College of Arts and Sciences
Dr. Bing Zhang Department of Statistics

UNIVERSITY OF KENTUCKY

College of Arts and Sciences
Dr. Bing Zhang Department of Statistics

DR. BING ZHANG DEPARTMENT OF STATISTICS

Predicting the best model for Bike Sharing Demand

Author:
Pronob Barman

July 30, 2021

1 Bike Sharing Data Analysis

Introduction

The data set used for this project is publicly available through the UCI Machine Learning Repository (Bike Sharing Data) and Fanaee-T and Gama (2013). This data set contains two types of dataset for bike one is hour and the other is day. We are interested in hour dataset in our analysis. Our interested dataset contains 17 variables. And this dataset contains sample size $n = 17379$. For our convenience purpose we confined our dataset only for six-month interval from 01 April 2012 to 30 September 2012. This leaves 4390 observations in the dataset. Below are the candidate variables that used in our analysis:

- casual: count of casual users, response variable.
- season(1:winter, 2:spring, 3:summer, 4:fall)
- mnth : month (1 to 12)
- hr : hour (0 to 23)
- holiday : weather day is holiday or not
- weekday : day of the week
- workingday : if day is neither weekend nor holiday is 1, otherwise is 0
- weathersit : 1 = Clear, Few clouds, Partly cloudy, Partly cloudy
  2 = Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  3 = Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  4 = Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp : Normalized temperature in Celsius
- atemp: Normalized feeling temperature in Celsius
- hum: Normalized humidity
- windspeed: Normalized wind speed

In the data analysis, we not only use `R`, but `SAS` to perform the data analysis so that we can present the results more comprehensive.

Data Analysis

1.1 Data Description

Before we perform the data analysis, the very important part is to do the data wrangling. Fortunately, the dataset is neat and using `R` package `dplyr` (Wickham et al. (2020)), we easily filter and subset the dataset we desired for the data analysis. Then we perform descriptive statistics on the outcome variable and on all candidates for explanatory variables. Here is our data set summary statistics for all continuous variables:

Variable	Min	1st Q	Median	Mean	3rd Q	Max	Sd
temp	0.28	0.56	0.64	0.6411	0.72	1.000	0.1279
atemp	0.2424	0.5303	0.6212	0.6013	0.6667	0.9242	0.1130
hum	0.16	0.46	0.62	0.6058	0.77	0.94	0.1868
windspeed	0.000	0.1045	0.1642	0.1824	0.2537	0.7164	0.1128
casual	0.000	10	40	58.01	81	361	63.5242
hr	-1.000	-0.707	0.000	-0.000	0.707	1.000	0.707

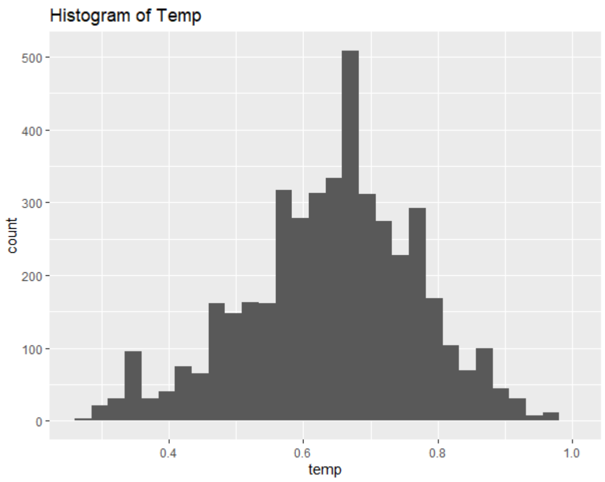
Table 1: Descriptive Statistics of quantitative variables

Frequency tables for all other categorical variables are given below:

