Uniwersytet Wrocławski

Uber app delivery time estimation

Maciej Szczutko Ewa Stebel

2024-01-29 22:29:56

Introduction

In our proposal we would like to describe the data we choose for Semiparametric regression final project. Course is conducted by Prof. J. Harezlak. Our goal is to explore the data and build a model which will help us answer several questions which we will propose in the latter part of this proposal.

Data description

START_DATE	END_DATE	CATEGORY	START	STOP	MILES	PURPOSE
7/18/2016 10:37	7/18/2016 10:49	Personal	Cary	Morrisville	4.1	Moving
4/16/2016 15:10	4/16/2016 15:26	Business	Morrisville	Cary	6.1	Meal/Entertain
12/22/2016 23:27	12/22/2016 23:32	Business	Lahore	Lahore	2.1	Customer Visit
08-07-2016 17:28	08-07-2016 17:43	Business	Edgehill Farms	Whitebridge	2.7	Customer Visit
1/27/2016 14:05	1/27/2016 14:13	Business	Raleigh	Raleigh	2.7	Customer Visit

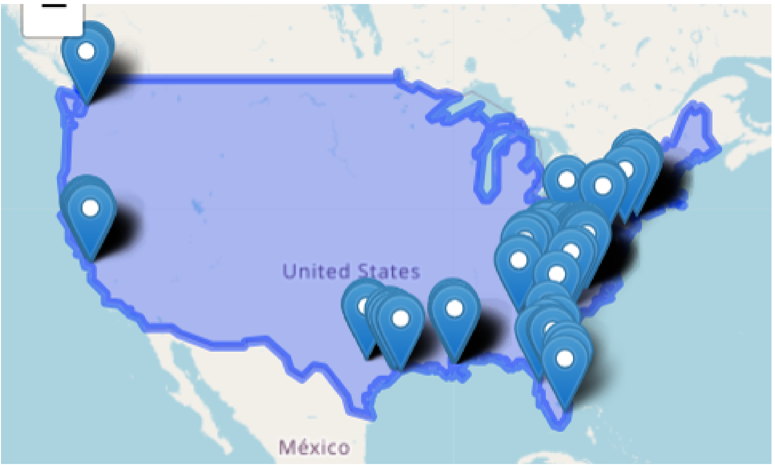
We choose UberDataset from Kaggle website. The dataset contains information about the provision of transportation services by Uber - a multinational transportation network company that operates a platform connecting riders with drivers through a mobile app. Data includes information on trips made in the USA in 2016 year. This dataset consists of 1156 observation of the following 7 columns:

1. start date - date and time of service start,
2. end date - date and time of service end,
3. category - categorical variable, division of the trip into private and business,
4. start - location - city or district - of the starting point,
5. end - location - city or district - of the final point,
6. miles - distance travelled in miles,
7. purpose - categorical variable, purpose of the transport.

Goal

The first thing we would like to do is standard data preprocessing to deeper understanding and find possible garbage in data (non realistic delivery time, NA values etc.). We want to built model (or models) to estimate delivery time. Delivery time variable will be additional column created by transformation of start date and end date variables. For the simplest model, with one predictor, we try add some smooth terms and compare results. As the data consist observation for various location we want to analyse them by some region e.g. New York should be consider separately from Carry district. Also we would like to test relation between daytime and time of service using models with interactions. Moreover we suspect there the category of service has no significant impact on delivery time. We will use statistical tests to confirm (or reject) this hypothesis.

