APSTA-GE 2123 Project

Joseph Marlo

May 19, 2020

Contents

Introduction

Citi Bike is the number three[1] mode of public transportation in New York City behind the subway and buses. There were 1,169,973 Citi ike trips in February 2020 equaling to 40,343 daily trips[2]. Unlike it's larger counterparts, Citi Bike daily ridership is difficult to predict as it is a mixture of commuter and leisure riders, the former is tied to weekdays and the latter is tied to the weather.

Research question

Can Citi Bike ridership be accurately predicted using only basic information on the day: day of the week and the weather?

Citi Bike data

Citi Bike publishes real-time and monthly datasets[3] detailing each bike trip taken since 2013. Information includes the date, start- and end-times, departure and arrival stations, subscriber status, rider sex and rider birth year. Ridership has been steadily growing with an average daily ridership of 26,238 in 2013 compared to 56,306 in 2019. To minimize the impact of this omitted growth variable while maximizing the size of the training set, only data from 2017, 2018, and 2019 will be included. The data is then aggregated and bike trip counts are calculated for each day[4]. A random sample of 80% is used to train the model and the remaining 20% is used for model prediction evaluation.

Weather data

The Citi Bike dataset does not contain weather data. Weather information is obtained from the National Oceanic and Atmospheric Administration (NOAA) for the Central Park weather station. Information includes the daily precipitation, average temperature, and maximum gust speed. The data is collected during the day (i.e. ex post). The final model will be used for prediction so a practical application would require a weather forecast (i.e. ex ante). This discontinuity is acceptable as one-day weather forecasts are quite accurate.

Final dataset

The final dataset[5] consists of 938 observations and 5 variables.

Table 1: Dataset description

Variable	Type	Description
Trip_count	continuous	Count of daily Citi Bike trips
Weekday	boolean	1 = weekday, 0 = weekend
Precipitation	boolean	1 = rain, 0 = no rain
Temp	integer	Average daily temperature in Fahrenheit
Gust_speed	continuous	Maximum gust speed in miles per hour

Model selection

The outcome variable is a simple count of how many bike trips occurred in a single day. Count data is frequently modeled using a Poisson model or the more flexible negative binomial model. First, I will fit a

[1] 2019 NYC Mobility Report
[2] Citi Bike February 2020 Monthly Report
[3] Citi Bike system data
[4] Aggregation script on Github
[5] Final dataset on Github

negative binomial then evaluate it against a Poisson.

The model form in R syntax:

$$\texttt{Trip count} \sim \texttt{Weekday + Precipitation + Temperature + Gust speed}$$

Drawing from the prior predictive distribution

Priors first need to be set for each coefficient, the intercept, and shape parameter. Passing the model specification to `brms::get_prior()` returns the priors that need to be set along with their default values.

Temperature and gust speed are not likely to have large effects on the outcome for each one unit increase. Temperature is in Fahrenheit and gust speed is in miles per hour. A one unit increase in either of these will unlikely to have a measurable effect on the trips. Both are set to $N(0, 0.01)$.

Weekday and precipitation, I believe, are more likely to have a larger effect since these are binary variables. However, their effects will be inverses of each other: weekdays have more commuter riders and rainy days have less overall riders The priors are set to $N(0.5, 0.1)$ and $N(-0.5, 0.1)$ respectively.

Table 2: Priors

prior	class	coef
normal(0, 0.01)	b	Gust_speed
normal(-0.5, 0.1)	b	Precipitation
normal(0, 0.01)	b	Temp
normal(0.5, 0.1)	b	Weekday
normal(10, 1)	Intercept	
exponential(10)	shape	

We can draw from this model using `brms::brm()` with the optional arguments `family = 'negbinomial'` and `sample_prior = "only"`. Then we compute the expected value using `brms::pp_expect()`. These draws from the prior distribution of the conditional expectation are reasonable. Examining the deciles shows that the middle 90% are covering a reasonable range of data. The values are little low but close to the actuals: the mean draw is 40,414 whereas the mean daily ridership for 2017-2019 is closer to 50,000.

