Part 3: Generalized Linear Models

Charles Julien, Chike Odenigbo, Atul Sharma, Gabriel Jobert

11/17/2023

Contents

Introduction

In the dynamic realm of urban mobility, the Bixi public cycling service plays a pivotal role in providing a sustainable and accessible transportation alternative. As consultants entrusted with a comprehensive analysis of Bixi's operational data, our approach integrates sophisticated statistical techniques, specifically Generalized Linear Models (GLM), to derive actionable insights. This report unfolds the findings derived from GLM applications, shedding light on critical aspects such as trip durations, ridership patterns, and the impact of external factors.

Our analytical scope encompasses a multifaceted examination of factors influencing trip durations, total trip counts, and the temporal dynamics of ridership. By applying GLM to address the identified research questions, we aim to unearth insights that are instrumental in shaping strategic decisions for Bixi's operational enhancements. Central to our methodology is the implementation of Generalized Linear Models, a statistical framework adept at capturing complex relationships within diverse datasets. Our application of GLM is tailored to respond to specific research questions, providing a granular understanding of the nuanced dynamics at play in Bixi's operational landscape.

The main focus of this analysis will be on number of rentals (total, AM, and PM), and on long trips (>15min).

Preliminary T-Test for Variable Selection

As a general hypothesis, prior to diving into our research questions we explored the target variables through variables of interest using t-tests. More specifically, we wanted to determine whether the holiday period and the weather had a significant impact on the targets.

Using a 2 sample t-test, we tested if the average trip duration was the same in holiday vs non holiday period. In all cases, the test statistic is the mean difference difference by the difference in standard deviations of both groups. Using a significance level of 5% we can reject the null hypothesis and conclude that the average trip duration is not the same in holiday and non-holiday periods though they differ by less than a minute using the sample means of the groups.

```
t.test(avg ~ holiday, data = df_main, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  avg by holiday
## t = -2.0716, df = 9998, p-value = 0.03833
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -1.77648134 -0.04909079
## sample estimates:
## mean in group 0 mean in group 1
##        15.28386        16.19665
```

On the other hand, looking at the total number of trips we can conclude that using a significance level of 5% that the trips do not differ from holiday periods to non-holiday periods.

```
t.test(n_tot ~ holiday, data = df_main, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  n_tot by holiday
## t = -0.51403, df = 9998, p-value = 0.6072
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -3.936157  2.300655
```

```
## sample estimates:
## mean in group 0 mean in group 1
##        19.91587        20.73362
```

Looking at membership, we can conclude that non-members take longer trips on average than members using a significance level of 5%. Non-members seem to take trips about 2 minutes longer than members.

```
t.test(avg ~ mem, data = df_main, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  avg by mem
## t = 18.664, df = 9998, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  2.168026 2.676864
## sample estimates:
## mean in group 0 mean in group 1
##        16.58042        14.15798
```

We can also conclude that members take more total trips than non-members using a significance level of 5%.

```
t.test(n_tot ~ mem, data = df_main, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  n_tot by mem
## t = -57.268, df = 9998, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -24.49591 -22.87448
## sample estimates:
## mean in group 0 mean in group 1
##        7.461977       31.147171
```

Looking at days when it rains, we can conclude that there is a significant difference in average trip duration on rainy days compared to non-rainy days with a 95% confidence level. Though the difference is by less than 1 minute.

```
t.test(avg ~ rain_ind, data = df_main, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  avg by rain_ind
## t = 9.1523, df = 9998, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group NoRain and group Rain is not equal to 0
## 95 percent confidence interval:
##  0.9713765 1.5008696
## sample estimates:
## mean in group NoRain   mean in group Rain
##             15.77795             14.54183
```

We can also conclude that the mean difference in number of trips in rainy days and non-rainy days is significant using a 5% significance level.

```r
t.test(n_tot ~ rain_ind, data = df_main, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  n_tot by rain_ind
## t = 3.9303, df = 9998, p-value = 8.543e-05
## alternative hypothesis: true difference in means between group NoRain and group Rain is not equal to 0
## 95 percent confidence interval:
##  0.963759 2.881639
## sample estimates:
## mean in group NoRain   mean in group Rain
##             20.67061             18.74791
```

Through the above t-test, we can expect holidays periods to play a less important role in our variables of interest and for weather as well as membership to play a more significant role. As such we can explore strategies such as dynamic pricing given the weather and membership incentives.