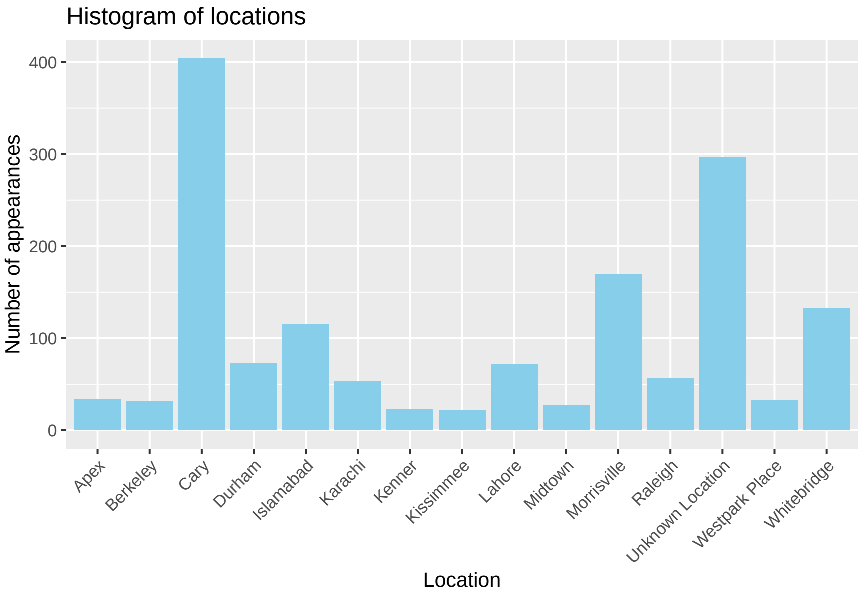
Preliminary analysis



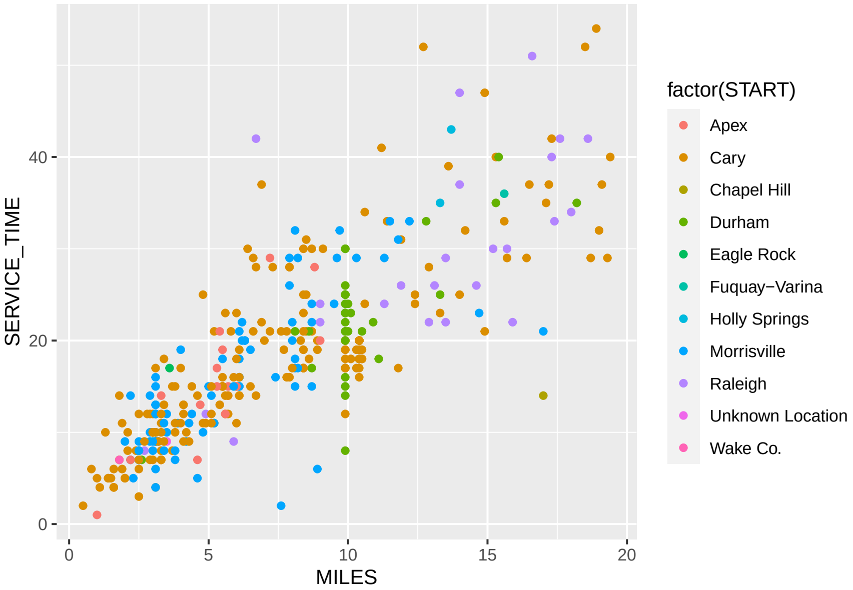
In order to better understand the data, we have put the locations from the dataset onto a map of the United States. On the map, we can see that the locations are divided into a pair of clusters, between which the average delivery time may also vary.

In order to gain a more thorough understanding of the data analyzed and to select models more effectively, we conducted a preliminary analysis. First, we created a graph below of the relationship between the miles variable and service time by initial destination.

Analyzing the locations from the start and stop variables, we see that the frequency of occurrence of a given location is various. Therefore, we selected the 15 most frequently occurring locations and presented their frequency in a histogram. In our analyzed dataset, we also filtered out the most frequent locations.



We see that the Cary location occurs much more frequently than the others. The Raleight, Durcham, Morrisville and Apex locations appear frequently, while the other locations appear sporadically.



From the chart, we can see that some locations form a distinct group on the graph. For example, service which has starting point in Raleight (violet points) is concentrated mostly above line with service time equal to 20 minutes. This suggests taking this variable into account when building the model. In addition, service which has starting point in Durham concentrates around the straight miles equals to 10, making us think that the variable miles and starting location may be correlated.

