Uniwersytet Wrocławski

Project proposal

Uber app delivery time estimation

Maciej Szczutko Ewa Stebel

2024-01-28 22:19:25

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

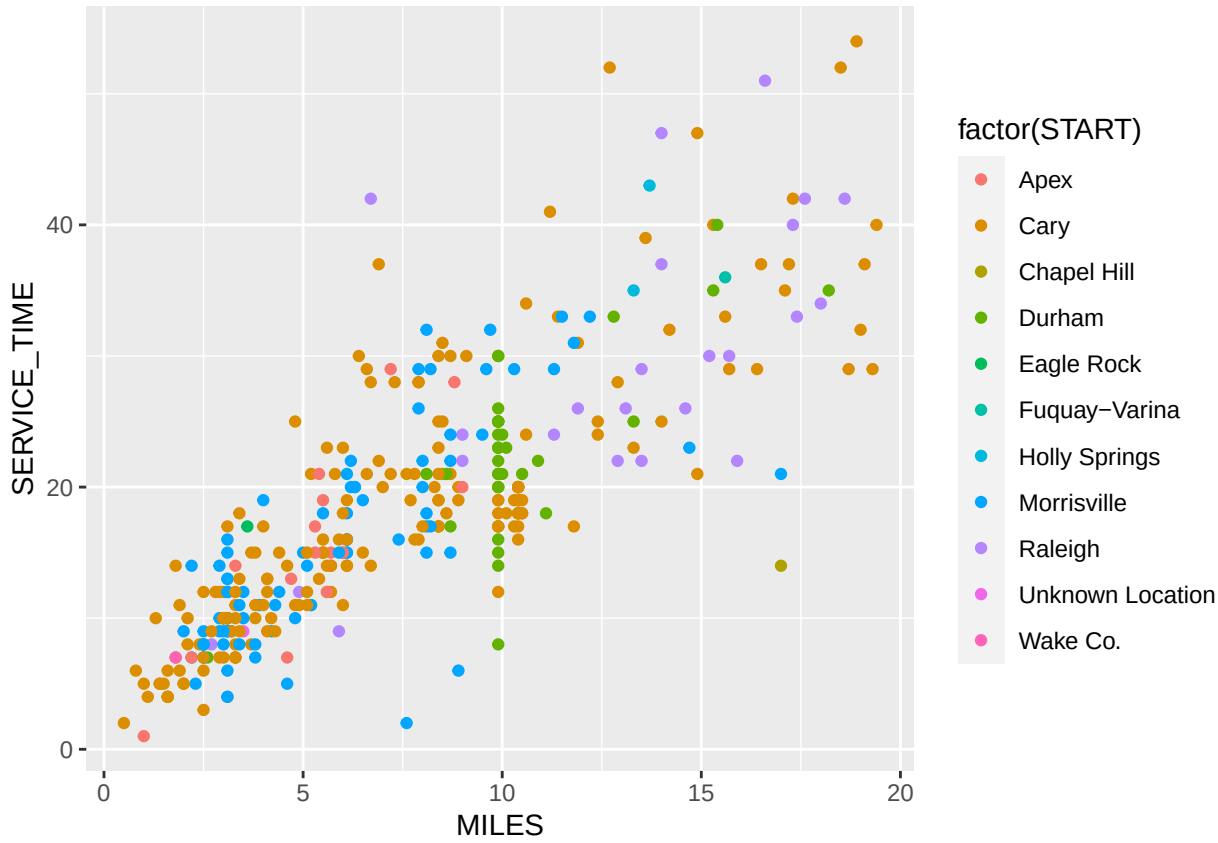
The first thing we would like to do is standard data preprocessing to deeper understanding and find possible garbage in data (non realistic delivery time, NA values etc.). We want to built model (or models) to estimate delivery time. Delivery time variable will be additional column created by transformation 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

