

DATA 621 Final

Brad Harbans

5/22/2022

Abstract

The NYC Department of Transportation (NYC DOT) collects a daily total of bike counts conducted monthly on the Brooklyn Bridge, Manhattan Bridge, Williamsburg Bridge, and Queensboro Bridge ((DOT) 2022). The data is used by the city for transportation planning. The dataset also contains the temperature and precipitation for the days. One would expect that days with extreme temperatures or that have precipitation should have lesser bicycle usage. It is the purpose of this paper to attempt to predict bicycle utilization using a variety of linear models. For this paper I will use a subset of data that is available on Kaggle (York 2017).

Key words: Regression, Poisson , predicting count values, general additive models

Introduction

In this paper I will attempt to predict the number of bicycle crossing across the east river bridges. As we are analyzing count data, a poisson distribution will be used. We can also use the presence of precipitation as a target for a binomial regression model. As a comparison I will also be using a generalized additive model for prediction.

Literature review

With rising urban populations cars create more traffic, pollution, noise, and green house emissions. Encouraging cycling has the ability to mitigate some of these factors. Encouraging cycling has a number of benefits for a city and its inhabitants. For the typical cyclist a check of the weather is the first thing that is done before leaving the house, if it is raining or too hot it is unlikely that the individual will choose to cycle(Neef, Bean, and Rojas 2021). A 2021 study aimed to answer this question by analyzing data from “forty Public Bicycle Sharing Programs located in forty cities (16 countries) across five different climate zones, spanning tropical to boreal climates” (Bean, Pojani, and Corcoran 2021).

Key findings from the study included “: (a) the most significant variable, particularly on weekdays, is the time of day, followed by precipitation; (b) in most cities, usage increases on weekdays and weekends up to a point around 27 to 28° C, before declining; (c) usage by hour usually follows a bimodal or trimodal daily pattern on weekdays, except for schemes which are too small to serve a commuter function (weekend and weekday usage is similar in small schemes); (d) weekend usage peaks at around 2 to 3 pm in most schemes, except those in hotter climates where the peak is around 5 pm; (e) precipitation negatively affects female ridership more than male ridership; and, (f) a changing climate is likely to affect cycling by boosting ridership in cold climates and lowering ridership in warm climates, but the effects will likely be small” (Bean, Pojani, and Corcoran 2021).

In particular the analysis of data from NY revealed that higher temperatures predict more cycling trips whereas rainy, humid, windy and especially snowy weather led to fewer cycling trips

(Bean, Pojani, and Corcoran 2021). Other interesting information that was revealed by the data is that “you’d think that people in the tropics would be particularly sensitive to cold weather, and that people in colder climates would be more willing to ride when it’s cold ... acclimatization doesn’t affect rider numbers as much as we might have thought.”(Neef, Bean, and Rojas 2021). Also the fact that the study used several cities as a comparison, it allowed the researchers to deduce that “safe cycling infrastructure is enough to encourage people to ride even if the weather is bad”(Neef, Bean, and Rojas 2021). There is even a discussion on how climate change will effect riding patterns.

The study utilized the generalized additive model:

$$usage(h) = s(u_h) + p_h + s(h) + s(j_h) + \epsilon$$

Where:

$usage(h)$ = number of trips within half an hour of hour

u_h = UTCI temperature at hour h

p_h = total precipitation in the previous hour

j_h = Julian date

ϵ = Error term

Generalized addtitive models (GAM) is a technique which links the preditors to the target using a smoothing functions rather than a coefficient. This the method proves to be useful in uncovering nonlinear covariate effects (Hastie and Tibshirani 1986). Since this was the method used in the Bean model, I will be applying this to the data from the NYC DOT.

Methodology

The sample dataset consits of 210 observations, running from the month of April 2016 with 9 predictors. The sample size is rather small, and the aim would be to build a modeling technique that can be adapted to use data from the NY DOT and another source for weather.

Experimentation and Results



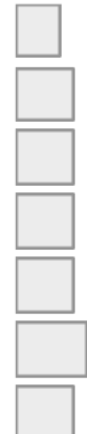
I will begin by displaying some summary staistics :