Data cleaning

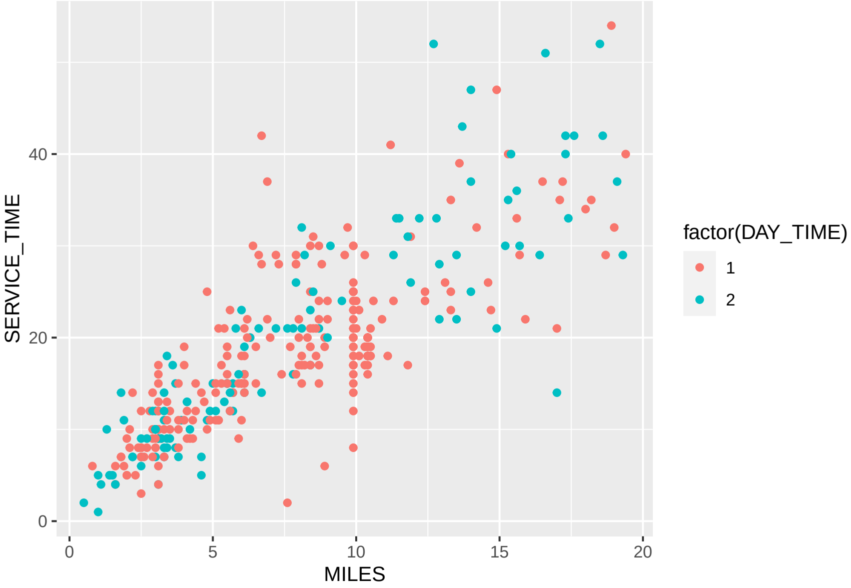
To prepare the data for modelling, we must first clean it. To do this, we replace the NA values in the Purpose column with "Unknown". Other columns do not show missing values, however the start and stop columns show some "Unknown location" values. We also remove the last row of data, which is the summary.

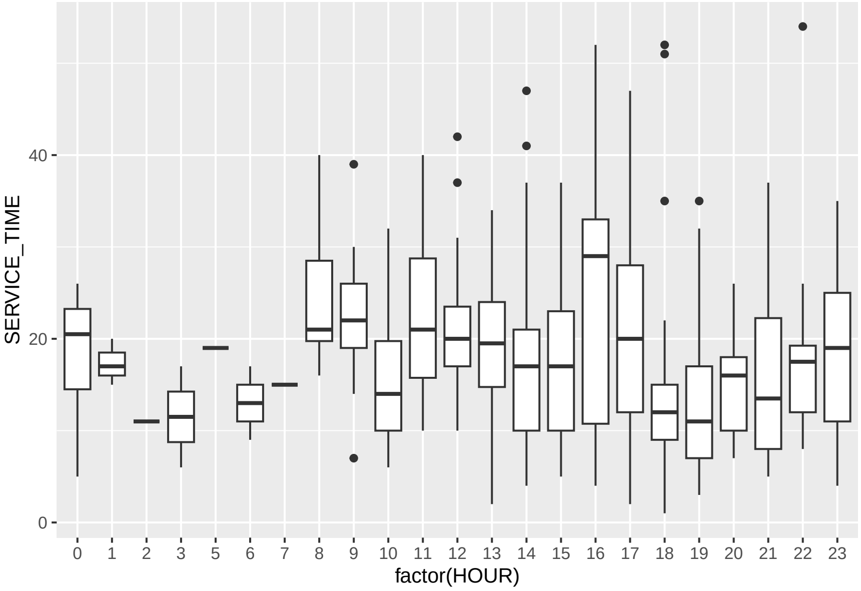
We added two columns to the raw data - Service time, which shows the duration of the service in minutes. To do this, we first standardized the format of the start and end date. Second added column is Daytime column, which indicates whether the service took place in rush hours (15-18) or regular hours (all others). We created this division based on the boxplot below. This division is also in line with our intuition. We believe this variable can be significant for modeling. The short summary of data after cleansing is presented below.

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

```
##    START_DATE                          END_DATE                  
##  Min.   :2016-01-07 13:27:00.00   Min.   :2016-01-07 13:33:00.00  
##  1st Qu.:2016-04-15 12:19:45.00   1st Qu.:2016-04-15 12:50:30.00  
##  Median :2016-07-05 21:23:30.00   Median :2016-07-05 21:44:00.00  
##  Mean   :2016-07-12 03:50:24.35   Mean   :2016-07-12 04:08:51.54  
##  3rd Qu.:2016-10-31 19:55:15.00   3rd Qu.:2016-10-31 20:22:00.00  
##  Max.   :2016-12-14 20:24:00.00   Max.   :2016-12-14 20:40:00.00  
##    CATEGORY            START               STOP               MILES       
##  Length:362         Length:362         Length:362         Min.   : 0.500  
##  Class :character   Class :character   Class :character   1st Qu.: 3.400  
##  Mode  :character   Mode  :character   Mode  :character   Median : 6.650  
##                                                           Mean   : 7.503  
##                                                           3rd Qu.:10.000  
##                                                           Max.   :19.400  
##    PURPOSE           SERVICE_TIME     DAY_TIME          
##   Length:362         Min.   : 1.00   Length:362
##   Class :character   1st Qu.:11.00   Class :character
##   Mode  :character   Median :17.00   Mode  :character
##                      Mean   :18.45
##                      3rd Qu.:23.00
##                      Max.   :54.00
```