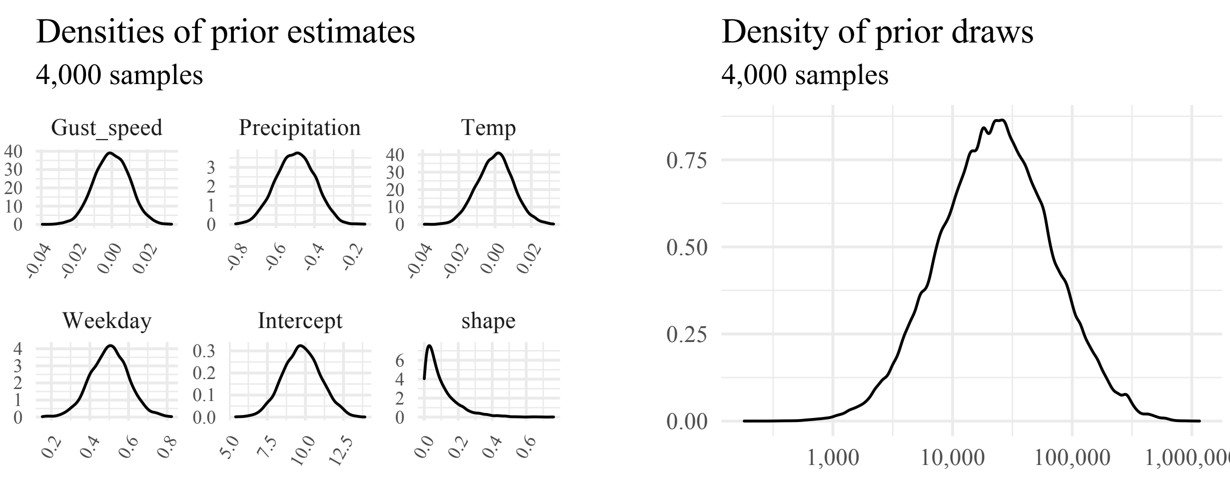

Figure 1: Density of prior estimates and draws

Table 3: Deciles of prior draws

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
181.1	5,586	9,082	12,924	17,393	22,851	29,905	40,256	56,646	92,107	1,159,991

Conditioning on the observed data

Since the priors are reasonable we can now condition on the data using `brms::update(..., sample_prior = "no")`.

After running this, we see the model converges and Rhat values are each 1.00. The effective-sample-sizes are all large as well, ranging from 2,800 to 5,200. We also see the estimates are close to the priors. The largest surprise is that precipitation has a larger effect on the outcome than weekday which may indicate that the hypothesized weekday commuters do not have as strong a positive effect as rain has a negative effect.

Table 4: Fixed effects

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	9.414	0.087	9.243	9.582	1.001	5,517	3,272
Weekday	0.244	0.027	0.189	0.296	1.001	4,190	2,960
Precipitation	-0.362	0.031	-0.423	-0.3	1	4,154	2,500
Temp	0.021	0.001	0.02	0.023	1.002	4,745	3,429
Gust_speed	-0.01	0.003	-0.017	-0.003	1.003	5,607	2,875

The posterior draws are well within range. The middle 90% [26,142, 82,711] fits the data well; the middle 90% of the actual data is [25,881, 75,358]. Additionally, the mean (52,616) and median (49,890) are close to the actual data (51,528 and 53,809 respectively).

