stat_learning_assignment_09

jshdmm

2024-04-23

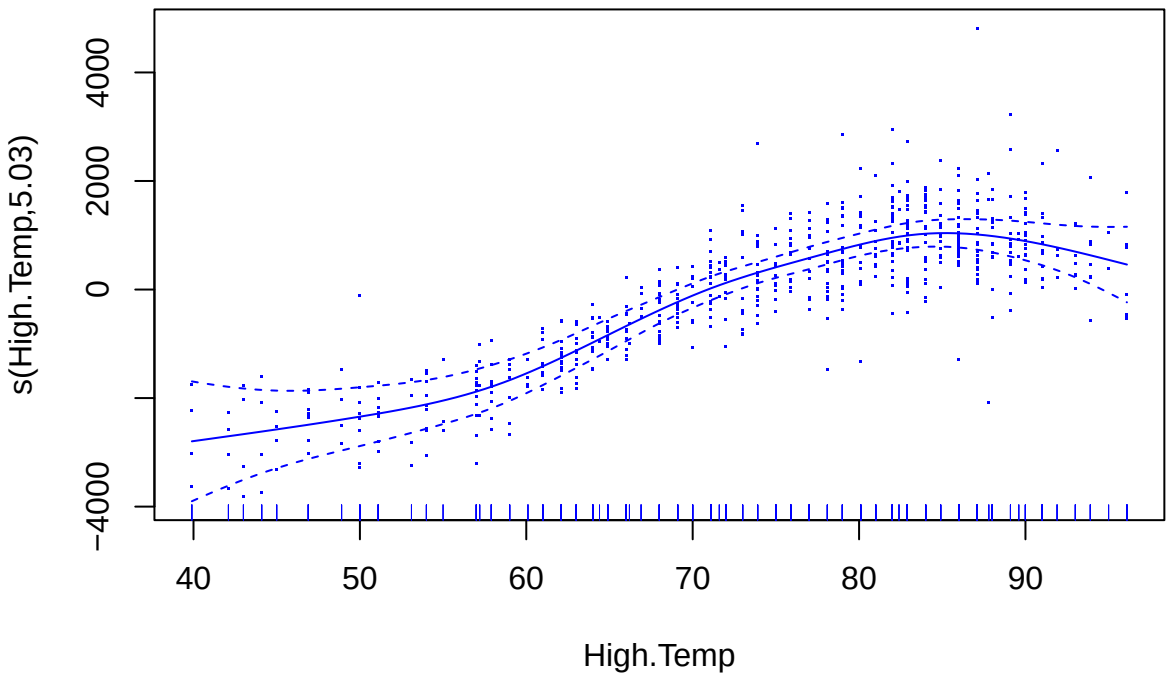
Question a

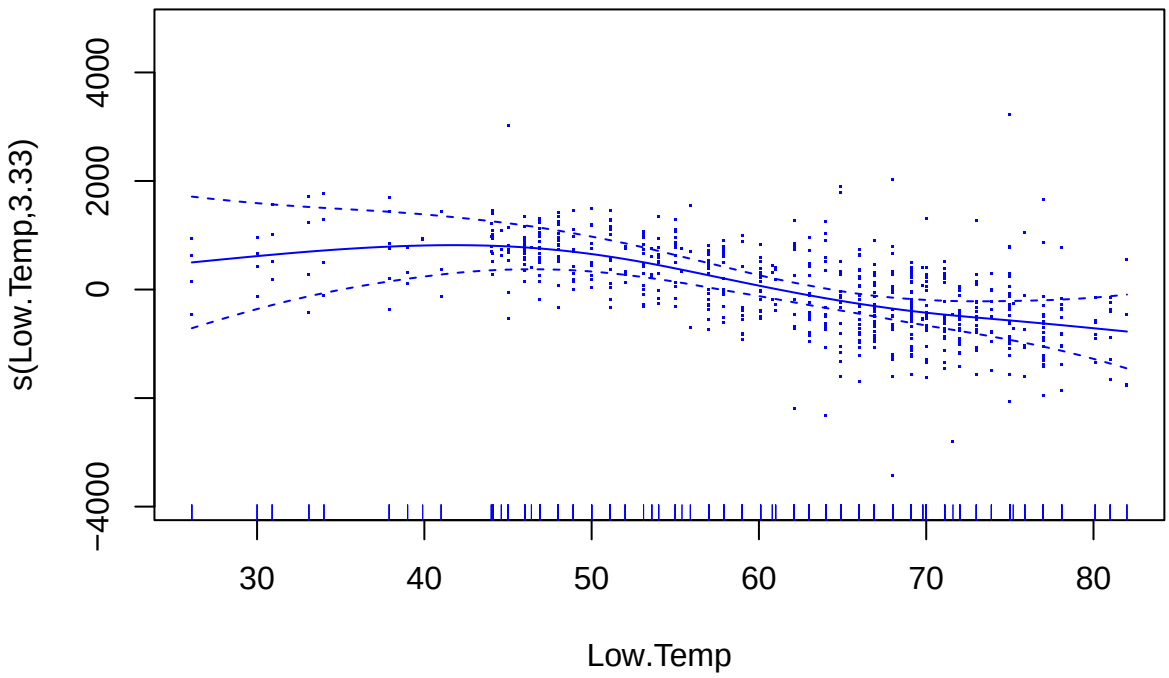
```
set.seed(4036018)

# Define indices for test data - sampling 200 observations for testing
test_indices <- sample(nrow(bike_dat), 200)
test_data <- bike_dat[test_indices, ]
train_data <- bike_dat[-test_indices, ]

# Define and fit the GAM model, factor: bridge, month, exclude: total, day
gam_model <- gam(counts ~ s(High.Temp, bs = "cr", k = 9) +
                   s(Low.Temp, bs = "cr", k = 9) +
                   s(prec, bs = "cr", k = 9) +
                   bridge +
                   month +
                   Day +
                   snow +
                   rain +
                   s(Date, bs = "re"),
                 data = train_data, method = "REML")
```