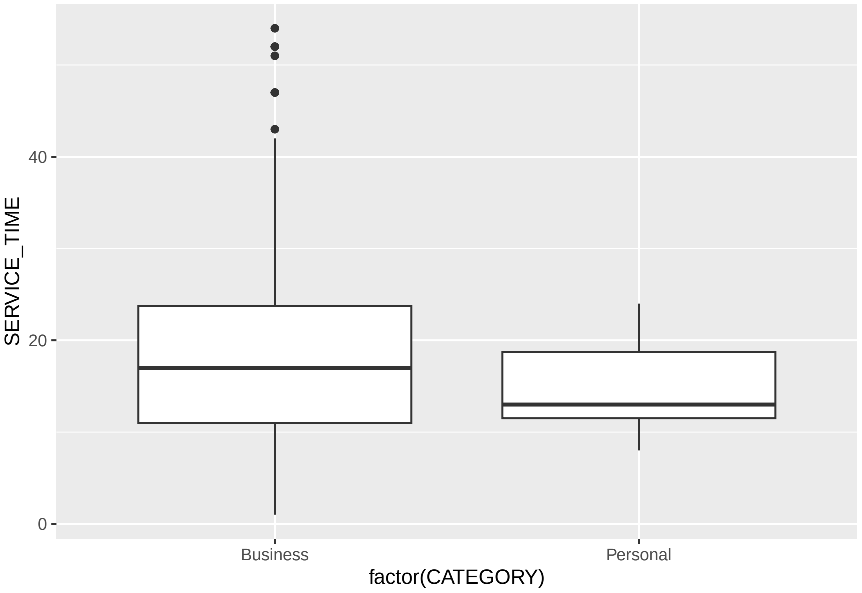
We also analyzed a newly created variable - hours and day time.

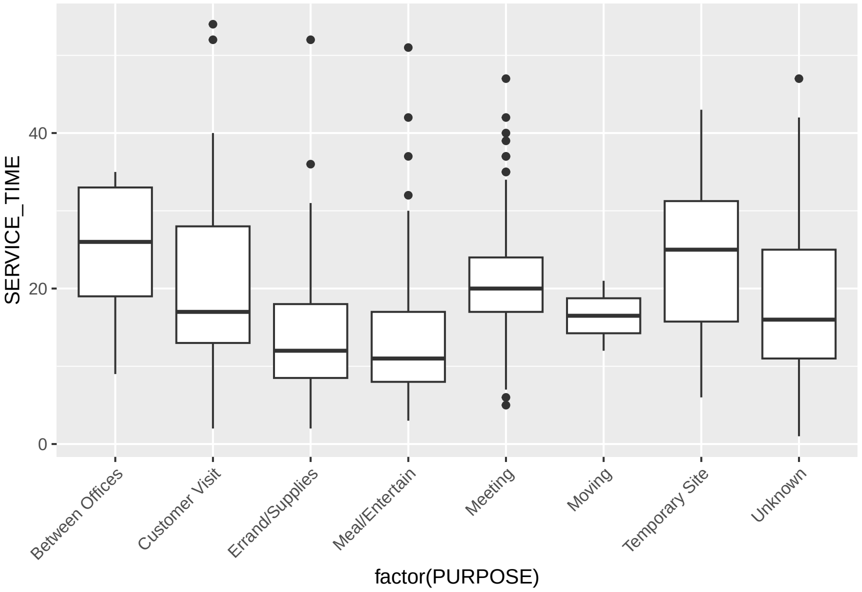




We are not able to see a clear division by time of day on the scatterplot, however based on the boxplot, we see variation in service time by hour of the day, which suggest the importance of this variable.

Then, we checked the other two variables included in the dataset - purpose and category and their impact on service time.





We do not see a significant difference in service time due to the breakdown by travel category. The average time and distance between quantiles for personal trips are slightly smaller than for buisness trips.

In the case of the purpose variable, we see differences between factors, so we will also consider this variable in later modeling.

