hw3

Xingwen Wei, Xin Hu, Liding Li

March 6, 2021

Question 1

The true model is a polynomial of degree 3.

```
m <- matrix(c('high', 'low', 'low', 'low', 'low', 'high'), ncol=2, byrow=FALSE)
colnames(m) <- c('Bias', 'Variance')
rownames(m) <- c('Linear regression', 'Polynomial regression with degree 3',
                 'Polynomial regression with degree 10')
as.table(m)
```

```
##                                      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.

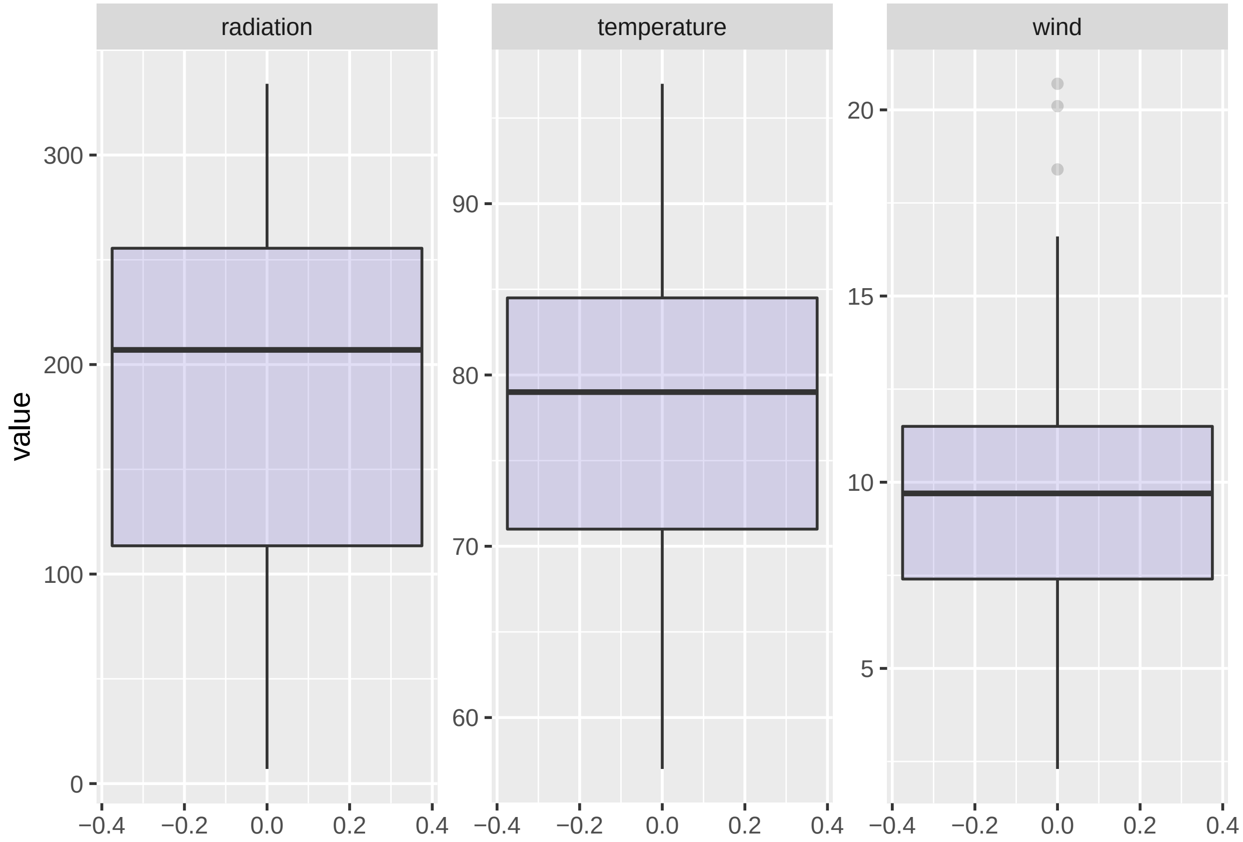
Question 3

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

$$\text{ozone}^{\frac{1}{3}} = 0.001 \times \text{radiation} + 0.056 \times \text{temperature} - 0.072 \times \text{wind} - 0.654$$

