Investigating the differences between Criminal Mischief and Miscellaneous Offenses

Zane Wolf

Abstract

The goal at the onset of this project was to model specific crime types (of which there were 26 individual types) as a function of time and location. Eventually, the aim was simplified to modeling "Criminal Mischief & Related Offenses" and "Miscellaneous Offenses," still with the goal of modeling the response variable as a function of time and location. Working towards this goal, I simulated data, made a model with this data, and compared to a sampling of the original data.

1 Data

Data was sourced from the NYC Open Data website (link). The original data contained 5,580,035 observations and 24 variables. Of these 24 variables, all were categorical. One was a report number, five regarded the time and date of the crime and the report, five regarded the type of crime committed, six described the location of the crime, another five gave exact location coordinates for the crime, and another cataloged which crime authority handled the crime and the precinct number. Due to the repetitious nature of the data and the immense amount of information, I cleaned the data and parsed the predictor and response variables down to something manageable.

1.1 Cleaning Methods

Due to the time required to run the complete the cleaning process, the code was segregated into a separate file, 'cleaning_NYC_Crime_data.R', which you may run with the original data file, 'NYPD _Complaint_Data_Historic.csv' if you choose. To summarize my steps:

- Selected only for completed crimes ('CRM_ATPT_CPTD_CD'), reducing my data to 5.48 million.
- Removed the following predictor variables: 'CRM_ATPT_CPTD_CD', 'ADDR_PCT_CD', 'CMPLNT_NUM', 'CMPLNT_TO_DT', 'CMPLNT_TO_TM', 'RPT_DT', 'X_COORD_CD', 'Y_COORD_CD', 'Lat_Lon'.

- Converted the following predictor variables to binary: 'PARKS_NM', 'HADEVELOPT', 'JURIS_DESC' (1 = NYPD, 0 = other).
- Filled in missing data concerning the location description of crimes, 'LOC_OF_OCCUR_DESC', using logical rules based on existing data and then converted to a binary indicator (1 = inside, 0 = outside).
- Selected for years 2006-2016, separated time variable (HH:MM:SS) into three separate columns , and the same with the date variable (MM:DD:YY).
- Created a Month:Day variable and vectors of holiday dates to create the binary indicator 'Holiday' (1 = holiday, 0 = not).
- Used the built-in function weekdays() to create a variable listing which day of the week the crime was committed, and then used vectors of week-days and weekend-days to create the binary indicator 'Weekend' (1 = Weekend, 0 = not).
- Deleted all incidences where 'both' response variables, 'LAW_CAT_CD' and 'OFNS_DESC' were NAs.
- Filled in missing data in the response variable ('OFNS_DESC') by finding previous incidences of the corresponding 'PD_DESC' variable and substituting the first non-NA value into the current NA cell.
- Parsed the variable 'OFNS_DESC' from 72 unique classifiers down to 26 by deleting the most infrequent crimes and combining crimes of similar natures.
- Selected only for the crimes "Criminal Mischief" or "Misc. Offenses", as they were committed in roughly equal frequencies throughout the years
- Selected only for the year 2010, as this year had the least amount of disparity between the numbers of each crime committed.

At this point, the cleaned data frame was saved to 'NYPD_Crime_Data_CLEAN_2010.csv', which can be used directly in the modeling code.

In the remaining sections, a randomly selected 10,000 data points of the 2010 data set was used to reduce computation time.

2 Preliminary Analysis

In order to evaluate which of the remaining predictors (Hour, PARK_NM, JURIS_DESC, LOC_OF_OCCUR_DESC, HADEVELOPT, Holiday, Weekend, BORO_NM) would be most useful in my analysis, I made several graphs (Fig. 1a,b, Fig. 2a) to visualize the potential patterns and differences in the data.[1] Ultimately, only Hour, LOC_OF_OCCUR_DESC ('Location'), and BORO_NM ('Boro') showed any potential as predictors. Hour is a categorical variable with values ranging from 00 to 23 in steps of 1, Location is a binary variable with 0 representing 'Outside' and 1 representing 'Inside', and Boro is also a categorical variable with the values Queens, Staten Island, Brooklyn, Bronx, and Manhattan.