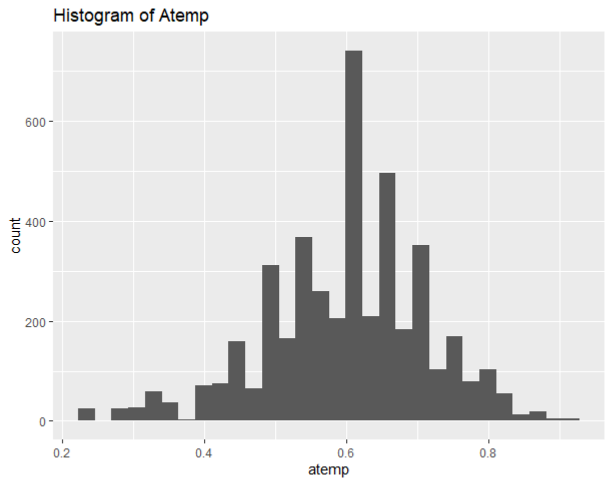
```
   season
  2    3    4
1942 2256  192

   mnth
  4   5   6   7   8   9
718 744 720 744 744 720
   hr
  0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
183 183 183 182 182 183 183 183 183 183 183 183 183 183 183 183 183 183 183 183 183
 21  22  23
183 183 183
   holiday
   0    1
4294   96
   weekday
 0    1    2    3    4    5    6
648  623  624  623  624  624  624
   workingday
   0    1
1368 3022
   weathersit
   1    2    3
3130  958  302
```

Next, we provide the histograms in Figure 1 of all continuous variables in our dataset. We are using `ggplot2` (Wickham (2016)) to perform the graphing.

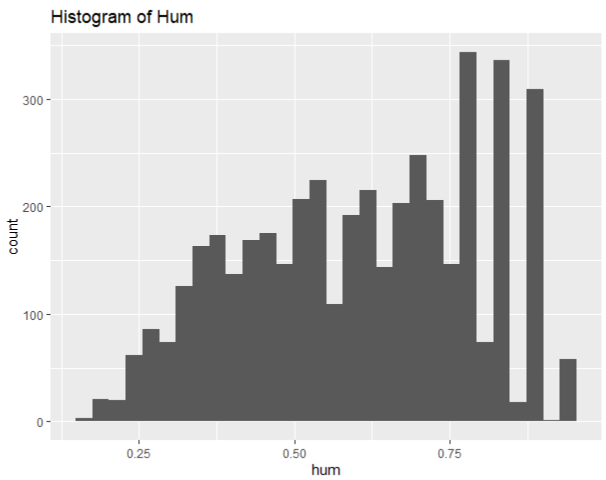
According to the scatter plots shown in Figure 2 and boxplots on Figure 3, there are a few things to notice. First, the correlation between temp and atemp is very high(0.943). Therefore, if one of two is included in the model, we don't need the other. Second, according to the boxplot of weekdays and casual, the day 0(Sunday) and day 6(Saturday) showed similar shape boxplot and day 1(Monday) through day 5(Friday) showed similar shape boxplot. So, we can consider combining those levels to simplify the model. Third, the variable `hr` has 24 levels and apparently not suitable to treat it as categorical. Instead of directly treat it as continuous variable, we consider a transformation. A common way to do it is to transform the variable using sine function to fit the model. Hence, `hr` we use in the model later will be treated as a sine transformed continuous variable.



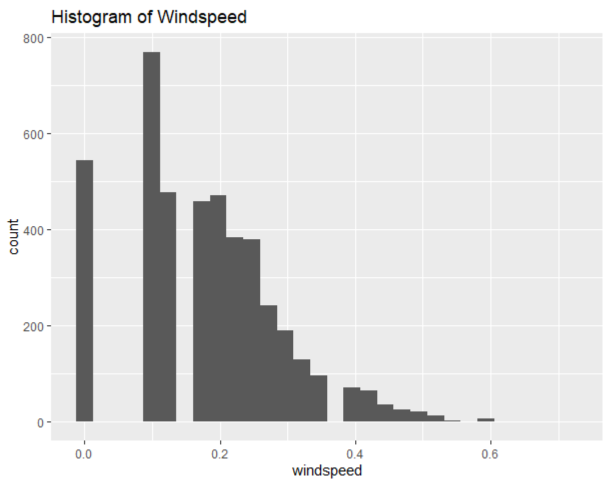
(a)