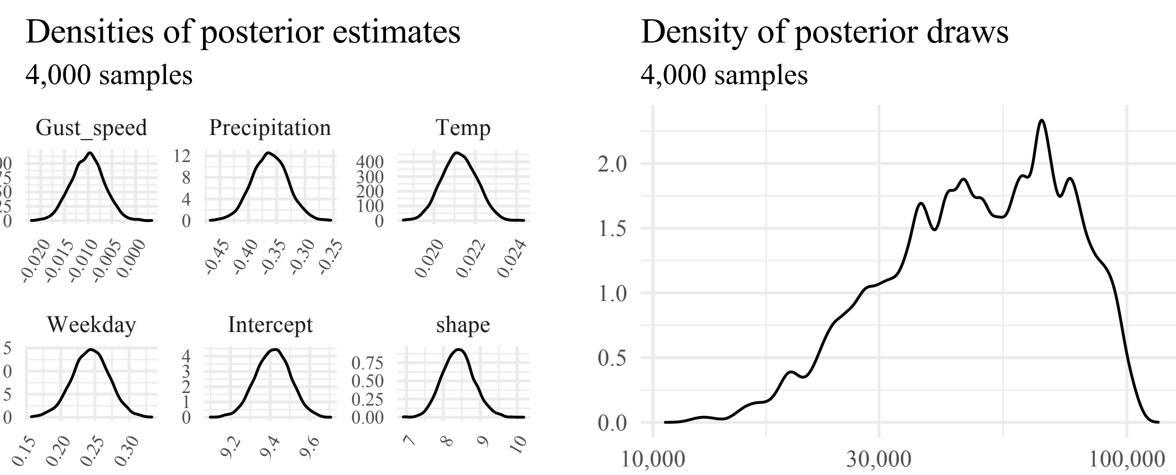

Figure 2: Density of posterior estimates and draws

Table 5: Deciles of posterior draws

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
10,619	26,142	32,623	38,170	43,914	49,890	57,540	64,868	72,437	82,711	116,487

Evaluating the negative binomial model

The model meets all the criteria. Executing leave-one-out cross-validation via `brms::loo()` we see the Pareto k estimates for the negative binomial model are fine with values less than 0.5. The expected log predictive density (ELPD) of the model is approximately -8,000.

Table 6: Negative binomial evaluation

	Estimate	SE
elpd__loo	-8,331	21.55
p__loo	5.784	0.759
looic	16,661	43.1

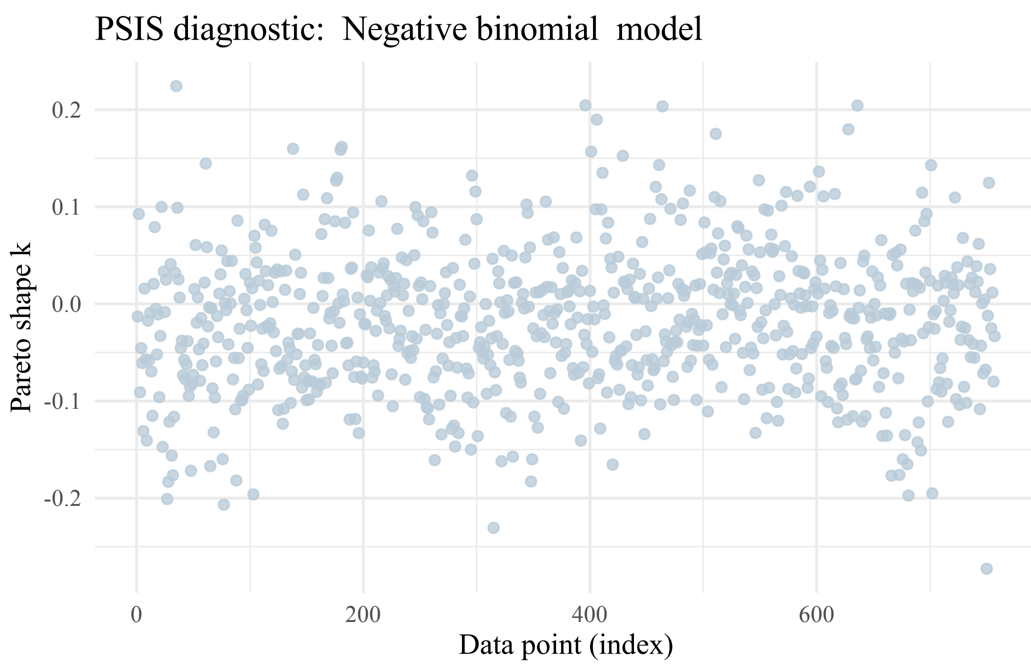

Figure 3: Negative binomial model does not contain large Pareto k values

The model estimates are in-line with the actual observations. The credible interval may be wide but the middle of the estimates are close to the actuals.