[1] At this point, all coding is done the 'modeling_Nyc_Crime_data.R' script.

There are a few particular trends I noted in these preliminary figures. The first is that the proportions of Mischievous crimes vs Miscellaneous crimes oscillates between boros (Fig. 1a). The second is that Misc. Offenses peak during work hours (9am-3pm) but Criminal Mischief offenses peak in the evening (Fig. 1b). This probably speaks to the natures of the crimes themselves. The final pattern of note is that Location, whether the crime occurred inside or outside, seemed to be a strong predictor of the type of crime committed at all hours (Fig. 1a).

3 Statistical Analysis

3.1 Model

Due to the nature of the variables (all categorical), a hierarchical multinomial logistic model was run using stan_glmer:

$$Crime \sim Hour + Location + Hour : Location + (1|Boro) \tag{1}$$

Models that lacked ther interaction term but were partially pooled and models that contained the interaction term but were completely pooled were also run, but were determined to be less accurate using loo comparisons[2]

3.2 Fake Data

As detailed in the modeling script, fake data was first created by making a matrix of zeros with 5000 rows (data points). Dummy variables for $J-1$ levels of the categorical predictors Hour and Location were made, and J dummy variables for the categorical predictor Boro. In this way, a coefficient could be applied to every value the categorical values could have.

The response variable, "Crimes", was then created using the following equation:

$$Crime \sim \mu_{[j]} * Hour_{[j]} + \mu_{in} * Location + \mu_{Loc-Hour_{[j]}} * Hour_{[j]} * Location + \mu_{[k]} * Boro_{[k]}$$

$$j = seq(0, 23, 1)$$

$$k = seq(0, 4, 1)$$

[2] Differences between these models ranged from 0.4 in 'ELPD Difference' to 2.0.

The data was then run through the Equation 1 using the categorical predictors created along side the dummy indicators in the fake data.

Comparing the figures I made during the preliminary analysis to the same figures made using the fake data (Figs. 1, 2), it appears that for the most part, my fake data emulates certain key aspects of the real data quite well. However, in Fig. 2, it can be seen quite easily that because the location predictor is so strong, it skews the data in a way that is not quite as realistic as the other graphs.

Comparison

Comparing the coefficient plots generated by the real-data model and the test-data model (Figs. 3, 4), it is apparent that something is amiss. While the graphs do capture the same patterns present in the data (e.g. 'Work Hours'), the magnitude of the coefficient values is quite different. The test-data model shows coefficients magnitudes larger. After attempting to simplify the test-data model to pinpoint the problem [3], I am left with the conclusion that the error lies in the equation used to create the crime predictor in my fake data.

On a final note, the posterior predictive checks of the models (Figs. 5) reveal that the models do a good job estimating the mean, min, and max of the data. However, it is unknown why the models do poorly when estimating the standard deviance, though the model made on simulated data does appear to do a better job.

4 Conclusion

There are problems either with my test data or my test-data model. While the model catches the patterns I encoded from the real data, it was incapable of returning the exact value of the coefficients. Additionally, one would need to investigate the reason the SD of the models was off. If these could be fixed, next steps would include trying to run the model on data sets from other years, to see if the patterns continue to exist. At some point, perhaps a third crime could be added to the model. Alternatively, it would be interesting to find other crimes that fluctuate solely by location.

[3] Simpler models did not fix the effect and failed at duplicating the same pattern in the data.