```
set.seed(2021)
ozone <- read.table("C:/Users/xingw/Desktop/503/stats503/hw3/ozone_data.txt", header=1)
train <- sample(nrow(ozone), floor(nrow(ozone)*0.7))
ozone$cbr <- ozone$ozone^(1/3)
lmod <- lm(cbr~radiation+temperature+wind, data=ozone[train, ])
summary(lmod)
```

```
##
## Call:
## lm(formula = cbr ~ radiation + temperature + wind, data = ozone[train,
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.13856 -0.36588 -0.01221  0.31766  1.17835
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.6544347  0.6524434  -1.003 0.319149
## radiation    0.0013104  0.0006516   2.011 0.048024 *
## temperature  0.0564556  0.0073096   7.724 4.65e-11 ***
## wind        -0.0724900  0.0190030  -3.815 0.000283 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4949 on 73 degrees of freedom
## Multiple R-squared:  0.6918, Adjusted R-squared:  0.6791
## F-statistic: 54.62 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 $\lambda = 3$. Based on the fitted splines plot of each variables, we find that temperature looks linear where radiation and wind does not.

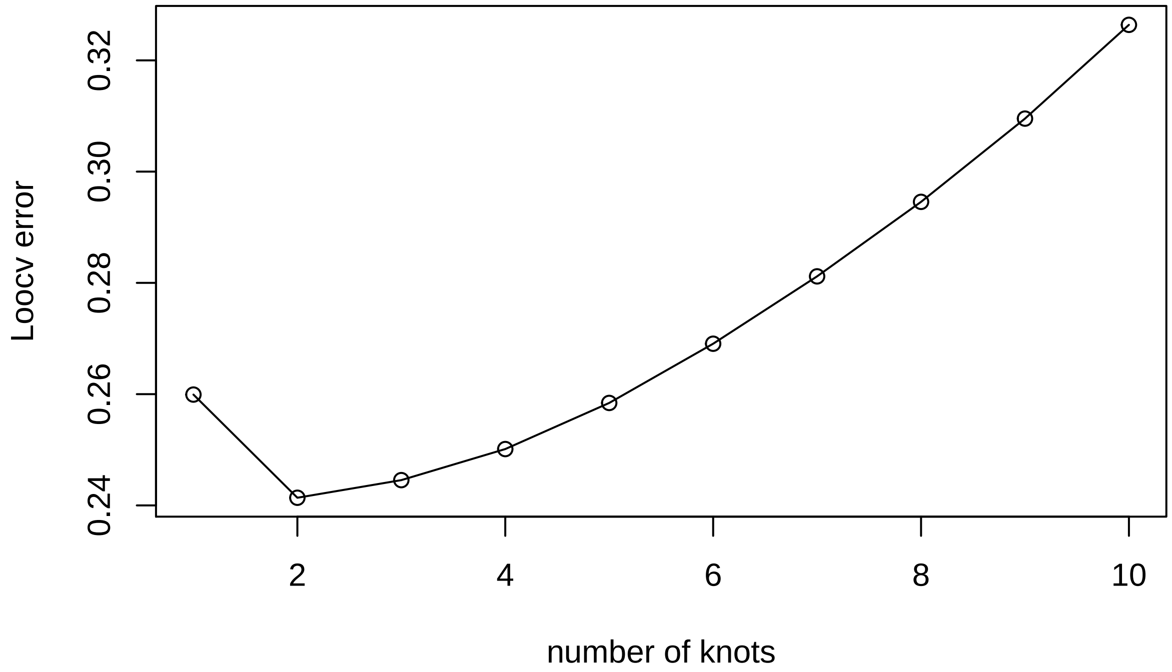
```
library(gam)
```

```
## Loading required package: splines
```