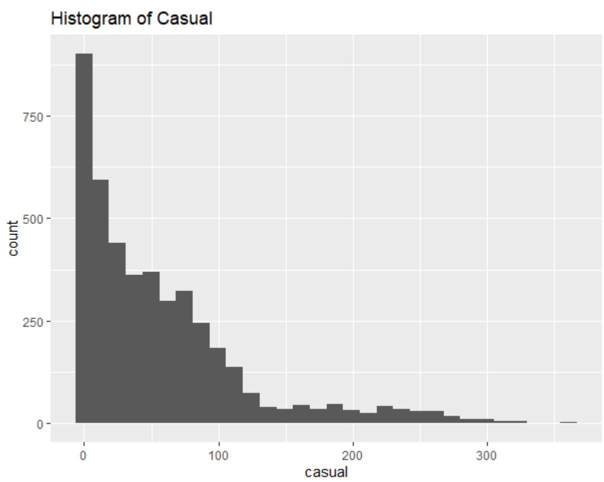
(b)



(c)



(d)



(e)

Figure 1: The histogram of (a): Temp. (b): Atemp. (c): Hum. (d): Wind. (e): Casual

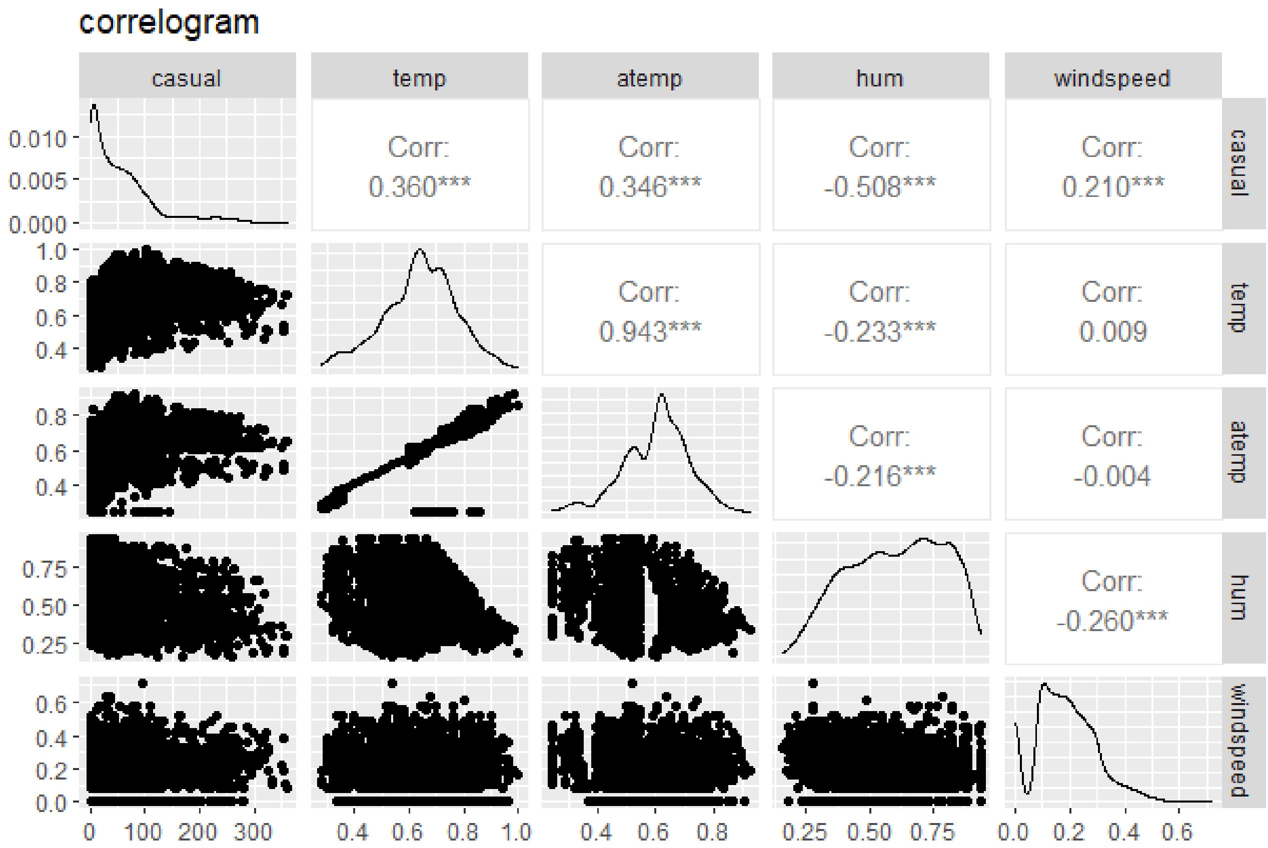


Figure 2: Scatterplots of five quantitative variables

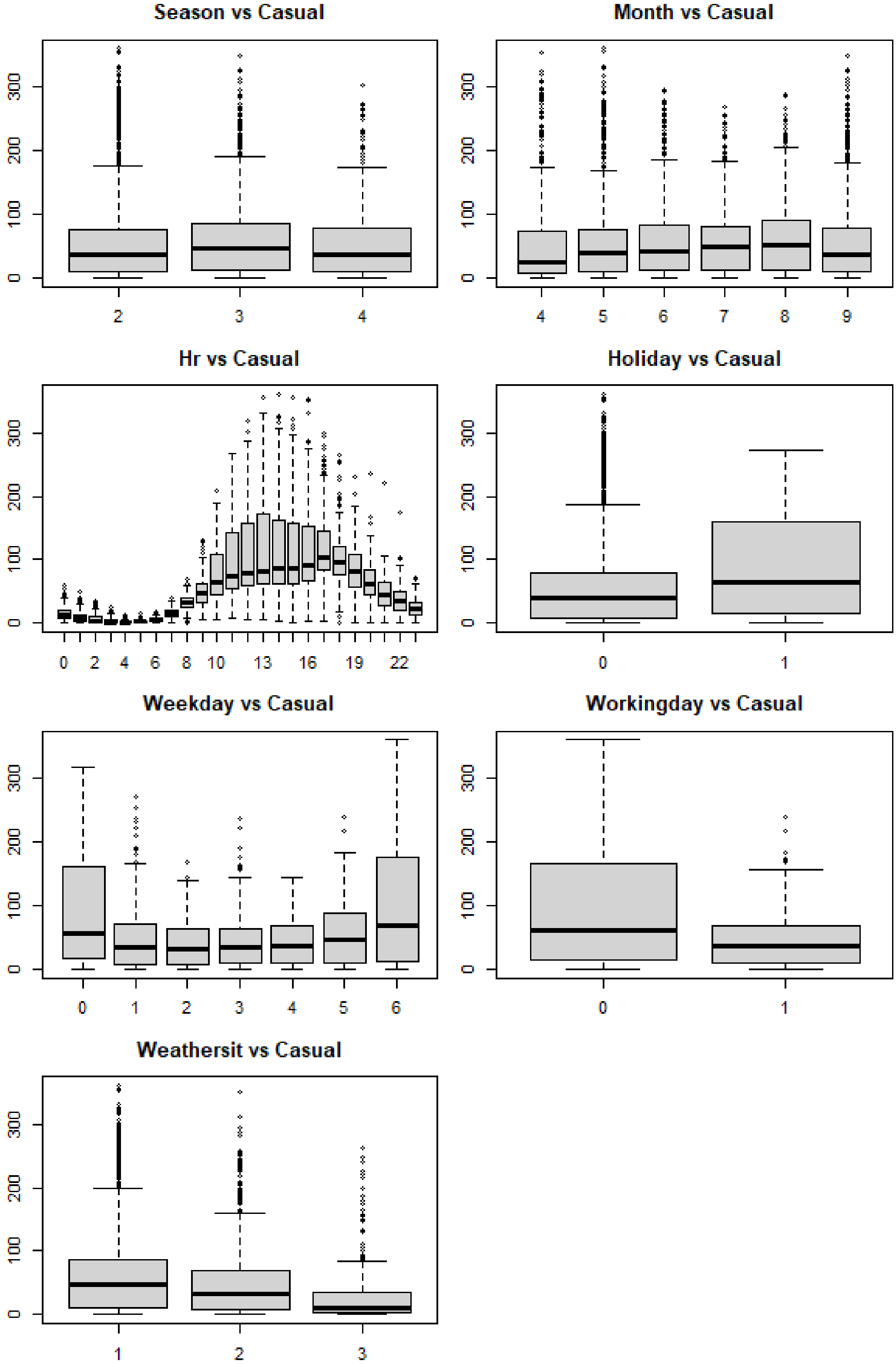


Figure 3: Boxplots of response variable by each categories

1.2 Model

Casual is a count variable, and as suggested in Figure 2, the response variable(casual) has a peak at count 0. Considering the characteristics, we employed Poisson regression, zero-inflated Poisson regression, negative binomial regression, and zero-inflated negative binomial to find our best model. We used forward selection[1] method to fit each model[2]. In order to fit the model, we use package `pscl` (Achim Zeileis (2008)) to help us fitting the zero-inflated model.