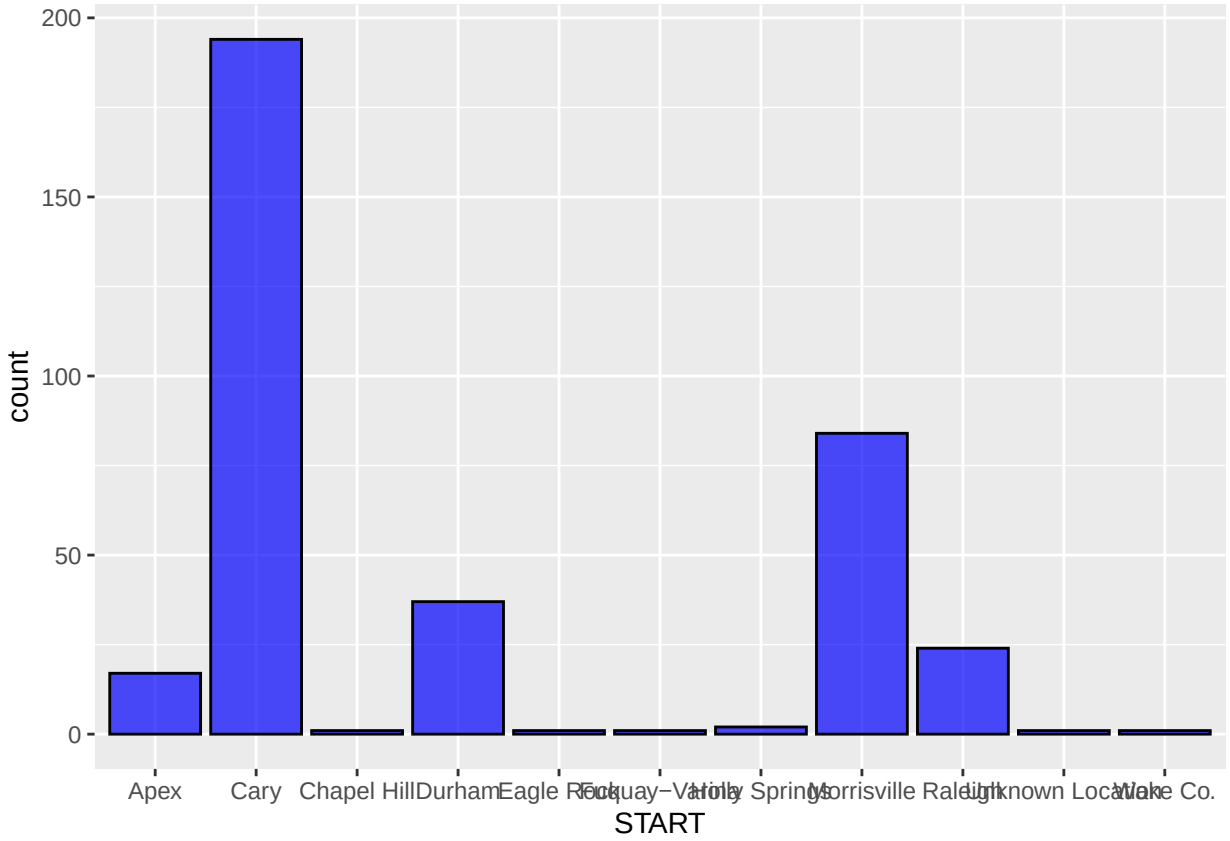
TODO SECTION: Add some basic plots e.g. location popularity chart etc.

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.

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.



We also see that some locations occur more frequently than others, so we also created a histogram of the popularity of a location.



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.

