Part 4: Linear Mixed Models

Charles Julien, Chike Odenigbo, Atul Sharma, Gabriel Jobert

10/20/2023

Contents

Introduction

Enhancing Urban Mobility Through Advanced Analytics: Unraveling Patterns in BIXI Data

BIXI, the public cycling service, has emerged as a pivotal player in urban transportation, offering an accessible and eco-friendly mode of transportation that has reshaped urban mobility. Our commitment to understanding and improving urban transportation systems has led our consultant team to conduct an extensive analysis of BIXI's operational data.

This report builds upon our previous exploration of BIXI's data, focusing on the application of linear mixed models to uncover nuanced insights. Our objective is to provide a comprehensive analysis of factors influencing BIXI's performance, extending our investigation to three specific research questions (RQs). These RQs delve into the impact of meteorological conditions, temporal patterns, and user classifications on BIXI's revenue generation.

In this journey, we leverage advanced statistical techniques, particularly linear mixed models, to unravel complex relationships within the dataset. R, a powerful statistical tool, serves as our primary instrument for data analysis and modeling. Our findings aim to not only deepen the understanding of BIXI's dynamics but also provide actionable insights for enhancing operational efficiency.

The central RQs explored in this report include the assessment of the seasonal impact on revenue, understanding the temporal patterns affecting trip duration, and examining the influence of user classifications on BIXI's performance. By addressing these questions, we aim to contribute valuable insights that can inform strategic decision-making for BIXI and serve as a reference for urban planners, researchers, and policymakers committed to creating sustainable and enjoyable urban environments.

The subsequent sections of this report will delve into the methodologies employed, share the findings derived from our analysis, and offer recommendations to support BIXI in continually improving its services.

Business/Research questions

- Research Question 1: How do seasonal factors impact trip revenue for BIXI Montréal?
- Research Question 2: How do daily and weekly patterns impact trip durations for BIXI Montréal?
- Research Question 3: What variables impact the average bixi trip duration?

Before jumping in, let's perform a quick exploration of our data.

```
df_explore = df_main %>%
  group_by(station, mem) %>%
  summarize(n = n())
```

```
## ‘summarise()‘ has grouped output by ’station’. You can override using the
## ‘.groups‘ argument.
```

```
summary(df_explore$n)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   4.000   6.000   6.545   9.000  17.000
```

The unique identifier of a line in our dataset is a combination of the station, the date and the membership status. On average, a station for a given membership status appears 6 times in our dataset.

Research Question 1: How do seasonal factors impact trip revenue for BIXI Montréal?

ADD INTERACTION TERM **Objective of Analysis:** This regression model is examining the impact of the month (`mm`), average daily temperature (`temp`), and total amount of rainfall (`rain`) and membershipt (`mem`) on the revenue (`rev`) generated by trips leaving from a specified station.

Model Linear regression

```
##
## Call:
## lm(formula = rev ~ mm + temp + rain + mem, data = df_main)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6946  -1388   -475    838  47246
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2091.216    172.311 -12.136  < 2e-16 ***
## mm5          1337.321    172.504   7.752 9.90e-15 ***
## mm6          1557.546    197.900   7.870 3.90e-15 ***
## mm7          1593.946    191.697   8.315  < 2e-16 ***
## mm8          1391.546    209.516   6.642 3.26e-11 ***
## mm9          1894.608    177.695  10.662  < 2e-16 ***
## mm10          816.977    162.684   5.022 5.21e-07 ***
## mm11          516.539    192.773   2.680  0.00738 **
## temp           57.589      9.949   5.789 7.31e-09 ***
## rain          -72.689      7.029 -10.341  < 2e-16 ***
## mem1         6134.595     72.237  84.923  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3599 on 9989 degrees of freedom
## Multiple R-squared:  0.4344, Adjusted R-squared:  0.4339
## F-statistic: 767.3 on 10 and 9989 DF,  p-value: < 2.2e-16
```

Model Linear Mixed model

Comment: compound symmetric correlation structure is not ideal for time series if I am not mistaken

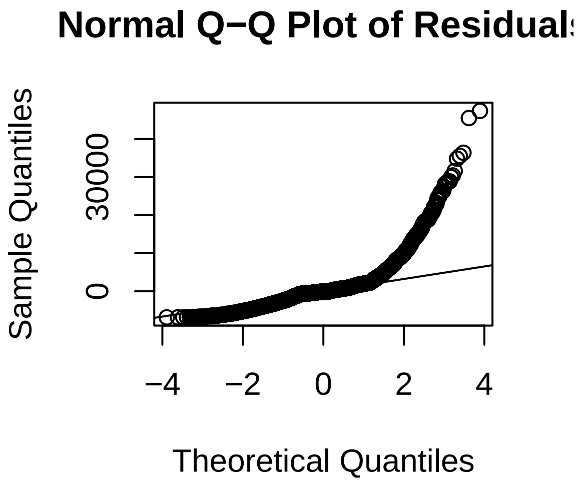
```
seasonal_effect_rev_gls <- gls(rev ~ temp + rain + mm + mem, correlation = corCompSymm(form = ~ 1 | statio

# Display model summary
summary(seasonal_effect_rev_gls)
```

```
## Generalized least squares fit by REML
##   Model: rev ~ temp + rain + mm + mem
##   Data: df_main
##        AIC      BIC    logLik
##   189095.7 189189.4 -94534.86
##
## Correlation Structure: Compound symmetry
##  Formula: ~1 | station
##  Parameter estimate(s):
##       Rho
## 0.3714602
##
## Coefficients:
##                 Value Std.Error   t-value p-value
## (Intercept) -2637.297 164.09836 -16.07144   0e+00
## temp           63.102   8.20592   7.68982   0e+00
## rain          -83.319   5.79387 -14.38050   0e+00
## mm5          1241.513 142.33492   8.72248   0e+00
## mm6          1576.754 163.95963   9.61672   0e+00
## mm7          1663.683 158.67860  10.48461   0e+00
## mm8          1488.772 173.79641   8.56619   0e+00
## mm9          2116.005 147.74396  14.32211   0e+00
## mm10         1112.919 136.12558   8.17568   0e+00
## mm11          648.125 159.95436   4.05194   1e-04
## mem1         6326.879  59.77901 105.83779   0e+00
##
##  Correlation:
##      (Intr) temp   rain   mm5    mm6    mm7    mm8    mm9    mm10   mm11
## temp -0.510
## rain -0.069 -0.006
## mm5  -0.382 -0.271  0.089
## mm6  -0.166 -0.557 -0.026  0.659
## mm7  -0.196 -0.530  0.019  0.671  0.757
## mm8  -0.108 -0.634  0.044  0.655  0.774  0.774
## mm9  -0.300 -0.402 -0.008  0.672  0.721  0.726  0.726
## mm10 -0.486 -0.130 -0.028  0.646  0.612  0.625  0.595  0.660
## mm11 -0.584  0.232 -0.002  0.457  0.326  0.348  0.285  0.417  0.530
## mem1 -0.233  0.012 -0.040  0.025  0.035  0.027  0.023  0.039  0.050  0.029
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -1.8926963 -0.3233412 -0.0298840  0.2913426 13.1127739
##
## Residual standard error: 3613.719
## Degrees of freedom: 10000 total; 9989 residual
```

```
# CHIKE: Note that using compound symmetric gives the same covariance to each observation
cov.matrix = getVarCov(seasonal_effect_rev_gls,individual = 4)
cov2cor(cov.matrix)
```

```
## Marginal variance covariance matrix
##         [,1]    [,2]    [,3]    [,4]    [,5]    [,6]    [,7]
## [1,] 1.00000 0.37146 0.37146 0.37146 0.37146 0.37146 0.37146
## [2,] 0.37146 1.00000 0.37146 0.37146 0.37146 0.37146 0.37146
## [3,] 0.37146 0.37146 1.00000 0.37146 0.37146 0.37146 0.37146
## [4,] 0.37146 0.37146 0.37146 1.00000 0.37146 0.37146 0.37146
## [5,] 0.37146 0.37146 0.37146 0.37146 1.00000 0.37146 0.37146
## [6,] 0.37146 0.37146 0.37146 0.37146 0.37146 1.00000 0.37146
## [7,] 0.37146 0.37146 0.37146 0.37146 0.37146 0.37146 1.00000
##   Standard Deviations: 1 1 1 1 1 1 1
```

```
##                   Value  Std.Error    t-value      p-value
## (Intercept) -2637.29741 164.098362 -16.071443 2.115404e-57
## temp           63.10207   8.205925   7.689818 1.612369e-14
## rain          -83.31878   5.793874 -14.380496 1.991983e-46
## mm5          1241.51337 142.334924   8.722479 3.155636e-18
## mm6          1576.75420 163.959627   9.616722 8.446031e-22
## mm7          1663.68338 158.678598  10.484611 1.380125e-25
## mm8          1488.77225 173.796405   8.566186 1.227668e-17
## mm9          2116.00461 147.743955  14.322106 4.545430e-46
## mm10         1112.91857 136.125584   8.175675 3.299495e-16
## mm11          648.12506 159.954359   4.051937 5.117986e-05
## mem1         6326.87864  59.779011 105.837793 0.000000e+00
```



