Multiple regression and smoothing splines for a Bayesian model: predicting daily ozone levels

Cimmaron Yeoman

2023-04-25

Introduction

This data set includes the environmental variables ozone, humidity, and temperature, along with day of the week and day of the year. Ozone has some level of a relationship with human disturbances, especially actions which alter temperature, such as vehicle use and and greenhouse gas outputs. This report will attempt to predict the daily ozone level, our response, with a multiple regression Bayesian framework. A smoothing spline will be implemented to help make predictions and identify patterns in the data set.

Data organization and cleaning

The data set contains categorical and index variables including days of the year, and days of the week. These values cannot be standardized before building models. I insured both of these variables were stored as integers. Ozone (**OZ**), humidity (**H**), and temperature (**FT**) in degrees Fahrenheit were all standardized and renamed.

Exploring the data

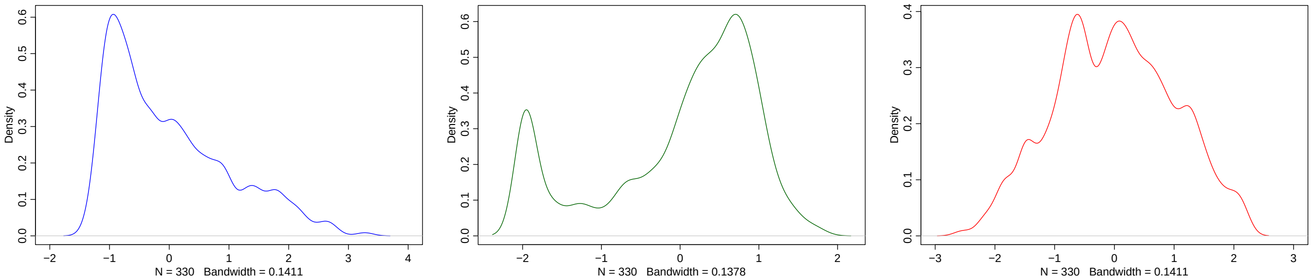
After standardizing, we can see from the data set summary that the average values for ozone, humidity, and temperature (unsquared) were at or around zero, and 75% of the observations fell below roughly 0.75 for each.

```
##        OZ                H                  FT          
##  Min.   :-1.3451   Min.   :-1.9698   Min.   :-2.54203  
##  1st Qu.:-0.8458   1st Qu.:-0.5603   1st Qu.:-0.74381  
##  Median :-0.2217   Median : 0.2955   Median : 0.01698  
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000  
##  3rd Qu.: 0.6521   3rd Qu.: 0.7485   3rd Qu.: 0.70860  
##  Max.   : 3.2734   Max.   : 1.7553   Max.   : 2.16101  
```

Density plots of ozone, humidty and temperature

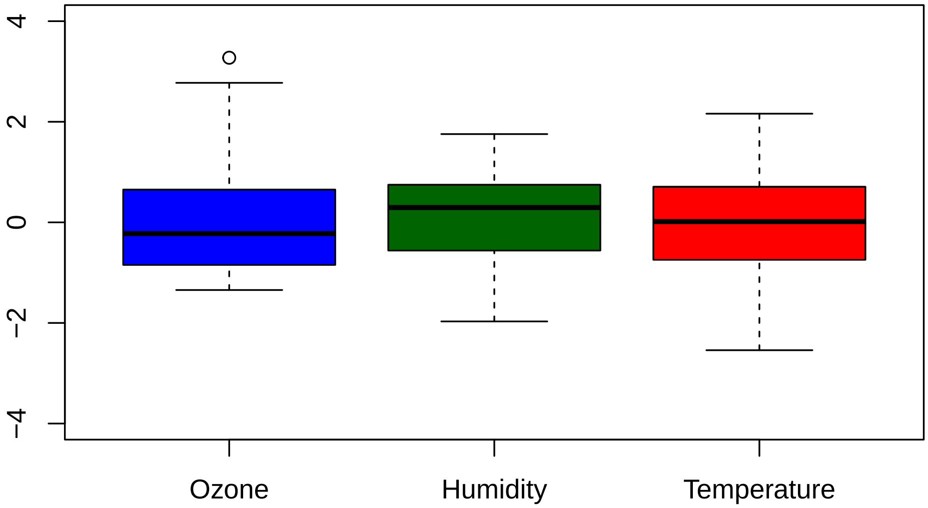
The density plots below show the distribution of the standardized variables. Ozone was skewed right and had a long tail, but if you compare this to the boxplot of ozone, the tail can be explained by some outliers. Humidity looked a bit odd with a bimodal appearance, although the shorter peak on the left was clearly much smaller than the peak on the right. Temperature had a double peak appearance as well. This can be explained by seasonal differences most likely, especially temperature.