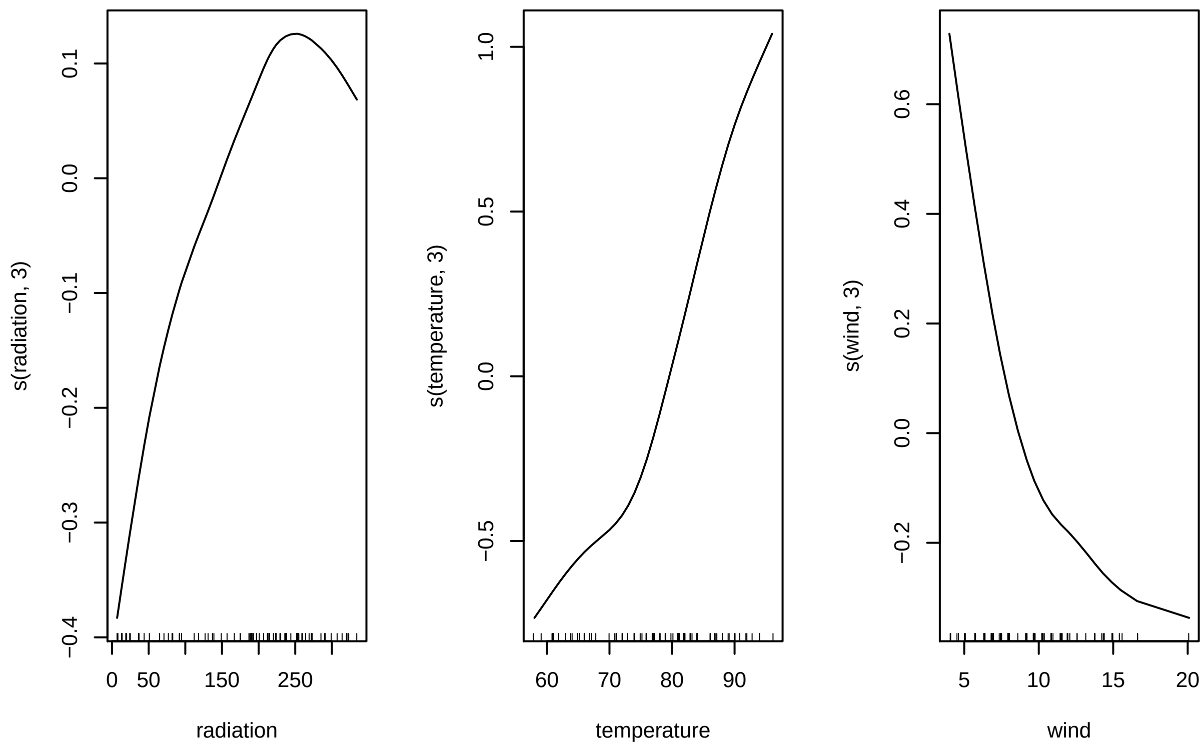
```
## Loading required package: foreach
```

```
## Loaded gam 1.20
```

```
train_loo <- ozone[train, ]
errors <- matrix(NA, nrow=nrow(train_loo), ncol=10)
ss <- 1:10
for(i in 1:nrow(train_loo)){
  train1 <- train_loo[-i, ]
  test1 <- train_loo[i, ]
  for(k in 1:10){
    gam_mod <- gam(cbr~s(radiation, ss[k])+s(temperature, ss[k])+s(wind, ss[k]), data=train1)
    pred <- predict(gam_mod, test1)
    errors[i, k] <- (test1$cbr-pred)^2
  }
}
plot(x=ss, y=apply(errors, 2, mean), 'o', xlab='number of knots', ylab='Loocv error')
```



```
gam_mod <- gam(cbr~s(radiation, 3)+s(temperature, 3)+s(wind, 3), data=ozone[train, ])
par(mfrow=c(1, 3))
plot.Gam(gam_mod)
```

