Part 3: Generalized Linear Models

Charles Julien, Chike Odenigbo, Atul Sharma, Gabriel Jobert

11/17/2023

Contents

Instructions

Part 3: Generalized linear models (due November 10 before 11:55 PM)

• Explore various generalized linear models for the response variables of interest, specifically, for the number of rentals (total, AM, and PM). In addition, create a new variable indicating whether the average daily trip duration exceeds 15 minutes, and explore models for this new variable.

• Be sure that your analyses allow you to answer well formulated business / research questions that you wish to address. The goal is to use generalized linear models to provide interesting and relevant insights from the data.

• Comment on findings and discuss the main takeaways from this analysis from a business perspective. Be sure to provide relevant model outputs that support your discussion.

• Discuss any shortcomings or limitations of the analyses carried out.

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.

Business/Research questions

The target variables is number of rentals (total, AM, and PM)

Preliminary T-tests

We decided to look at t-tests as a means of feature selection when assessing the importance of our variables.

Using a two sample t test, we further assessed the impact of membership on average trip duration, the total number of trips and the delta in trips from morning to afternoon for members in comparison to non-members. The goal is to determine whether members behave differently from non-members prior to doing a deep dive on membership. In all cases, the null hypothesis was that the average of the aforementioned variables did not change between members and non-members. Through this analysis, we discovered that average trip duration for non-members is 16.58 minutes in comparison to members at 14.16 minutes, members take on average 31 trips per station in comparison to 7 per station for non-members and both take more trips in the PM than the AM however that difference is larger for members than non-members. Furthermore, we also tested the binary outcome of membership's impact on a station having an average trip duration greater than 15 minutes and we noted that membership again is significant using the McNemar test.

```r
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
```

```r
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
```

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

```
##
##  Two Sample t-test
##
## data:  n_AM_PM_delta by mem
## t = 41.168, 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:
##  7.698121 8.467868
## sample estimates:
## mean in group 0 mean in group 1
##       -2.935995      -11.018990
```

```r
mcnemar.test(df_main$mem , df_main$avg_15_ind)
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  df_main$mem and df_main$avg_15_ind
## McNemar's chi-squared = 256.53, df = 1, p-value < 2.2e-16
```

Moving on to analyzing ridership data during holidays vs non holidays as well as during the week vs non weekend we can arrive to similar conclusions using a 5% significance level. The data shows that on holidays in comparison to non holidays, and on weekends in comparison to weekdays riders take slightly longer trips and more afternoon trips during holidays as well as on weekends. This difference is statistically significant. When comparing to the total number of trips, in both cases of weekends and holidays, we did notice a very high p-value and thus failed to reject the null hypothesis that there is a significant difference in total number of trips in both cases. As such, with regards to holidays and weekends, we recommend strategies focusing on the time of day and average duration of rides as opposed to total number of trips which would be harder to predict.

```r
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
```

```r
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
```

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

```
##
##  Two Sample t-test
##
## data:  n_AM_PM_delta by holiday
## t = 2.0304, df = 9998, p-value = 0.04235
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  0.04971299 2.82768600
## sample estimates:
## mean in group 0 mean in group 1
##       -7.159554       -8.598253
```

```r
mcnemar.test(df_main$holiday , df_main$avg_15_ind)
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  df_main$holiday and df_main$avg_15_ind
## McNemar's chi-squared = 3570.8, df = 1, p-value < 2.2e-16
```

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

```
##
##  Two Sample t-test
##
## data:  avg by wknd_ind
## t = -12.067, df = 9998, p-value < 2.2e-16
```

```
## alternative hypothesis: true difference in means between group Weekday and group Weekend is not equal t
## 95 percent confidence interval:
##  -2.022257 -1.457083
## sample estimates:
## mean in group Weekday mean in group Weekend
##              14.79939              16.53906
```