Research Question 1: How does membership and holidays affect likelihood of longer trips (average duration exceeds 15 minutes) ?

Objective of Analysis: Understand member behavior in terms of rental duration to tailor membership benefits and pricing strategies.

Variables Selection

To evaluate how the membership and holidays affect likelihood of longer trips, we need to create a model that accounts for the membership and holidays variable and to use the new variable created to know if the trip is above 15 minutes as the interest variable. The goal would be to quantify the relationship between these factor and the "longer trips" variable.

Model

Here we use a logistic regression model, which is a type of GLM suitable for binary outcomes, with a binomial distribution and a logit link function.

```
##
## Call:
## glm(formula = avg_15_ind ~ mem + holiday + wknd_ind, family = binomial,
##     data = df_main)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3847  -0.9898  -0.7854   1.2351   1.6288
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.13473    0.03251  -4.144 3.41e-05 ***
## mem1            -0.88328    0.04214 -20.962  < 2e-16 ***
## holiday1         0.60985    0.13784   4.424 9.68e-06 ***
## wknd_indWeekend  0.55915    0.04593  12.173  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 13472  on 9999  degrees of freedom
## Residual deviance: 12861  on 9996  degrees of freedom
## AIC: 12869
##
## Number of Fisher Scoring iterations: 4
```

Interpretation

- **Intercept (`-0.13473`)**: This is the log-odds of a trip exceeding 15 minutes when all predictors (mem, holiday, and wknd_ind) are at their reference level (likely non-members, non-holiday, and on weekdays).**Odds (Intercept)**: $e^{-0.13473} = 0.87395$.**Probability (Intercept)**: $\frac{0.87395}{1+0.87395} = 0.4664$ (approximately 46.64%). This means when all independent variables are at their reference levels, the probability of a trip exceeding 15 minutes is approximately 46.64%.
- **Membership (`mem1, -0.88328`)**: The log-odds of a trip exceeding 15 minutes is 0.88328 units lower for members compared to non-members, holding other variables constant (Non-holidays and weekdays). **Odds**: $e^{-0.88328} = 0.41342$. **Probability**: $\frac{0.41342}{1+0.41342} = 0.2925$ (approximately 29.25%). This means that the odds of a trip exceeding 15 minutes for members are about 0.41 times the odds for non-members, translating to a probability of approximately 29.25%.
- **Holiday (`holiday1, 0.60985`)**: The log-odds of a trip exceeding 15 minutes is 0.60985 units higher during holidays compared to non-holidays, holding other variables constant. **Odds**: $e^{0.60985} = 1.84016$. **Probability**: $\frac{1.84016}{1+1.84016} = 0.6479$ (approximately 64.79%). This indicates that the odds of a trip exceeding 15 minutes on holidays are about 1.84 times higher than on non-holidays, resulting in a probability of approximately 64.79%.
- **Weekend (`wknd_indWeekend, 0.55915`)**:The log-odds of a trip exceeding 15 minutes is 0.55915 units higher during weekends compared to weekdays, holding other variables constant (during week-end on non-holidays times). **Odds**: $e^{0.55915} = 1.74919$. **Probability**: $\frac{1.74919}{1+1.74919} = 0.6363$ (approximately 63.63%). This means that the odds of a trip exceeding 15 minutes on weekends are about 1.75 times higher than on weekdays, translating to a probability of approximately 63.63%.
- **Statistical Significance**: In this model, all the p-values are under the 5% level of significance, indicating that the relationship between these variables and the likelihood of taking a longer trip are statistically significant.
- **Model Fit**: The AIC of the model is 12869, which can be used for model comparison purposes. The lower the AIC, the better the model fits the data while penalizing for complexity.

Business implications

- Since members are less likely to take longer trips, membership benefits and pricing could be adjusted to encourage more extended use, or to better cater to the frequent, shorter trips that members seem to prefer.
- The increase in longer trips during holidays and weekends indicates potential opportunities for targeted marketing and promotions to encourage bike usage during these periods.
- The significant increase in the likelihood of longer trips during long weekends, especially on the weekend days, suggests that there might be a need for increased bike availability and maintenance during these times to accommodate the higher demand for leisurely rides.

