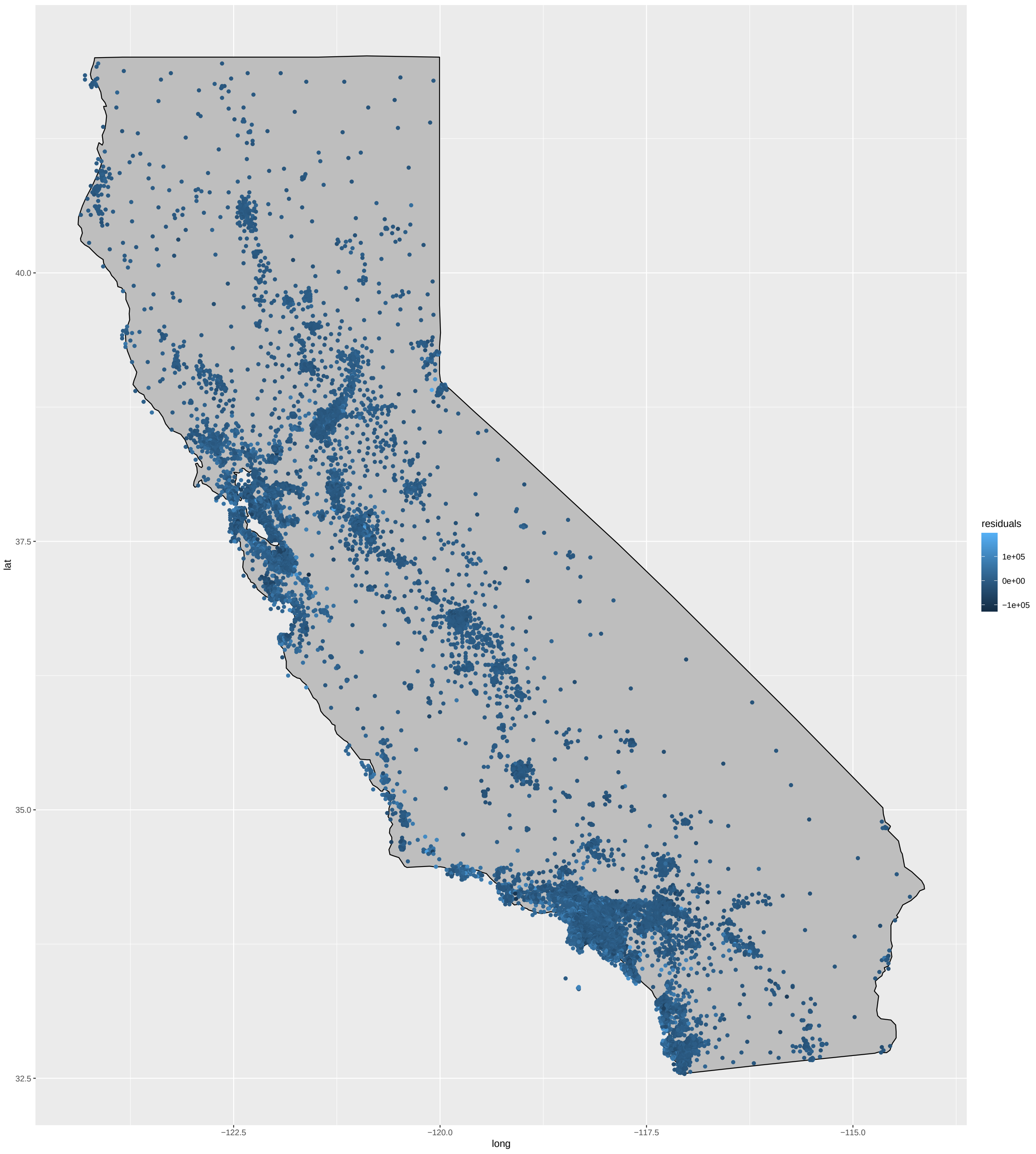


Figure 3.4 Residuals from prediction of mean house value using Random Forest


40.0
37.5
35.0
32.5
lat
−122.5
−120.0
−117.5
−115.0
long
residuals
1e+05
0e+00
−1e+05