The results are presented in Table 2, Table 3, Table 4, and Table 5. According to the result, BIC improves enormous amount when the model changes from Poisson to Negative Binomial. Zero-inflated setting also improved BIC significant amount.

We chose our best model based on BIC. The model is Zero-inflated Negative Binomial with five explanatory variables(hr, hum, workingday, temp, mnth). The result is presented in Figure 4 and also model visualization in Figure 5.

Model	Explanatory Variables	BIC
P_1	hum	212177.5
P_2	hum, workingday	172361.4
P_3	hum, workingday, hr	141589.5
P_4	hum, workingday, hr, atemp	133779.7
P_5	hum, workingday, hr, atemp, mnth	128351.1

Table 2: Poisson Regression Models: Result of Forward Selection

Model	Explanatory Variables	BIC
ZIP_1	hum	197676.1
ZIP_2	hum, workingday	160985.8
ZIP_3	hum, workingday, hr	134473.3
ZIP_4	hum, workingday, hr, atemp	127696.8
ZIP_5	hum, workingday, hr, atemp, mnth	122659.4

Table 3: Zero-inflated Poisson Regression Models: Result of Forward Selection

Model	Explanatory Variables	BIC
NB_1	hr	42839.37
NB_2	hr, hum	42148.69
NB_3	hr, hum, workingday	41506.99
NB_4	hr, hum, workingday, temp	41182.99
NB_5	hr, hum, workingday, temp, mnth	41117.43

Table 4: Negative Binomial Models: Result of Forward Selection

Model	Explanatory Variables	BIC
$ZINB_1$	hr	42847.52
$ZINB_2$	hr, hum	42147.63
$ZINB_3$	hr, hum, workingday	41487.46
$ZINB_4$	hr, hum, workingday, temp	41171.90
$ZINB_5$	hr, hum, workingday, temp, mnth	41103.63

Table 5: Zero-inflated Negative Binomial Models: Result of Forward Selection

[1] Backward elimination could leave out the most influential variable first if there is high level of collinearity between the candidate explanatory variables. Since some of explanatory variables have high correlation(atemp and temp, workingday and holiday), forward selection is safer way to fit the model to guarantee that the model includes the most influential variables.