Anova test

Furthermore, we also performed an Anova test between the full model and a reduced model containing all full model variables except for membership data. The goal here is to determine the addition of membership significantly improves the model's fit.

```
glm_membership_reduced <- glm(avg_15_ind ~ holiday + wknd_ind, family = binomial, data = df_main)
anova(glm_membership_reduced,glm_membership,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: avg_15_ind ~ holiday + wknd_ind
## Model 2: avg_15_ind ~ mem + holiday + wknd_ind
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      9997      13311
## 2      9996      12861  1   450.23 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Model 1 has 9997 degrees of freedom. Model 2 has 9996 degrees of freedom, which is one less than Model 1, indicating that one additional parameter (here, mem) was included in Model 2.
- Model 1 has a residual deviance of 13311. Model 2 has a residual deviance of 12861, which is lower, suggesting a better fit to the data than Model 1.
- The deviance reduction from Model 1 to Model 2 is 450.23. This is the difference in the Resid. Dev between the two models, representing the improvement in fit due to including the mem variable.
- The p-value associated with this deviance reduction is less than 2.2e-16, which is extremely small and indicates that the improvement in the model fit by including mem is statistically significant.

The addition of the mem variable to the model significantly improves the model's ability to predict whether a trip will last longer than 15 minutes. The very low p-value associated with the deviance reduction upon adding mem to the model confirms that membership status has a statistically significant effect on the likelihood of a trip exceeding 15 minutes, beyond what is explained by holiday and weekend indicators alone.

Research Question 2: How do the seasonal variation in weather affect the total number of trip ?

Objective of Analysis: Evaluate how seasonal weather patterns influence rental numbers to inform seasonal staffing and maintenance schedules.

Variables Selection

To evaluate how seasonal weather patterns influence rental numbers, we need to create a model that accounts for the various factors that can vary with seasons, such as temperature, rainfall and specific time of year (season). The goal would be to quantify the relationship between these factors and the number of rentals, which can then inform decisions on staffing and maintenance schedules.

Model

Rental numbers are count data, so a Poisson or negative binomial GLM would both be suitable. Here, we first fit a Poisson regression model and check for over-dispersion. Since the value is significantly greater than 1, we then fit a Negative Binomial model that will be more appropriate for this particular case.

```
##
##  Overdispersion test
##
## data:  glm_seasonal_weather
## z = 26.588, p-value < 2.2e-16
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   26.76113
```

```
##
## Call:
## glm.nb(formula = n_tot ~ temp + rain + season, data = df_main,
```

```
##     init.theta = 0.9901851844, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9974  -1.1527  -0.4834   0.3163   4.5051
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.581791   0.034032  75.863  < 2e-16 ***
## temp          0.031157   0.002213  14.077  < 2e-16 ***
## rain         -0.015916   0.002009  -7.922 2.33e-15 ***
## seasonSpring -0.265988   0.028781  -9.242  < 2e-16 ***
## seasonSummer -0.113913   0.030220  -3.769 0.000164 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.9902) family taken to be 1)
##
##     Null deviance: 11566  on 9999  degrees of freedom
## Residual deviance: 11135  on 9995  degrees of freedom
## AIC: 79919
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.9902
##           Std. Err.:  0.0135
##
##  2 x log-likelihood:  -79906.6310
```

Interpretation

- **Theta**: The estimate for theta ($\theta = 0.9902$) is close to 1. This indicates that the variance is slightly greater than the mean, which is consistent with some overdispersion in the data. This overdispersion justifies the use of the Negative Binomial model over the Poisson model.
- **Model Fit**: The AIC of the model is 79919, which can be used for model comparison purposes. The lower the AIC, the better the model fits the data while penalizing for complexity.
- **Intercept**: The intercept ($\beta = 2.581791$) represents the log count of the total number of trips when all other variables are zero (which is not possible for temperature or season, but serves as a reference point). The IRR for the intercept cannot be interpreted in the same way because it relates to the situation where all predictor variables are zero, which may not be meaningful for variables like temperature.
- **Temperature (`temp`)**: The incidence rate ratio (IRR) for temperature is $e^{0.031157}$, which is approximately 1.032. This means that for each one-degree Celsius increase in temperature, the expected number of total trips increases by a factor of 1.032, or 3.2%.
- **Rainfall (`rain`)**: The IRR for rainfall is $e^{-0.015916}$, which is approximately 0.984. This indicates that for each additional millimeter of rainfall, the expected number of trips decreases by a factor of 0.984, or 1.6%. So, if rainfall increases by 1 mm, the model predicts a 1.6% decrease in the number of trips.
- **Season (`seasonSpring`, `seasonSummer`)**:
  - For `seasonSpring`: The IRR is $e^{-0.265988}$, approximately 0.767. This suggests that, all else being equal, the expected number of trips in spring is 76.7% of the number in the baseline season (`seasonFall`), which is a 23.3% decrease.
  - For `seasonSummer`: The IRR is $e^{-0.113913}$, approximately 0.892. This means that in summer, the expected number of trips is 89.2% of the number in the baseline season (`seasonFall`), a 10.8% decrease.

Business implications

- The bike-sharing service is likely to see increased demand on warmer, drier days. This can guide the allocation of bikes across stations and the scheduling of staff for redistribution and customer service.
- During rainy days, demand is expected to drop, which could be a good time for scheduling maintenance work.
- The unexpected decrease in trips during spring and summer compared to the baseline season suggests that additional factors might need to be considered, or specific marketing strategies might be implemented to boost ridership during these seasons.

Anova test

