Airbnb Data Analysis: Effects on Airbnb Ratings and Pricing at Boston

Camille Tianyue Wang

1.Introduction

Airbnb is known as an online platform enabling people to rent or lease lodgings for a short term. Established in 2008, Airbnb now has over 150 million of users worldwide. Boston, known for its inhospitable rents pricing and hotel prices, is growing into a popular place for Airbnb hosts.

For this project, I am interested in studying the effects of geographies, satisfactory ratings and violent crime rates on the pricing of Airbnb listings. Also, I am interested in the effects on ratings of Airbnb listings. The goal for this project is to understand what are the issues affecting clients satisfactory ratings and pricing for Airbnb listings in Boston.

2. Datasets

For this project, I mainly used year 2016 Airbnb datasets generated from Airbnb listings located in boston. The Airbnb datasets were collected 11 times in 2016, including time period from late January,2016 till late November, 2016(http://tomslee.net/airbnb-data-collection-get-the-data). After merging the 11 datasets together, I removed duplicated observations and took average of variables including price, overall satisfaction, minimum length of stay and number of reviews.

To better understand the effects of satisfaction ratings towards Airbnb listings, I took consideration of violent crime rates in Boston divided by Boston Police Districts. In order to compare the violent crime rates by different districts, I created a new variable named "District" by dividing boston neighborhoods into 11 Boston Police Districts, which is based on the City of Boston 2016 Crime data(accessible here: https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system). I defined violent crimes according to the definition of violent crimes by FBI Uniform Crime Reporting (UCR) Program. And the calculatng formula of violent crime rates for each district is showing as below:

Violent Crime Rate(by district)=(Number of Violent Crimes/ Population) * 1,000

The codes and details for data cleaning are included in another file named "Housekeeping".