Figures

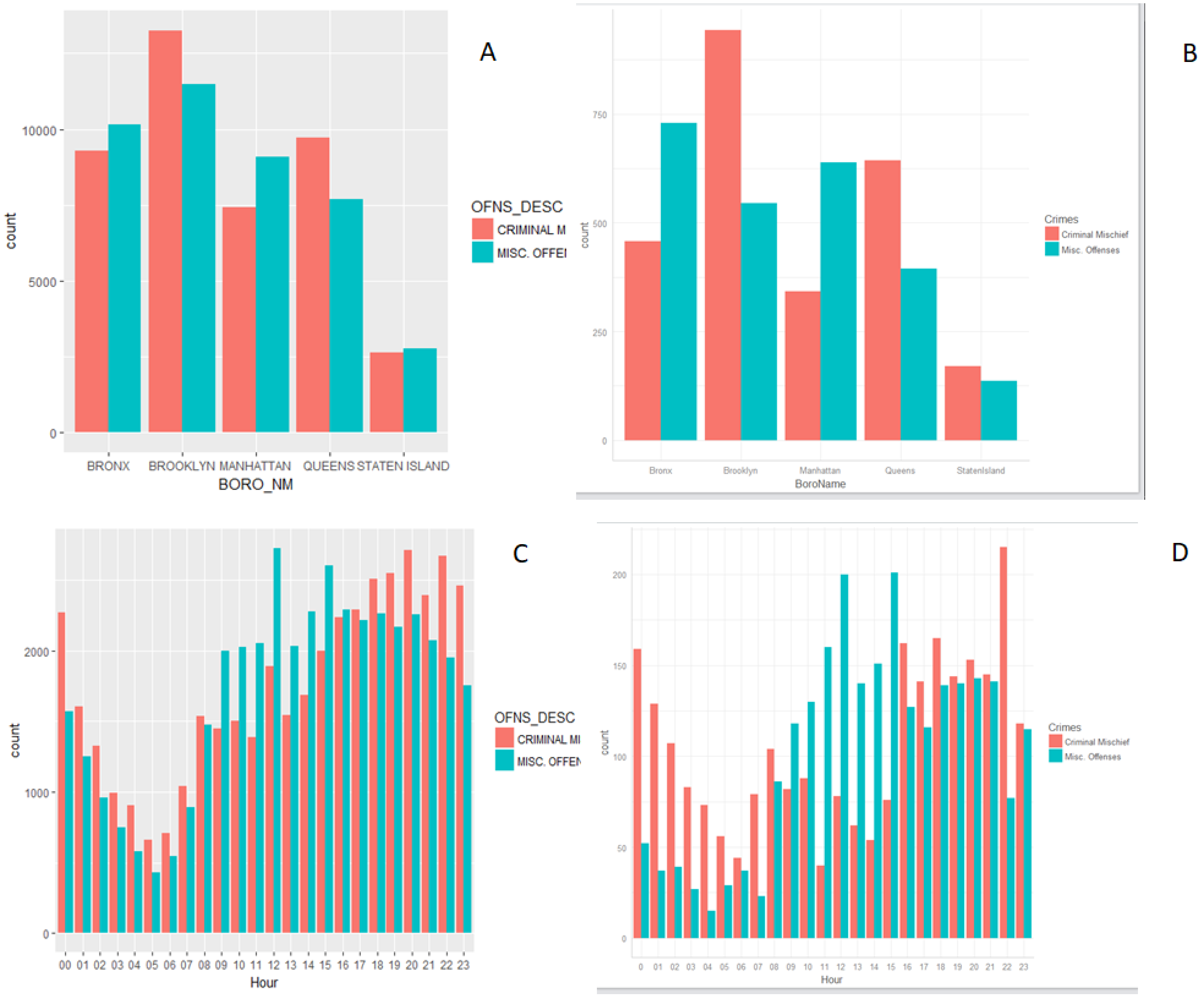

Figure 1: This figure compares the graphs of real data (A, C) to the fake data (B,D). When graphing by Boro or Hour, it appears that the fake data replicates the real data to a certain degree.