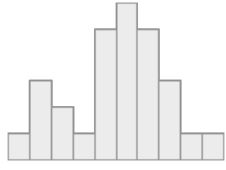
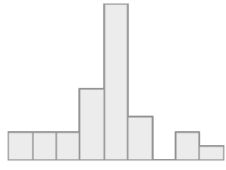
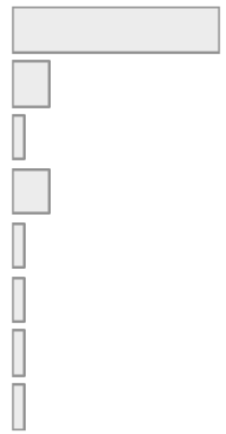
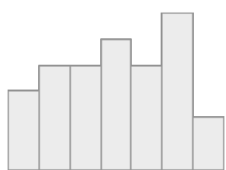
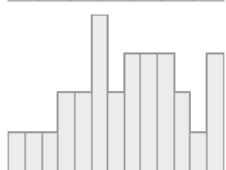
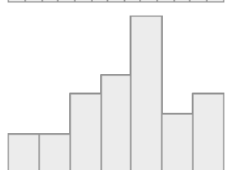

```
##           X           Date           Day           High.Temp...F.
## Min.      : 0.00   Length:210   Length:210   Min.      :39.90
## 1st Qu.: 52.25   Class :character Class :character 1st Qu.:55.00
## Median :104.50   Mode  :character Mode  :character Median :62.10
## Mean    :104.50                                Mean    :60.58
## 3rd Qu.:156.75                                3rd Qu.:68.00
## Max.     :209.00                                Max.     :81.00
## Low.Temp...F. Precipitation Brooklyn.Bridge Manhattan.Bridge
## Min.      :26.10   Length:210   Min.      : 504   Min.      : 997
## 1st Qu.:44.10   Class :character 1st Qu.:1447   1st Qu.:2617
## Median :46.90   Mode  :character Median :2380   Median :4165
## Mean    :46.41                                Mean    :2270   Mean    :4050
## 3rd Qu.:50.00                                3rd Qu.:3147   3rd Qu.:5309
## Max.     :66.00                                Max.     :3871   Max.     :6951
## Williamsburg.Bridge Queensboro.Bridge Total
## Min.      :1440   Min.      :1306   Min.      : 4335
## 1st Qu.:3282     1st Qu.:2457   1st Qu.: 9596
## Median :5194     Median :3477   Median :15292
```

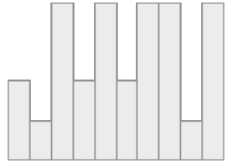
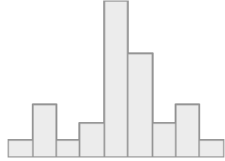
```
## Mean      :4862      Mean      :3353      Mean      :14534
## 3rd Qu.   :6030      3rd Qu.   :4192      3rd Qu.   :18315
## Max.      :7834      Max.      :5032      Max.      :23318
```

Before displaying summary statistics, I will create a few auxiliary variables. For one, I will introduce a mean temperature, remove the index and Day columns (as this is identical to date). I will also coerce the precipitation column to a numeric, this introduces NAs (which I will drop), as it includes two repeated rows.

Let us look at some summary statistics.

Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
X [integer]	Mean (sd) : 104.7 (60.8) min < med < max: 0 < 104.5 < 209 IQR (CV) : 104.5 (0.6)	196 distinct values		0 (0.0%)
Date [character]	1. 2016-04-01 00:00:00 2. 2016-04-02 00:00:00 3. 2016-04-03 00:00:00 4. 2016-04-05 00:00:00 5. 2016-04-06 00:00:00 6. 2016-04-07 00:00:00 7. 2016-04-08 00:00:00 8. 2016-04-09 00:00:00 9. 2016-04-10 00:00:00 10. 2016-04-11 00:00:00 [18 others]	7 (3.6%) 7 (3.6%) 7 (3.6%) 7 (3.6%) 7 (3.6%) 7 (3.6%) 7 (3.6%) 7 (3.6%) 7 (3.6%) 7 (3.6%) 126 (64.3%)		0 (0.0%)
Day [factor]	1. Monday 2. Tuesday 3. Wednesday 4. Thursday 5. Friday 6. Saturday 7. Sunday	21 (10.7%) 28 (14.3%) 28 (14.3%) 28 (14.3%) 28 (14.3%) 35 (17.9%) 28 (14.3%)		0 (0.0%)