Data cleaning

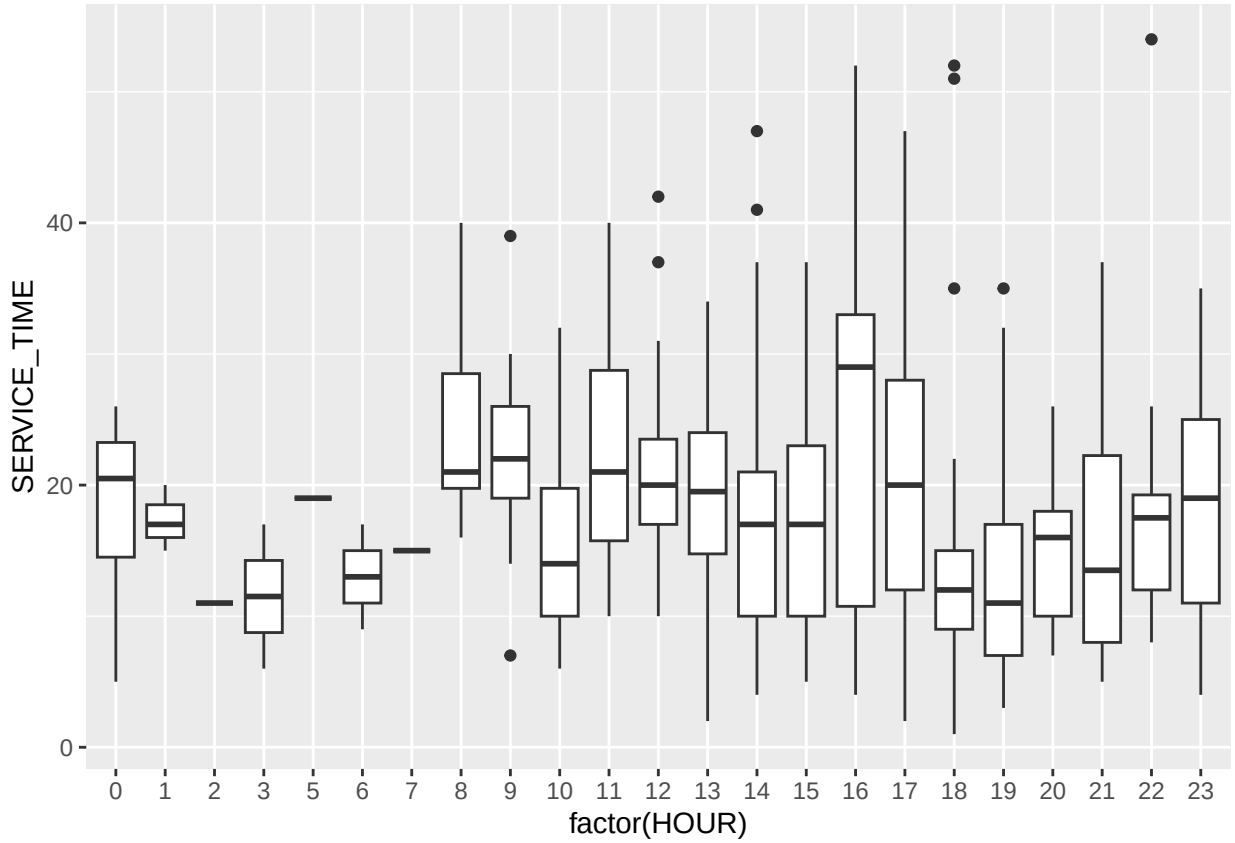
To prepare the data for modelling, we must first clean it. To do this, we replace the NA values in the Purpose column with "Unknown". Other columns do not show missing values, however the start and stop columns show some "Unknown location" values. We also remove the last row of data, which is the summary.