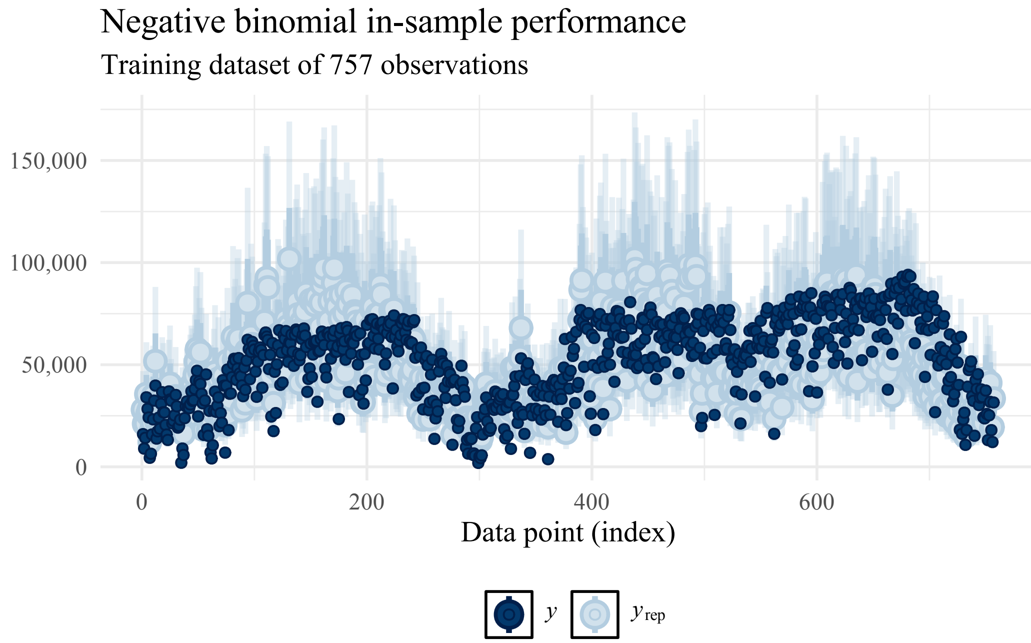

Figure 4: Negative binomial in-sample performance

Predicting new data

The goal of the model is accurate prediction. The data was originally split 80% (757 observations) for training and 20% (181 observations) for testing.

Posterior predictions are made using `brms::posterior_predict()` with argument `newdata = Citibike_test` data. Similar to the in-sample estimates in Figure 5, the out-of-sample estimates in Figure 5 fit the data well but with a wide credible interval.

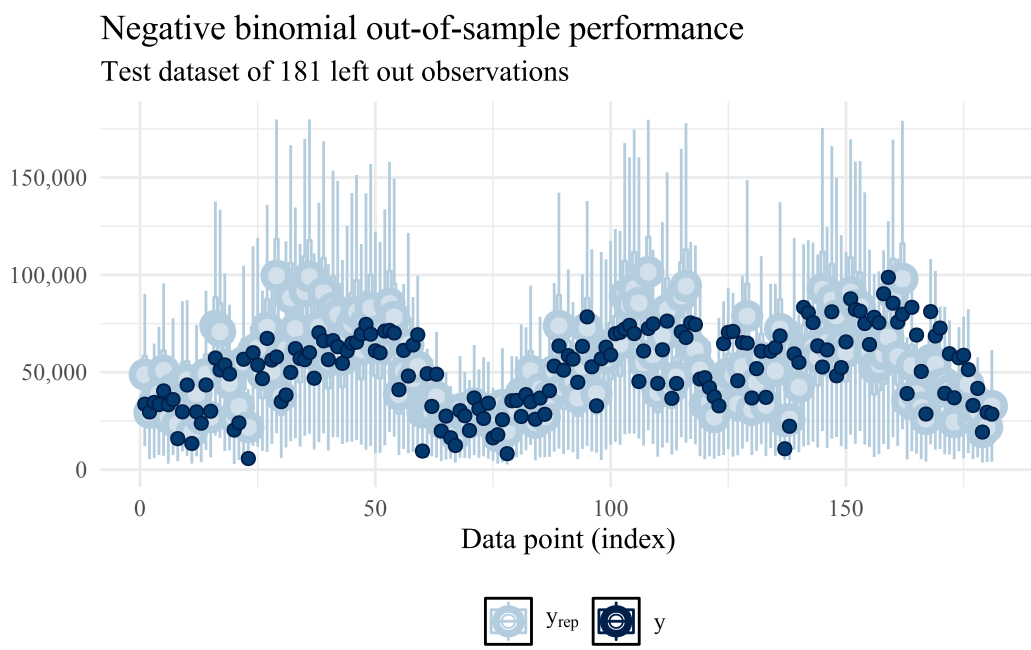

Figure 5: Negative binomial out-of-sample performance

An alternative model: Poisson

Count data is most frequently associated with poisson models, which are a special case of negative binomial. Below, the negative binomial model is refit as a poisson using the same model form.