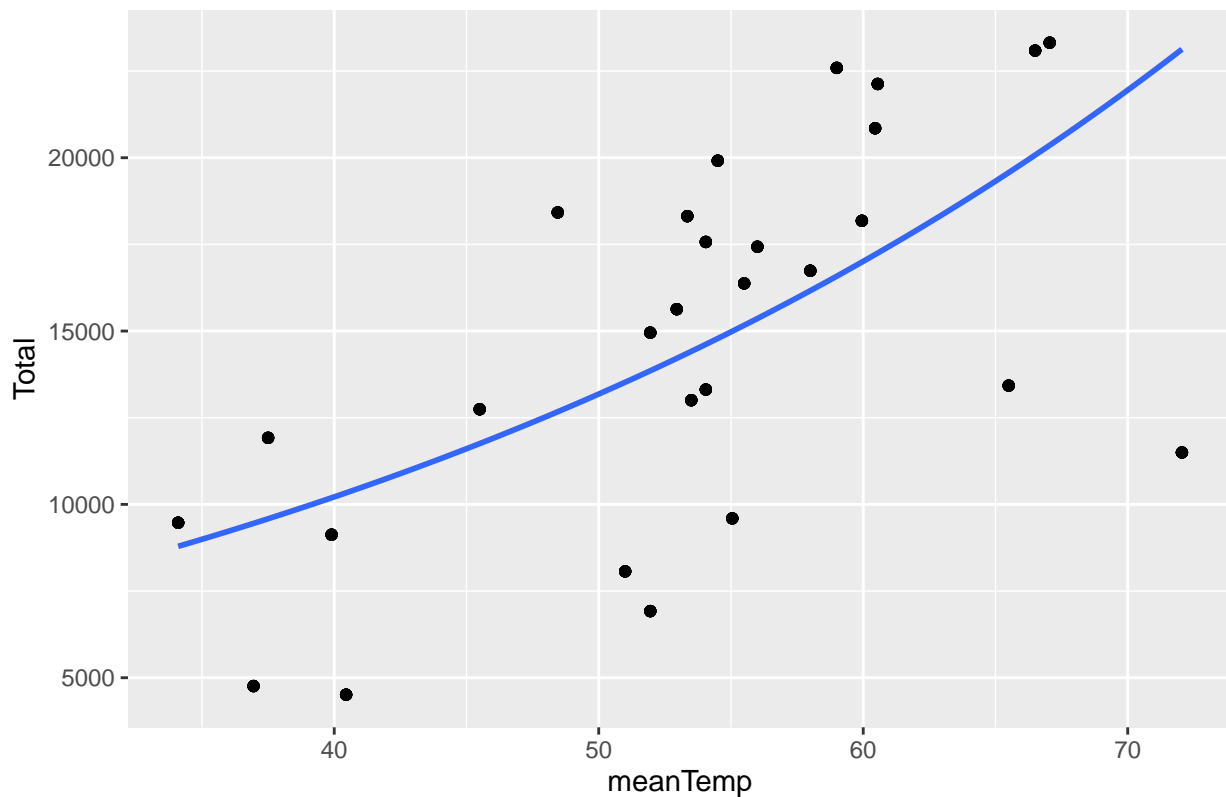
Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
High.Temp...F. [numeric]	Mean (sd) : 60.5 (10.6) min < med < max: 39.9 < 62.1 < 81 IQR (CV) : 11.5 (0.2)	20 distinct values		0 (0.0%)
Low.Temp...F. [numeric]	Mean (sd) : 46.3 (9) min < med < max: 26.1 < 46.9 < 66 IQR (CV) : 5.9 (0.2)	20 distinct values		0 (0.0%)
Precipitation [numeric]	Mean (sd) : 0 (0.1) min < med < max: 0 < 0 < 0.2 IQR (CV) : 0.1 (1.7)	0.00 : 119 (60.7%) 0.01 : 21 (10.7%) 0.05 : 7 (3.6%) 0.09 : 21 (10.7%) 0.15 : 7 (3.6%) 0.16 : 7 (3.6%) 0.20 : 7 (3.6%) 0.24 : 7 (3.6%)		0 (0.0%)
Brooklyn.Bridge [numeric]	Mean (sd) : 2306.9 (950.2) min < med < max: 504 < 2379.5 < 3871 IQR (CV) : 1520.2 (0.4)	28 distinct values		0 (0.0%)
Manhattan.Bridge [integer]	Mean (sd) : 4125.4 (1662.3) min < med < max: 997 < 4165 < 6951 IQR (CV) : 2331.8 (0.4)	28 distinct values		0 (0.0%)
Williamsburg.Bridge [numeric]	Mean (sd) : 4940.8 (1745.3) min < med < max: 1507 < 5194 < 7834 IQR (CV) : 2291.2 (0.4)	28 distinct values		0 (0.0%)
Queensboro.Bridge [numeric]	Mean (sd) : 3407.8 (1064.2) min < med < max: 1306 < 3477 < 5032 IQR (CV) : 1665 (0.3)	28 distinct values		0 (0.0%)

Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
Total [integer]	Mean (sd) : 14780.8 (5390.3) min < med < max: 4510 < 15292.5 < 23318 IQR (CV) : 7320 (0.4)	28 distinct values		0 (0.0%)
meanTemp [numeric]	Mean (sd) : 53.4 (9.3) min < med < max: 34.1 < 54 < 72 IQR (CV) : 8.9 (0.2)	26 distinct values		0 (0.0%)

As expected, higher temperatures seem to correlate to a higher number of cyclists. Additionally, precipitation seems to be negatively correlated with the number of cyclists. We can also see that there are more cyclist on Wednesday and Thursdays.