[2] We used BIC as the criteria

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	3.0156	0.1529	2.7162	3.3157	388.98	<.0001
hour		1	-0.5460	0.0221	-0.5894	-0.5027	609.47	<.0001
hum		1	-2.3393	0.0917	-2.5191	-2.1596	651.23	<.0001
workingday	0	1	0.7496	0.0287	0.6935	0.8060	682.54	<.0001
workingday	1	0	0.0000	0.0000	0.0000	0.0000	.	.
temp		1	3.0305	0.1890	2.6604	3.4015	257.05	<.0001
mnth	4	1	0.0326	0.0581	-0.0813	0.1466	0.32	0.5743
mnth	5	1	0.1129	0.0461	0.0225	0.2034	6.00	0.0143
mnth	6	1	-0.2399	0.0469	-0.3319	-0.1479	26.14	<.0001
mnth	7	1	-0.3536	0.0504	-0.4525	-0.2549	49.25	<.0001
mnth	8	1	-0.1519	0.0490	-0.2480	-0.0559	9.62	0.0019
mnth	9	0	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion		1	0.7392	0.0167	0.7073	0.7727		

Note: The negative binomial dispersion parameter was estimated by maximum likelihood.

Figure 4: Fit for Zero-inflated Negative Binomial Model

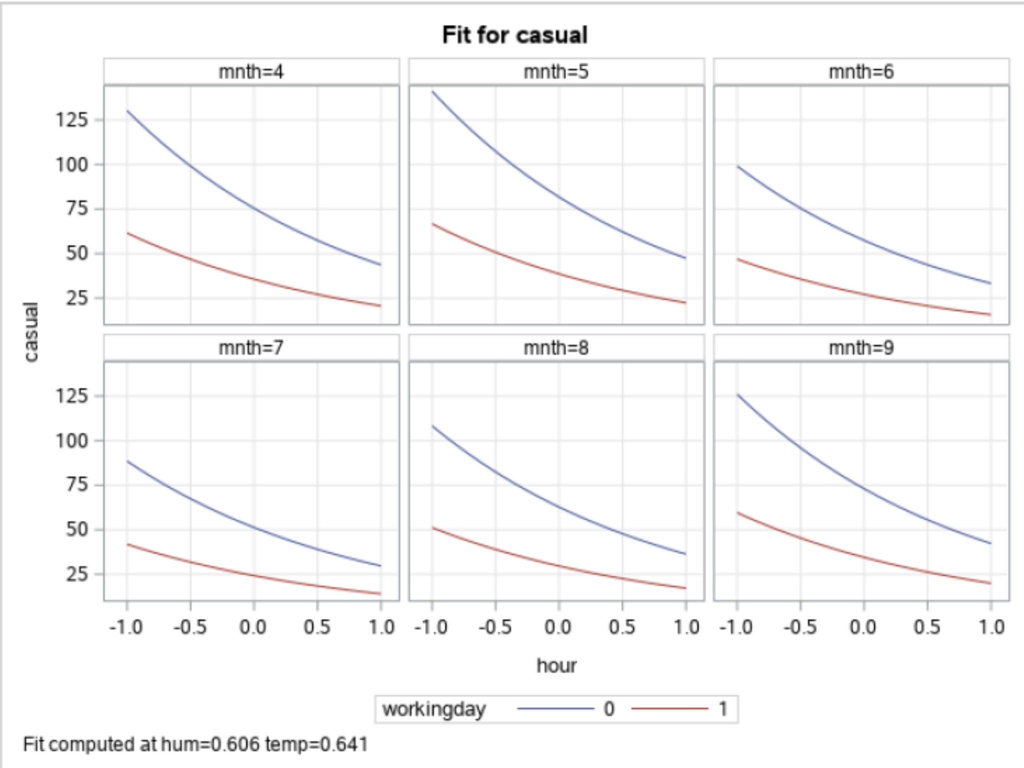


Figure 5: Visualization of Zero-inflated Negative Binomial Model

1.3 Interpretation

Figure 4 summarizes the result of the fitted model. The most influential variables were hum, hr, and workingday. The more humid the weather is, people rent less bicycles. Compared to the driest day, the average of response variable drops 2.33 in the most humid day. Hour was also one of the most important variable. According to the result, people rented 1.09 more bicycles in the busiest hour compared to idle hour. In addition, people rent 0.75 less bicycles on working days. Temperature was also a significant factor, compared to the coldest day, 3.03 more bicycles were rent in the hottest day.[3]