Models

Once the data has been cleaned and subjected to preliminary analysis, we can begin to build models that predict service times. The first basic model we will create is a linear regression model with an explained variable - service time and an explanatory variable - miles. The linear regression model has the form

$$f(X) = \beta_0 + \sum_{j=1}^{p} x_j \beta_j,$$

where the β_j ' are unknown parameters or coefficients.

This is a basic method, but allows us an easy interpretation of regressors effects.

Simple linear model

```
##
## Call:
## lm(formula = SERVICE_TIME ~ MILES, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.7761  -3.1744  -0.7904   2.5396  25.0112
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.76794    0.58185   8.194 4.46e-15 ***
## MILES        1.82401    0.06674  27.329  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.638 on 360 degrees of freedom
## Multiple R-squared:  0.6748, Adjusted R-squared:  0.6739
## F-statistic: 746.9 on 1 and 360 DF,  p-value: < 2.2e-16
```

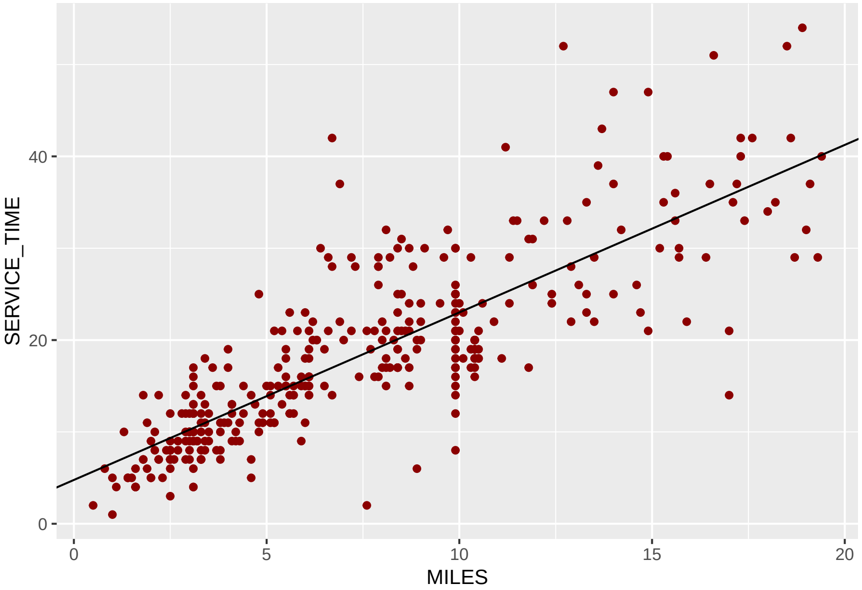


Figure 1: Ordinary linear model

The regression line captures the general trend of the data well, but we can see that a straight line is not the best method of prediction in this case.

Generalized Additive Model

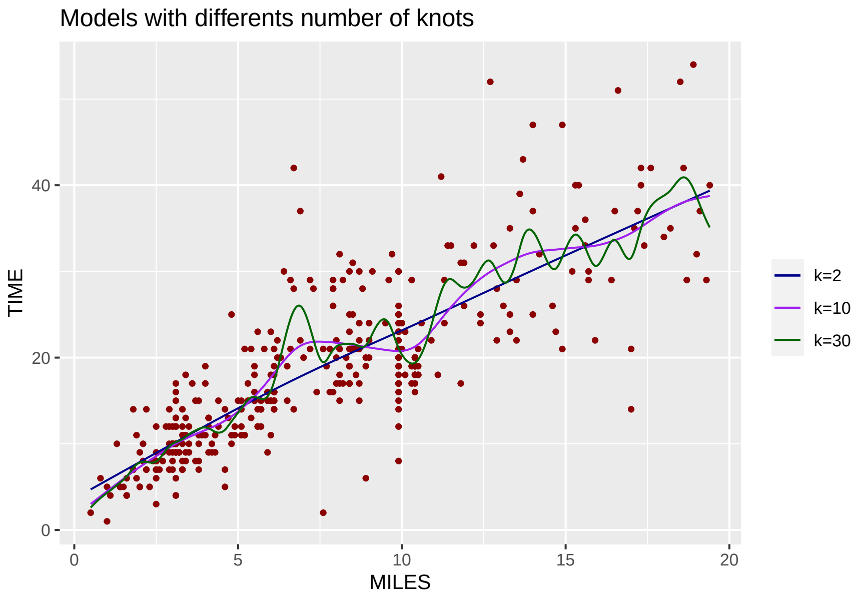
Next, we use gam function from **mgcv** package to create Generalized Additive Model. It is a Generalized Linear Model (GLM) in which the linear predictor is given by a specified sum of smooth functions of the covariates plus a conventional parametric component of the linear predictor.We can define with regression splines using an equation