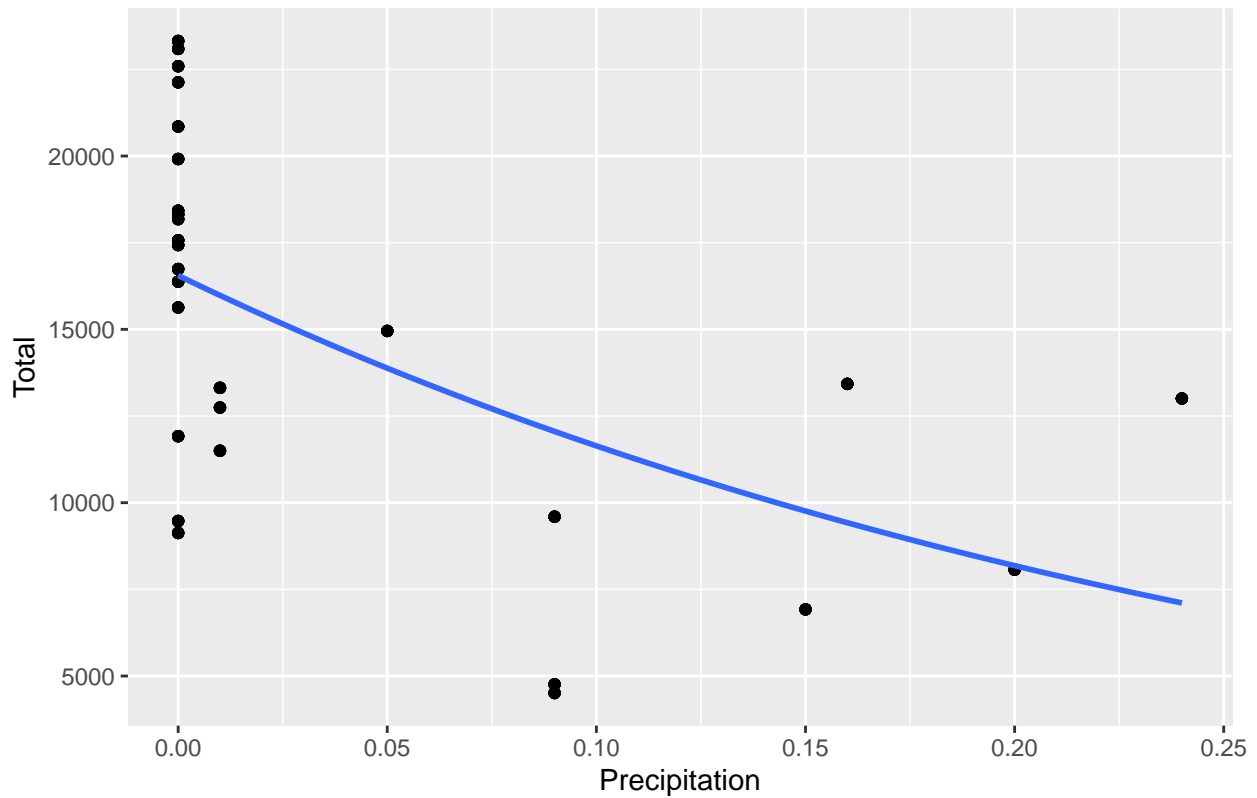
```
## `geom_smooth()` using formula 'y ~ x'
```

Mean Temperature vs. Total Number of Cyclists

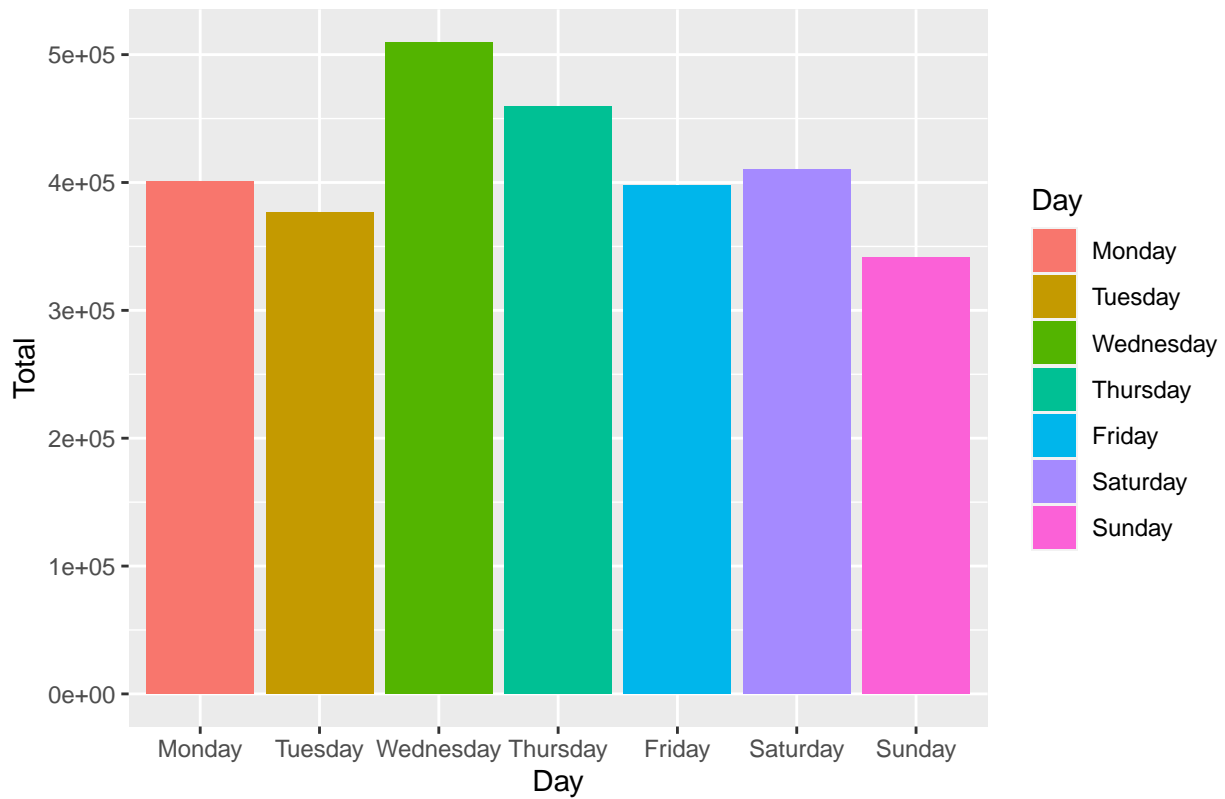


```
## `geom_smooth()` using formula 'y ~ x'
```