Figure 1



Boxplots of ozone, humidity, and temperature

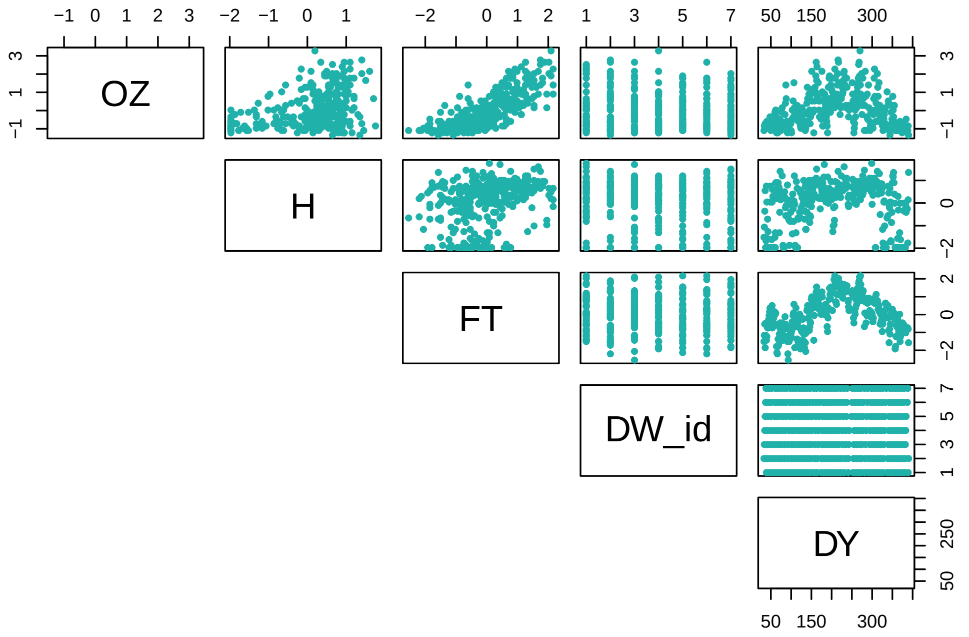
Figure 2



You can see in the scatter plots why it is necessary to standardize varibles used in models, first. Without their varying scales, they can be compared more easily against each other. The potential outlier is visible in the ozone boxplot.

Scatter plots and relationships, spline candidates

Figure 3



Ozone had a positive linear relationship with temperature (**FT**), and some level of overlapping observations for humidity and days of the year. The days of the week variable (**DW_ID**) needs to be properly indexed into any models, so we can actually observe its relationship to ozone. Days of the week produces a concave shape when paired with different variables. These patterns could work for building the smooths.

Creating smoothing splines

I knew in advance there were no missing cases, but I stored the data set **d** as a data frame using the complete.cases function, and renamed it **d2**. Ruling out missing cases must be done before modelling. The final choice of knots for the smooths were decided after trying a few different values. Degrees was kept at 3 for cubic splines. One other version like the spline set up below was also created for a second smooth, but was hidden to save space. Only the knots changed in this version.

```
library(splines)
d2 <- d[complete.cases(d), ]
num_knots <- 15
knot_list <- quantile(d$DY, probs = seq(0, 1, length.out = num_knots))
```

```
B <- bs(d2$DY,
        knots = knot_list[-c(1, num_knots)],
        degree = 3,
        intercept = TRUE
        )
```