3.Exploratory Descriptive Analysis

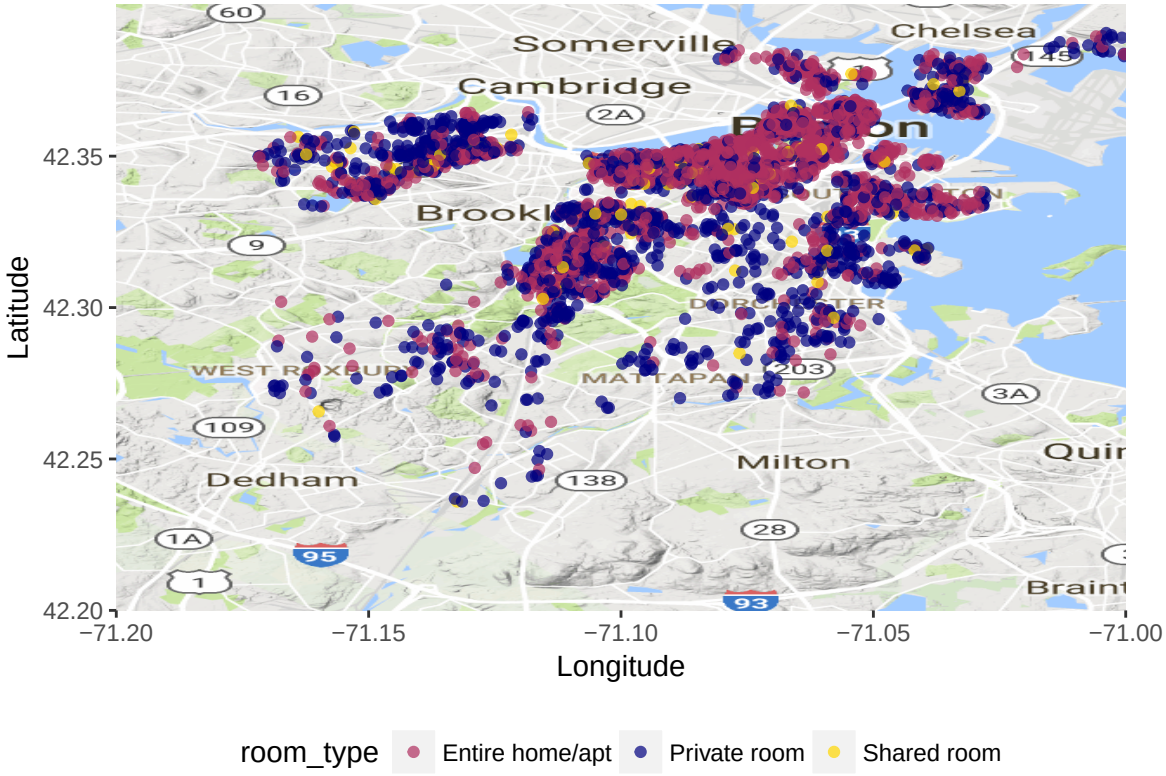
Boston Airbnb listings Map

After data cleaning, there are 4092 unique Airbnb listings in Boston which have been continously reviewed by guests since January 2016 till November 2016. The distribution of these listings is demonstrated in the map below: we can see that for year 2016, most listings are listed as "Entire home/apartment" and "Private room" while "shared room" has the least number of listings. The downtown district has the most number of unqie listings being active for the whole year of 2016.

```
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=Boston&zoom=11&size=640x640&scal
```

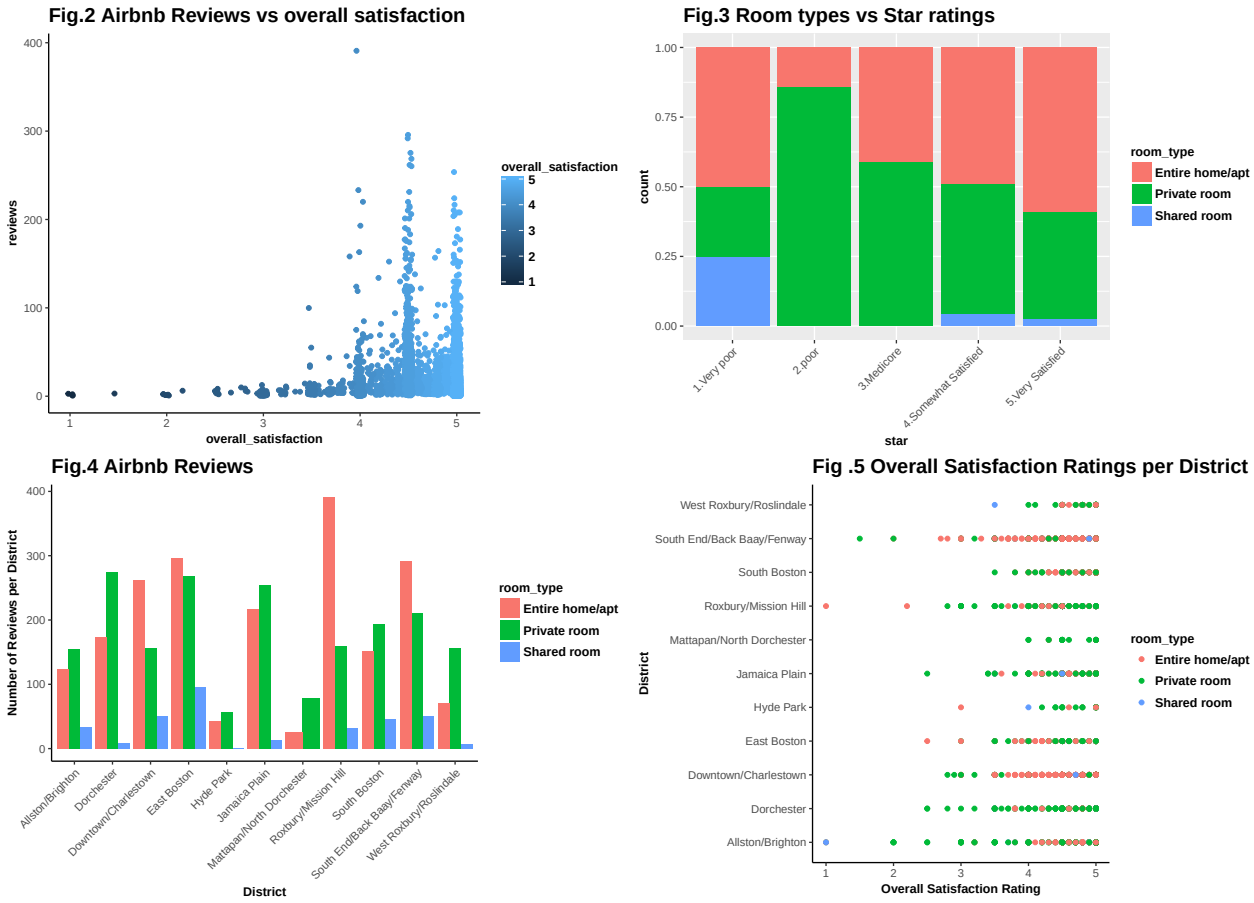
```
## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?address=Boston&sensor=false
```