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
```



Based on the boxplots, we see variation in service time by hour of the day, suggesting the importance of this variable

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.

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 explanatory variable - service time and an explanatory variable - miles. The linear regression model has the form

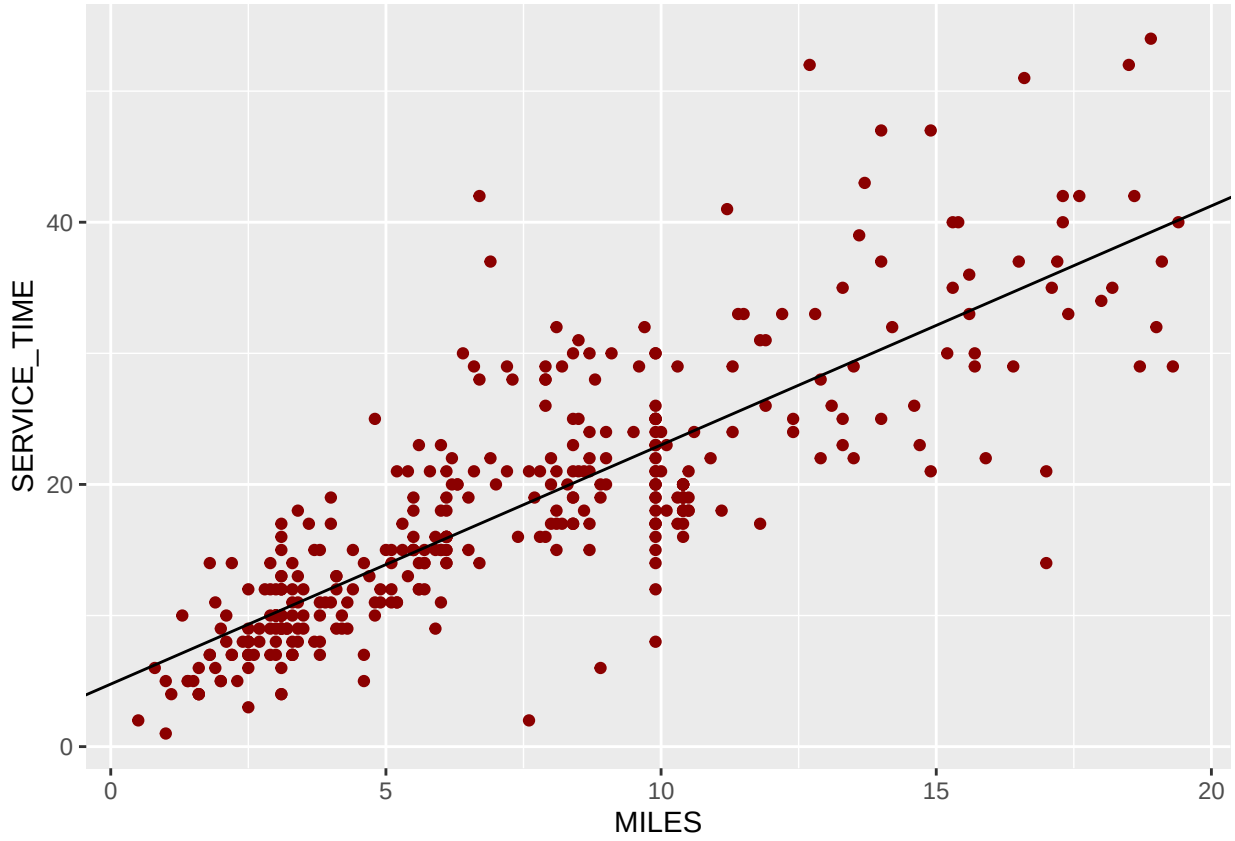
$$f(X) = \beta_0 + \sum_{j=1}^{p} x_j \beta_j,$$

where the β_j ' are unknown parameters or coefficients.

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