##Strength of the model

The Normal Q-Q plot provides valuable insights into the distribution of residuals in the BIXI data model. The near-perfect alignment of residuals with the reference line from -4 to +2 suggests that a substantial portion of the residuals adheres to a normal distribution. However, the major upward deviation observed from +2 to +4 indicates the presence of extreme positive residuals that do not align with the expected normal distribution. This discrepancy highlights a limitation in the model, signaling the existence of outliers or influential observations that could significantly impact the model's accuracy. These outliers may be indicative of unaccounted-for factors or unexpected events that contribute to revenue variations beyond the model's current specifications. Addressing this limitation may involve further exploration of the data to identify the sources of these extreme residuals, potential model refinements, or the consideration of alternative modeling approaches to better capture the underlying patterns in the BIXI revenue data.

Interpretation

Intercept (14.27): The expected revenue in April is, on average, 14.27 $ when temperature (temp) is zero, there is not rain). It is difficult to interpret at practically, temperature would not be zero in April.

Temperature (0.15): A one-unit increase in temperature is associated with a 0.15 $ increase, on average, in revenue, keeping other variables constant. Higher temperatures are positively correlated with increased revenue.

Rainfall (-0.22): Rainfall leads a 0.22 unit decrease in revenue keeping other variables constant. Higher rainfall is negatively correlated with revenue, suggesting potential negative effects on Bixi usage.

May (mm5 - 2.44): Revenue is expected to increase by 2.44 $, on average, in May compared to April (reference month) keeping other variables constant. May is associated with higher revenue compared to April.

June (mm6 - 6.99): Revenue is expected to increase by 6.99 $, on average, in June compared to April keeping other variables constant. June has a significant positive impact on revenue.

July (mm7 - 13.73): Revenue is expected to increase by 13.73 $, on average, in July compared to April keeping other variables constant. Interpretation: July has the most significant positive impact on revenue among the months.

August (mm8 - 14.69): Revenue is expected to increase by 14.69 $, on average, in August compared to April keeping other variables constant. Interpretation: August has a substantial positive impact on revenue.

September (mm9 - 15.82): Revenue is expected to increase by 15.82 $, on average, in September compared to April keeping other variables constant. September has substantial positive impact on revenue.

October (mm10 - 9.19): Revenue is expected to increase by 9.19 $, on average, in October compared to April keeping other variables constant. Interpretation: October has a positive impact on revenue.

November (mm11 - 3.29): Revenue is expected to increase by 3.29 $, on average, in November compared to April keeping other variables constant. November has a modest positive impact on revenue.

##Business Implications:

Temperature: Bixi can capitalize on warmer temperatures by promoting increased ridership during favorable weather conditions.

Rainfall: Strategies to mitigate the negative impact of rainfall on revenue may include targeted marketing during rainy periods or offering promotions to incentivize usage.

Seasonal Variation: Understanding the seasonal variation allows Bixi to allocate resources effectively, focusing on peak months like July, August, and September for marketing and service enhancements.

Month-specific Strategies: Tailoring marketing campaigns or promotional offers based on the impact of each month on revenue can optimize Bixi's overall financial performance.

Planning and Resource Allocation: Knowledge of specific months with higher revenue can guide resource allocation, such as increasing bike availability and marketing efforts during peak months.

Operational Adjustments: Bixi can make operational adjustments, such as increasing staff or bikes, during months with the most significant positive impact on revenue.

Verification of Assumptions

NEED A MORE COMPLETE ASSUPTION VALIDATION ### Normality of residuals

1. **Histogram of Residuals**:
2. **Normal Q-Q Plot**:

Overall Interpretation:

Research Question 1: Autoregressive Structure (CHIKE)

Linear Mixed Models (LMMs) - Model Comparisons (joshuawiley.com)

```
#library(ggplot2)
# tous / all id
#ggplot(data = df_main, aes(x = week_num, y = rev_imputed, group = station)) +
#geom_line(alpha = 0.2) + scale_x_continuous(expand = c(0,
#0), limits = c(1, 5))
```

```
seasonal_effect_rev.ar <- gls(rev ~ temp + rain + mm + mem, correlation = corAR1(form = ~ 1 | station), da

# Display model summary
#summary(seasonal_effect_rev.ar)
```

```
getVarCov(seasonal_effect_rev.ar,individual = 4)
```

```
## Marginal variance covariance matrix
##          [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]
## [1,] 12853000  5024100  1963900   767690   300090   117300    45854
## [2,]  5024100 12853000  5024100  1963900   767690   300090   117300
## [3,]  1963900  5024100 12853000  5024100  1963900   767690   300090
## [4,]   767690  1963900  5024100 12853000  5024100  1963900   767690
## [5,]   300090   767690  1963900  5024100 12853000  5024100  1963900
## [6,]   117300   300090   767690  1963900  5024100 12853000  5024100
## [7,]    45854   117300   300090   767690  1963900  5024100 12853000
##   Standard Deviations: 3585.1 3585.1 3585.1 3585.1 3585.1 3585.1 3585.1
```

As expected with an auto-regressive correlation structure, the further back the time period, the lower the correlation.