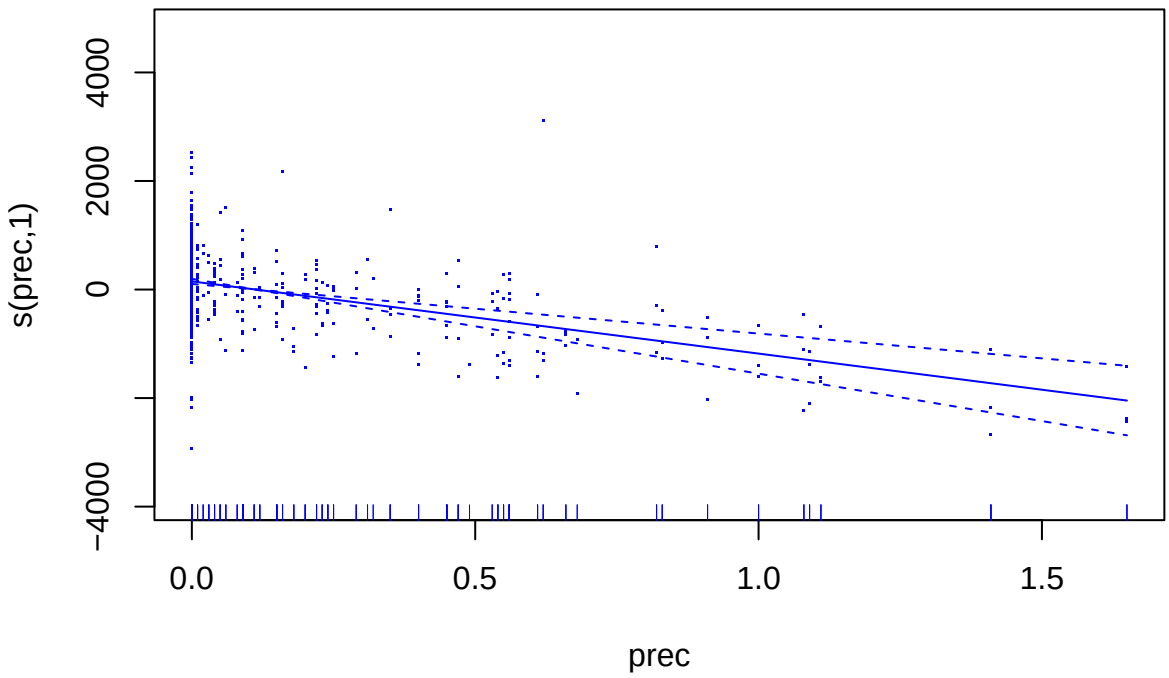
Question b

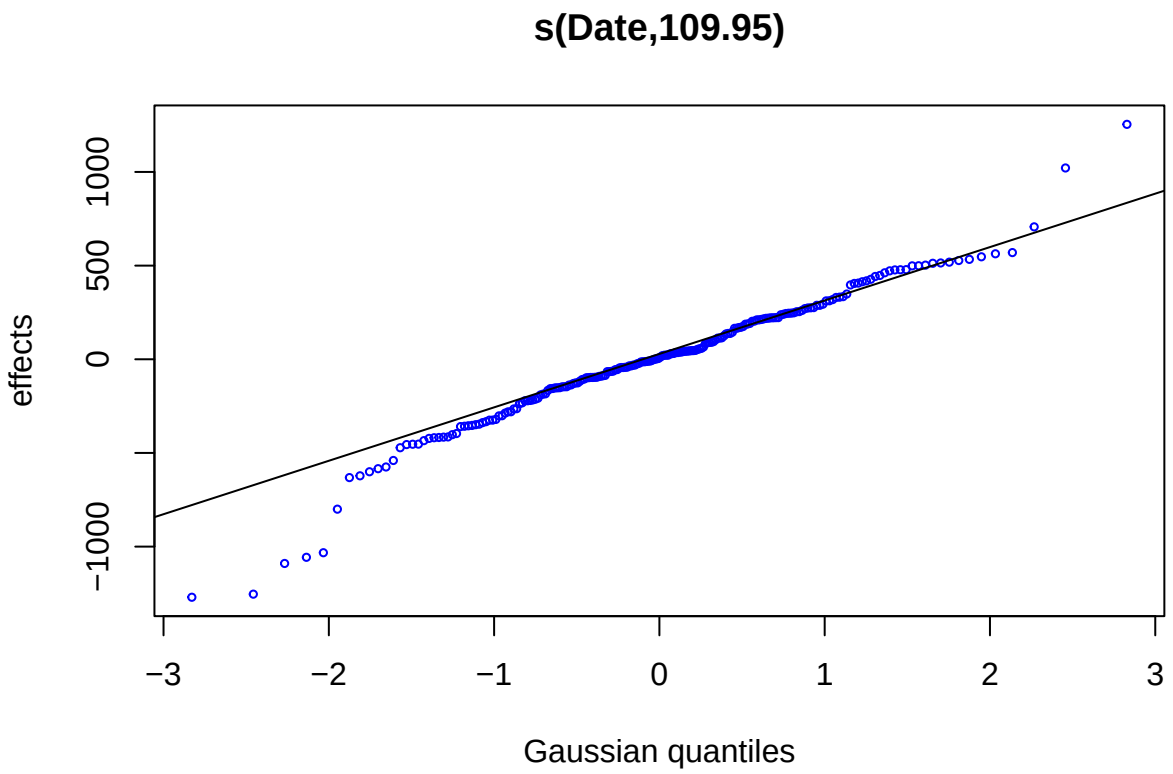
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## counts ~ s(High.Temp, bs = "cr", k = 9) + s(Low.Temp, bs = "cr",
##     k = 9) + s(prec, bs = "cr", k = 9) + bridge + month + Day +
##     snow + rain + s(Date, bs = "re")
##
## Parametric coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2901.73     200.99  14.438  < 2e-16 ***
## bridgeManhattan.Bridge      1978.65      80.41  24.608  < 2e-16 ***
## bridgeWilliamsburg.Bridge   3056.18      81.70  37.408  < 2e-16 ***
## bridgeQueensboro.Bridge     1257.36      79.96  15.724  < 2e-16 ***
## monthAug                     376.07     243.66   1.543  0.12334
```