$$f(x) = \beta_0 + \beta_1 + x + \sum_{k=1}^{K} b_k(x - x_k)$$

We would like to compare a simple linear regression model and a model with an additional hours variable and applied smoothing on the miles variable. AIC value of simple model is equal to 2 283.44, while AIC value of second model equals 2 269.84. With this comparison, we can conclude that the use of splines and day_time variable made the model better fit the data. First, we constructed model using only miles as explanatory variable. In our second model we combined miles and hours as explanatory variables. The best model for these two variables is the model where we used the miles and smoothing term. In this model we also testes the influence of different number of knots. As we know selection of number of knots is crucial, so we compared models with 2, 10 and 30 number of knots. As we see first model is slightly different from the linear model, second seems to adjusts to the punctures correctly. However last model may present overfitting. The conclusion is to delegate the decision about knots number for model as we don't have reason to do it manually.

```
##
## Call: gam(formula = df$SERVICE_TIME ~ s(df$MILES) + df$DAY_TIME)
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -21.8450  -3.1572  -0.5689   2.4478  23.8131
##
## (Dispersion Parameter for gaussian family taken to be 30.2786)
##
##     Null Deviance: 35179.7 on 361 degrees of freedom
## Residual Deviance: 10779.19 on 355.9999 degrees of freedom
## AIC: 2269.841
##
## Number of Local Scoring Iterations: NA
##
## Anova for Parametric Effects
##              Df  Sum Sq Mean Sq  F value  Pr(>F)
## s(df$MILES)   1 23737.9 23737.9 783.9820 < 2e-16 ***
## df$DAY_TIME   1    97.5    97.5   3.2217 0.07352 .
## Residuals   356 10779.2    30.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##             Npar Df Npar F     Pr(F)
## (Intercept)
## s(df$MILES)       3 6.4664 0.0002846 ***
## df$DAY_TIME
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Then, we would like to test the influence of different number of knots in model with only miles as smoothing variable. As we know selection of number of knots i crucial, so we compared models with 2, 10 and 30 number of knots. As we see first model is slightly different from the linear model,so, as we noted earlier, the model only captures the general trend. Second model seems to adjusts to the punctures correctly. Hovewer last model may present overfitting, because is sensitive to the impact of single observations.



Linearity test

Now we will perform test for linearity of f function. Recall: We assume that relation is in the form

$$y_i = f(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{ind.}}{\sim} N\left(0, \sigma_\varepsilon^2\right)$$

and we consider following hypothesis

$$H_0 : f \text{ is linear versus } H_1 : f \text{ is a smooth non-linear function.}$$

We will use F test.

```
linear_model_using_gam <- gam(y~x)
anova(linear_model_using_gam, gam_model, test="F")$"Pr(>F)"[2]
```

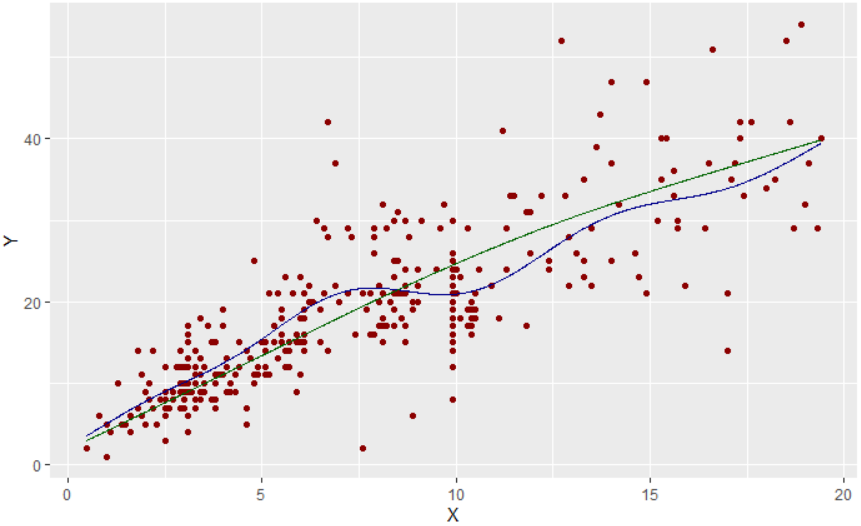
```
## [1] 0.01993326
```

Based on p value, using standard $\alpha = 0.05$ confidence level, we reject H_0. The relation between SERVICE_TIME and MILES isn't strictly linear and using smoothing term should lead us to better model. In next section we will try to add interaction with previously developed feature - DAY_TIME.

Gam with interactions beetwen day time and distance.