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

```
##
##  Two Sample t-test
##
## data:  n_tot by wknd_ind
## t = 1.5311, df = 9998, p-value = 0.1258
## alternative hypothesis: true difference in means between group Weekday and group Weekend is not equal t
## 95 percent confidence interval:
##  -0.2248752  1.8298580
## sample estimates:
## mean in group Weekday mean in group Weekend
##              20.16772              19.36523
```

```
t.test(n_AM_PM_delta ~ wknd_ind, data = df_main, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  n_AM_PM_delta by wknd_ind
## t = 4.8254, df = 9998, p-value = 1.418e-06
## alternative hypothesis: true difference in means between group Weekday and group Weekend is not equal t
## 95 percent confidence interval:
##  0.6682959 1.5827257
## sample estimates:
## mean in group Weekday mean in group Weekend
##             -6.865539             -7.991050
```

```
mcnemar.test(df_main$wknd_ind , df_main$avg_15_ind)
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  df_main$wknd_ind and df_main$avg_15_ind
## McNemar's chi-squared = 303.23, df = 1, p-value < 2.2e-16
```

Finally, we observed the impact of the weather of the different outcomes. More specifically, we looked at the difference between raining periods and warmer weather and non raining periods and colder weather respectively on average trip duration, the total number of rides and taking more rides in the afternoon. As expected in colder periods or rainy days had shorter average ride duration, less total number of rides and less and less afternoon rides than their respective counter parts in the t test.

```
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
```

```
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
```

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

```
##
##  Two Sample t-test
##
## data:  n_AM_PM_delta by rain_ind
## t = -6.3094, df = 9998, p-value = 2.919e-10
## alternative hypothesis: true difference in means between group NoRain and group Rain is not equal to 0
## 95 percent confidence interval:
##  -1.8001066 -0.9467256
## sample estimates:
## mean in group NoRain   mean in group Rain
##            -7.718244            -6.344828
```

```
mcnemar.test(df_main$rain_ind , df_main$avg_15_ind)
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  df_main$rain_ind and df_main$avg_15_ind
## McNemar's chi-squared = 6.6801, df = 1, p-value = 0.00975
```

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

```
##
##  Two Sample t-test
##
## data:  avg by hot_weather_ind
## t = -8.1796, df = 9998, p-value = 3.194e-16
## alternative hypothesis: true difference in means between group Cold and group Hot is not equal to 0
## 95 percent confidence interval:
##  -1.401399 -0.859569
## sample estimates:
## mean in group Cold  mean in group Hot
##           14.56441           15.69489
```

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

```
##
##  Two Sample t-test
##
## data:  n_tot by hot_weather_ind
## t = -14.745, df = 9998, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Cold and group Hot is not equal to 0
## 95 percent confidence interval:
##  -8.272874 -6.331434
## sample estimates:
## mean in group Cold  mean in group Hot
##           15.15242           22.45457
```

```
t.test(n_AM_PM_delta ~ hot_weather_ind, data = df_main, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  n_AM_PM_delta by hot_weather_ind
## t = 11.577, df = 9998, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Cold and group Hot is not equal to 0
## 95 percent confidence interval:
##  2.130441 2.998907
## sample estimates:
## mean in group Cold  mean in group Hot
##          -5.512895          -8.077569
```

```
mcnemar.test(df_main$hot_weather_ind , df_main$avg_15_ind)
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  df_main$hot_weather_ind and df_main$avg_15_ind
## McNemar's chi-squared = 1313.7, df = 1, p-value < 2.2e-16
```

With this information at hand, we can perform multiple regression accounting for other variables of interest while also focusing on membership, time of the week and weather. We can further recommend Bixi to explore dynamic pricing with regards to weather because trips go up significantly in warmer periods.