1.4 Diagnostic

To see the model fit, we use the randomized quantile residual plot. Since the generalized linear model does not have the assumption that the error term has to follow normal distribution, we cannot simply use QQ-plot here. Instead, we can use randomized quantile residual plot, which by theory, needs to follow normal distribution. In `R`, instead of using `pscl` package to fit the zero-inflation model, we have to use `glmmTMB` package (Brooks et al. (2017)). The result, however, is staying the same no matter which package we used. Figure 6 gives the plot. And we can see that ZINB is quite a good fit. Definitely, we have a little bit of non linear shape in the left tail. But the quadratic pattern is not very obvious since the same behavior does not appear in the right end. One possible way to fix it is to consider different zero-inflated models that could potentially solve the issue. However, based on my experience, it is hard to beat ZINB in the zero-inflation dataset. Regardless, the overall pattern is pretty good in our opinion.

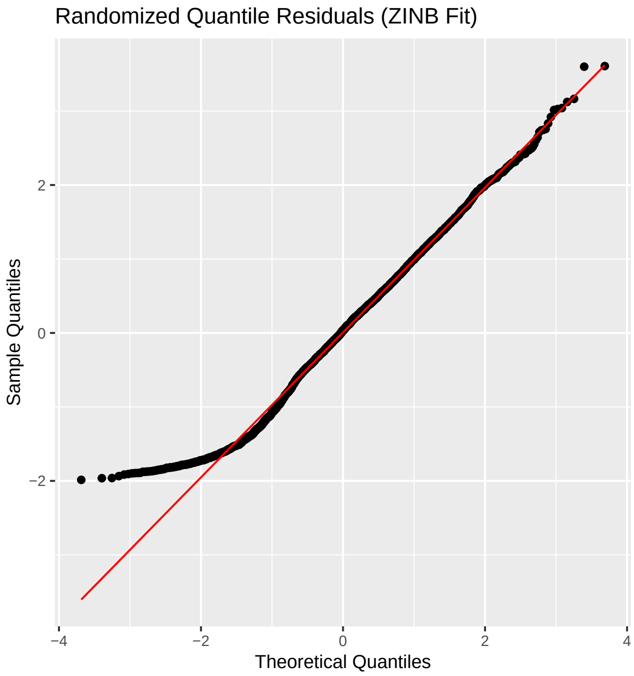


Figure 6: Randomized Quantile Residual

Conclusion

If we were actually explain to the bike rental shop owner, based on the Figure 5, we could say that under the same humidity and temperature. First, workingday is definitely a factor that influence the rentals as our common sense. Second, Spring months have higher rental numbers. Summer months, however, are not boosting

[3]Considering the fact people don't go out when the weather is too hot, this is counter-intuitive. We suspect that the indicator variables for June, July and August have captured the effect of hot weather since the coefficients of all variables(June: -0.24, July: -0.35, August: -0.15) are negative.

the rentals. But September has similar rental pattern like April and May. Lastly, the peak of a day falls on 6 pm.

References

Simon Jackman Achim Zeileis, Christian Kleiber. *Regression Models for Count Data in R.*, 2008. URL `http://www.jstatsoft.org/v27/i08/`.

Mollie E. Brooks, Kasper Kristensen, Koen J. van Benthem, Arni Magnusson, Casper W. Berg, Anders Nielsen, Hans J. Skaug, Martin Maechler, and Benjamin M. Bolker. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9(2):378–400, 2017. URL `https://journal.r-project.org/archive/2017/RJ-2017-066/index.html`.

Hadi Fanaee-T and Joao Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, pages 1–15, 2013.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org`.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2020. URL `https://CRAN.R-project.org/package=dplyr`. R package version 1.0.2.