It is important to note that the time distance between these observations is not constant, in some cases, it can even be null. For example, when a station is observed on the same day for members and non-members. Having a covariance that considers the time between observation would be the best. (ARH1)

Research Question 1: Random Intercept (CHIKE)

```r
#df_main
seasonal_effect_rev.rand_int <- lme(rev ~ temp + rain + mm + mem, random = ~ 1 | station, data = df_main)

# Display model summary
#summary(seasonal_effect_rev.rand_int)
```

```r
# Expected to be True comparing fixed effects coefficients between base model and random effects model
isTRUE(all.equal(coef(seasonal_effect_rev_model), fixef(seasonal_effect_rev.rand_int)))
```

```
## [1] FALSE
```

```r
getVarCov(seasonal_effect_rev.rand_int, type = "random.effects")
```

```
## Random effects variance covariance matrix
##             (Intercept)
## (Intercept)     4850900
##   Standard Deviations: 2202.5
```

```r
seasonal_effect_rev.ar_rand_int <- lme(rev ~ temp + rain + mm + mem, random = ~1 |
station, correlation = corAR1(form = ~1 | station), data = df_main)
#summary(seasonal_effect_rev.ar_rand_int)
```

```r
getVarCov(seasonal_effect_rev.ar_rand_int, type = "random.effects")
```

```
## Random effects variance covariance matrix
##             (Intercept)
## (Intercept)     4742600
##   Standard Deviations: 2177.7
```

```r
AIC(seasonal_effect_rev.ar, seasonal_effect_rev.rand_int, seasonal_effect_rev.ar_rand_int )
```

```
##                                 df      AIC
## seasonal_effect_rev.ar          13 190445.2
## seasonal_effect_rev.rand_int    13 189095.7
## seasonal_effect_rev.ar_rand_int 14 189076.3
```

Research Question 1: Random Slope (CHIKE)

```r
#Doesnt make sense in this context to use random effects on the coefficient for #temperature because tempe
#Using rain random effect does not work
seasonal_effect_rev.rand_coef <- lme(rev ~ temp + rain + mm + mem, random = ~1 + temp |
station, data = df_main)
#summary(seasonal_effect_rev.rand_coef)
```

```r
summary_season_rev.base = tidy(seasonal_effect_rev_model)
colnames(summary_season_rev.base) <- c('Covariates','Value','Std.Error','t-value','p-value')
summary_season_rev.base$type = 'Base'
#summary_season_rev.base
```

Using a 5% significance threshold, we can conclude that each of the covariates used to predict revenue had a significant impact.

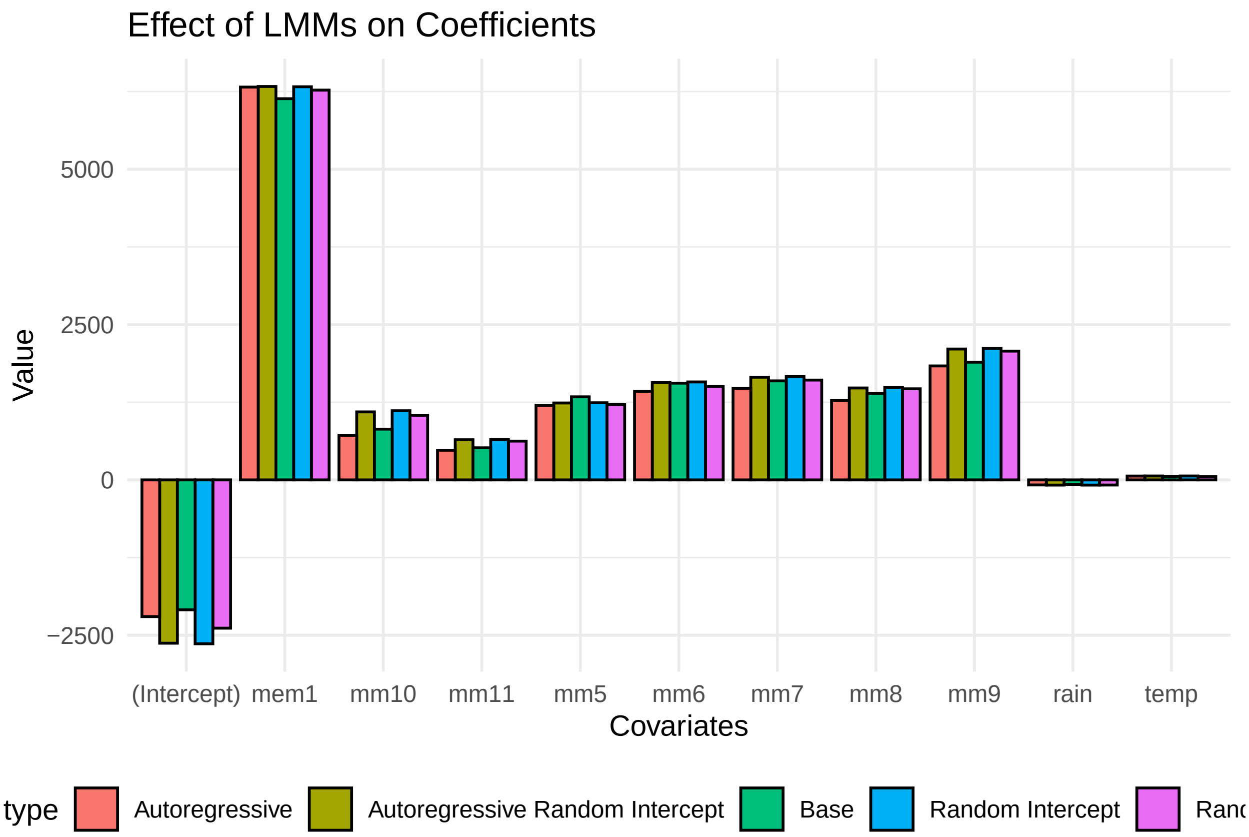
```
knitr::kable(season_summary_combined %>%
  group_by(Covariates) %>%
  summarise(significant = sum(significance == "Significant Feature"),
            not_significant = sum(significance == "Not Significant")),caption = 'Feature Significance (5%)
```

Table 1: Feature Significance (5%)

Covariates	significant	not_significant
(Intercept)	5	0
mem1	5	0
mm10	5	0
mm11	5	0
mm5	5	0
mm6	5	0
mm7	5	0
mm8	5	0
mm9	5	0
rain	5	0
temp	5	0

The Model estimates are similar despite different correlation structures. In this sense, we can see that the size and direction of each covariate is similar across each model.

```
ggplot(season_summary_combined, aes(fill=type, y=Value, x=Covariates)) +
  geom_bar(colour="black",position='dodge', stat='identity')  + ggtitle ("Effect of LMMs on Coefficients")
```



###Interpretation * Intercept: This is the average revenue when all values are set at 0. In our case, it would be that for non members, in the month of April, with no rain and temperature at 0 degrees, the expected revenue is roughly -($2,500) across all the models. This number is unrealistic as Bixi revenue is a strictly positive number. It would have been more interpretable if revenue was allowed to be a negative number by accounting for cost.

- mem1: Members contribute roughly $6,000 in additional revenue for a given station compared to non-members holding other variables constant.
- mm5: Rides in the month of May contribute about $1,200 in additional revenue for a given station compared to the month of April holding all other variables constant.
- mm6: Rides in the month of June contribute about $1,250 in additional revenue for a given station compared to the month of April holding all other variables constant.
- mm7: Rides in the month of July contribute about $1,280 in additional revenue for a given station compared to the month of April holding all other variables constant.
- mm8: Rides in the month of August contribute about $1,250 in additional revenue for a given station compared to the month of April holding all other variables constant.
- mm9: Rides in the month of May contribute about $2,200 in additional revenue for a given station compared to the month of April holding all other variables constant.
- mm10: Rides in the month of May contribute about $1,000 in additional revenue for a given station compared to the month of April holding all other variables constant.
- mm11: Rides in the month of May contribute about $700 in additional revenue for a given station compared to the month of April holding all other variables constant.
- rain: A 1 unit increase in rain contributes to a $100 decrease in revenue for a given station holding all other variables constant.
- temp: A 1 unit increase in temperature contributes to a $100 increase in revenue for a given station holding all other variables constant.