```
##
## 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
```



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. Below is presented an example of the model GAM.

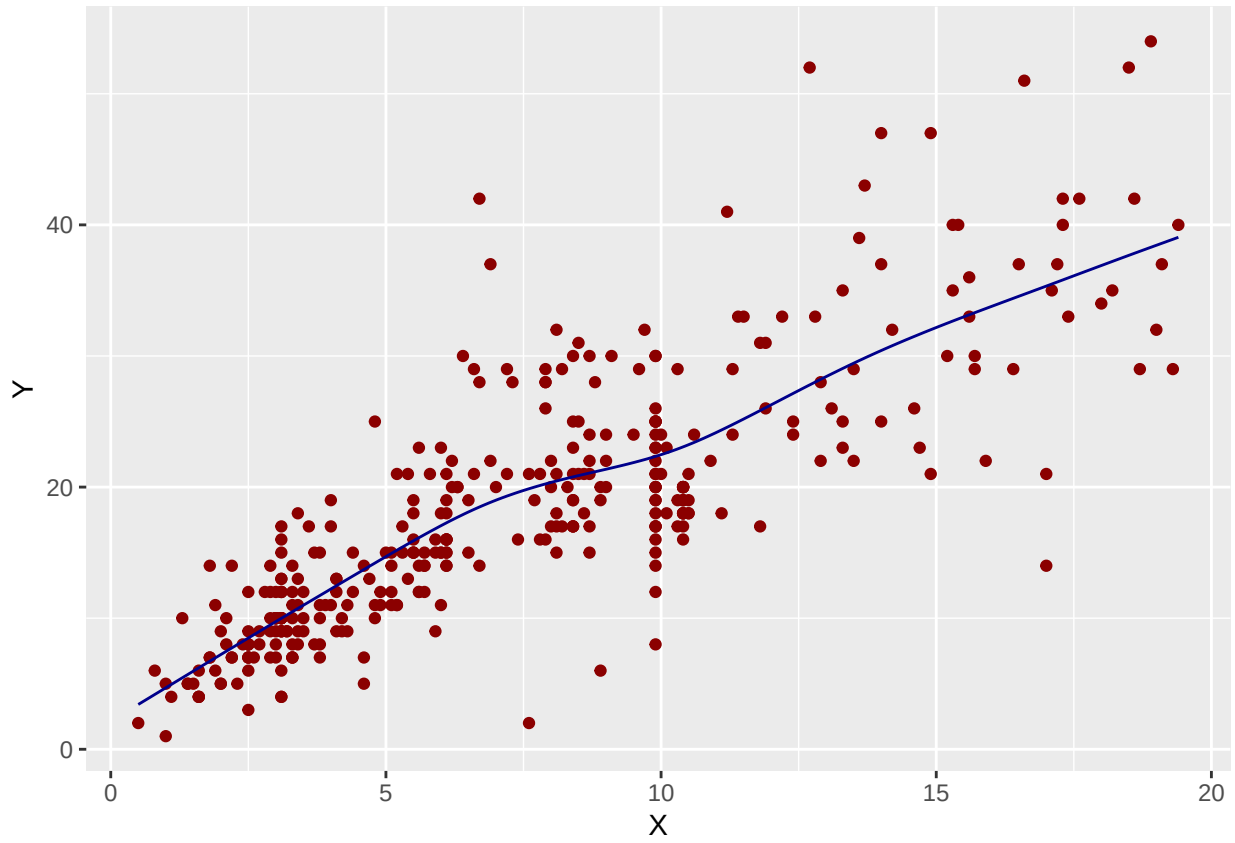
$$log\{E(y_i\} = \alpha + f_1(x_{1i}) + f_2(x_{2i}),$$

where f_1 and f_2 are smooth functions of covariates x_1 and x_2. The log is an example of a link function.

First, we constructed GAM 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 ad smoothing term.

```
##
## Call: gam(formula = df$SERVICE_TIME ~ s(df$MILES) + df$HOUR)
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -21.3633  -2.9940  -0.7069   2.2459  24.1253
##
## (Dispersion Parameter for gaussian family taken to be 30.3951)
##
##     Null Deviance: 35179.7 on 361 degrees of freedom
```

```
## Residual Deviance: 10820.65 on 355.9999 degrees of freedom
## AIC: 2271.231
##
## Number of Local Scoring Iterations: NA
##
## Anova for Parametric Effects
##              Df  Sum Sq Mean Sq F value Pr(>F)
## s(df$MILES)   1 23737.9 23737.9 780.979 <2e-16 ***
## df$HOUR       1    37.9    37.9   1.247 0.2649
## Residuals   356 10820.6    30.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##             Npar Df Npar F     Pr(F)
## (Intercept)
## s(df$MILES)       3 6.5751 0.0002456 ***
## df$HOUR
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Call: gam(formula = y ~ s(x))
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -21.3313  -3.0928  -0.7304   2.4192  24.2072
##
## (Dispersion Parameter for gaussian family taken to be 30.3988)
##
##     Null Deviance: 35179.7 on 361 degrees of freedom
## Residual Deviance: 10852.38 on 356.9999 degrees of freedom
## AIC: 2270.291
##
## Number of Local Scoring Iterations: NA
##
## Anova for Parametric Effects
##            Df Sum Sq Mean Sq F value    Pr(>F)
## s(x)        1  23738 23737.9  780.88 < 2.2e-16 ***
## Residuals 357  10852    30.4
## ---
## 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)              3 6.4628 0.0002858 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