Fig.1 Geographical distributions of Boston Airbnb listings



Ratings

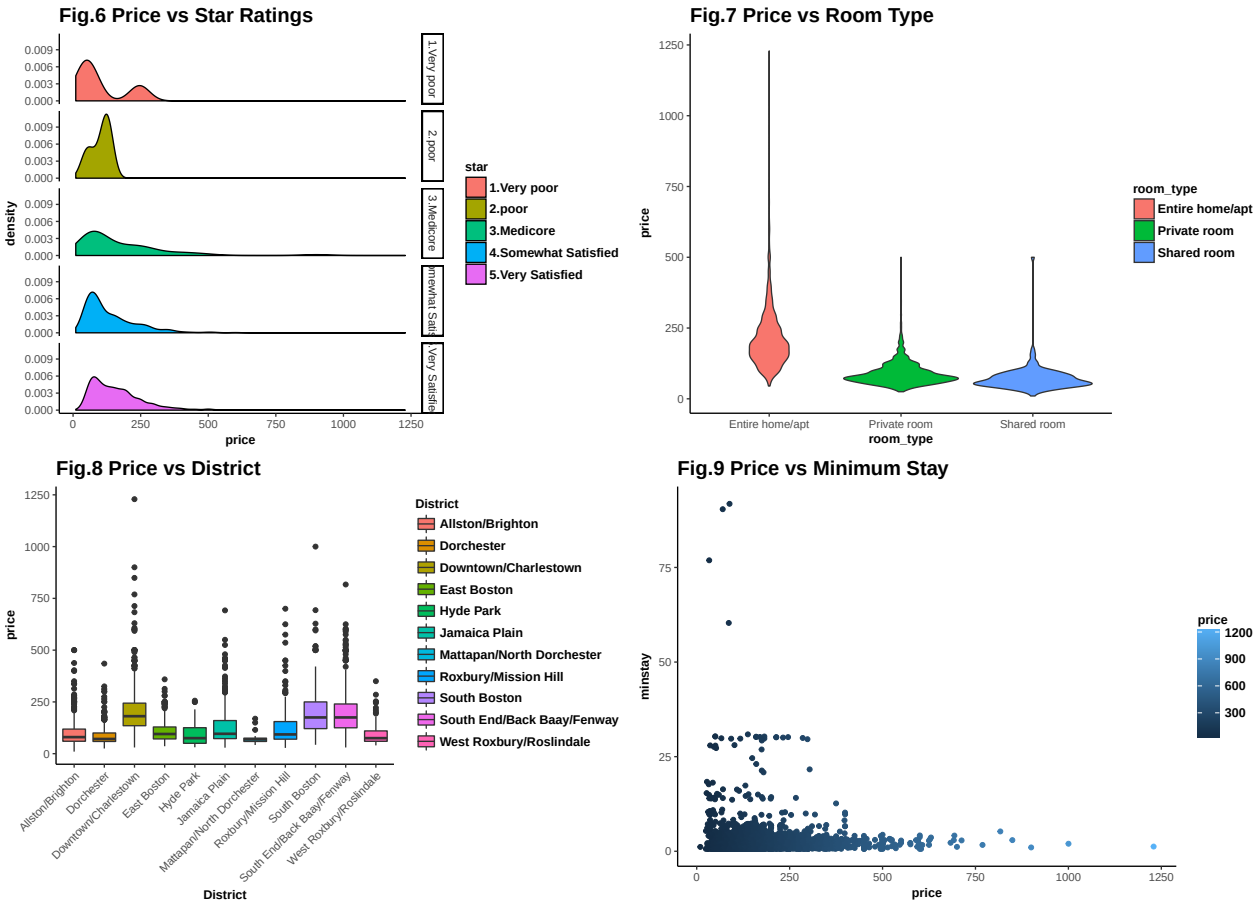
The following figures shows variation of ratings among selected features. Figure 2 indicates the distribution of number of reviews corresponding to overal satisfaction ratings, which is listings that have higher ratings tend to have more number of reviews. According to Figure 3, we can tell the distributions of star ratings in case of 3 different room types(Entire home, private room & shared room). Among the 5 stars rated listings, Entire home listings is the majority.