Using AIC and BIC metrics, the model incorporating a random slope as well as the model incorporating an autoregressive correlation structure with a random intercept perform best when predicting revenue using season, membership and period data. This was assessed by the fact that they both have the lowest BIC and AIC metrics of all the models considered. It is also worth noting that the basic linear model that does not account for autocorrelation in the data fits the data the least optimally.

```
performance.df <- data.frame("Type"=numeric(),"AIC"=numeric(),"BIC"=numeric(),"LL"=numeric())
performance.df[nrow(performance.df) + 1,] = c("Base",AIC(seasonal_effect_rev_model),BIC(seasonal_effect_re

performance.df[nrow(performance.df) + 1,] = c("Autoregressive",AIC(seasonal_effect_rev.ar),BIC(seasonal_ef

performance.df[nrow(performance.df) + 1,] = c("Autoregressive Random Intercept",AIC(seasonal_effect_rev.ar

performance.df[nrow(performance.df) + 1,] = c("Random Intercept",AIC(seasonal_effect_rev.rand_int),BIC(sea

performance.df[nrow(performance.df) + 1,] = c("Random Slope",AIC(seasonal_effect_rev.rand_coef),BIC(seasona

knitr::kable(performance.df %>% arrange(BIC),caption='Model Performance')
```

Table 2: Model Performance

Type	AIC	BIC	LL
Random Slope	188783.341277982	188891.47987448	-94376.6706389911
Autoregressive Random Intercept	189076.260516278	189177.189873009	-94524.130258139
Random Intercept	189095.718549919	189189.438666884	-94534.8592749593
Autoregressive	190445.22721716	190538.947334125	-95209.61360858
Base	192159.782535936	192246.3066204	-96067.891267968

```
anova(seasonal_effect_rev.rand_coef,seasonal_effect_rev.rand_int)
```

```
##                               Model df      AIC      BIC    logLik   Test
## seasonal_effect_rev.rand_coef     1 15 188783.3 188891.5 -94376.67
## seasonal_effect_rev.rand_int      2 13 189095.7 189189.4 -94534.86 1 vs 2
##                                L.Ratio p-value
## seasonal_effect_rev.rand_coef
## seasonal_effect_rev.rand_int  316.3773  <.0001
```

```
# Testing to see if anova command produces same result for LRT test
#D <- -2 * (seasonal_effect_rev.rand_int$logLik - seasonal_effect_rev.rand_coef$logLik)
#print(D)
#pchisq(D, df = 13, lower.tail = FALSE)/2
```

We further performed likelihood ratio tests between each of the linear mixed models and the base linear model in order to assess whether the full model fits the data significantly better than the nested model. In our analysis, the nested model consisted simply of the linear model and the full model accounted for the addition of parameters relating to the correlation structure of random effects. In each case, we concluded that the full model performed significantly better using a 1% significance threshold. This effectively means that making changes to the model structure by accounting for autocorrelation leads to a significant improvement in fit relative to a linear model.

```
anova(seasonal_effect_rev.rand_int,seasonal_effect_rev_model)
```

```
##                              Model df      AIC      BIC    logLik   Test
## seasonal_effect_rev.rand_int     1 13 189095.7 189189.4 -94534.86
## seasonal_effect_rev_model        2 12 192048.3 192134.8 -96012.13 1 vs 2
##                               L.Ratio p-value
## seasonal_effect_rev.rand_int
## seasonal_effect_rev_model    2954.539  <.0001
```

```
anova(seasonal_effect_rev.rand_coef,seasonal_effect_rev_model)
```

```
##                               Model df      AIC      BIC    logLik   Test
## seasonal_effect_rev.rand_coef     1 15 188783.3 188891.5 -94376.67
## seasonal_effect_rev_model         2 12 192048.3 192134.8 -96012.13 1 vs 2
##                                L.Ratio p-value
## seasonal_effect_rev.rand_coef
## seasonal_effect_rev_model     3270.916  <.0001
```

```
anova(seasonal_effect_rev.ar_rand_int,seasonal_effect_rev_model)
```

```
##                                 Model df      AIC      BIC    logLik   Test
## seasonal_effect_rev.ar_rand_int     1 14 189076.3 189177.2 -94524.13
## seasonal_effect_rev_model           2 12 192048.3 192134.8 -96012.13 1 vs 2
##                                  L.Ratio p-value
## seasonal_effect_rev.ar_rand_int
## seasonal_effect_rev_model       2975.997  <.0001
```

```
anova(seasonal_effect_rev.ar,seasonal_effect_rev_model)
```

```
##                           Model df      AIC      BIC    logLik   Test L.Ratio
## seasonal_effect_rev.ar        1 13 190445.2 190539.0 -95209.61
## seasonal_effect_rev_model     2 12 192048.3 192134.8 -96012.13 1 vs 2 1605.03
##                           p-value
## seasonal_effect_rev.ar
## seasonal_effect_rev_model  <.0001
```

```
getVarCov(seasonal_effect_rev.rand_coef, type = "random.effects")
```

```
## Random effects variance covariance matrix
##              (Intercept)  temp
## (Intercept)      485560 59384
## temp              59384  8691
##   Standard Deviations: 696.82 93.226
```

Using information criterion and likelihood ratio tests, we compared the 4 linear mixed models together.

```
anova(seasonal_effect_rev.ar, seasonal_effect_rev.ar_rand_int,seasonal_effect_rev.rand_coef,seasonal_effec
```

```
##                                  Model df      AIC      BIC    logLik   Test
## seasonal_effect_rev.ar               1 13 190445.2 190539.0 -95209.61
## seasonal_effect_rev.ar_rand_int      2 14 189076.3 189177.2 -94524.13 1 vs 2
## seasonal_effect_rev.rand_coef        3 15 188783.3 188891.5 -94376.67 2 vs 3
## seasonal_effect_rev_gls              4 13 189095.7 189189.4 -94534.86 3 vs 4
##                                    L.Ratio p-value
## seasonal_effect_rev.ar
## seasonal_effect_rev.ar_rand_int 1370.9667  <.0001
## seasonal_effect_rev.rand_coef    294.9192  <.0001
## seasonal_effect_rev_gls          316.3773  <.0001
```

```
AIC(seasonal_effect_rev_model)
```

```
## [1] 192159.8
```

```
#anova(seasonal_effect_rev_model,seasonal_effect_rev.ar)
```

```
anova(seasonal_effect_rev.rand_coef,seasonal_effect_rev_gls)
```

```
##                                Model df      AIC      BIC    logLik   Test
## seasonal_effect_rev.rand_coef     1 15 188783.3 188891.5 -94376.67
## seasonal_effect_rev_gls           2 13 189095.7 189189.4 -94534.86 1 vs 2
##                                 L.Ratio p-value
## seasonal_effect_rev.rand_coef
## seasonal_effect_rev_gls       316.3773  <.0001
```

```
lrtest(seasonal_effect_rev.ar,seasonal_effect_rev.rand_coef)
```

```
## Warning in modelUpdate(objects[[i - 1]], objects[[i]]): original model was of
## class "gls", updated model is of class "lme"
```

```
## Likelihood ratio test
##
## Model 1: rev ~ temp + rain + mm + mem
## Model 2: rev ~ temp + rain + mm + mem
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1  13 -95210
## 2  15 -94377  2 1665.9  < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
lrtest(seasonal_effect_rev_gls,seasonal_effect_rev.rand_coef)
```

```
## Warning in modelUpdate(objects[[i - 1]], objects[[i]]): original model was of
## class "gls", updated model is of class "lme"
```

```
## Likelihood ratio test
##
## Model 1: rev ~ temp + rain + mm + mem
## Model 2: rev ~ temp + rain + mm + mem
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1  13 -94535
## 2  15 -94377  2 316.38  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(seasonal_effect_rev_gls,seasonal_effect_rev.ar_rand_int)
```

```
## Warning in modelUpdate(objects[[i - 1]], objects[[i]]): original model was of
## class "gls", updated model is of class "lme"
```

```
## Likelihood ratio test
##
## Model 1: rev ~ temp + rain + mm + mem
## Model 2: rev ~ temp + rain + mm + mem
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1  13 -94535
## 2  14 -94524  1 21.458  3.617e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(seasonal_effect_rev.ar)
```

```
## [1] 190445.2
```

```
nobs(seasonal_effect_rev_model)
```

```
## [1] 10000
```

```
nobs(seasonal_effect_rev.ar_rand_int)
```

```
## [1] 10000
```

```
#widyr::pairwise_cor(df_main, station, rain, c)
#df_main %>% select(station:rain) %>% modelsummary::datasummary_correlation()
```

Research Question 2: How do daily and weekly patterns impact trip durations for BIXI Montréal?