```
# Poisson model
glm_seasonal_weather_full <- glm.nb(n_tot ~ mem + holiday + temp + rain + season, data = df_main)
anova(glm_seasonal_weather,glm_seasonal_weather_full,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: n_tot ~ temp + rain + season
## Model 2: n_tot ~ mem + holiday + temp + rain + season
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      9995     204541
## 2      9993      10578  2   193964 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Model 1: Predicts n_tot using temp, rain, and season as predictors.
- Model 2: Predicts n_tot using mem, holiday, temp, rain, and season.
- Degrees of Freedom (Df): Model 1 has 9995 residual degrees of freedom, and Model 2 has 9993, which suggests that two parameters were added in Model 2 (mem and holiday).
- Residual Deviance: It is a measure of unexplained variance by the model. Model 1 has a residual deviance of 204541, and Model 2 has a significantly lower residual deviance of 125397, indicating that Model 2 fits the data much better.
- Deviance Difference: The difference in deviance between the two models is 79144, which is highly significant ($p < 2.2e\text{-}16$), indicating that the predictors added in Model 2 (mem and holiday) significantly improve the model.

In conclusion, the analysis strongly suggests that including mem and holiday improves the model's ability to predict n_tot. Given that information, some additional research on the member variable impact on the number of trips will be made. Additionally, the presence of overdispersion justifies the use of the Negative Binomial model over the Poisson model for this data.

Research Question 3: What variables impacts the porportion of trips in the morning versus in the evening and in what way ?

Objective of Analysis: The goal of this analysis is to understand what variables influence the repartition of the trips throughout the day. Knowing this would help to better forecast the demand for bikes across the bixi system.

Before starting the analysis, it is important to know that our datasets has 57.76% of its trips in the afternoon.

```
sum(df_main$n_PM)/sum(df_main$n_tot)
```

```
## [1] 0.5775887
```

Variables Selection

Some variables that would be interesting to investigate are the following:

`mem` : Membership indicator

`wknd_ind` : Indicator of weekend

`season` : Categorical variable with autumn, summer and fall

`temp` : Temperature in degrees celcius

`rain` : Precipitation in mm

`North_South` : Indicator of cardinality compared to parc lafontaine

`West_East`: Indicator of cardinality compared to parc lafontaine

`Metro_ind` : Indicator of metro station nearby

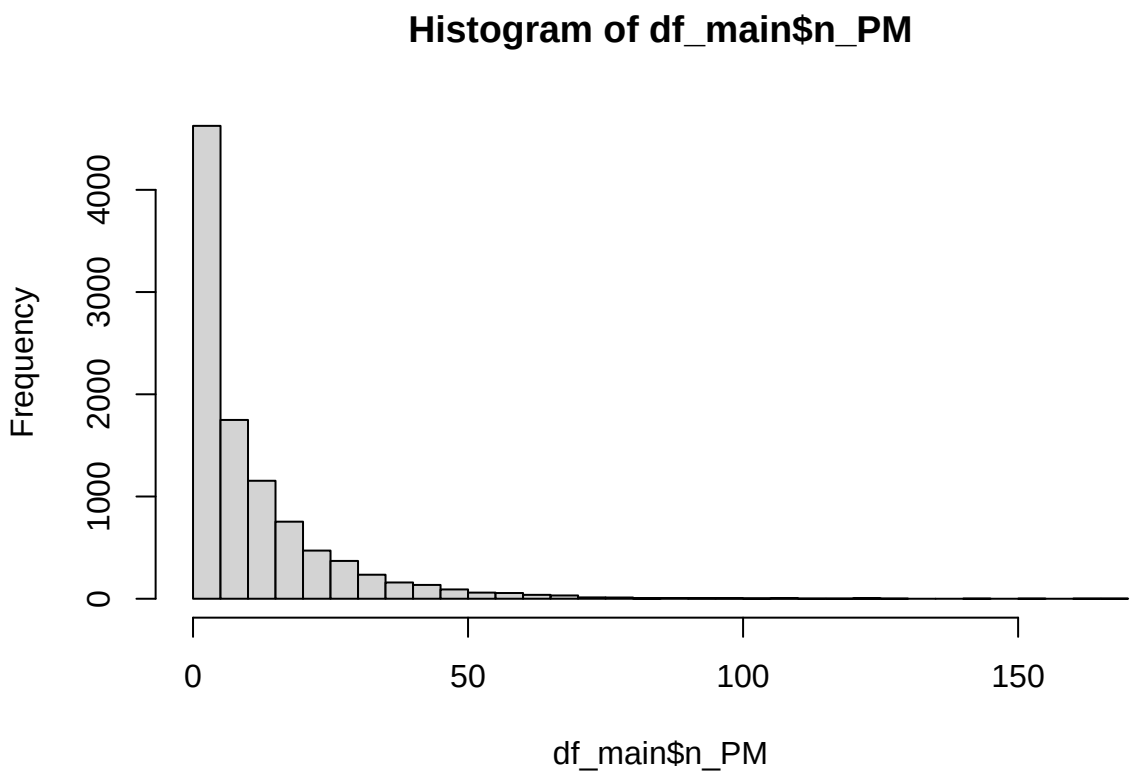
Correlation:

The correlation between the variables selected above was already tested via multicollinearity tests like VIF in previous analysis and did not show source of concern.

Interest variable:

Our interest variable for this question is the proportion of trips in the afternoon compared to the total number of trips. Since `n_PM` is a count we expect a poisson like distribution. To obtain a rate we will use the variable `n_tot` as an offset.

```
hist(df_main$n_PM, breaks = 30)
```




```
mean(df_main$n_PM)
```

```
## [1] 11.514
```

```r
var(df_main$n_PM)
```

```
## [1] 203.9574
```

We observe some big disparities between the mean and variance of the variable `n_PM` which could lead to some overdispersion in our model. Some formal test will be explored in the model part.

Model

```
##
## Call:
## glm(formula = n_PM ~ mem + wknd_ind + season + temp + rain +
##     North_South + West_East + Metro_ind + offset(log(n_tot)),
##     family = poisson(link = "log"), data = df_main)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6371  -0.6189  -0.0042   0.5520   3.4319
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.6050360  0.0140507 -43.061  < 2e-16 ***
## mem1              0.0390626  0.0078914   4.950 7.42e-07 ***
## wknd_indWeekend   0.0222846  0.0065671   3.393  0.00069 ***
## seasonSpring      0.1812232  0.0086055  21.059  < 2e-16 ***
## seasonSummer      0.0053612  0.0083533   0.642  0.52100
## temp             -0.0015968  0.0006609  -2.416  0.01569 *
## rain             -0.0015055  0.0006500  -2.316  0.02055 *
## North_SouthSouth  0.0414373  0.0061343   6.755 1.43e-11 ***
## West_EastWest    -0.0179495  0.0063071  -2.846  0.00443 **
## Metro_ind1        0.0090906  0.0098543   0.922  0.35627
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 8274.5  on 9999  degrees of freedom
## Residual deviance: 7575.7  on 9990  degrees of freedom
## AIC: 43343
##
## Number of Fisher Scoring iterations: 4
```

```r
# Deviance based
mod.poi$deviance/mod.poi$df.residual
```

```
## [1] 0.7583251
```

```r
# Based on Pearon X2 statistic
sum(residuals(mod.poi, type = "pearson")^2)/mod.poi$df.residual
```

```
## [1] 0.6835404
```