Table 7: Poisson fixed effects

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	9.654	0.001	9.651	9.656	1	3,852	3,243
Weekday	0.189	0	0.188	0.189	1.003	1,121	1,511
Precipitation	-0.29	0	-0.291	-0.289	1.003	1,376	1,526
Temp	0.018	0	0.018	0.018	1	4,427	3,495
Gust_speed	-0.01	0	-0.01	-0.01	1	3,894	2,603

Compared to the negative binomial model it has a worse ELPD, is more complicated (considerably larger `p_loo` value), and has a substantial number of large Pareto k. 542 or 72% of the observations have Pareto k values large than 0.5 indicating the posterior distribution is sensitive. 384 or 51% of the observations have values over 1.

Table 8: Model comparison

	Poisson Estimate	Poisson SE	Negative binomial Estimate	Negative binomial SE
elpd_loo	-1,236,683	64,667	-8,331	21.5
p_loo	11,480	495.4	5.8	0.8
looic	2,473,366	129,335	16,662	43.1

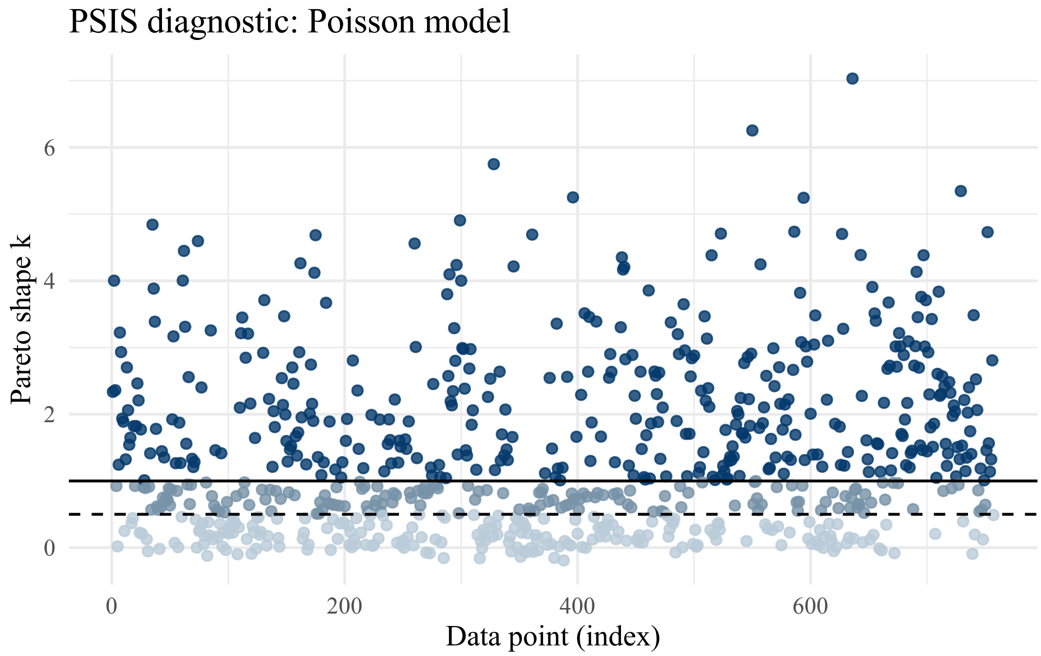

Figure 6: Poisson model contains many large Pareto k values

The poisson model is estimating the data well but is severely overfitting.