Objective of Analysis: This regression model is examining the impact of the day of the month (`dd`), day of the week (`wday`), and holidays (`holiday`) on the revenue (`rev`) generated by trips leaving from a specified station.

Model

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: dur ~ holiday + wknd_ind + wknd_ind * mem + (1 | district/station)
##    Data: df_main
##
## REML criterion at convergence: 136021.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.3968 -0.4977 -0.0481  0.3757 15.6285
##
## Random effects:
##  Groups           Name        Variance Std.Dev.
##  station:district (Intercept) 23795    154.3
##  district         (Intercept) 12617    112.3
##  Residual                     40235    200.6
## Number of obs: 10000, groups:  station:district, 793; district, 21
##
## Fixed effects:
##                      Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)           -34.453     26.802   20.666  -1.285 0.212854
## holiday1               51.945     13.936 9433.782   3.727 0.000195 ***
## wknd_indWeekend        67.896      6.607 9419.592  10.276  < 2e-16 ***
## mem1                  313.065      4.949 9436.954  63.261  < 2e-16 ***
## wknd_indWeekend:mem1  -68.763      9.137 9429.454  -7.526 5.73e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) holdy1 wknd_W mem1
## holiday1    -0.019
## wknd_ndWknd -0.076  0.071
## mem1        -0.110  0.011  0.396
## wknd_ndWk:1  0.054 -0.006 -0.718 -0.536
```

The model explores the relationship between trip duration (`dur`) and factors like holidays (`holiday`), weekend indicator (`wknd_ind`), membership status (`mem`) and its interaction with weekend indicator, considering the nested structure of stations within districts.

Random Effects - Station:District Variability: The significant variance in the random intercepts for stations within districts (Variance = 23,795, Std. Dev. = 154.3) suggests considerable differences in baseline trip durations across stations, depending on their district. - **District-Level Variability**: There is also notable variability between districts (Variance = 12,617, Std. Dev. = 112.3), indicating that the district a station belongs to influences trip duration. - These results highlight the importance of accounting for the hierarchical structure of the data (stations nested within districts).

Fixed Effects (Significance & Interpretation) - Intercept: The negative intercept (-34.453) may not be meaningful by itself, as it represents the expected trip duration when all other variables are at their reference levels. In this context, an intercept of -34.453 would mean that when it's a non-holiday weekday, and the rider is not a member, the model predicts a trip duration of -34.453 units. Since negative trip duration is not possible, this result might initially seem nonsensical. - **Holiday (significant)**: On average, holding other variables constant, total trip durations on holidays are 51.945 minutes longer compared to non-holidays. This reflects a tendency for longer trips during holidays. This effect is statistically significant ($p < 0.001$). - **Weekend Indicator (significant)**: On average, with other factors held constant, total trip durations on weekends are 67.896 minutes longer than on weekdays. This indicates a preference or tendency for longer trips during weekends. This is highly significant ($p < 0.001$). - **Membership Status (significant)**: Holding other variables at their reference levels, on average, members have a total trip durations that are 313.065 minutes longer compared to non-members. This might indicate different usage patterns, such as members taking longer trips., a highly significant effect ($p < 0.001$). - **Interaction: Weekend and Membership (significant)**: On average, and with other variables held constant, the interaction effect suggests that the increased total trip duration associated with membership is reduced by 68.763 minutes on weekends. This indicates that the distinction in trip duration between members and non-members is less pronounced on weekends. This is also statistically significant (p < 0.001).

Correlations of Fixed Effects - The correlation matrix shows the relationships between the different fixed effects in the model. High correlations can indicate potential multicollinearity issues, which might affect the interpretation of coefficients. However, in the model, these correlations seem relatively moderate.

Overall Interpretation - The model indicates that both the day of the week (weekend vs. weekday) and membership status significantly impact trip durations, with an interesting interaction effect on weekends for members. - The significant random effects imply that both the specific station and the district it's in are important factors influencing trip durations. - The model appears to be a good fit for the data, capturing key variability both within and between groups (stations and districts).

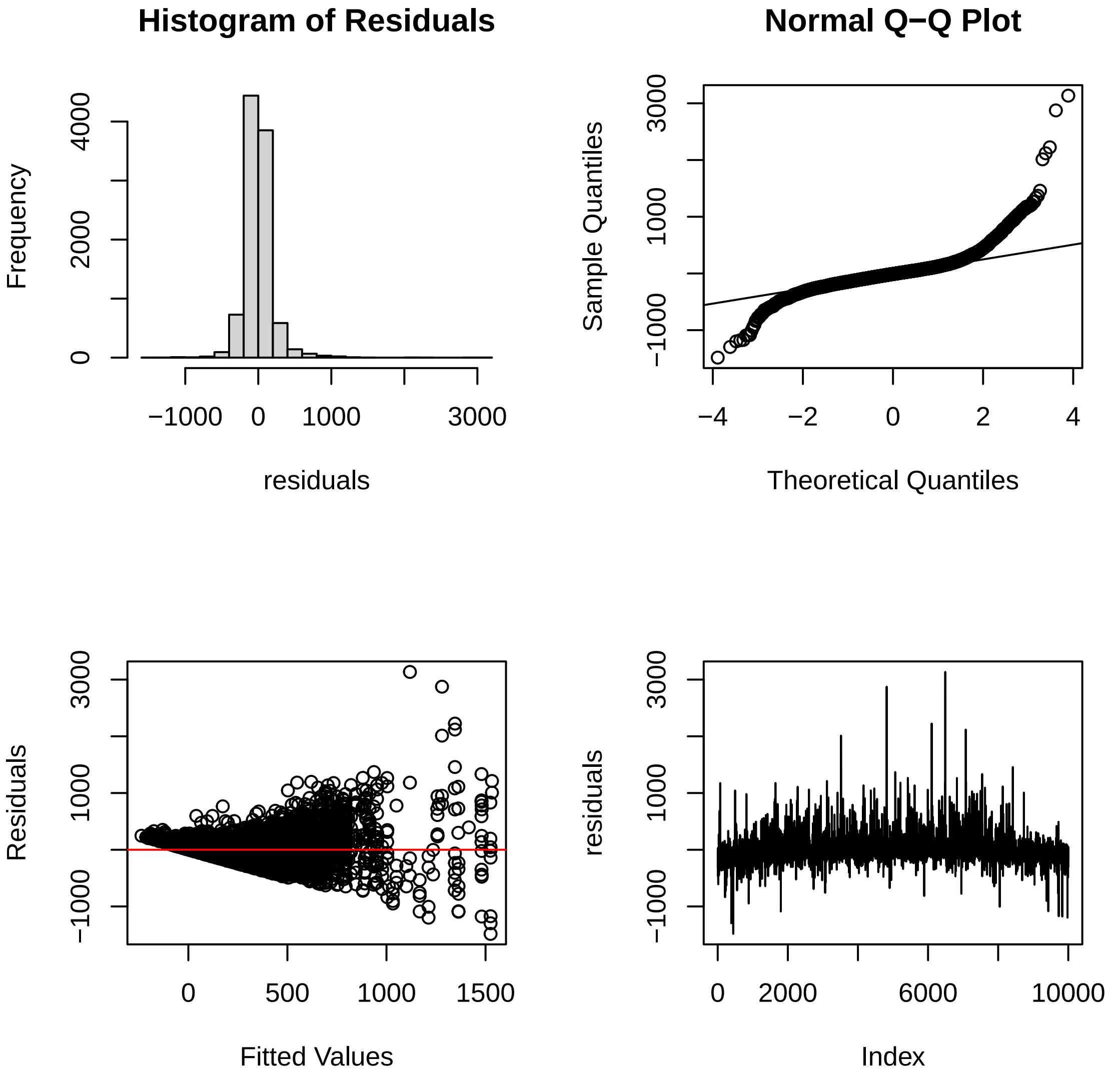
```
null_model <- lmer(dur ~ 1 + (1 | district/station), data = df_main)

