hw3

Xingwen Wei, Xin Hu, Liding Li

March 6, 2021

Question 1

The true model is a polynomial of degree 3.

```
##                                      Bias Variance
## Linear regression                    high low
## Polynomial regression with degree 3  low  low
## Polynomial regression with degree 10 low  high
```

Question 2

a

As $\lambda \to \infty$, $\hat{g_1}$ will have all $g^{(3)}(x) = 0$ and $\hat{g_2}$ will have all $g^{(4)}(x) = 0$. So this is similar to constraining $\hat{g_1}$ to have degree less than 3 and $\hat{g_2}$ less than 4. Thus, $\hat{g_2}$ will always have smaller or equal training error than $\hat{g_1}$.

b

On one hand, if the true curve has degree higher than or equal to 3, $\hat{g_1}$ will not be able to capture it at all, while $\hat{g_2}$ can capture it. So $\hat{g_2}$ will have the smaller test error in this case. On the other hand, if the true curve has degree smaller than 3, $\hat{g_2}$ may pick some noise up as signal and overfits the training data, while $\hat{g_1}$ will not. So $\hat{g_1}$ will have the smaller test error in this case.

c

For $\lambda = 0$, $\hat{g_1} = \hat{g_2}$. So they will have the same training and test error.

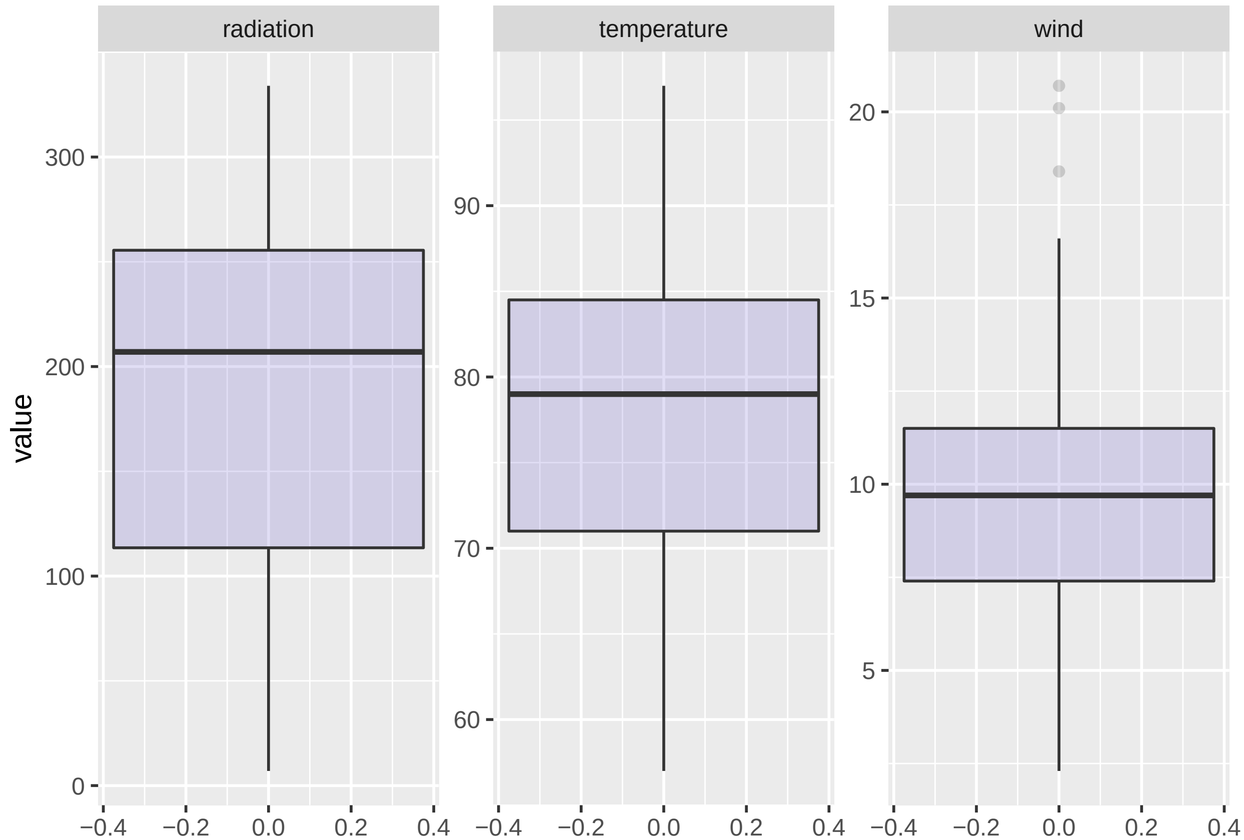
Question 3

a

Exploratory Data Analysis

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
##      ozone         radiation      temperature         wind
##  Min.   :  1.0   Min.   :  7.0   Min.   :57.00   Min.   : 2.300
##  1st Qu.: 18.0   1st Qu.:113.5   1st Qu.:71.00   1st Qu.: 7.400
##  Median : 31.0   Median :207.0   Median :79.00   Median : 9.700
##  Mean   : 42.1   Mean   :184.8   Mean   :77.79   Mean   : 9.939
##  3rd Qu.: 62.0   3rd Qu.:255.5   3rd Qu.:84.50   3rd Qu.:11.500
##  Max.   :168.0   Max.   :334.0   Max.   :97.00   Max.   :20.700
```



According to the model summary, the linear model we choose is

$$\text{ozone}^{\frac{1}{3}} = 0.001 \times \text{radiation} + 0.058 \times \text{temperature} - 0.066 \times \text{wind} - 0.852$$

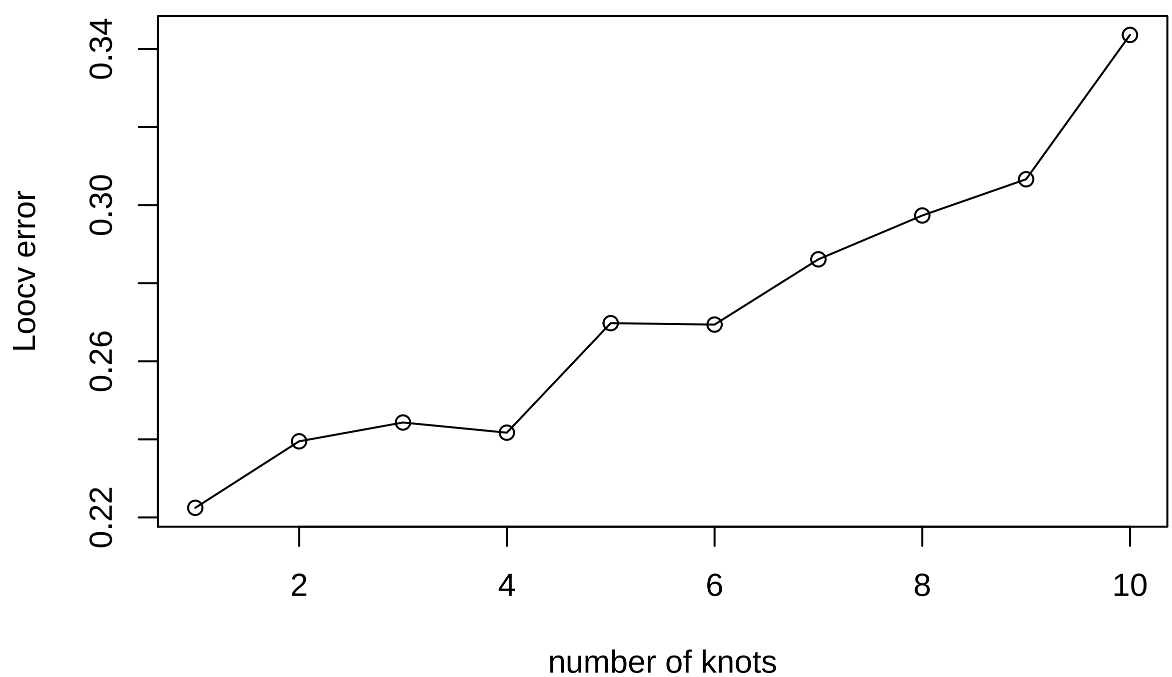
This is not a very satisfactory model as the $R^2 = 0.71$.