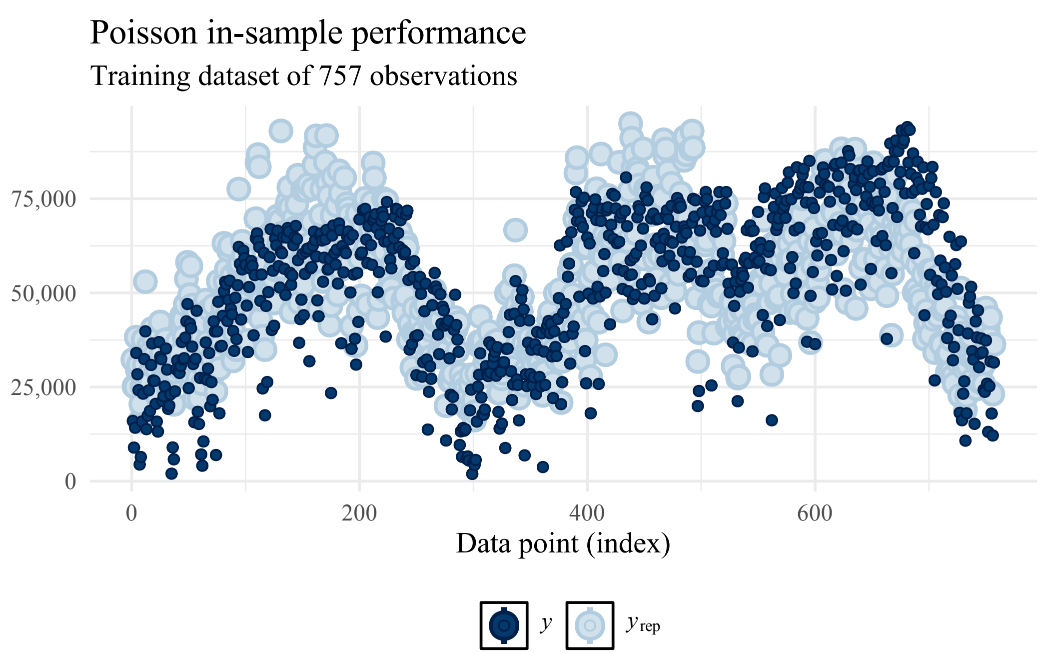

Figure 7: Poisson model in-sample performance

Prediction comparison

The point estimates of each model are similar. Using the mean estimates for each out-of-sample observation, the root-mean-squared-error (RMSE) of the negative binomial and the poisson are close However, the severe overfitting of the poisson relative to the negative binomial is evident in figure 8.