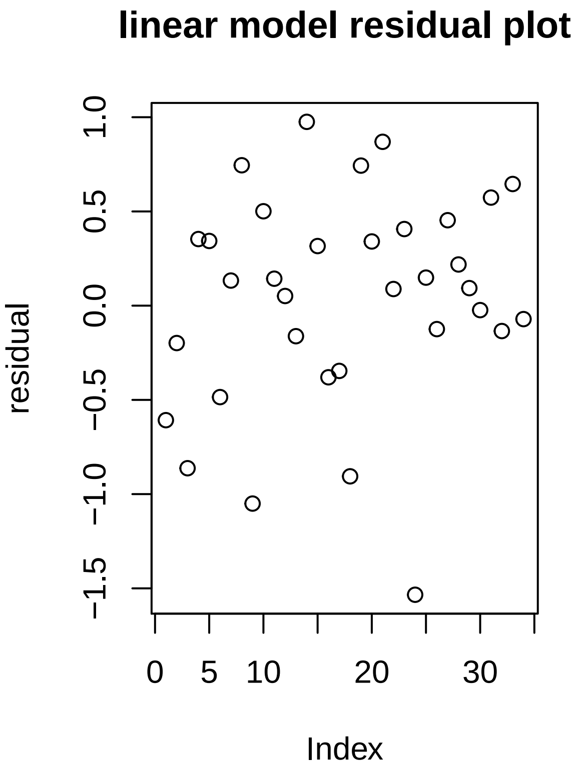


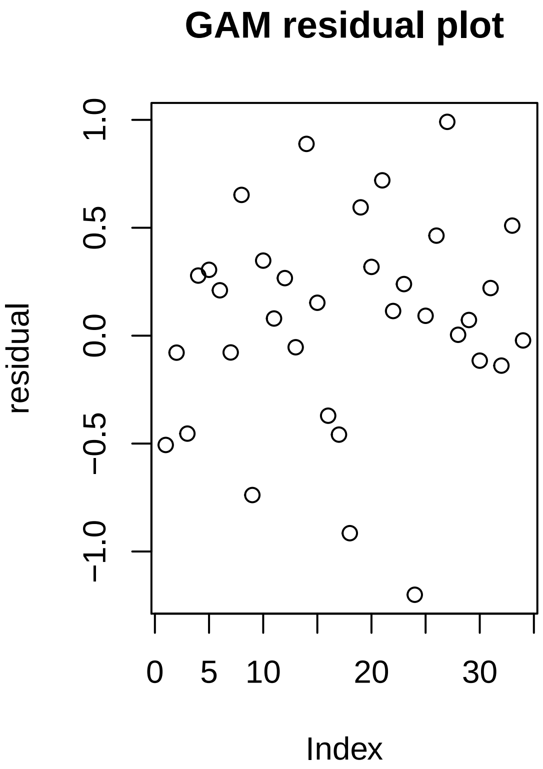
c

The mean test error for linear model is 0.316. On the other hand, we get a test error of 0.232 with GAM. According to both residual plots, the error terms are roughly normally distributed around 0 with no apparent pattern. We believe the nonlinearity is better captured by GAM and resulted in a better test error than the linear model. We suspect that the non-linearity we observed in fitted spline plots above may be a result of small sample size.

```r
par(mfrow=c(1, 2))
pred <- predict(lmod, ozone[-train, ])
lm_mse <- sum((pred-ozone[-train, 'cbr'])^2)/nrow(ozone[-train, ])
plot(pred-ozone[-train, 'cbr'], ylab='residual', main='linear model residual plot')

gam_pred <- predict(gam_mod, ozone[-train, ])
gam_mse <- sum((gam_pred-ozone[-train, 'cbr'])^2)/nrow(ozone[-train, ])
plot(gam_pred-ozone[-train, 'cbr'], ylab='residual', main='GAM residual plot')
```