Precipitation vs. Total Number of Cyclists



Day of Week vs Total Number of Cyclists



Prediction using GLM(Poisson)

I have already split my data into a test and training data set. I will now create a model using the precipitation, mean temperature, and the day of week as predictors.

```
##
## Call:
## glm(formula = Total ~ Precipitation + meanTemp + Day, family = poisson(link = log),
##      data = training)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -66.376  -14.095   1.325   14.501   40.645
##
## Coefficients:
##              Estimate Std. Error  z value Pr(>|z|)
## (Intercept)  8.373e+00  5.309e-03 1577.182  <2e-16 ***
## Precipitation -2.909e+00  1.350e-02 -215.500  <2e-16 ***
## meanTemp      2.483e-02  8.106e-05  306.394  <2e-16 ***
## DayTuesday    9.069e-02  2.857e-03  31.739  <2e-16 ***
## DayWednesday  1.782e-01  2.574e-03  69.231  <2e-16 ***
## DayThursday  -5.280e-03  2.638e-03  -2.001   0.0453 *
## DayFriday     -2.222e-01  2.612e-03 -85.069  <2e-16 ***
## DaySaturday  -1.562e-01  2.739e-03 -57.033  <2e-16 ***
## DaySunday     -1.633e-01  2.843e-03 -57.441  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 315354  on 147  degrees of freedom
## Residual deviance:  88607  on 139  degrees of freedom
## AIC: 90303
##
## Number of Fisher Scoring iterations: 4
```

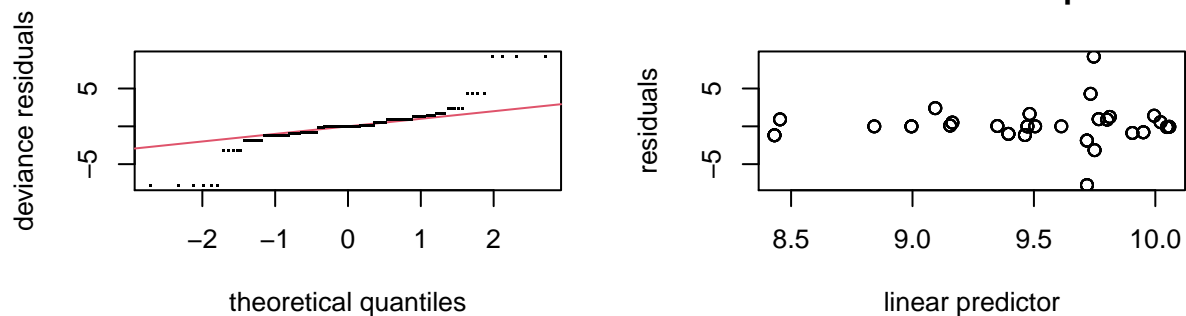
Prediction using GAM

```
##
## Family: poisson
## Link function: log
##
## Formula:
## Total ~ s(Precipitation, k = 8) + s(meanTemp) + Day
##
## Parametric coefficients:
##              Estimate Std. Error  z value Pr(>|z|)
## (Intercept)  9.510626  0.003313 2870.587  < 2e-16 ***
## DayTuesday    0.032897  0.004206   7.822 5.21e-15 ***
## DayWednesday  0.120731  0.004687  25.758  < 2e-16 ***
## DayThursday   0.040646  0.003701  10.982  < 2e-16 ***
## DayFriday     0.011110  0.005192   2.140  0.0324 *
## DaySaturday  -0.114878  0.005763 -19.934  < 2e-16 ***
## DaySunday     -0.130021  0.004560 -28.511  < 2e-16 ***
## ---
```