#Compare full model to null model (refitting using mle)
anova(time_pattern_dur_model_mixed, null_model)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: df_main
## Models:
## null_model: dur ~ 1 + (1 | district/station)
## time_pattern_dur_model_mixed: dur ~ holiday + wknd_ind + wknd_ind * mem + (1 | district/station)
##                              npar    AIC    BIC logLik deviance  Chisq Df
## null_model                      4 140073 140102 -70032   140065
## time_pattern_dur_model_mixed    8 136069 136127 -68027   136053 4011.6  4
##                              Pr(>Chisq)
## null_model
## time_pattern_dur_model_mixed  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The significant Chi-square test (p < 0.001) suggests that the fixed effects included in the full model (related to holidays, weekends, and membership status) contribute meaningfully to explaining the variability in trip durations.
- The lower AIC and BIC values for the full model compared to the null model further support that the full model provides a better fit to the data.
- This analysis strongly indicates that the factors of holidays, weekends, and membership status, along with their interactions, are important predictors of trip duration in the context of the bike-sharing data.

Assumptions



Histogram of Residuals The histogram shows the distribution of residuals. It suggests that the residuals are fairly symmetrically distributed around zero, indicating that the assumption of normality might be reasonably met. However, the distribution appears slightly leptokurtic (having a peak higher than a normal distribution), suggested by the tall center of the histogram.

Normal Q-Q Plot The Q-Q plot compares the quantiles of the residuals to the quantiles of a normal distribution. If the residuals were perfectly normally distributed, the points would lie on the 45-degree reference line. In the Q-Q plot, the points deviate from the line at the ends, indicating potential heavy tails in the distribution of residuals. This could suggest some departure from normality, particularly with potential outliers or extreme values.

Residuals vs Fitted Values Plot The residuals should be randomly scattered around the horizontal line at zero, with no clear pattern. In the plot, there seems to be a slight “funnel” shape, where the variance of the residuals increases with the fitted values, which could indicate heteroscedasticity.

Residuals vs Index Plot This plot displays residuals against the observation index. It’s useful for detecting patterns that may indicate violation of independence. The residuals appear randomly scattered, suggesting no obvious violation of independence. However, there are some visible outliers, which should be investigated further.

ACF Plot of Residuals The autocorrelation function (ACF) plot is used to check for autocorrelation in the residuals at different lags. The bars represent correlations at different lag values. If most of them are within the blue dashed lines (representing confidence intervals), it suggests little to no autocorrelation. The ACF plot shows that autocorrelation is not a concern as the correlations are within the bounds.

Q-Q Plot of Random Effects This plot should show whether the random effects are normally distributed. The random effects (intercepts for `district/station` in the model) should fall along the reference line if they’re normally distributed. There's some deviation from normality, but it's not extreme.

Predicted vs Actual Values Plot This plot compares the predicted values from the model to the actual values. Ideally, the points should fall around the 45-degree line, indicating good model fit. The plot shows a reasonable alignment along the line, although it seems to diverge for higher values, suggesting the model might not predict as well in that range.

Interpretation Summary The model assumptions are not strictly violated, but there are indications of potential issues:

- The residuals are roughly normally distributed but show signs of leptokurtosis.
- There might be some heteroscedasticity, as indicated by the Residuals vs Fitted Values plot.
- There are outliers in the data that could be influential points worth investigating.
- The assumption of independence seems to be met based on the Residuals vs Index and ACF plots.
- The random effects may slightly deviate from normality, but not severely.

Limitation of the model Given these observations, the following improvement could be made :

- Transforming the response variable or using robust regression techniques to handle non-normality and heteroscedasticity.
- Investigating and potentially addressing outliers.

```
time_pattern_dur.ar <- gls(dur ~ dd + wday + holiday, correlation = corAR1(form = ~ 1 | station), data = d

# Display model summary
summary(time_pattern_dur.ar)
```

```
## Generalized least squares fit by REML
##   Model: dur ~ dd + wday + holiday
##   Data: df_main
##        AIC      BIC    logLik
##   141131.6 141210.9 -70554.79
##
## Correlation Structure: AR(1)
##  Formula: ~1 | station
##  Parameter estimate(s):
##       Phi
## 0.4023917
##
## Coefficients:
##                  Value Std.Error   t-value p-value
## (Intercept)  283.03014  9.083027 31.160331  0.0000
## dd            -0.18235  0.306744 -0.594462  0.5522
## wdayMonday   -45.08756  9.932037 -4.539609  0.0000
## wdaySaturday  11.35289  9.759227  1.163298  0.2447
## wdaySunday   -13.42787  9.799037 -1.370326  0.1706
## wdayThursday -22.79589  9.790562 -2.328353  0.0199
## wdayTuesday  -36.13164  9.912338 -3.645118  0.0003
## wdayWednesday -29.46184  9.820837 -2.999931  0.0027
## holiday1      60.77485 18.146140  3.349189  0.0008
##
##  Correlation:
##              (Intr) dd     wdyMnd wdyStr wdySnd wdyThr wdyTsd wdyWdn
## dd           -0.509
## wdayMonday   -0.523 -0.027
## wdaySaturday -0.543 -0.010  0.501
## wdaySunday   -0.537 -0.010  0.496  0.508
## wdayThursday -0.525 -0.024  0.514  0.499  0.497
## wdayTuesday  -0.527 -0.018  0.494  0.502  0.494  0.494
## wdayWednesday -0.528 -0.022  0.494  0.505  0.502  0.499  0.498
## holiday1      -0.028  0.050 -0.180  0.004  0.005 -0.101  0.005  0.001
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -1.0290197 -0.6311301 -0.3020733  0.3448169 12.9772479
##
## Residual standard error: 305.1222
## Degrees of freedom: 10000 total; 9991 residual
```

```
# holiday not working as random coefficient
# time_pattern_dur.rand_coef <- lme(dur ~ dd + wday + holiday, random = ~1 + dd |
# station, data = df_main)
# summary(time_pattern_dur.rand_coef)

## Does not converges...
```

Business interpretation

From a business perspective, the findings from this analysis offer valuable insights for strategic planning, marketing, operational adjustments, and potential policy development. Here are the main takeaways:

Holidays and Weekends promotion The model indicates longer trip durations during holidays and weekends. This suggests higher usage or leisurely rides during these periods. There could be an opportunity to increase bike availability or introduce special promotions during holidays and weekends to cater to this demand. The interaction effect suggests that members' increased trip duration is less pronounced on weekends. This could imply that members use the service differently on weekends compared to weekdays. Design weekend-specific promotions or services for members. Understanding why this pattern occurs (leisure vs. commuting) can help tailor these offerings.

Membership pricing strategy Members tend to have significantly longer trip durations compared to non-members. This highlights the importance of members to the system. There should have a focus on member retention strategies and consider special offers or loyalty programs to encourage repeat usage. Additionally, analyzing non-member behavior to tailor services and promotions effectively would be pertinent.