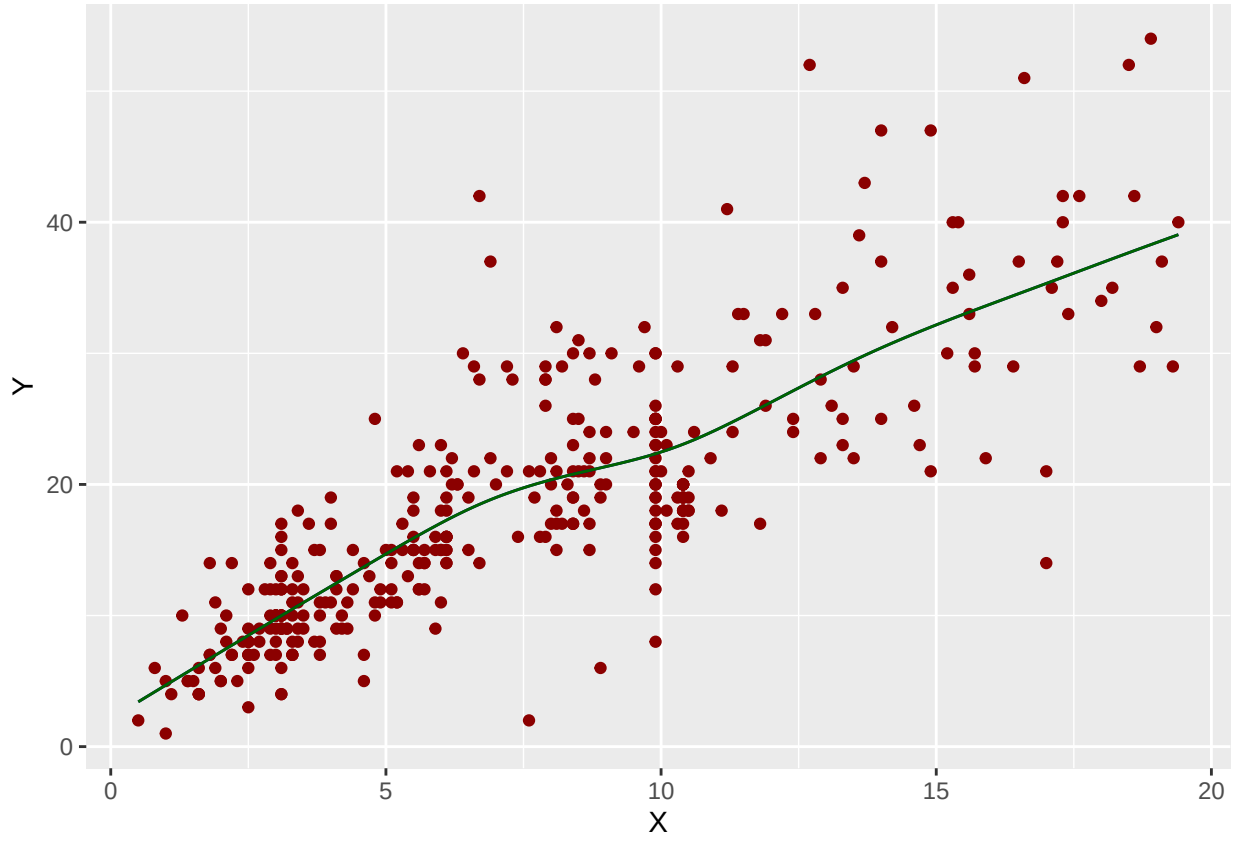
Linearity test

```
## [1] 0.0002857651
```

Gam with interactions beetwen day time and distance.

```
##
## Call: gam(formula = y ~ s(x))
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -21.3313  -3.0928  -0.7304   2.4192  24.2072
##
## (Dispersion Parameter for gaussian family taken to be 30.3988)
##
##     Null Deviance: 35179.7 on 361 degrees of freedom
## Residual Deviance: 10852.38 on 356.9999 degrees of freedom
## AIC: 2270.291
##
## Number of Local Scoring Iterations: NA
##
## Anova for Parametric Effects
##            Df Sum Sq Mean Sq F value    Pr(>F)
## s(x)        1  23738 23737.9  780.88 < 2.2e-16 ***
## Residuals 357  10852    30.4
## ---
## 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)              3 6.4628 0.0002858 ***
## ---
## 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
```