```
## monthJul                    661.64     243.42   2.718  0.00679 **
## monthJun                    622.51     216.81   2.871  0.00426 **
## monthMay                    228.87     185.76   1.232  0.21848
## monthOct                    335.87     183.56   1.830  0.06786 .
## monthSep                    328.98     216.37   1.520  0.12900
## DayMonday                   191.23     168.98   1.132  0.25830
## DaySaturday                -821.80     164.48  -4.996 8.00e-07 ***
## DaySunday                  -1119.42     163.66  -6.840 2.24e-11 ***
## DayThursday                 410.11     165.70   2.475  0.01364 *
## DayTuesday                  486.41     165.92   2.932  0.00352 **
## DayWednesday                721.81     168.37   4.287 2.16e-05 ***
## snowyes                    -1188.30     664.35  -1.789  0.07425 .
## rainyes                     -531.56     116.42  -4.566 6.22e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                   edf  Ref.df      F p-value
## s(High.Temp)   5.028   5.454 19.130 < 2e-16 ***
## s(Low.Temp)    3.331   3.701  4.577 0.00155 **
## s(prec)        1.004   1.005 40.635 < 2e-16 ***
## s(Date)      109.954 196.000  1.360 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.868   Deviance explained = 89.6%
## -REML =   5201  Scale est. = 4.9751e+05  n = 656


##
## Family: gaussian
## Link function: identity
##
## Formula:
## counts ~ s(High.Temp, bs = "cr", k = 9) + s(Low.Temp, bs = "cr",
##     k = 9) + s(prec, bs = "cr", k = 9) + bridge + month + Day +
##     snow + rain + s(Date, bs = "re")
##
## Parametric Terms:
##        df       F  p-value
## bridge  3 495.652  < 2e-16
## month   6   2.123   0.0493
## Day     6  35.134  < 2e-16
## snow    1   3.199   0.0743
## rain    1  20.847 6.22e-06
##
## Approximate significance of smooth terms:
##                   edf  Ref.df      F p-value
## s(High.Temp)   5.028   5.454 19.130 < 2e-16
## s(Low.Temp)    3.331   3.701  4.577 0.00155
## s(prec)        1.004   1.005 40.635 < 2e-16
## s(Date)      109.954 196.000  1.360 < 2e-16
```


s(High.Temp,5.03)
−4000
0
2000
4000
40
50
60
70
80
90
High.Temp


s(Low.Temp,3.33)
4000
2000
0
−4000
30
40
50
60
70
80
Low.Temp


s(prec,1)
4000
2000
0
−4000
0.0
0.5
1.0
1.5
prec



As one can infer from the coefficients July is the busiest month ($„_{July} = 661.64$) and Wednesday was the busiest weekday ($„_{Wednesday} = 721.81$). In our model snow does not significantly influence bike counts (p = 0.07425). However, this still shows a trend and $„_{Snow} = -1188.30$ suggests a strong negative impact of snowing days on bicycle counts with some statistical uncertainty. Rain on the other hand does negatively influence the response (ß(rain) = -531.56, $p < .001$). The coefficient for snowing days is stronger and therefore one can expect snow days to more heavily influence the decrease in cycling both compared to dry and rain days, with some statistical uncertainty.

Both lowest and highest temperature are statistically significant (both $p < 0.05$). Precipitation seems to have a very linear negative effect on the counts both by visual inspection and through the edf value of 1.004. The highest temperature peaks have a strongly non-linear positive effect on the counts (edf = 5.028), and lowest temperature has a weaker non-linear effect on counts (edf = 3.331).

When looking at the plot of the random intercept one can see that the middle range of the data is quite normally distributed. When looking at the S-shape of the plot however, this indicates heavy tails of the distribution and more extreme values on the outer ends of the distribution and deviations from normality. So there might be times in the year that have a bigger impact on counts than the rest of the year.

Question c

The MSE on the test data is 607821. The variance of the counts in the test data is a lot higher (3340766), suggesting a good fit of the fitted GAM model.

```
## [1] "Mean Squared Error on Test Set: 607821.218797663"
```

```
## [1] "Variance of the counts in the test data: 3340766.70288945"
```