Geographic optimization Significant variability in trip durations across different stations and districts indicates diverse usage patterns in different areas. Optimize bike and dock availability based on specific district and station demands. Targeted investments in high-usage areas could improve service efficiency.

Potential Policy Implications Understanding how different areas and demographics use the bike-sharing system can inform urban planning and public transport policies. Promoting bike-sharing effectively can contribute to environmental goals by reducing reliance on motorized transportation. # Research Question 3: What variables impact the average bixi trip duration?

Research question 3 : Environmental factors that impact non-members avg trip length

The objective is to identify environmental factors that impact average trip length of non-members. As we know, non-members' revenue is generated from a fixed cost per trip and varying cost proportional to the duration of the trip. Hence, evaluating properly the effect of the environmental conditions on the average trip length of non-members is primordial.

Variables Selection

Variables that make business sense to include:

- Temperature in degrees celcius (`temp`)
- Rainfall in mm (`rain`)
- Part of the week i.e. weekend or weekday (`wknd_ind`)

- Location of the bixi station compared to Parc Lafontaine, a landmark in the middle of the bixi station system (cardinality)
- If the station name contains the word 'metro' (Metro_ind)
- Season (season)

Data preparation

```
df_nomem <- df_main[df_main$mem == 0, ]

start_date = min(df_nomem$date)
df_nomem$day_num = as.integer(df_nomem$date - start_date)

df_nomem =df_nomem[order(df_nomem$date),]
```

We first filter the dataframe so that only non-members are included. Then we assigned a measure of distance in between observations of the same station. This distance is the number of days in between observations. These can vary from one station to another. Finally we ordered the dataset with dates ascending. This way,observation in auto regressive structure will be in the correct order.

Model

We will now build a first reference model. This model is not valid as it does not take into consideration the correlation intra station. The result will yield higher level of confidence on the estimate of the coefficient. This model will be refered as model 0.

```
mod0 <- gls(avg ~ temp + rain + wknd_ind + cardinality + Metro_ind + season, data = df_nomem)
```

We will procede to build a model that takes into consideration the structure of correlation of the errors. This way, intra station correlation will be captured. As we observe the same subject over time, auto-regressive structure is an appropriate choice. Also, knowing that time interval are not constant between observations, we will specify the time dimension.

```
mod1 <- gls(avg ~ temp + rain + wknd_ind + cardinality + Metro_ind + season, correlation = corAR1(form = ~
```

As model 0 is nested in model 1, we will perform a lrt to compare both.

```
##      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## mod0     1 11 32325.26 32396.32 -16151.63                        
## mod1     2 12 32321.27 32398.79 -16148.64 1 vs 2 5.986377  0.0144
```

The likelihood ratio tells us that the ordinary linear regression is not an adequate simplification of the model with AR1 structure, at a 5% confidence level. It is interesting to note that BIC criteria would have suggested using model 0.

Let's look at the correlation structure that was estimated in our AR1 model.

```
## Marginal variance covariance matrix
##            [,1]       [,2]       [,3]       [,4]
## [1,] 1.0000e+00 2.3261e-01 6.8101e-04 5.8086e-47
## [2,] 2.3261e-01 1.0000e+00 2.9277e-03 2.4971e-46
## [3,] 6.8101e-04 2.9277e-03 1.0000e+00 8.5294e-44
## [4,] 5.8086e-47 2.4971e-46 8.5294e-44 1.0000e+00
##   Standard Deviations: 1 1 1 1
```

We observe that the correlation is very small in between observations of the same station.

Exploring the ARH1 structure would have been interesting, but computing limitation do not allow us to test for this, as the model takes too long to fit.

Instead, we will jump right into models with random intercept.

```
mod2 <- lme(avg ~ temp + rain + wknd_ind + cardinality + Metro_ind + season, random = ~1 | station, data =
```

Let's compare this model with the simple linear regression model.

```
##      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## mod0     1 11 32325.26 32396.32 -16151.63
## mod2     2 12 32181.53 32259.05 -16078.76 1 vs 2 145.7281  <.0001
```

We observe that the null model is not a adequate simplificaiton of the random intercept model.

Let's continue by combining the random intercept and the error with auto regressive structure.

```
mod3 <- lme(avg ~ temp + rain + wknd_ind + cardinality + Metro_ind + season, random = ~1 |
station, correlation = corAR1(form = ~day_num | station), data = df_nomem)
```

```
##      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## mod0     1 11 32325.26 32396.32 -16151.63
## mod3     2 13 32182.22 32266.21 -16078.11 1 vs 2 147.0324  <.0001
```

Here again, We observe that the null model is not a adequate simplificaiton of the random intercept model plus the AR1 covariance structure on the errors.

Furthermore, as all the models fitted so far have the same fixed effect, it is possible to compare them using AIC and BIC obtained from the REML method.

```
##      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## mod0     1 11 32325.26 32396.32 -16151.63
## mod1     2 12 32321.27 32398.79 -16148.64 1 vs 2 5.986377  0.0144
## mod2     3 12 32181.53 32259.05 -16078.76
## mod3     4 13 32182.22 32266.21 -16078.11 3 vs 4 1.304320  0.2534
```

According to both AIC and BIC, the best model would be model 2 which is the one using only random intercept. This may indicate that the auto-regressive structure was not performing so well.

Let's print the necessary information for interpretation:

```
## Linear mixed-effects model fit by REML
##   Data: df_nomem
##        AIC      BIC    logLik
##   32181.53 32259.05 -16078.76
##
## Random effects:
##  Formula: ~1 | station
##         (Intercept) Residual
## StdDev:    2.738247 6.850674
##
## Fixed effects:  avg ~ temp + rain + wknd_ind + cardinality + Metro_ind + season
##                           Value Std.Error   DF   t-value p-value
## (Intercept)           12.973358 0.5186294 3982 25.014695  0.0000
## temp                   0.142608 0.0225509 3982  6.323822  0.0000
## rain                  -0.092315 0.0215184 3982 -4.290051  0.0000
## wknd_indWeekend        2.407109 0.2269664 3982 10.605574  0.0000
## cardinalityNorth-West -0.700548 0.4781136  742 -1.465234  0.1433
## cardinalitySouth-East  0.250369 0.5189772  742  0.482428  0.6296
## cardinalitySouth-West -0.308207 0.4645493  742 -0.663454  0.5072
## Metro_ind1            -0.472810 0.5333470  742 -0.886497  0.3756
## seasonSpring           3.921520 0.2981397 3982 13.153295  0.0000
## seasonSummer           0.805814 0.3033980 3982  2.655964  0.0079
##  Correlation:
##                       (Intr) temp   rain   wknd_W crdN-W crdS-E crdS-W Mtr_n1
## temp                  -0.561
## rain                  -0.080 -0.016
## wknd_indWeekend       -0.137  0.039 -0.050
## cardinalityNorth-West -0.631 -0.011 -0.008  0.000
## cardinalitySouth-East -0.581 -0.006 -0.016 -0.006  0.635
## cardinalitySouth-West -0.645 -0.018 -0.007 -0.002  0.710  0.653
## Metro_ind1            -0.105 -0.021  0.000 -0.003  0.044  0.027  0.038
## seasonSpring          -0.171 -0.041  0.152 -0.068  0.002  0.001  0.000 -0.004
## seasonSummer           0.176 -0.643  0.058 -0.039  0.002  0.001  0.009  0.026
##                       ssnSpr
## temp
## rain
## wknd_indWeekend
## cardinalityNorth-West
## cardinalitySouth-East
## cardinalitySouth-West
## Metro_ind1
## seasonSpring
## seasonSummer           0.356
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -2.7607307 -0.5797848 -0.1245663  0.3938619  6.0235870
##
## Number of Observations: 4734
## Number of Groups: 747
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: avg
##                Chisq Df Pr(>Chisq)
## temp         39.9907  1  2.552e-10 ***
## rain         18.4045  1  1.786e-05 ***
## wknd_ind    112.4782  1  < 2.2e-16 ***
## cardinality   5.5138  3     0.1378
## Metro_ind     0.7859  1     0.3753
## season      177.7389  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Correlation of individual 3:

```
##           1         2         3         4         5         6
## 1 1.0000000 0.1377556 0.1377556 0.1377556 0.1377556 0.1377556
## 2 0.1377556 1.0000000 0.1377556 0.1377556 0.1377556 0.1377556
## 3 0.1377556 0.1377556 1.0000000 0.1377556 0.1377556 0.1377556
## 4 0.1377556 0.1377556 0.1377556 1.0000000 0.1377556 0.1377556
## 5 0.1377556 0.1377556 0.1377556 0.1377556 1.0000000 0.1377556
## 6 0.1377556 0.1377556 0.1377556 0.1377556 0.1377556 1.0000000
```

We observe that the marginal correlation structure ressembles a lot that of a compound symmetric.

Interpretation of model with random intercept

Overall Model - The model was fit using the rectified maxmimum likelihood method. It had an AIC of 32181 and BIC of 32259. It was found to be the best model according to these criteria. The random intercept standard deviation is of 2.7 quite an important variation relative to the fixed part of the intercept (12.97). We see that the model identified 747 groups aka stations in the dataset and a total of 4734 obsevations. Hence there would be on average 4734/747 =6.33 observations of each station of non-members.

Intercept : The intercept can be interpeted as the expected average trip duration of non-members in the specific case where temperature is 0, there is no rain, it is a weekday, the trip departure is in the north-east of parc lafontaine, the station is not near a metro and the season is fall. In this case, we expect trip to be on average 12.97 minutes long.

Season: The reference level is fall. We can see that on average trip duration during spring and summer are respectively 3.9 and 0.8 minutes longer than in fall holding everything else constant.

Temperature: The coefficient of temperature is 0.14 which means that an increase in temperature of 1 degree celcius corresponds to an increase of average trip duration of 0.14 minutes on average holding all else constant.

Rainfall: The coefficient for rain is -0.09 which means that an increase in rainfall of 1 mm corresponds to a decrease of average trip duration of 0.09 minutes on average holding all else constant.

Cardinality: Their coefficients can be interpreted as the change in expected average trip duration compared to the reference departure point of north-east when all other variables are held constant. The effect of this variable is not deemed significant according to the analysis of deviance for a 5% confidence level.

Metro Indicator : Metro indicator's coefficient is -0.47 which means that the expected value for average trip length decreases by 0.47 minutes when a bixi station is near a metro acces point, holding all else constant. This would suggest that user who rent bikes after making a metro ride are closer to their final destination than in other cases. Although this effect is deemed no significative according to the analysis on deviance for a confidence level of 5%.

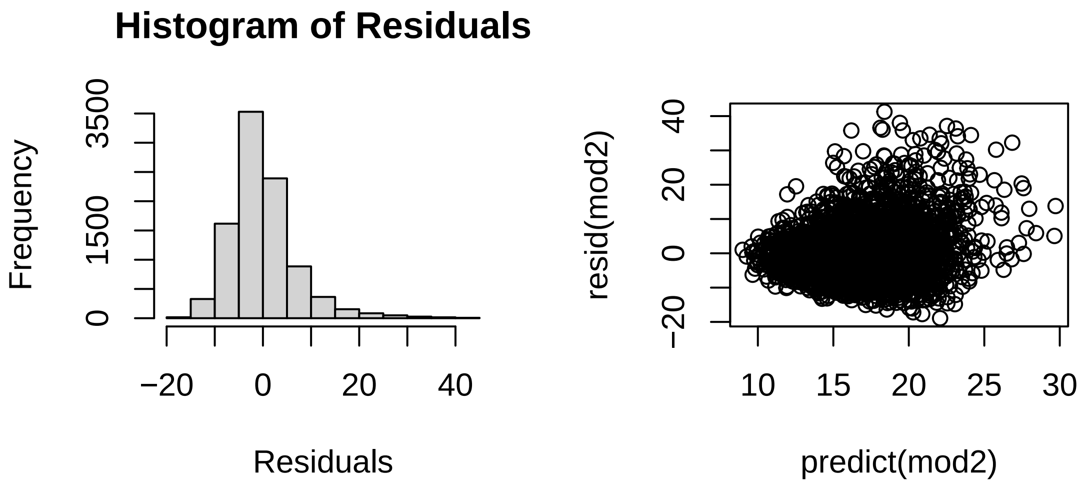
Weekend Indicator : This coefficient means that during weekends, non-members are expected to have longer average trip length by 2.4 minutes when all other covariates are kept constant.

Business Implications:

What we uncoverd is that given a season, temperature increases promotes longer trip for non-members. This will in turn generate more revenue and increase the demand on the bixi system. Also, rainfall as the exact opposite effect. Furthermore, weekend will generally imply significantly longer trips. Finaly, the usage during the spring season seems to be quite different than in other seasons as the coefficient for that specific month is quite higher and that for a constant temperature.

Verification of assumptions

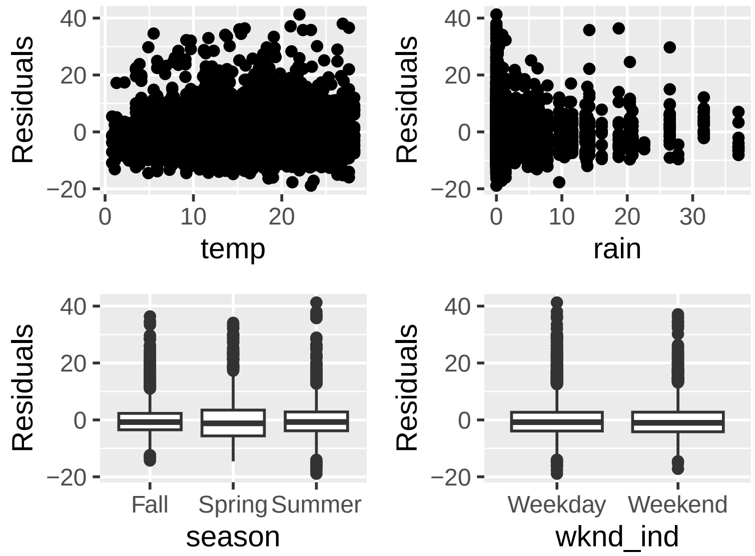
Verification of Normality of Residuals and model correctly specified



We observe a slightly longer tail on the right, but noting alarming.

There may be some evidences of small heteroscedasticity in the second plot.

Verification of Heteroscedasticity



No major problem of heteroscedasticity were detected except some slightly higher variance where there is no rain.

Limitations and shortcomings

- Presence of outliers in the dataset. These have impacted the fit of some of our model and may resut in heteroscedasticity.
- Difficulty regarding having a proper measure of members revenue. Having a fixed cost is hard to reflect in the revenue per day.
- The Dataset is already aggregated so getting proper individual measure is quite challenging.
- Causation vs. Correlation: The regression model captures relationships but does not establish causation.

- Data Exclusions: The data only considers trips under 60 minutes, which might exclude a segment of users who use BIXI for longer journeys.

Conclusion

In conclusion, several key operational and strategic considerations have emerged from the data analysis of BIXI bike rentals:

Contribution

Charles Julien :Research question 3 and limitations

Gabriel Jobert :

Chike Odenigbo:

Atul Sharma: Introduction