# 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

# Business/Research questions

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

# 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.

## 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.

## Research Question 3: What variables . . . ?

**Objective of Analysis:**

### Variables Selection

**Correlation:**

Let's take a quick look at the correlation between our numerical variables to estimate the effect of collinearity.

```
##                    avg        temp        rain        n_tot   percent_AM
## avg         1.00000000  0.09639054 -0.10619900 -0.215866274 -0.107387372
## temp        0.09639054  1.00000000 -0.02794911  0.139997362 -0.078110564
## rain       -0.10619900 -0.02794911  1.00000000 -0.054717667  0.013211523
## n_tot      -0.21586627  0.13999736 -0.05471767  1.000000000 -0.008953075
## percent_AM -0.10738737 -0.07811056  0.01321152 -0.008953075  1.000000000
```

### Model

### Interpretation

**Overall Model** R squared, F-stat

**Intercept** :

### Business Implications:

1. **Promotion and Marketing**:
2. **Resource Allocation**:
3. **Pricing Strategy**:

## Limitations and shortcomings

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

## Conclusion

## Contribution

Charles Julien :

Gabriel Jobert :

Chike Odenigbo :

Atul Sharma :