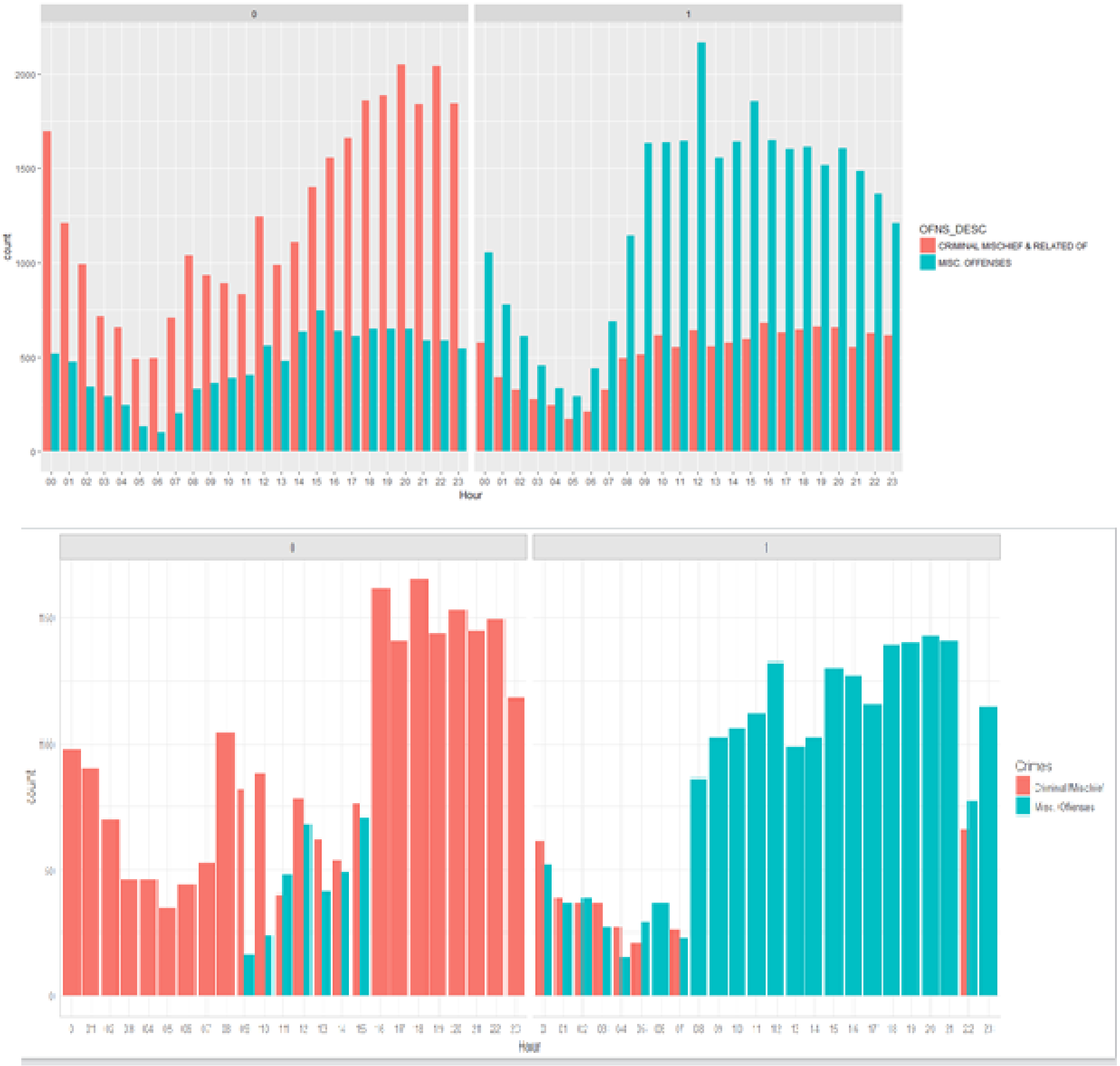

Figure 2: When comparing the real data, with crime graphed by hour and grouped by location, and the fake data graphed the same way, it is easy to discern the complications with the fake data. The location coefficient, though it has the same value as the real data's coefficient for location, appears to be incredibly strong, skewing it so that almost no Inside crimes can be Mischief and vice versa.