Once the covariates and offset are taken into consideration, we seem to be more in a case of underdispersion since theta is smaller than 1. For this reason, we will explore a quasipoisson distribution in order to increase flexibility and allow the mean to be different from the variance.

```r
mod.quasi <- glm(n_PM ~ mem + wknd_ind +season + temp + rain + North_South  + West_East + Metro_ind + offs
summary(mod.quasi)
```

```
##
## Call:
## glm(formula = n_PM ~ mem + wknd_ind + season + temp + rain +
##     North_South + West_East + Metro_ind + offset(log(n_tot)),
##     family = quasipoisson, data = df_main)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6371  -0.6189  -0.0042   0.5520   3.4319
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.6050360  0.0116166 -52.084  < 2e-16 ***
## mem1              0.0390626  0.0065244   5.987 2.21e-09 ***
## wknd_indWeekend   0.0222846  0.0054294   4.104 4.09e-05 ***
## seasonSpring      0.1812232  0.0071147  25.472  < 2e-16 ***
## seasonSummer      0.0053612  0.0069062   0.776 0.437596
## temp             -0.0015968  0.0005464  -2.922 0.003481 **
## rain             -0.0015055  0.0005374  -2.802 0.005095 **
## North_SouthSouth  0.0414373  0.0050716   8.170 3.45e-16 ***
## West_EastWest    -0.0179495  0.0052145  -3.442 0.000579 ***
## Metro_ind1        0.0090906  0.0081472   1.116 0.264539
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 0.6835423)
##
##     Null deviance: 8274.5  on 9999  degrees of freedom
## Residual deviance: 7575.7  on 9990  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

```r
Anova(mod.quasi, type = 3)
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: n_PM
##             LR Chisq Df Pr(>Chisq)
## mem            36.15  1  1.831e-09 ***
## wknd_ind       16.79  1  4.165e-05 ***
## season        711.07  2  < 2.2e-16 ***
## temp            8.54  1  0.0034798 **
## rain            7.92  1  0.0048854 **
## North_South    66.99  1  2.729e-16 ***
## West_East      11.82  1  0.0005845 ***
## Metro_ind       1.24  1  0.2650880
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation

Overall Model We observe that the deviation parameter of the model is estimated to be 0.68 which means that the means given the covariates is bigger than its variance. According to the Anova command, all variables included in the model are significant, once adjusted for all the covariates, except for `metro_ind`

Intercept : -0.605 can be interpreted as the log average rate of trip in the afternoon when all covariates have values of zero.

In other words, this is the average rate of trip in the afternoon for non-member, during the weekday, in fall, when temperature is zero degrees celcius, no precipitation, with rental location north east of parc lafontaine and at an acces point that as no metro station nearby which is exp(-0.605) = 54.6%.

Membership : 0.039 is a positive coefficient hence membership increase the rate of trip in the afternoon, on average, when all else is being held constant by a factor of exp(0.039) = 1.039 which means an increase of about 4% .

Weekend: 0.022 is a positive coefficient hence weekend increase the rate of trip in the afternoon, on average, when all else is being held constant by a factor of exp(0.022) = 1.022 which means an increase of about 2% .

Season: Spring's coefficient is 0.181 and summer's is 0.005 meaning that they both have an increase in rate of trip in the afternoon on average compared to fall when all else is being held constant. This increase is of a factor of exp(0.181) = 1.198 and exp(0.005) = 1.005 respectively for both season.

Temperature: -0.002 which means that a one degree celcius increase in temperature results in a decrease in rate of afternoon trips on average when all else is being held constant. This decrease is of a factor of exp(-0.002) = 0.998.

Precipitation :- rain coefficient is also -0.002 hence its interpretation is the same as for temperature except that the decrease is for each additional milliliters of rain.

Cardinality North-South: 0.041 is a positive coefficient hence departure from bixi station south of parc lafontaine have a higher rate of trip in the afternoon, on average, when all else is being held constant by a factor of exp(0.041) = 1.041 which means an increase of about 4% compared to northern departure.

Cardinality West-East: -0.017 is a positive coefficient hence departure from bixi station west of parc lafontaine have a lower rate of trip in the afternoon, on average, when all else is being held constant by a factor of exp(-0.017) = 0.983 which means a decrease of about 2% compared to eastern departure.

Metro station nearby: coefficient is 0.009 which although not being significatively differrent from zero can be interpreted as having a metro station nearby increase the rate of trip in the afternoon, on average, when all else is being held constant by a factor of exp(0.009) = 1.009 .