Pre-processing

Imputation (might have to remove)

Revenue for members is missing since they do not pay a usage fee, but rather a fixed cost.

```
imputation_model <- lm(rev ~ dur + avg + n_tot, data = df_main)
df_main$rev_pred = predict(imputation_model, df_main)
```

To impute revenue for members, we make the assumption that they would bring in as much revenue as non-members for the same usage.Thus we consider the same formula of revenue used for non-members. This unknown deterministic function is most likely a linear combination of usage variables like `dur`, `avg` and `n_tot`. We try to approximate this function and use it to impute members revenue. The imputation model has an r-squared of 1 on non-members data.

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 + long_wknd_ind, family = binomial,
##     data = df_main)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -0.72009    0.28766  -2.503   0.0123 *
## mem1                 -0.88294    0.04215 -20.949  < 2e-16 ***
## holiday1              0.98469    0.23099   4.263 2.02e-05 ***
## long_wknd_indWeekday  0.58520    0.28583   2.047   0.0406 *
## long_wknd_indWeekend  1.14434    0.28950   3.953 7.73e-05 ***
## ---
## 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: 12857  on 9995  degrees of freedom
## AIC: 12867
##
## Number of Fisher Scoring iterations: 4
```

Interpretation

- **Intercept (`-0.72009`)**: This is the log odds of the average trip duration exceeding 15 minutes for a non-member (because `mem` is 0 for non-members), on a day that is neither a holiday (`holiday` is

  0) nor part of a long weekend (`long_wknd_ind` is 0). The negative value of the intercept suggests that, for this reference group, the probability of having a trip duration exceeding 15 minutes is less than 50%. To convert the log odds to probability, you would use the logistic function :$P(\text{avg duration} > 15|\text{non-member, non-holiday, non-long weekend}) = \frac{1}{1+e^{0.72009}}$We can calculate the exact probability:

  $$P = \frac{1}{1+e^{-(-0.72009)}} \approx \frac{1}{1+e^{0.72009}} \approx \frac{1}{1+2.054} \approx \frac{1}{3.054} \approx 0.3275$$

  So, the estimated probability of a non-member taking a trip longer than 15 minutes on a regular day (not a holiday and not a long weekend) is approximately 32.75%.

- **Membership (`mem1, -0.88294`)**: The negative coefficient for `mem1` indicates that BIXI members are less likely to have trips that exceed 15 minutes on average, compared to non-members. The odds ratio can be calculated as $e^{-0.88294} \approx 0.414$, which means the odds of members taking longer trips are about 41.4% of the odds for non-members. Members are therefore less likely to take longer trips, which could indicate that members are using the service for shorter, more regular commutes.
- **Holiday (`holiday1, 0.98469`)**: The positive coefficient for `holiday1` indicates that on holidays, the odds of trips exceeding 15 minutes are higher. The odds ratio is $e^{0.98469} \approx 2.677$, suggesting that the likelihood of longer trips is about 167.7% higher on holidays than on non-holidays.

- **Long Weekend Weekday (`long_wknd_indWeekday, 0.58520`)**: The positive coefficient here suggests that on weekdays that are part of a long weekend, the likelihood of longer trips increases. The odds ratio is $e^{0.58520} \approx 1.796$, indicating that the odds are about 79.6% higher on these days compared to regular weekdays.
- **Long Weekend Weekend (`long_wknd_indWeekend, 1.14434`)**: Thisvcoefficient is even more substantial, with an odds ratio of $e^{1.14434} \approx 3.140$, suggesting that the odds of longer trips on weekends that are part of a long weekend are more than three times higher compared to a non-long weekend day.
- **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 12867, 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 (can change the sub categories)

- 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 long 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.

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. As we can see below in the reduced model, the coefficients and significance levels do not change much when removing membership which suggests no multicollinearity in the data. The Anova further confirms the benefit of membership in predicting stations with average trip durations greater than 15 minutes, by rejecting at a 5% significance level the null hypothesis that that the performance between the full and nested model is the same.