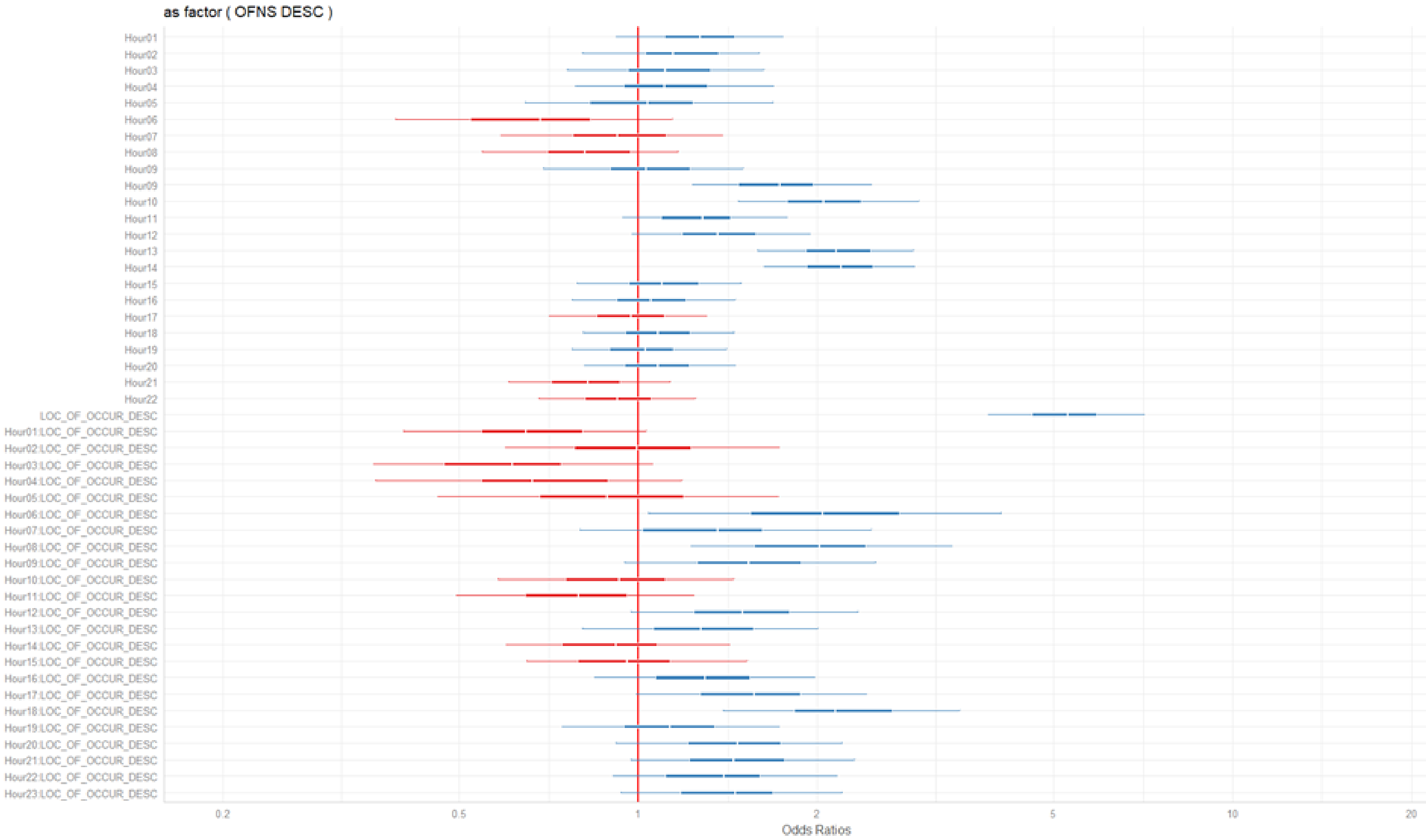

Figure 3: The coefficients of the categorical level predictors using the real data.