```
##
## Call:
## lm(formula = cbr ~ radiation + temperature + wind, data = ozone[train,
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94503 -0.40230 -0.00071  0.27566  1.50475
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.8521380  0.6449384  -1.321 0.190538
## radiation    0.0016477  0.0006398   2.575 0.012037 *
## temperature  0.0579790  0.0071750   8.081 9.93e-12 ***
## wind        -0.0656469  0.0180658  -3.634 0.000516 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4999 on 73 degrees of freedom
## Multiple R-squared:  0.7111, Adjusted R-squared:  0.6992
## F-statistic:  59.9 on 3 and 73 DF,  p-value: < 2.2e-16
```

b

We use LOOCV to find to optimal number of knots. According to the plot below, we find that we get the best result when number of knots is 2.

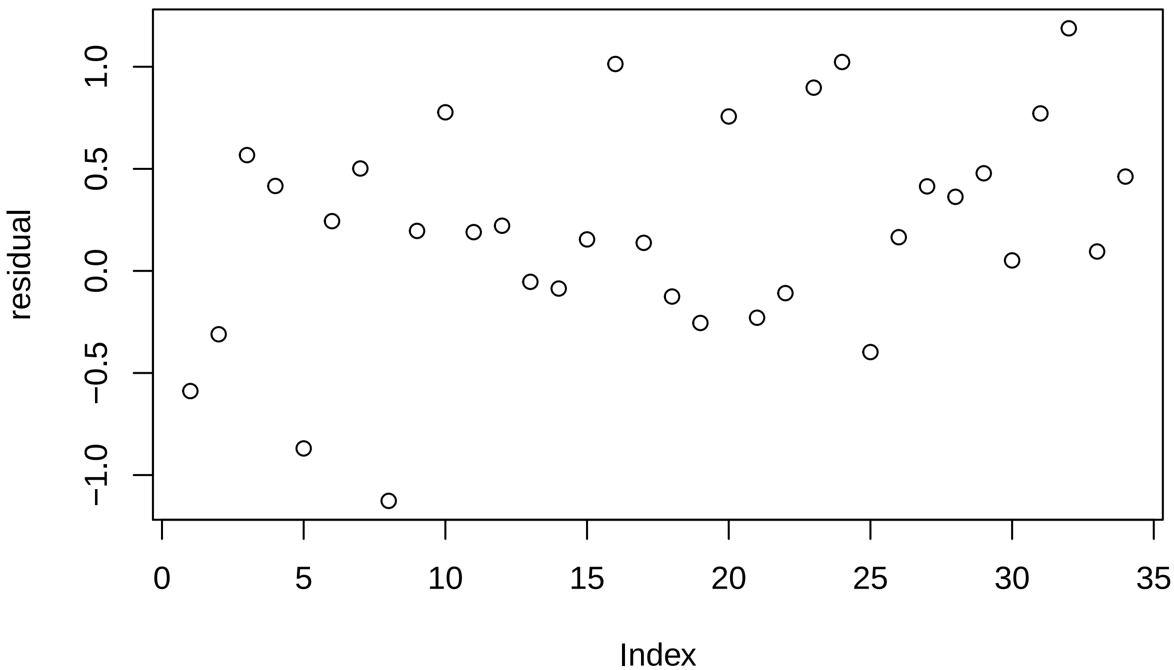
```
## Warning: package 'gam' was built under R version 3.6.3