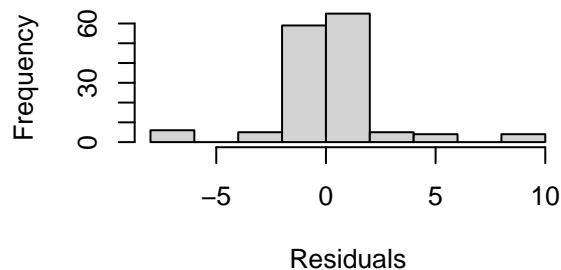
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##              edf Ref.df Chi.sq p-value
## s(Precipitation) 6.998      7  62294 <2e-16 ***
## s(meanTemp)      8.995      9  77182 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.996   Deviance explained = 99.7%
## UBRE = 5.7793  Scale est. = 1          n = 148
```

Lets perform a check on our model using the `gam.check()` function. The model converges, however, the small p value for the second smooth indicates that residuals are not randomly distributed. This often means there are not enough basis functions.

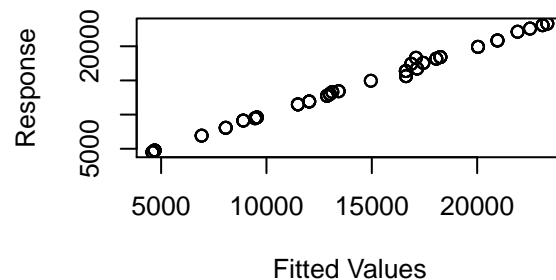