From Figure 4, we can see that for each district, entire home/apartment and private room listings are more popular among the three room types. The following Figure 5 shows the distribution of satisfaction ratings by district and room type. Guests are more likely to give higher ratings as showing in the plot.

Price

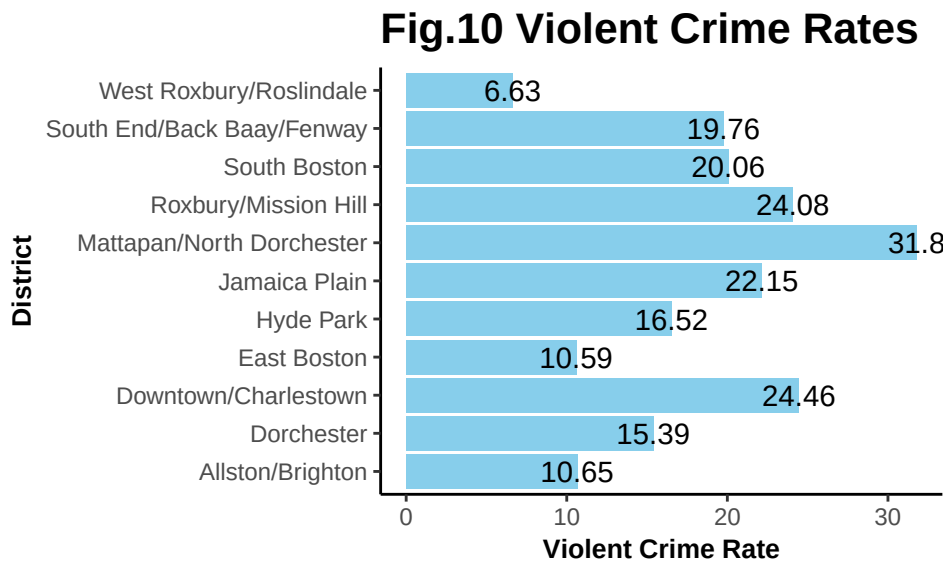
The below four figures demonstrate the relationship between price and other features including rating, room type, district and minimum length of stay. According to Figure 6, listings with ratings which are somewhat satisfied and very satisfied have very similar distributions as majority of them are listed under $200. Figure 7 is a violin plot showing distributions of prices in different room types, private and shared room listings have smaller price range while entire home listings have a large price range.



The variations of price in different districts are shown in Figure 8, listings located in Downtown, South Boston and South End districts have the highest listed prices while the Mattapan/North Dorchester district listings have the lowest prices among the 11 districts. Figure 9 is a scatter plot demonstrating the relationship between price and minimum length of stay, indicating lower priced listings require higher minimum length of stay.

Violent Crime Rate

For many Airbnb travelers, safety is a major concern when they are planning for accomodations before travelling. I organized boston crime data and calculated violent crime rates by each district. As showing in Figure 10, the safest district in boston in 2016 is West Roxbury/Roslindale with 6.63 violent crime rate, which means there are approximately 7 violent crimes in 1000 people in this area. In contrast, the most dangerous district is Mattapan/North Dorchester in year 2016.



4. Regression Analysis

Mixed effect regression

I utilized mixed effect regression to study how ratings, minimum length of stay, room types, districts and violent crime rate affect the pricing of Airbnb listings. For the response, I used log transformed price to reduce the effect of large variance. I also used 1/reviews as weight to rectify heteroscedastivity of residuals. I included Districts as the random effect and other predictors including room rypes, star ratings and violent crime rate as fixed effects.

Model

After comparing and testing multiple models, I have selected a mixed effect model as below:

Fixed effects:

Varible	Coefficient Estimate
Intercept	4.95
Minstay	-0.01
Rate	0.01
Private Room	-0.71
Shared Room	-1.08
2 Stars	0.08
3 Stars	0.26
4 Stars	-0.05
5 Stars	0.02

Random effect: Districts

According to the model, a 1 star rated Entire House/Apartment listing located in a safe area without requiring minimum length of stay has an average price at $141. Compared to Entire House/Apartment listings, Private room costs 51% less and shared rooms cost 66% less. Price would decrease by 1% if minimum length of stay required increases by 1 day. In reverse, price increases by 1% if violent crime rate increases by 1 unit. Also,

the star ratings do not change price significantly.