```
##
## Call: gam(formula = y ~ s(x, 2))
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -21.2826  -3.0558  -0.6964   2.5025  24.5693
##
## (Dispersion Parameter for gaussian family taken to be 31.3933)
##
##     Null Deviance: 35179.7 on 361 degrees of freedom
## Residual Deviance: 11270.18 on 359 degrees of freedom
## AIC: 2279.966
```

```
##
## Number of Local Scoring Iterations: NA
##
## Anova for Parametric Effects
##             Df Sum Sq Mean Sq F value    Pr(>F)
## s(x, 2)      1  23738 23737.9  756.15 < 2.2e-16 ***
## Residuals 359  11270    31.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##              Npar Df Npar F   Pr(F)
## (Intercept)
## s(x, 2)            1 5.4666 0.01993 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Test framework to automatic feature selection

To see if the other variables in the data set affect service time we used the function step. Gam from the gam library. This method creates all possible models based on possible variables and calculates the **AIC** for each model. The result is a model that uses the variables hour, start, stop and miles as smoothing term.

```
## Start:  y ~ df$MILES + df$HOUR + df$CATEGORY + df$START + df$STOP + df$PURPOSE; AIC= 2241.767
## Step:1 y ~ df$HOUR + s(df$MILES, 2) + df$CATEGORY + df$START + df$STOP +      df$PURPOSE ; AIC= 2228
## Step:2 y ~ df$HOUR + s(df$MILES, 2) + df$START + df$STOP + df$PURPOSE ; AIC= 2226.194
## Step:3 y ~ df$HOUR + s(df$MILES, 2) + df$START + df$STOP ; AIC= 2225.208
```

```
## [1] "df$HOUR"        "s(df$MILES, 2)" "df$START"       "df$STOP"
```

```
##
## Call: gam(formula = y ~ df$HOUR + s(df$MILES, 2) + df$START + df$STOP)
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -17.1976  -2.9199  -0.4793   2.2676  26.2396
##
## (Dispersion Parameter for gaussian family taken to be 25.5231)
##
```

```
##     Null Deviance: 35179.7 on 361 degrees of freedom
## Residual Deviance: 8626.8 on 338 degrees of freedom
## AIC: 2225.208
##
## Number of Local Scoring Iterations: NA
##
## Anova for Parametric Effects
##                 Df  Sum Sq Mean Sq  F value    Pr(>F)
## df$HOUR          1   202.1   202.1   7.9174 0.0051827 **
## s(df$MILES, 2)   1 23596.7 23596.7 924.5239 < 2.2e-16 ***
## df$START        10   791.8    79.2   3.1023 0.0008476 ***
## df$STOP         10  1911.4   191.1   7.4888 8.905e-11 ***
## Residuals      338  8626.8    25.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                Npar Df Npar F     Pr(F)
## (Intercept)
## df$HOUR
## s(df$MILES, 2)       1  13.01 0.0003566 ***
## df$START
## df$STOP
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Metric evaluation for different models