## Loading required package: splines

## Loading required package: foreach

## Warning: package 'foreach' was built under R version 3.6.3

## Loaded gam 1.20
```

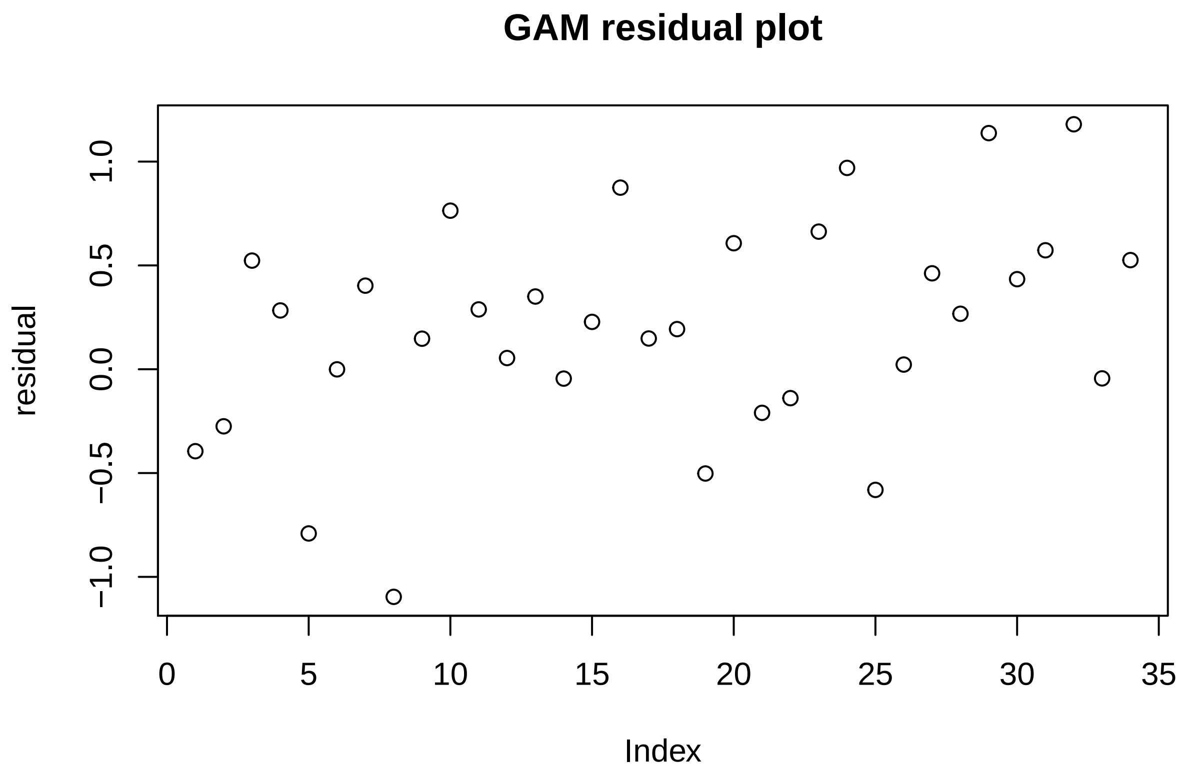


c

The mean test error for linear model is 0.25. On the other hand, we get a test error of 0.23 with GAM. According to both residual plots, the error terms are roughly normally distributed around 0 with no apparent pattern. We believe the similar result from both methods indicates the additional knots may not be too helpful.

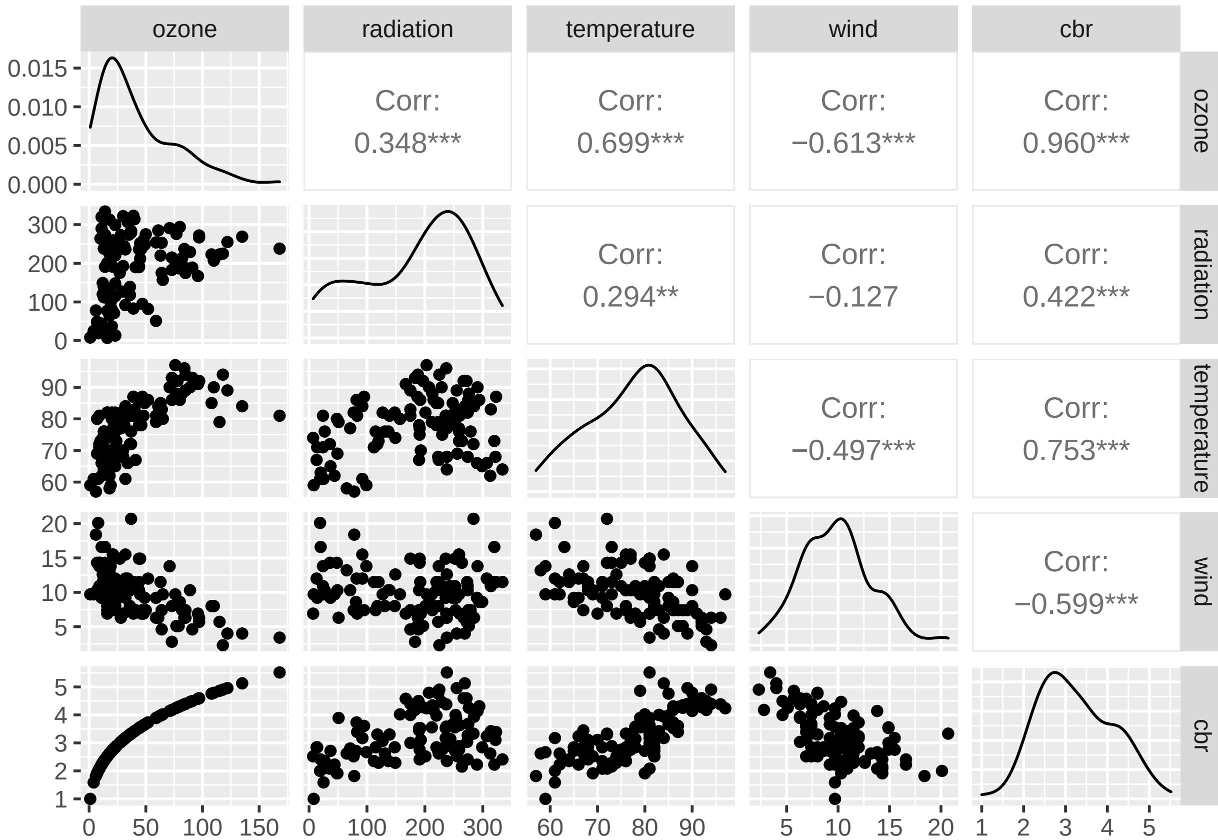
linear model residual plot



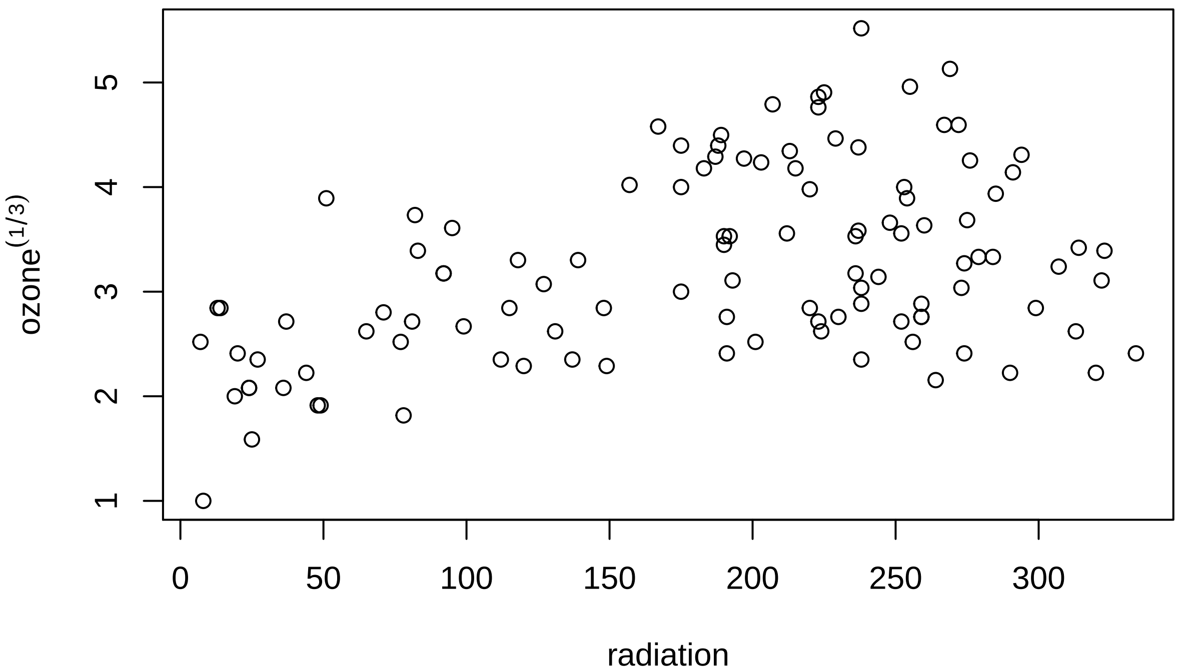


d

According to the pairwise scatterplot, we find there is a rather strong linear relationship between temperature and wind and the cubic root of ozone, as corroberated by the correlation coefficients 0.75 and -0.6 respectively. One can argue that the radiation does not have a strong linear relationship with the response variable, with correlation coefficient of 0.42. However, it is clear that lower values of radiations are associated with lower values of cubic root of ozone and higher values of radiations are associated with higher values of cubic root of ozone. Therefore, we would not apply GAM in this dataset with about 100 observations without further strong evidence for nonlinearity.



ozone
radiation
temperature
wind
cbr
Corr:
0.348***
Corr:
0.699***
Corr:
−0.613***
Corr:
0.960***
Corr:
0.294**
Corr:
−0.127
Corr:
0.422***
Corr:
−0.497***
Corr:
0.753***
Corr:
−0.599***



ozone$^{(1/3)}$
1
2
3
4
5
0
50
100
150
200
250
300
radiation