Diagnosis

The residul plot looks okay considering I am using two categorial variables. The qqplot also looks okay. For the random effect, I considered a likelihood ratio test.(See Appendix for detailed codes)

Ordinal Logistic Regression

Here I used an odinal logistic regression model to compare the star ratings for Airbnb listings in boston. The reponse variable I used is the star ratings. Also, I am using 1/reviews as weights to rectify heteroscedastivity. For predictors, I have included log transformed price, violent crime rate, minimum length of stay and room type.

Model

After testing multiple models, I selected the below model:

Varible	Coefficient Estimate
logprice	0.05
Minstay	-0.02
Rate	0.03
Private Room	-0.33
Shared Room	-0.41

According to the model, if log transformed price increases, the odds of getting a higher level of star rating is 1.05. The odds for the host to get one more star in ratings is 0.98 if minimum length of stay required increases by 1 day. If violent crime rate increases by 1 unit,the odds for the host to get a better level of rating is 1.03. If the room is listed as a private room, the odds of getting a higher level of rating from the guest is 0.72. If the room is listed as a shared room, the odds of getting a higher level of rating from the guest is 0.66.

The probability for hosts to have 1 star rating is 0.5%. The hosts are 1.2% likely to have 2 stars in rating and 5.2% likely to have 3 stars. The probability to have 4 stars is 20.5%. Thus, guests are mostly likely to give 5 stars.

Diagnosis

I used binned residuals plots to check the fitness of the model, which the result turned out that the model is a good fit(see Appendix).

5.Discussion

According to the previous analysis in pricing of Airbnb listings, we can tell that violent crime rate has positive effect on price and minimum length of stay has reverse effect on price. For limit-budget travelers, shared rooms and private rooms are better options compared to entire apartments. While ratings do not affect pricing too much, price actually is taken into consideration while guests are giving ratings for their hosts. Privacy is actually another big issue, listings that offer better privacy tends to have higher ratings.

I am also interested in whether the Airbnb rating system is efficient enough in indicating different levels of physical features and geographies of listings. So I performed a linear discriminant analysis to check if different features of Airbnb listings can charaterize the 5 star rating system.

Linear Discriminant Analysis

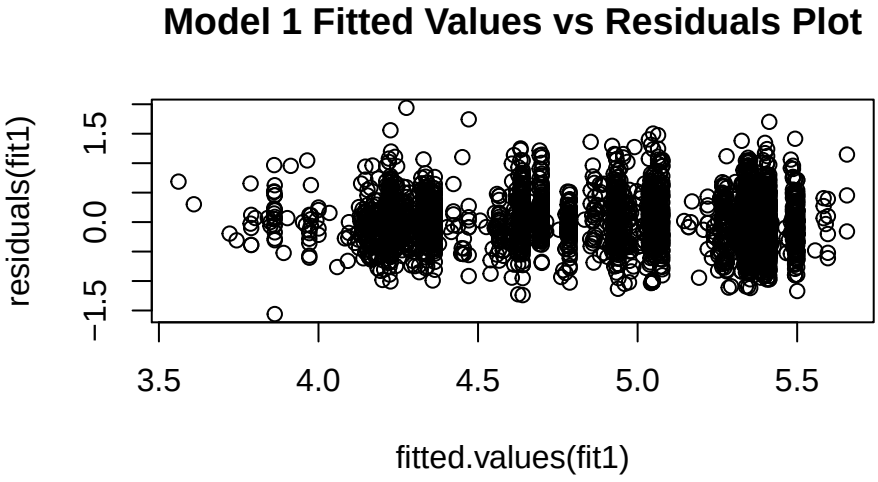
I used star rating as the response and roomtype, minimum length of stay, log transformed price and violent crime rate as predictors for the LDA model. After fitting a LDA model, I made predictions based on the model. And then I check in each linear discrimnant if there are obviously differed pattern for different levels of ratings by plotting a stacked histogram of predicited values for each discrimnant.