First model

This model predicts mean ozone using an indexed days of the week variable (**DW_id**), coefficients for humidity (**H**) and temperature (**T**), and the spline created from the days of the year variable (**DY**).

```
set.seed(41)
flist1 <- alist(ozone ~ dnorm(mu, sigma),
               mu <- a[DW_id] + bH*H + bFT*FT + B %*% w,
               a[DW_id] ~ dnorm(0, 1),
               bH ~ dnorm(0, 1),
               bFT ~ dnorm(0, 1),
               w ~ dnorm(0, 1),
               sigma ~ dexp(1)
              )
mod1 <- quap(flist = flist1, data = list(ozone = d2$OZ,
                                         B = B, H = d2$H, FT = d2$FT,
                                         DW_id = d2$DW_id),
             start = list(w = rep(0, ncol(B)))
            )
precis(mod1, depth = 2, prob = 0.93)
```

```
##               mean         sd         3.5%      96.5%
## w[1]   0.451964527 0.36361207 -0.206868060 1.11079711
## w[2]  -0.368542903 0.39449038 -1.083324227 0.34623842
## w[3]  -0.335644552 0.37562431 -1.016242253 0.34495315
## w[4]   0.624508201 0.34262345  0.003705113 1.24531129
## w[5]   0.343281003 0.31409134 -0.225824456 0.91238646
## w[6]   0.322016224 0.32030814 -0.258353515 0.90238596
## w[7]   0.504390622 0.32842865 -0.090692759 1.09947400
## w[8]  -0.456830799 0.30803120 -1.014955817 0.10129422
## w[9]   0.285755209 0.32002074 -0.294093793 0.86560421
## w[10] -0.509313972 0.32751800 -1.102747335 0.08411939
## w[11]  0.075836336 0.31751899 -0.499479716 0.65115239
## w[12] -0.470020092 0.31970029 -1.049288468 0.10924828
## w[13] -0.026571503 0.31321723 -0.594093140 0.54095013
## w[14] -0.234068962 0.33440959 -0.839989266 0.37185134
## w[15] -0.016139909 0.37429484 -0.694328722 0.66204890
## w[16]  0.059036565 0.40124414 -0.667981976 0.78605511
## w[17] -0.299180814 0.36858980 -0.967032611 0.36867098
## a[1]   0.024547451 0.21828699 -0.370969069 0.42006397
## a[2]   0.054748242 0.21789969 -0.340066535 0.44956302
## a[3]  -0.056307418 0.21835110 -0.451940103 0.33932527
## a[4]  -0.052796266 0.21854061 -0.448772336 0.34317980
## a[5]   0.053204701 0.21862893 -0.342931394 0.44934080
## a[6]   0.007721129 0.21836213 -0.387931549 0.40337381
## a[7]  -0.080642660 0.21848564 -0.476519115 0.31523379
## bH     0.210946968 0.03825749  0.141627816 0.28026612
## bFT    0.810592837 0.05094025  0.718293649 0.90289202
## sigma  0.545086760 0.02120404  0.506666939 0.58350658
```

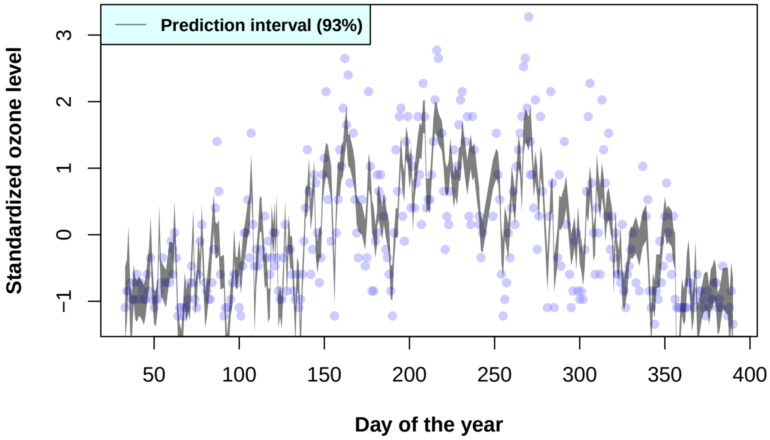
The values for w were forced to start at 0 and the model selected the values for the posterior. The spline matrix, **B** represents 330×17. Matrix multiplication was performed with the %*% operator, multiplying it with 17×1 vector, producing an object with 330 observations. Statistical Rethinking describes this as a sum linear predictor for each day of the year.

The indexed day of the week (**DW_id**) variable provides an estimate of the ozone levels during different days of the week. Tuesdays and Fridays (days 2, 5) seem to have higher ozone levels, and Sundays (day 7) have the lowest ozone levels.

Mean posterior predictive intervals

This is a plot of the 93% posterior predictive interval for mean ozone (**OZ**), plotted against day of the year (**DY**). The shade includes the smooth, the day of the week (**DW_id**) index, humidity (**H**), and temperature (**FT**). It looks very scraggly and chaotic, although we can see it roughly follows the more densely clustered areas of ozone level plotted with day of the year.