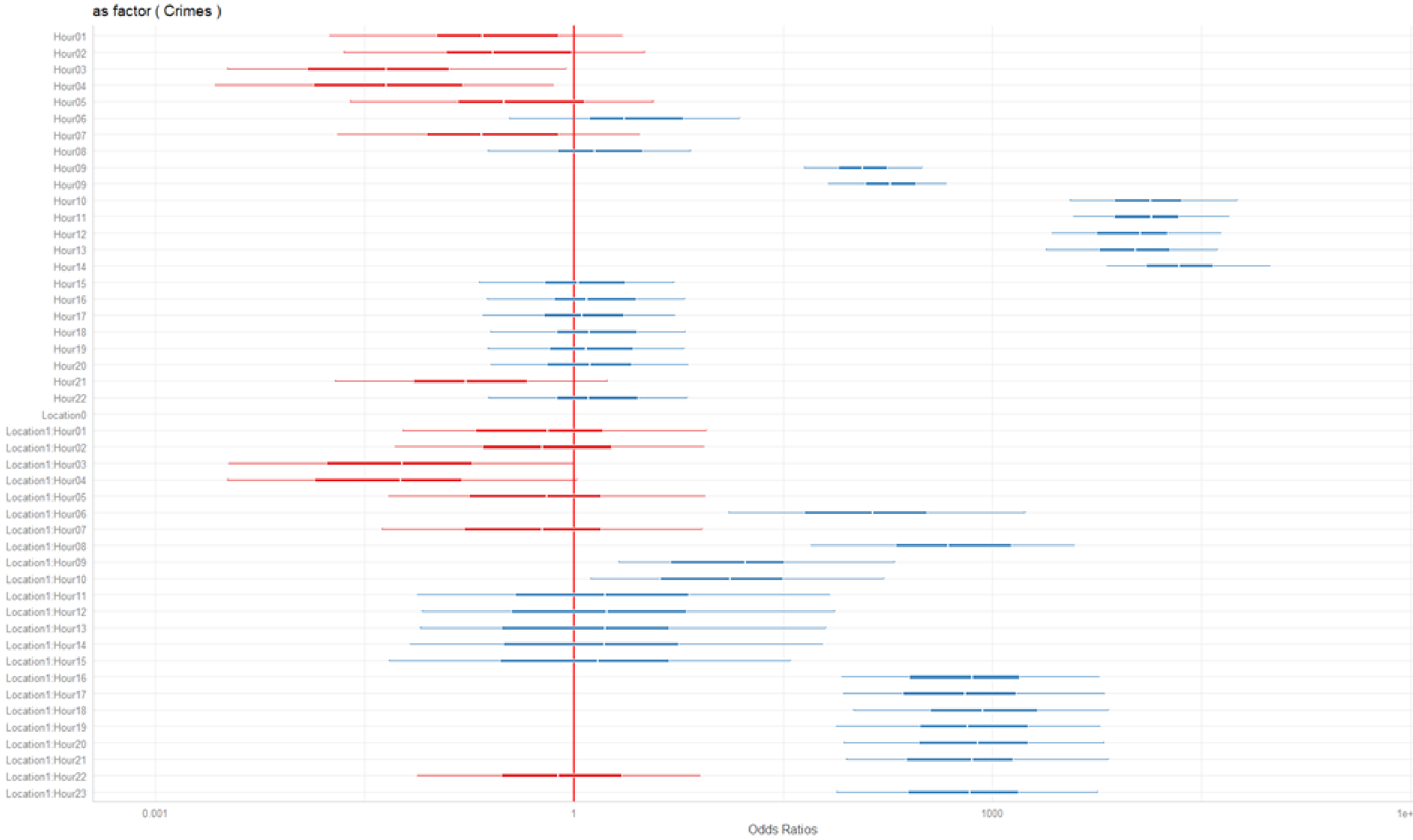

Figure 4: The coefficients of the categorical level predictors using the simulated data.