According to the histograms(see Appendix), there are no clear boundaries or patterns for 5 levels of ratings, which means there are no specific patterns in violent crimes, price, minimum length of stay and room types for different level of satisfactory ratings. However, since the predictors in this dataset are limited, there should be other issues taking into considerations when determining ratings for Airbnb listings.

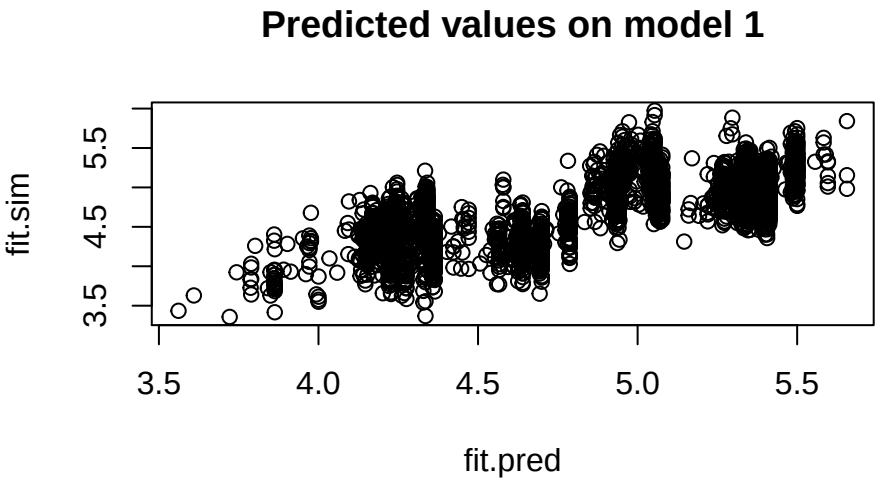
For the ratings part, the analysis could be compeleted in the future by taking more predictors into considerations. Also, sentimental feeling is a big part when guests give ratings to their hosts. In the fure, I will add on text analysis.

6. Appendix

Mixed Effect Regression



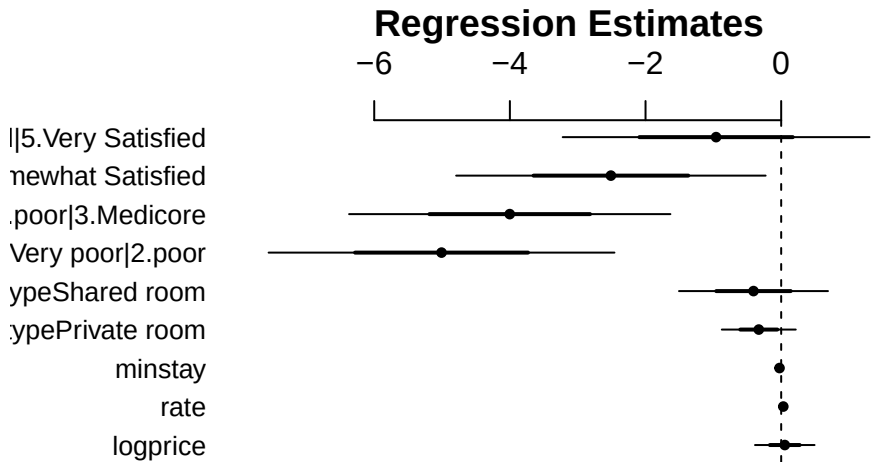

```
## Analysis of Random effects Table:
##          Chi.sq Chi.DF p.value
## District  -2543      1       1
```