```
glm_membership_reduced <- glm(avg_15_ind ~ holiday + long_wknd_ind, family = binomial, data = df_main)
summary(glm_membership_reduced)
```

```
##
## Call:
## glm(formula = avg_15_ind ~ holiday + long_wknd_ind, family = binomial,
##     data = df_main)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -1.1863     0.2803  -4.232 2.32e-05 ***
## holiday1               0.9939     0.2257   4.404 1.06e-05 ***
## long_wknd_indWeekday   0.6079     0.2792   2.177   0.0295 *
## long_wknd_indWeekend   1.1580     0.2828   4.095 4.22e-05 ***
## ---
## 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: 13306  on 9996  degrees of freedom
## AIC: 13314
##
## Number of Fisher Scoring iterations: 4
```

```
anova(glm_membership_reduced,glm_membership,test="LRT")
```

```
## Analysis of Deviance Table
```

```
##
## Model 1: avg_15_ind ~ holiday + long_wknd_ind
## Model 2: avg_15_ind ~ mem + holiday + long_wknd_ind
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      9996      13306
## 2      9995      12857  1   449.67 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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)
##
## 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 ($\theta = 0.9902$) suggests that the variance is slightly less than the mean, which 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.

```
# Poisson model
glm_seasonal_weather_full <- glm(n_tot ~ mem + holiday + temp + rain + season, family = poisson, data = df
dispersiontest(glm_seasonal_weather)
```

```
##
##  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
```

```r
# Negative Binomial model
glm_seasonal_weather_2_full <- glm.nb(n_tot ~ mem + holiday + temp + rain + season, data = df_main)
summary(glm_seasonal_weather_2)
```

```
##
## Call:
## glm.nb(formula = n_tot ~ temp + rain + season, data = df_main,
##     init.theta = 0.9901851844, link = log)
##
## 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
```

```r
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     125397  2    79144 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(glm_seasonal_weather_2_full,glm_seasonal_weather_2,test="LRT")
```

```
## Warning in anova.negbin(glm_seasonal_weather_2_full, glm_seasonal_weather_2, :
## only Chi-squared LR tests are implemented

## Likelihood ratio tests of Negative Binomial Models
##
## Response: n_tot
##                                   Model     theta Resid. df    2 x log-lik.
## 1                 temp + rain + season 0.9901852      9995       -79906.63
## 2 mem + holiday + temp + rain + season 1.6643316      9993       -74385.52
##     Test    df LR stat. Pr(Chi)
## 1
## 2 1 vs 2     2 5521.107       0
```

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

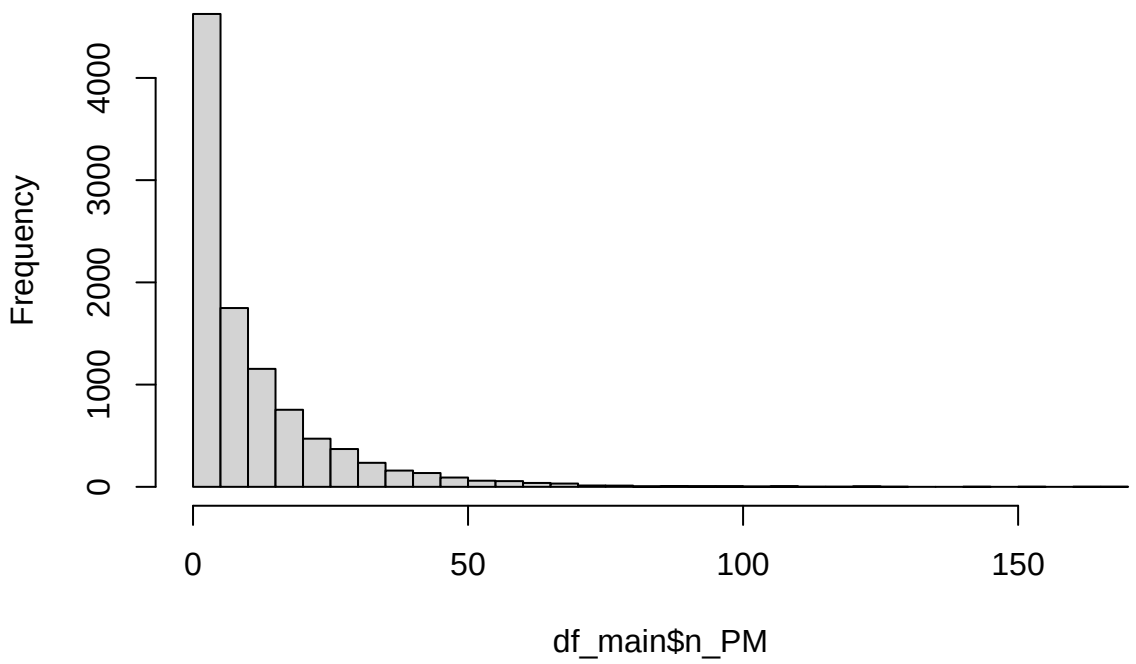
Correlation:

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
```

```
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)
##
## 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
```

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

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

```
# 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.

```
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)
##
## 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.

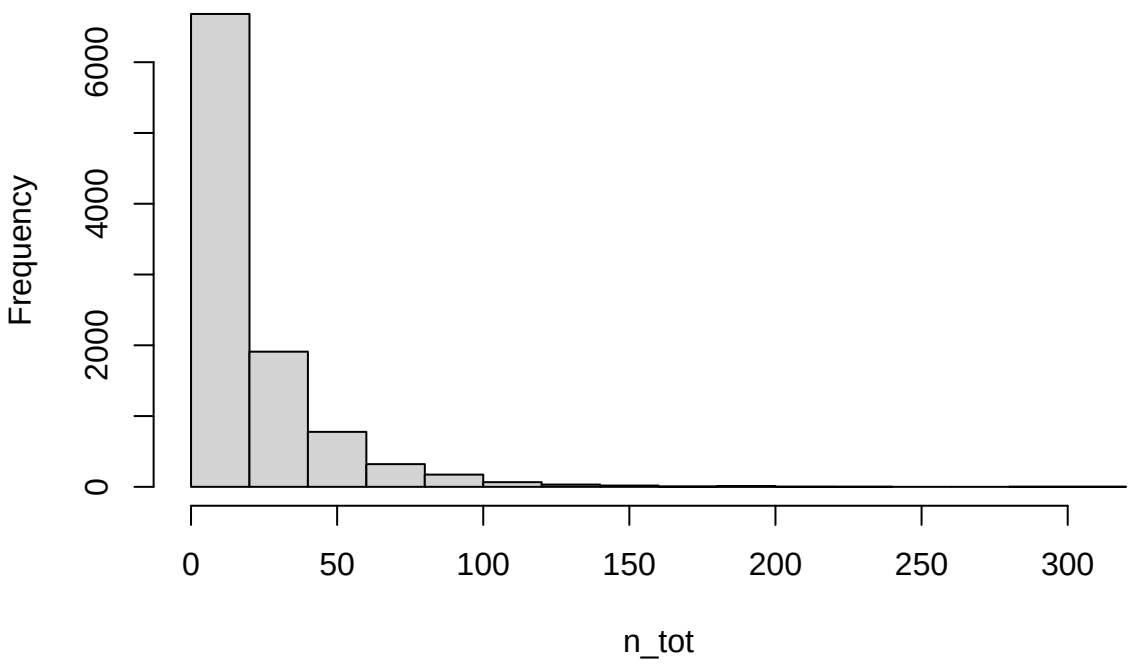
Research Question 4: Are there significant differences in bike trips counts between weekdays and weekends?

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)
##
## 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.

Business Question 5:

Limitations and shortcomings

Autocorrelation of data: the observations are not independant, as seens in our previous analysis.

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.

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 :

Atul Sharma :