Resids vs. linear pred.



Histogram of residuals



Response vs. Fitted Values

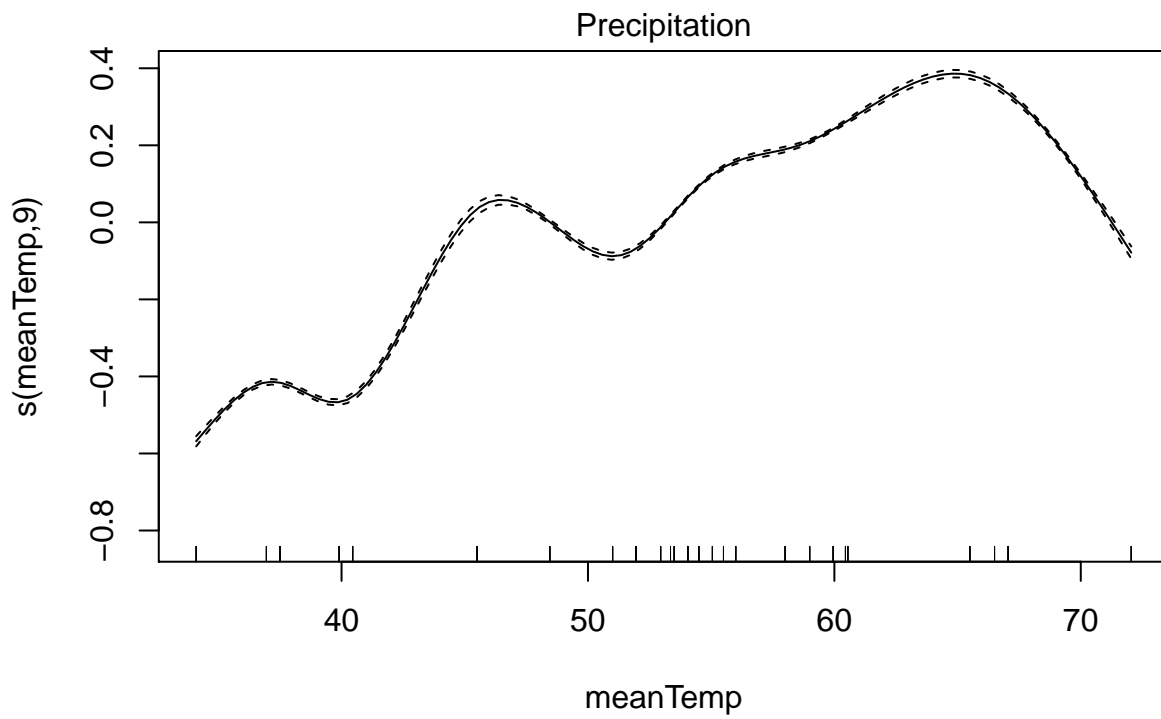
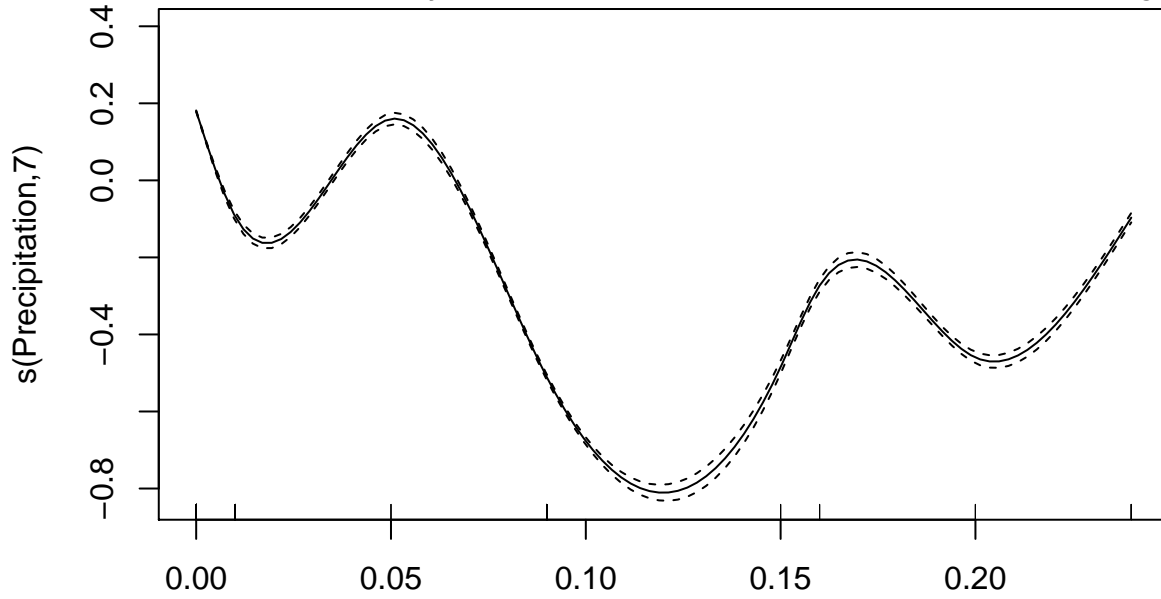


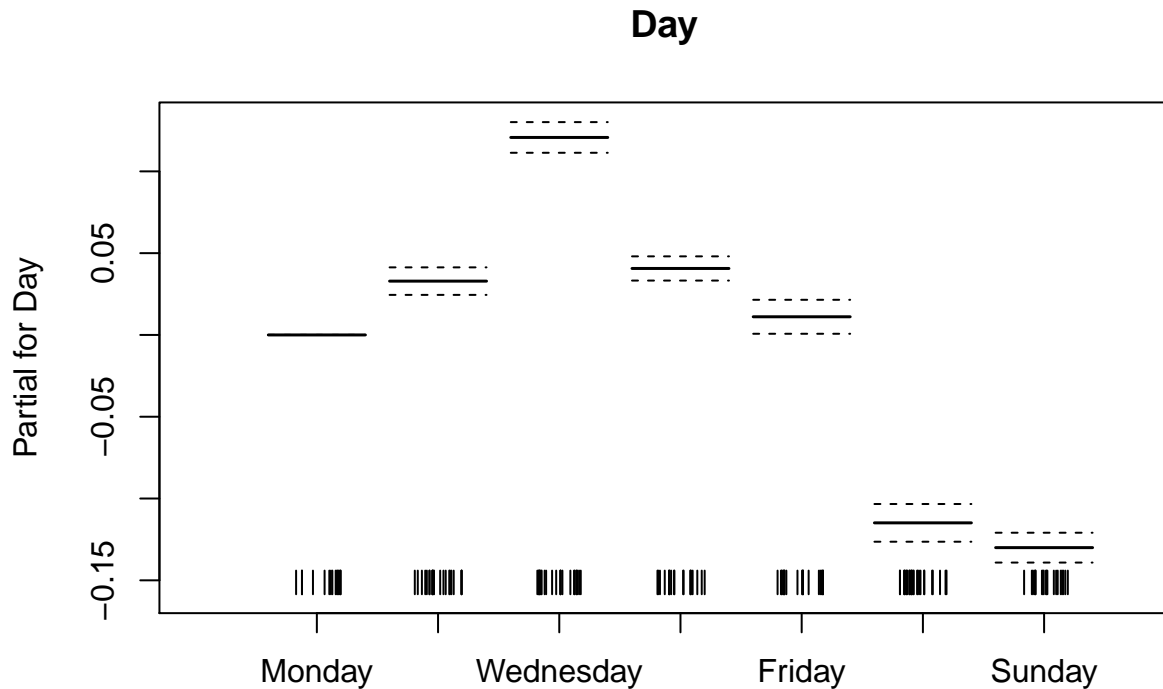
```
##
## Method: UBRE   Optimizer: outer newton
## full convergence after 15 iterations.
## Gradient range [5.077184e-07,5.729671e-06]
## (score 5.779325 & scale 1).
## Hessian positive definite, eigenvalue range [1.032638e-05,6.810228e-05].
## Model rank = 23 / 23
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##              k' edf k-index p-value
## s(Precipitation) 7  7    1.09  0.84
## s(meanTemp)      9  9    0.32 <2e-16 ***
## ---
```