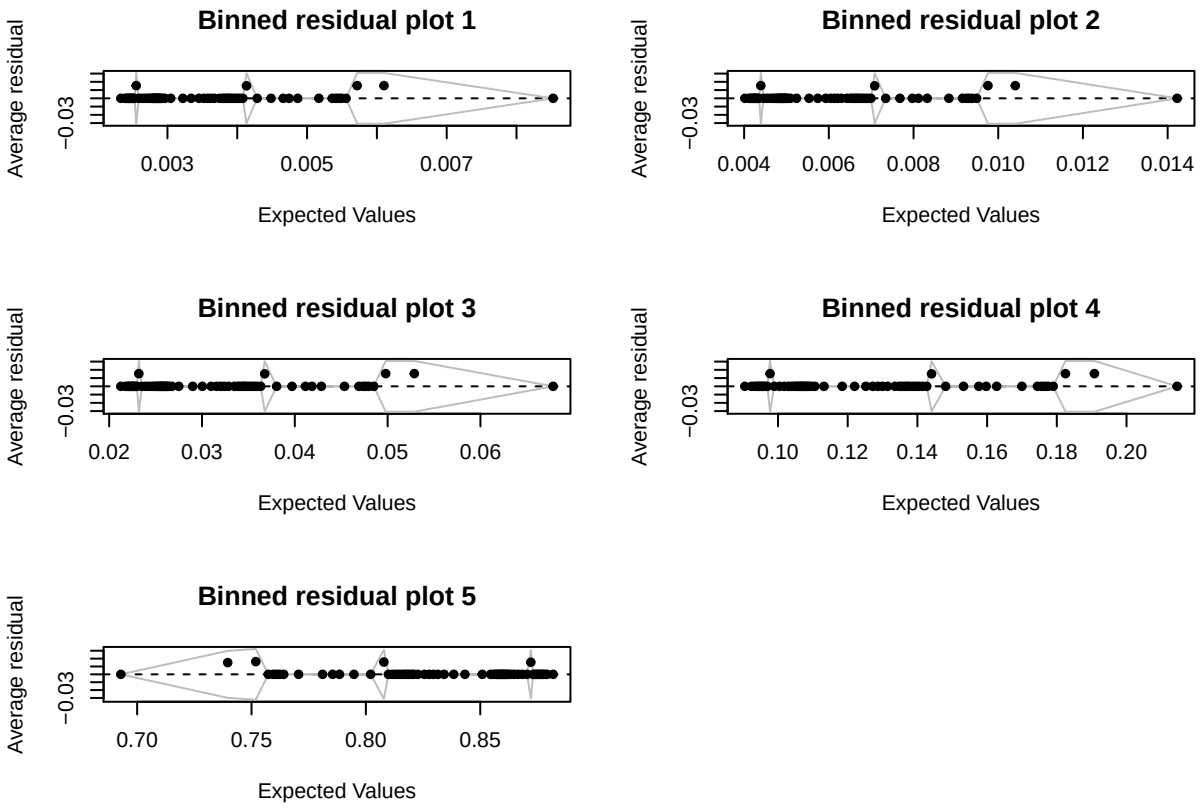


Multinomial Logit Regression

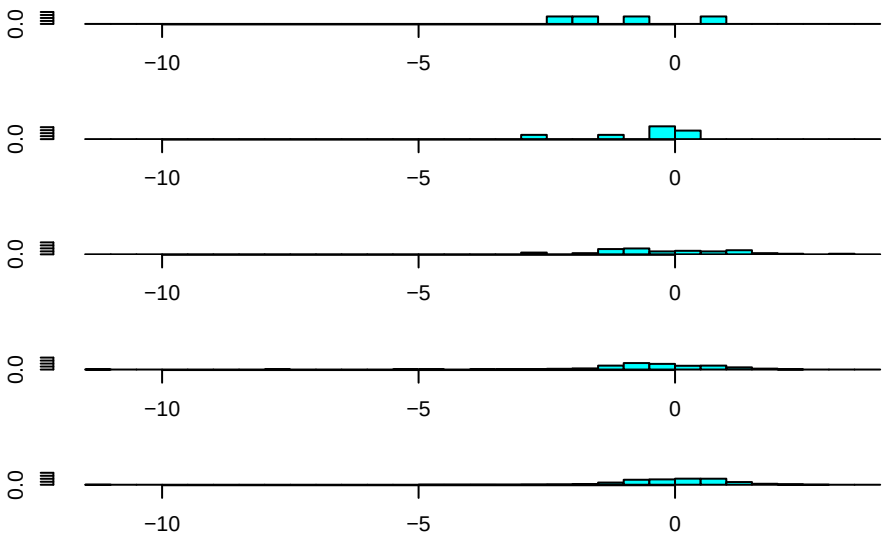
```
## Warning in eval(family$initialize): non-integer #successes in a binomial
## glm!
```