Table 9: RMSE

	RMSE
Negative binomial	13,984
Poisson	12,605

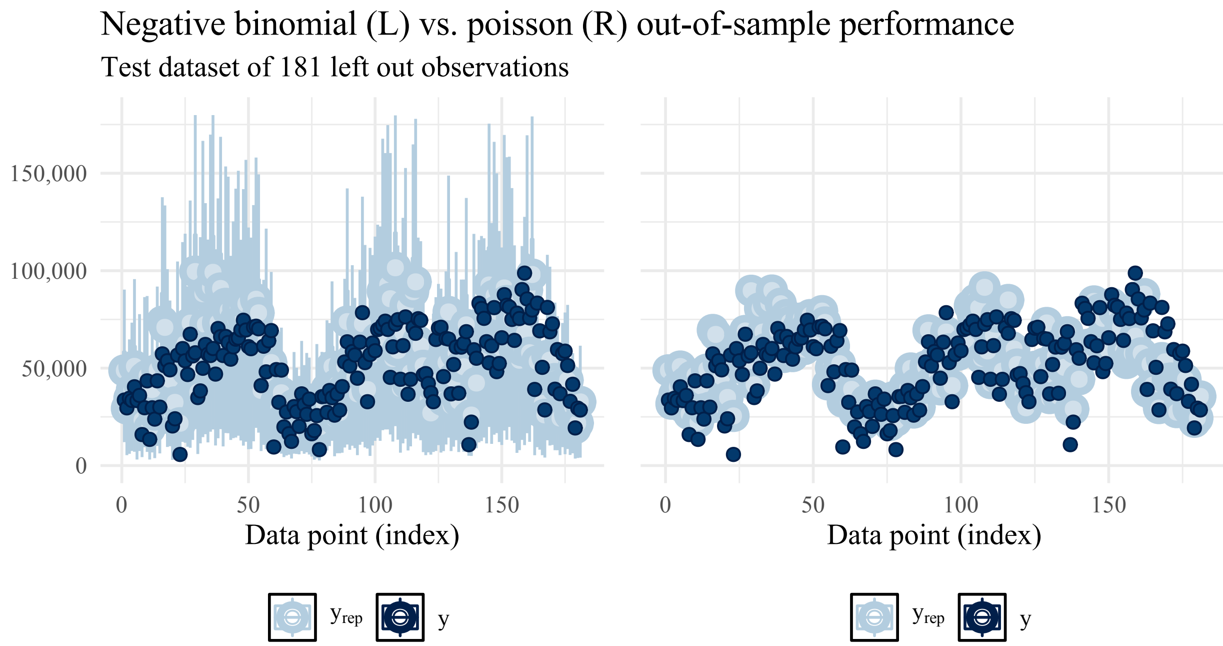

Figure 8: Out-of-sample comparison

Conclusion

Between these two models the negative binomial is superior. Negative-binomial models allow the variance of the distribution to be larger than the mean. This is important in the Citi Bike data as it is over-dispersed: the mean is 51,824 and the variance 418,294,720.

Overall model performance is mediocre. The model captures much of the variation due to weather and day of the week. However, the RMSE of 13,984 is rather large so it may not be an effective model in practice.

Appendix

Data characteristics

The trip count is roughly normally distributed but slightly left-skewed distribution with a mean of 51,824. There are more trips on weekdays and less for rainy days. It's positively correlated (0.76) with temperature, and negatively correlated (-0.44) with gust speed. Temperature is left skewed and gust speed is right skewed.

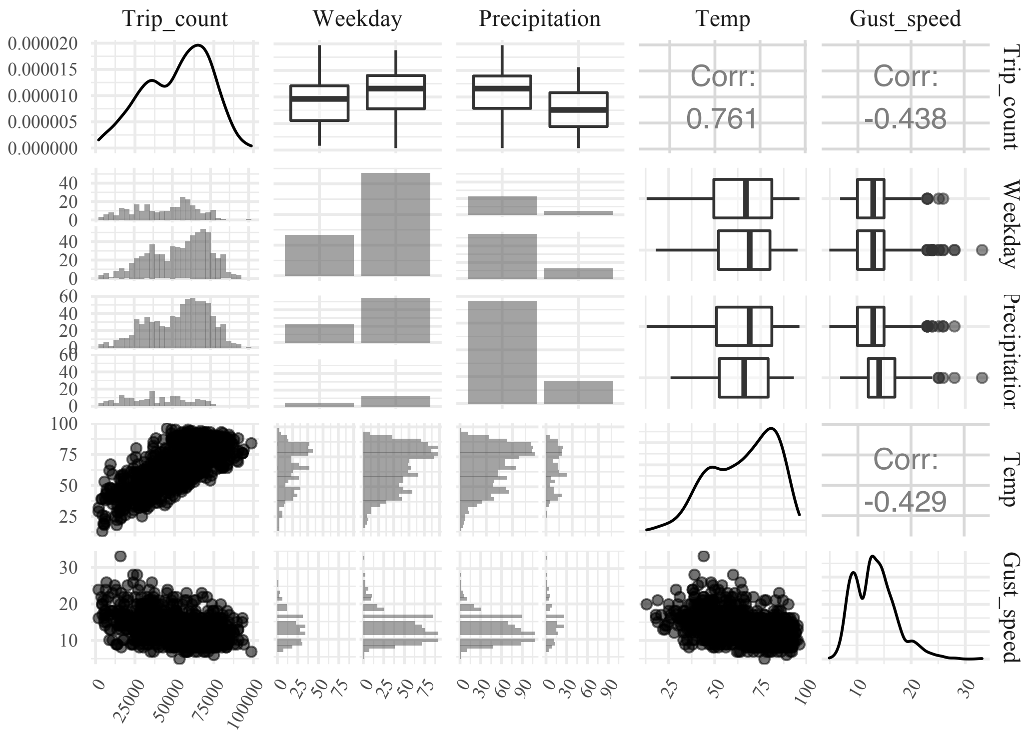

Figure 9: Pairs plot of the data