We use k-folds cross-validation to evaluate RMSE, MAE and R^2. In this approach, we split previously filtered data set, into k disjoint subset with (almost) equal size. The model is then trained k times, each time using $k-1$ folds as the training set and the remaining fold as the test set. This ensures that every data point is used for testing exactly once. For ordinary linear model we have ready to use solution, but for gam model we need to implement such functionality (see helpers.R for implementation). We compare 4 models:

Model name	Model formula
LM	SERVICE_TIME ~ MILES
LM with interactions	SERVICE_TIME ~ MILES * DAY_TIME
GAM	SERVICE_TIME ~ s(MILES)
GAM with interactions	SERVICE_TIME ~ s(MILES * DAY_TIME)

Metrics definitons

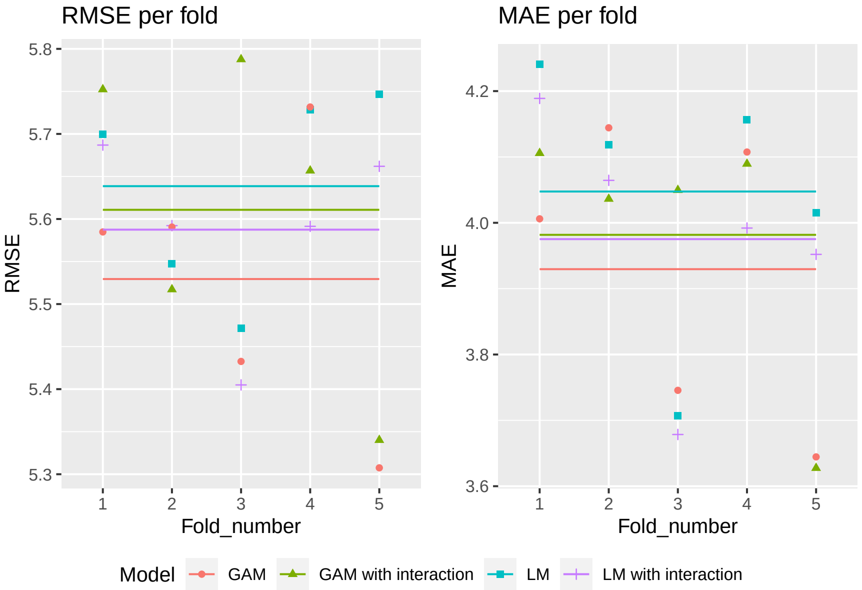
$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

$$\mathrm{R}^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$

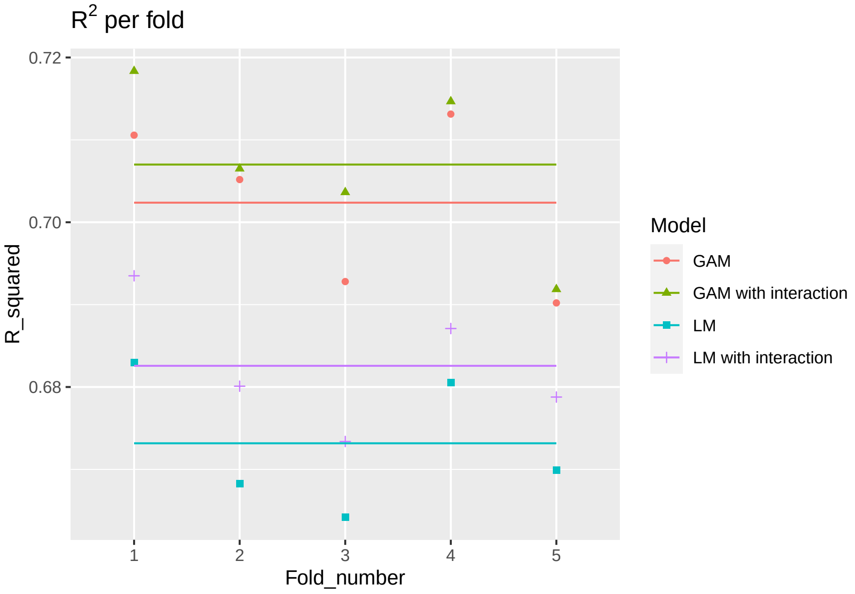
It’s not official R^2 formula, rather intuition that we should have during evaluation of model performance. First two metrics inform us about predictive power of our model (but in slightly different way). We want to minimize them both. R^2 tell how many percent of variance in data we explain (higher=better), but it shouldn't be use as base criterium for model comparison with different numbers of features (always increase when number of features increase), but still can be useful.

Here we use $k = 5$ folds. Models are evaluated on same folds.



On plots we see the GAM has lower RMSE and MAE on average. Adding interaction to GAM seems to be pointless. They contribute nothing to the model. I suspect the number of parameters might be a bit too high or the way we developed *DAY_TIME* predictor it's not optimal. By contrast, we observe the ordinary LM model gain if we add interaction (despite increasing numbers of parameters). So maybe the idea was not so bad.

From the R^2 perspective the both GAM models has significantly better score. However, the number of parameters used in models is notably greater. In fact they ordered by number of parameters. That's only reassures me to not use R^2 as reference here.



To sum up, we would leave idea with adding interaction term to the model as it seems to be redundant and choose non linear one as best.