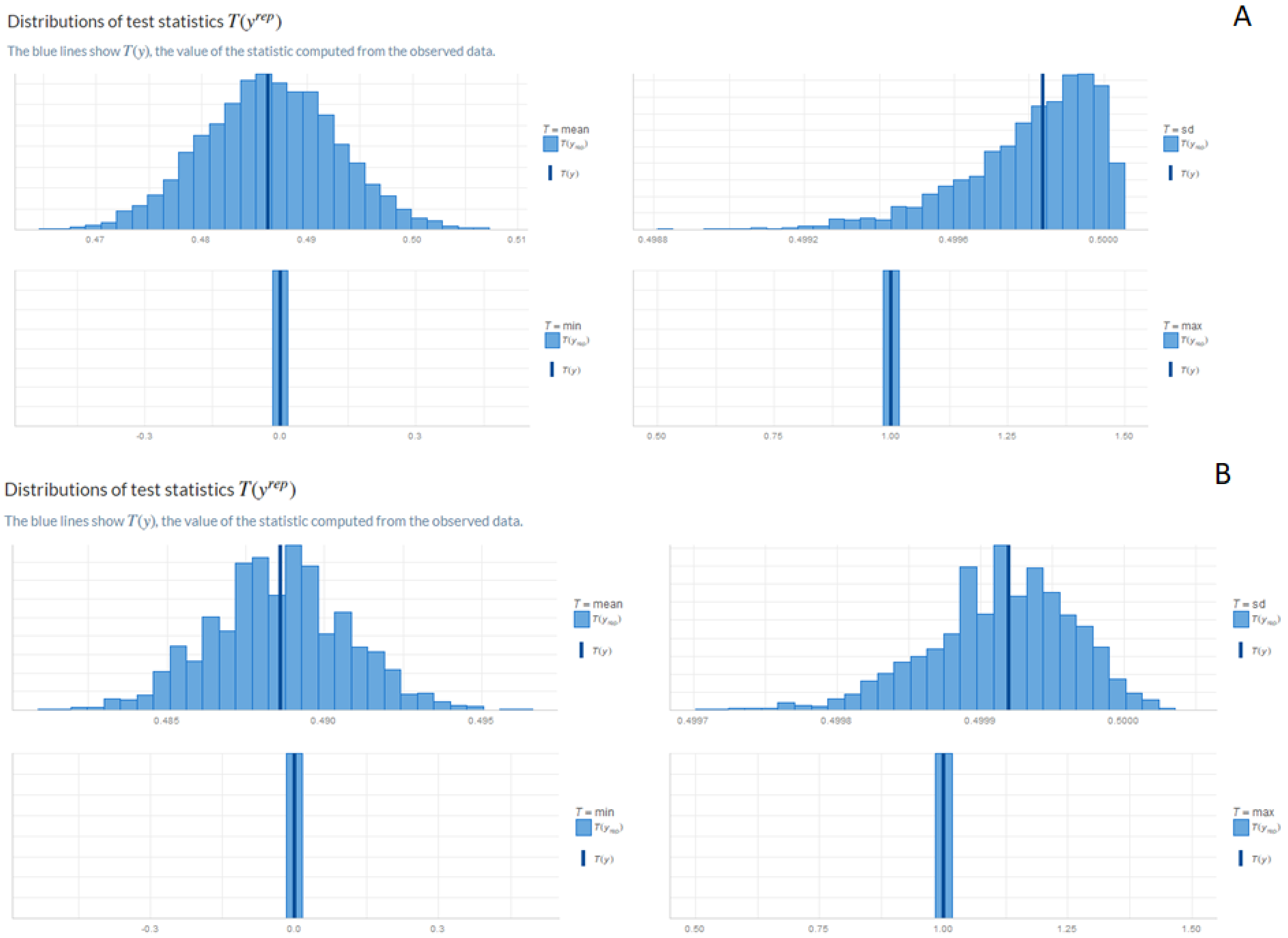

Figure 5: A) The model does seem to have some trouble picking up the true value of the standard deviation of my data, as the simulations tend to skew to the right. B) The test-data model does appear to do a better job with the standard deviation, though it also has an extended left tail.