Business Implications:

The main takeaways from this model are:

- Members have a higher rate of trips in the afternoon. Knowing that members account for most of the trips, it can explain why there is more trip in the afternoon in general.

```
## # A tibble: 2 x 2
##   mem    n_tot
##   <fct>  <int>
## 1 0      35325
## 2 1     164021
```

- There will be an increase demand on the system in the afternoon during the weekend and an increase demand on the system in the morning during the weekdays. This could reflect the usage of people using bixi to commute to work.
- There is a strong increase in rate of trips in the afternoon during the season of spring, this could be seen as an eagerness for bike after winter since afternoon trips seems to be more associated with leisure than commuting. Another hypothesis would be that during spring the mornings are too cold to bike most often.
- Following the above hypothesis, as temperature increase, there seems to be an increase of rate of trip in the morning. Keep in mind that this relation is only true for a given season.
- Finally concerning the general flow of trips, there seems to be a higher rate of departure from stations North West to Parc Lafontaine in the morning than in the afternoon. Similarly, we have the inverse relation for station in the South East. This means that from an operational standpoint, there might be some displacement of bikes required from stations to stations depending on the moment of the day to keep a balanced fleet of bikes all over the system.

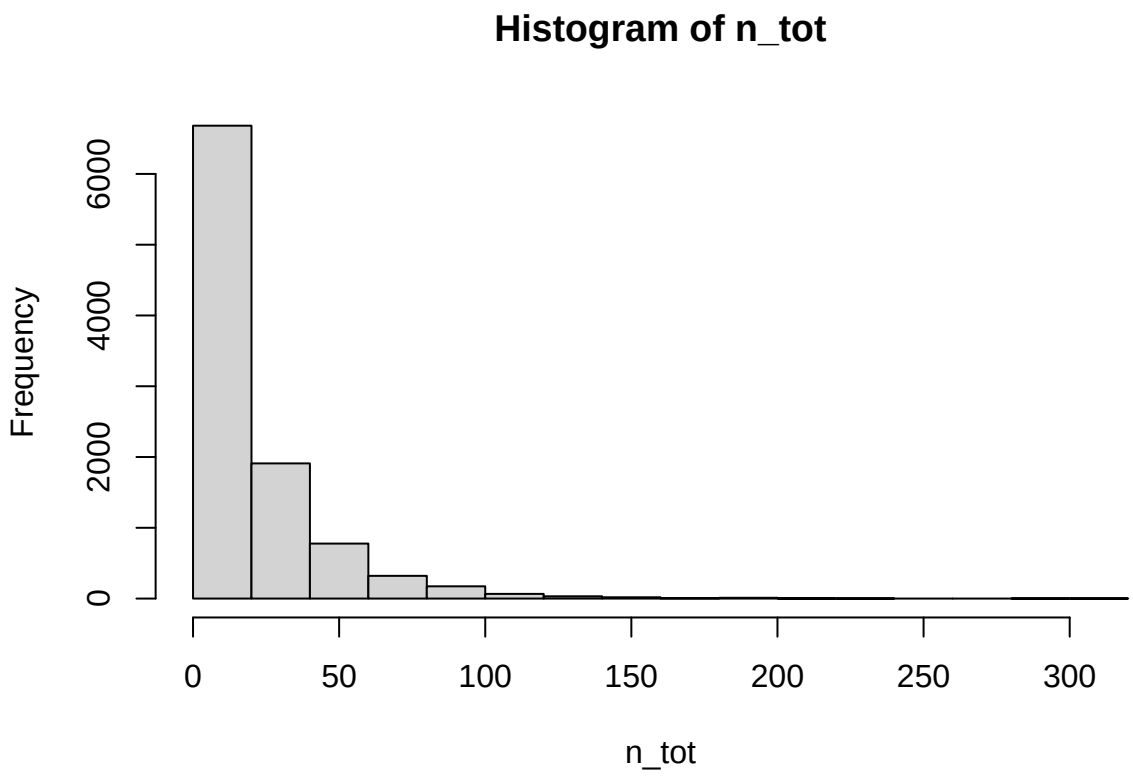
Research Question 4: Are there significant differences in bike trips counts between weekdays and weekends?

Objective of Analysis: The objective of the analysis is to quantify and assess the differences in ridership during weekend and weekdays. The analysis would generate insights into the patterns of ridership that could be useful for operational planning, resource allocation, or service improvements.