```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The smooths seem to accurately model the data, we should be careful over overfitting though.





```

##                                edf   Ref.df   Chi.sq p-value
## s(Precipitation) 6.998439 6.999998 62293.90      0
## s(meanTemp)      8.995170 8.999990 77182.24      0

```

Model Comparison

Note that using the `anova()` function to compare these models shows that the gam model has a more parsimonious fit of the data. The low p-value indicates that the second model is statistically significantly better at capturing the data than the linear model.

```

## Analysis of Deviance Table
##
## Model 1: Total ~ Precipitation + meanTemp + Day
## Model 2: Total ~ s(Precipitation, k = 8) + s(meanTemp) + Day
##   Resid. Df Resid. Dev      Df Deviance   Pr(>Chi)
## 1      139.00      88607
## 2      125.01       957 13.994    87649 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

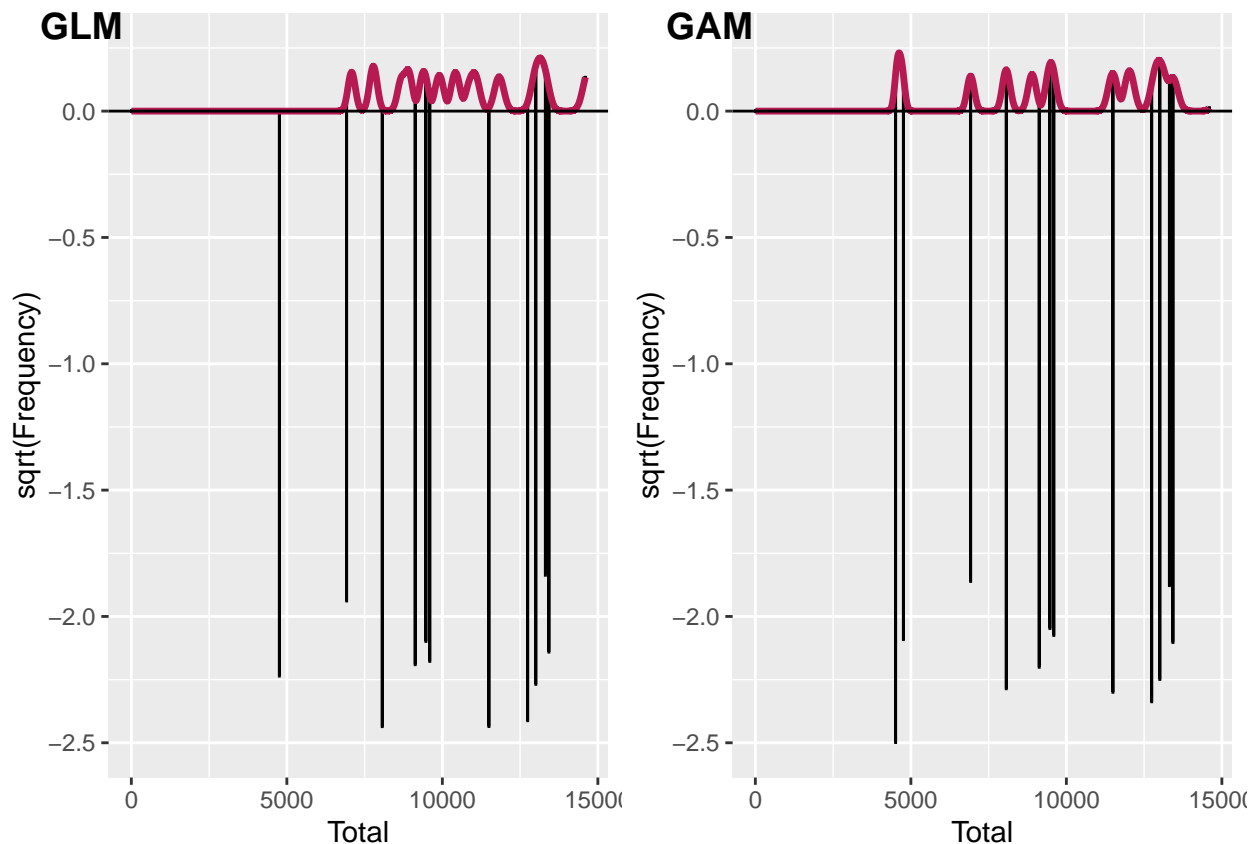
I will also look at the root mean square errors for the models. Note that the RMSE is lower for the GAM model and pearsons χ^2 , dispersion statistic, and AIC is lower.

	pearson.chi2	dispersion	RMSE	AIC
GLM	84668.59	609.1266	16264.57	90303.29
GAM	958.6343	7.668682	16264.54	2681.812

“The rootogram is a graphical tool associated with the work of J. W. Tukey that was originally used for assessing goodness of fit of univariate distributions. Here we extend the rootogram to regression models and show that this is particularly useful for diagnosing and treating issues such as overdispersion and/or excess zeros in count data models”. [Kleiber_2016]. If a bar

doesn't reach the zero line then the model over predicts a particular count bin, and if the bar exceeds the zero line it under predicts.(Simpson 2016).

In our case, our models both under predicts the counts, however, the GAM does a better job at fitting the data.



Discussion and Conclusions

My sample size is limited and has been confined to the Month of April. In order to get a better understanding of how the weather and day of week explains the number of cyclist, one should get a larger dataset from the NYC Dot and combine that with weather data. It would be interesting to see how the total number of cyclists change over winter months or how the changing weather patterns effect the number of cyclist.

The general additive model provided the best fit of the data. Nonetheless, there is no single coefficient that we can make inferences from. As a result we will need to look at the other model to deduce the effects of the variables. In this case, it would appear that precipitation has a highest influence on the number of cyclists, which makes intuitive sense. What is surprising is that the day of the week in general tends to be a stronger predictor than the temperature.

References

- Bean, Richard, Dorina Pojani, and Jonathan Corcoran. 2021. "How Does Weather Affect Bike-share Use? A Comparative Analysis of Forty Cities Across Climate Zones." *Journal of Transport Geography* 95: 103155. <https://doi.org/https://doi.org/10.1016/j.jtrangeo.2021.103155>.
- (DOT), Department of Transportation. 2022. "Bicycle Counts: NYC Open Data." *Bicycle Counts | NYC Open Data*. <https://data.cityofnewyork.us/Transportation/Bicycle-Counts/uczf-rk3c>.

- Hastie, Trevor, and Robert Tibshirani. 1986. "Generalized Additive Models." *Statistical Science* 1 (3): 297–310. <https://doi.org/10.1214/ss/1177013604>.
- Neef, Matt de, Dr Richard Bean, and Juan Rojas. 2021. "How Cyclists Respond to Bad Weather: Lessons Learnt from 100 Million Rides." *CyclingTips*. <https://cyclingtips.com/2021/10/how-cyclists-respond-to-bad-weather-lessons-learnt-from-100-million-rides/>.
- Simpson, Gavin. 2016. "Rootograms." *From the Bottom of the Heap*. From the Bottom of the Heap. <https://fromthebottomoftheheap.net/2016/06/07/rootograms/>.
- York, City of New. 2017. "New York City - East River Bicycle Crossings." *Kaggle*. <https://www.kaggle.com/datasets/new-york-city/nyc-east-river-bicycle-crossings>.