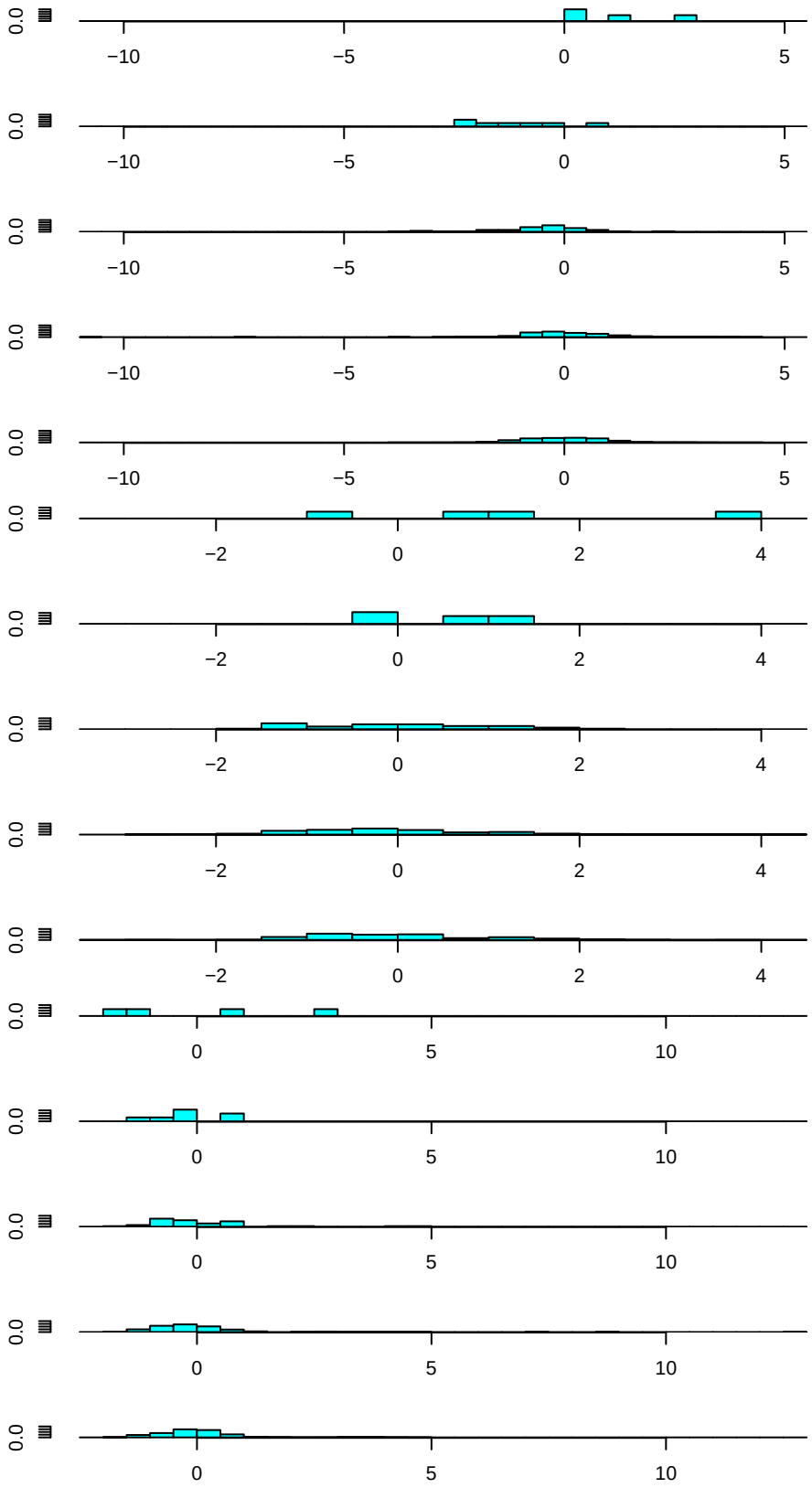
```
##
## Re-fitting to get Hessian
##
##
## Re-fitting to get Hessian
```





Linear Discriminant Analysis




0.0
−10
−5
0
5
0.0
−10
−5
0
5
0.0
−10
−5
0
5
0.0
−10
−5
0
5
0.0
−10
−5
0
5
0.0
−2
0
2
4
0.0
−2
0
2
4
0.0
−2
0
2
4
0.0
−2
0
2
4
0.0
−2
0
2
4
0.0
0
5
10
0.0
0
5
10
0.0
0
5
10
0.0
0
5
10
0.0
0
5
10