Variables Selection

We can also see from the histogram that the response variable is skewed to the right. It's clear that linear regression would not be appropriate in this context. We can fit the Poisson regression model using the glm function, specifying that the distribution is poisson. We'll start by fitting the model which includes all explanatory variable:

```
hist(df_main$n_tot, breaks = 20, main = "Histogram of n_tot", xlab = "n_tot")
```



Model

```
##
## Call:
## glm(formula = n_tot ~ weekday_weekend, family = poisson, data = df_main)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -5.686  -4.256  -2.071   1.446  33.873
##
## Coefficients:
##                          Estimate Std. Error  z value Pr(>|z|)
## (Intercept)              3.004083   0.002644 1136.377  < 2e-16 ***
## weekday_weekendWeekend -0.040604   0.004976   -8.159 3.37e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 212631  on 9999  degrees of freedom
## Residual deviance: 212564  on 9998  degrees of freedom
## AIC: 254754
##
## Number of Fisher Scoring iterations: 5
```

Interpretation

Overall Model R squared, F-stat

Intercept :

The intercept (3.004083) represents the log count of the total number of trips on weekdays.In other words, the number of trips on weekdays is exp(3.004083)=20.

Weekend :

weekend = -0.040604: number of trips is higher on weekdays than weekend. More specifically, the mean number of trips on weekend is exp(-0.040604) = 0.96 times those on weekdays. In other words, the mean number of trips on weekend is 4% lower than that in weekdays.The p-value is 3.37e-16 which is less that any acceptable value of alpha which represent that the difference in trip number during weekend and weekdays is statistically different.

Business Implications:

Resource Allocation: Given the higher demand during weekdays, businesses or services dependent on these trips may need to allocate more resources, such as bikes, maintenance, or staff, during these periods.

Pricing Strategies: If the service is fee-based, adjusting pricing strategies to account for the difference in demand between weekdays and weekends could be considered. For instance, offering promotions or discounts during lower demand days (weekends) to attract more users.

Marketing Efforts: Tailoring marketing campaigns or efforts to encourage more weekend usage could be explored. Promoting special events, family packages, or leisure-oriented offers during weekends might help in increasing weekend ridership.

Operational Optimization: During weekends, operational adjustments could be made to enhance the user experience. For instance, ensuring bike availability, adjusting operating hours, or implementing user-friendly initiatives to attract more weekend users.

Service Enhancements: Understanding the differences in usage patterns can guide service improvements. Addressing any barriers that might discourage weekend ridership, such as safety concerns, parking availability, or service accessibility, could be a focus for enhancements.

Limitations and shortcomings

Autocorrelation of data: The observations in the dataset are not independent, as seen in previous analyses. This autocorrelation can impact the validity of statistical tests and models, potentially leading to biased or inaccurate conclusions.

External Factors and Generalizability: The analysis primarily focuses on internal variables within the dataset, overlooking potential external influences such as changes in city infrastructure or broader economic conditions. This limits the generalizability of the results to broader contexts.

Temporal Dynamics and Long-Term Trends: The study's insights are confined to the timeframe for year 2021, potentially missing long-term shifts in user behavior or external factors. The temporal dynamics of the bike-sharing service may evolve beyond the study period.

Model Specificity: The use of Generalized Linear Models (GLMs) is tailored for the specific dataset and research questions at hand. While GLMs are versatile, they might not capture all the nuances or complexities of the data, such as non-linear relationships or interactions between variables that a different modeling approach could reveal.

Sensitivity to Parameter Choices: GLMs involve decisions about which link function to use and how to structure the model. Different choices can lead to different interpretations, and the report does not discuss the sensitivity of the results to these choices.

Assumptions of the Modeling Approach: Every statistical model, including GLMs, comes with underlying assumptions (e.g., about the distribution of errors). If these assumptions are violated, it can lead to biased results. The report does not explicitly discuss how these assumptions were tested or met.

Conclusion

In conclusion, the examination of Bixi's operational data utilizing Generalized Linear Models (GLM) has unveiled pivotal insights that can strategically reshape the bike-sharing service. The analysis has pinpointed the profound impact of weather conditions, seasonal variations, and membership dynamics on ridership behavior. This newfound understanding positions Bixi to implement targeted strategies, optimizing resource allocation, and addressing specific user preferences to enhance the overall service experience.

Strategically, Bixi is poised to benefit from dynamic resource allocation informed by weather patterns, ensuring optimal bike distribution and efficient staff scheduling. The identification of seasonal ridership nuances prompts tailored marketing initiatives, providing an opportunity to counteract dips in spring and summer ridership. Moreover, the strategic refinement of membership structures aligns with user preferences, enhancing engagement and satisfaction. Capitalizing on the surge in longer trips during holidays and extended weekends through targeted campaigns further positions Bixi to maximize user engagement and solidify its position in the competitive urban mobility sector.

Contribution

Charles Julien :

Gabriel Jobert :

Chike Odenigbo : t-test, anova, added to feature engineering, part of conclusion

Atul Sharma :