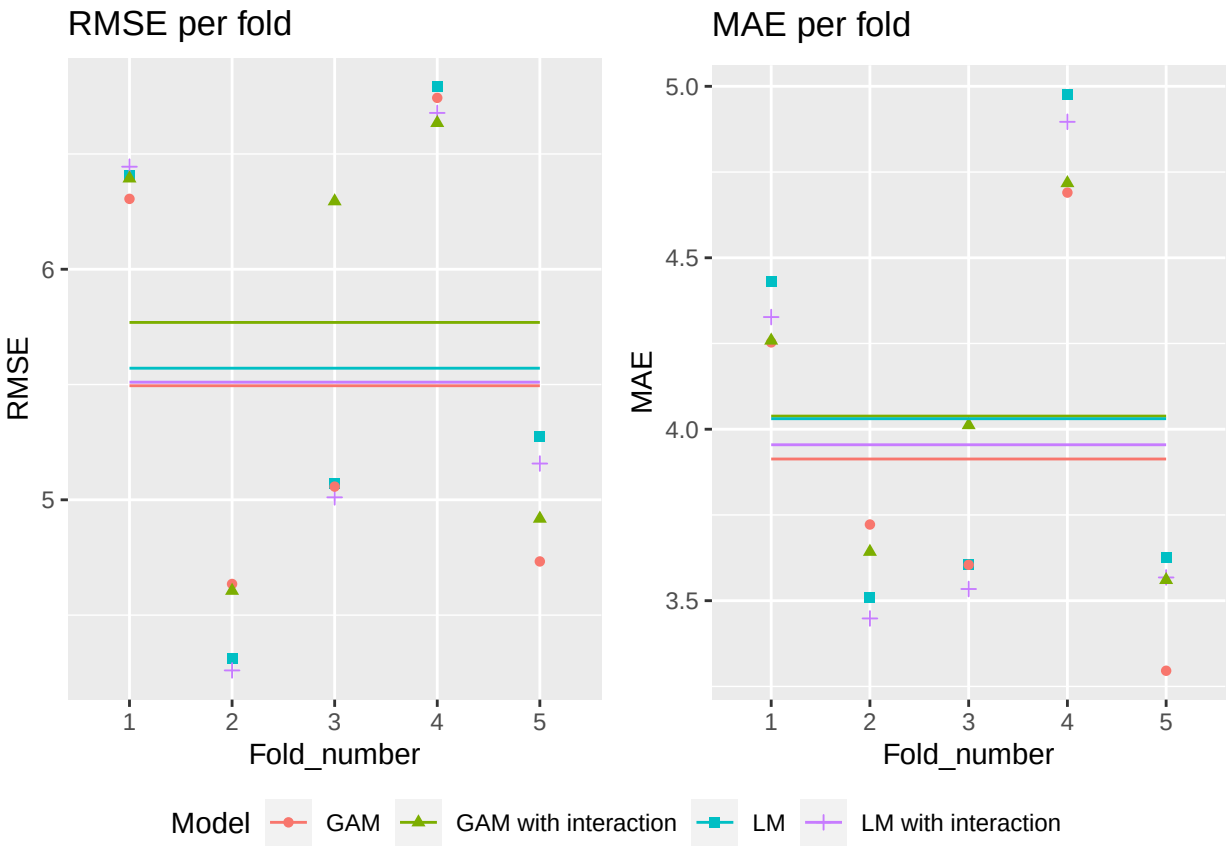
```
## [1] "df$HOUR"        "s(df$MILES, 2)" "df$START"       "df$STOP"
```