Figure 4



Thoughts on the model

The model predicting the response of ozone may potentially be too complex or over-fitting. The interval combines many variables though which I assume is why it is more narrow as well. We at least get a rough idea of the pattern of ozone levels against the days of the year. Ozone levels appear higher during warmer months as they tend to peak about halfway through each year.

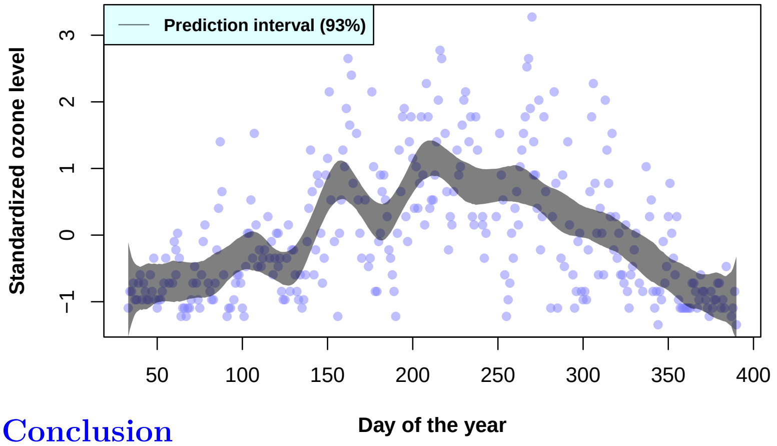
A simpler model and smooth

This model only uses an intercept term and the B spline variable of day of the year (**DY**) to predict ozone. While it does not include nearly as much useful information, it managed to produce a relatively smooth shade that follows the distribution of ozone. The shade provides a wider prediction interval that is less specific, but still follows the clusteres of ozone level measurements across each day of the year

```
set.seed(41)
flistSS1 <- alist(ozone ~ dnorm(mu, sigma),
               mu <- b0 + B2 %*% w,
               b0 ~ dnorm(0, 1),
               w ~ dnorm(0, 1),
               sigma ~ dexp(1)
              )
modSS1 <- quap(flist = flistSS1, data = list(ozone = d2$OZ,
                                        B2 = B2),
             start = list(w = rep(0, ncol(B2)))
            )
precis(modSS1, depth = 2, prob = 0.93)
```

```
##              mean         sd        3.5%        96.5%
## w[1]  -0.7032022 0.44217730 -1.50438794  0.097983582
## w[2]  -0.6567995 0.47989414 -1.52632483  0.212725809
## w[3]  -0.5057260 0.46174932 -1.34237451  0.330922548
## w[4]  -0.6250522 0.41004674 -1.36802024  0.117915891
## w[5]   0.2544504 0.38772850 -0.44807902  0.956979776
## w[6]  -0.9242183 0.38190511 -1.61619621 -0.232240331
## w[7]   1.7971669 0.39438155  1.08258277  2.511751057
## w[8]  -0.3673697 0.37877323 -1.05367291  0.318933595
## w[9]   1.7847010 0.37635320  1.10278266  2.466619429
## w[10]  0.6268467 0.39702202 -0.09252175  1.346215113
## w[11]  1.1163951 0.38495519  0.41889070  1.813899546
## w[12]  0.4032846 0.38861466 -0.30085047  1.107419622
## w[13]  0.2608381 0.38613424 -0.43880267  0.960478832
## w[14] -0.2360876 0.41207614 -0.98273271  0.510557602
## w[15] -0.7264371 0.45660766 -1.55376940  0.100895186
## w[16] -0.8199309 0.48329937 -1.69562615  0.055764412
## w[17] -0.8169919 0.45329465 -1.63832135  0.004337468
## b0    -0.1381324 0.23966843 -0.57239022  0.296125353
## sigma  0.7388255 0.02884747  0.68655644  0.791094513
```

Figure 5



Conclusion

Using variables such as humidity, temperature, and days of the year in particular, can create a model which predicts daily ozone levels. Using a multiple regression Bayesian model we can create a narrow 94% prediction interval smooth combined with selected predictor variables. When we simplify the model and decrease the number of predictors, we can make a wider 93% prediction interval that is more general, but follows the pattern of observed ozone levels. Ozone appeared to be seasonally influenced, with higher levels in the summer months, as ozone requires about 17 ◦ C to form. Humidity may be useful but appears less important than temperature or day of the year. Ozone levels may be higher during weekdays, than weekends, possibly explained by less traffic and commuting to work. Temperature could also increase depending on the day of the week in urban areas with large concentrations of vehicles.