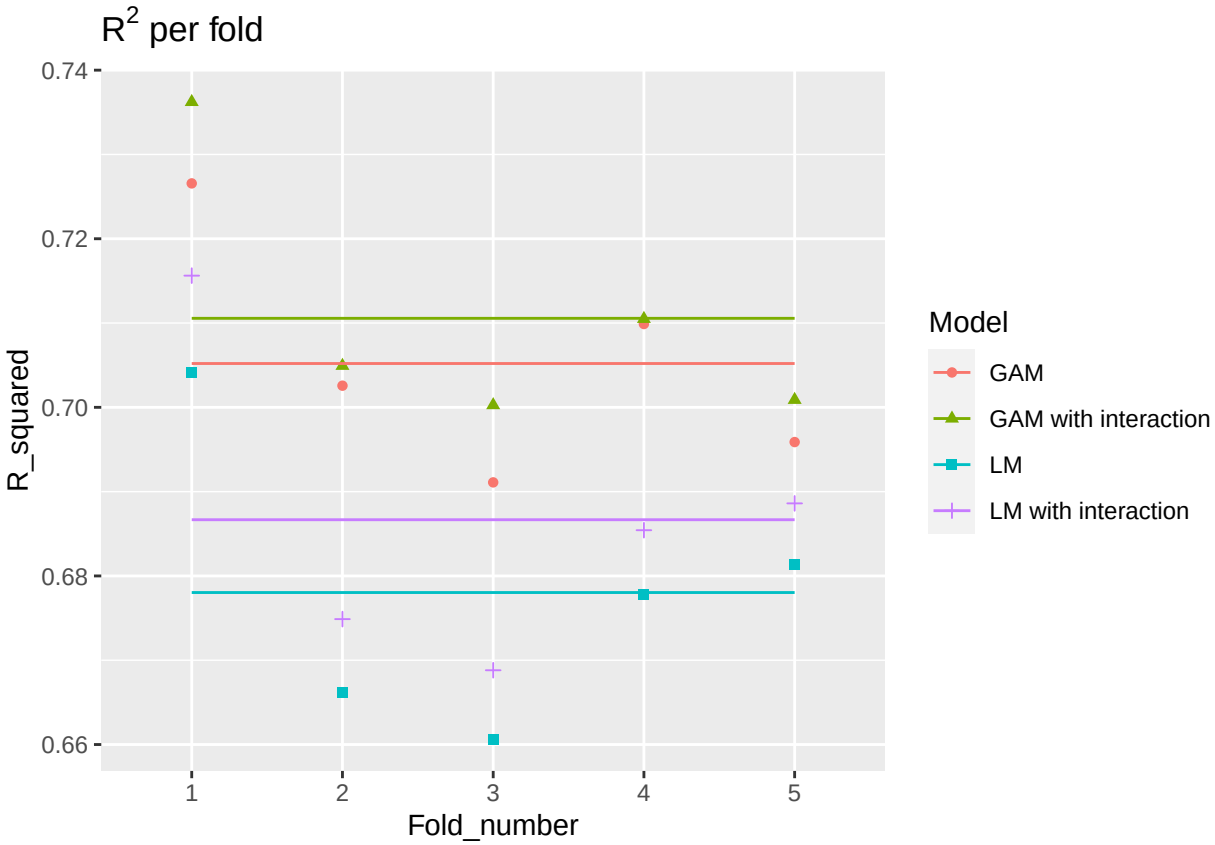
```
##
## Call: gam(formula = y ~ df$HOUR + s(df$MILES, 2) + df$START + df$STOP)
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -17.1976  -2.9199  -0.4793   2.2676  26.2396
##
## (Dispersion Parameter for gaussian family taken to be 25.5231)
```

```
##
##     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 differents model

We use k-folds cross-validation to evaluate **RMSE**, R^2, and their average. 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).


RMSE per fold
RMSE
6
5
1
2
3
4
5
Fold_number
MAE per fold
MAE
5.0
4.5
4.0
3.5
1
2
3
4
5
Fold_number
Model
GAM
GAM with interaction
LM
LM with interaction


R^2 per fold
0.74
0.72
0.70
0.68
0.66
R_squared
1
2
3
4
5
Fold_number
Model
GAM
GAM with interaction
LM
LM with interaction