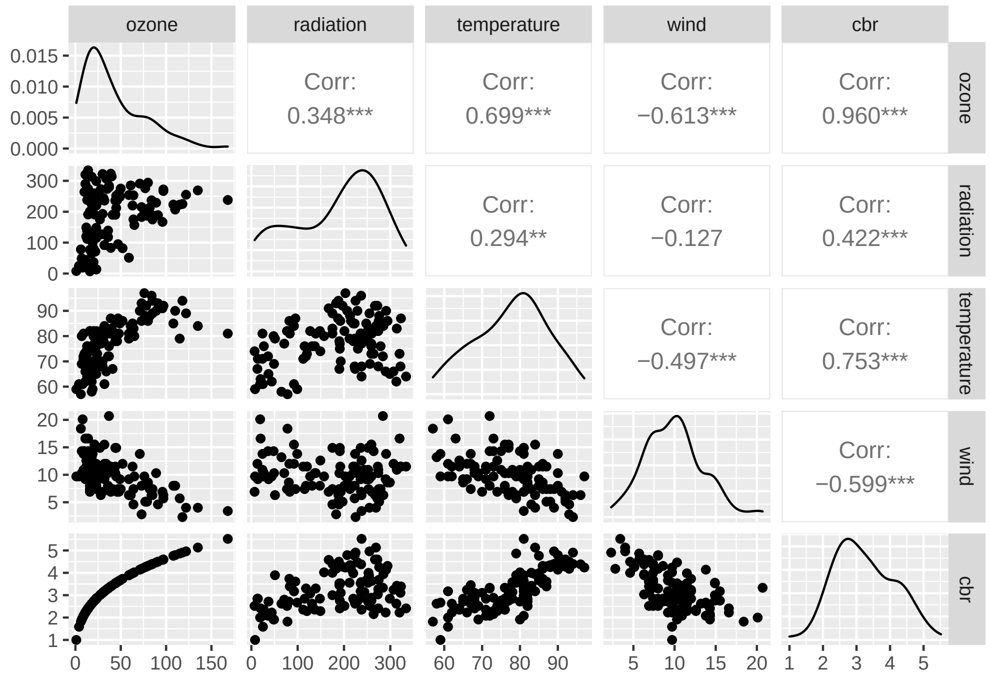




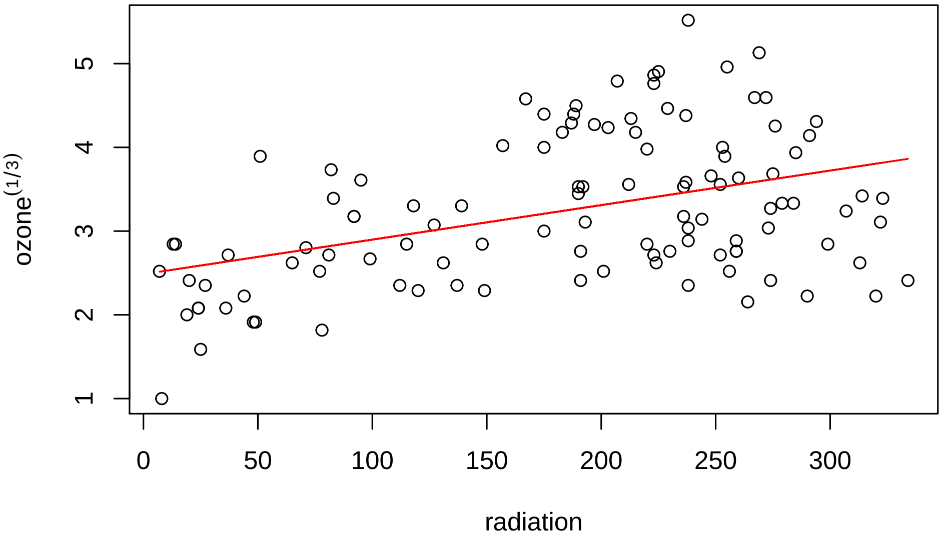
d

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 corroborated 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. Another perspective is provided by the GAM summary where there is significant statistical evidence to show the nonlinearity in temperature and wind, while not for radiation. We believe this inconsistency is a result of the small sample size. Therefore, we would not apply GAM in this dataset with about 100 observations without further strong evidence for nonlinearity.

```r
ggpairs(ozone)
```



```
plot(ozone$radiation, ozone$cbr, xlab='radiation', ylab=expression(ozone^(1/3)))
ts <- lm(ozone$cbr~ozone$radiation)
pred1 <- predict(ts, ozone['radiation'])
lines(x=ozone$radiation, y=pred1, col='red')
```



```
summary(gam_mod)
```

```
##
## Call: gam(formula = cbr ~ s(radiation, 3) + s(temperature, 3) + s(wind,
##     3), data = ozone[train, ])
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07101 -0.25712 -0.05684  0.32912  1.13315
##
## (Dispersion Parameter for gaussian family taken to be 0.2114)
##
##     Null Deviance: 58.0208 on 76 degrees of freedom
## Residual Deviance: 14.1621 on 66.9998 degrees of freedom
## AIC: 110.1379
##
## Number of Local Scoring Iterations: NA
##
## Anova for Parametric Effects
##                   Df  Sum Sq Mean Sq F value    Pr(>F)
## s(radiation, 3)    1  8.0607  8.0607  38.135 4.386e-08 ***
## s(temperature, 3)  1 23.6578 23.6578 111.923 6.233e-16 ***
## s(wind, 3)         1  3.3609  3.3609  15.900 0.0001676 ***
## Residuals         67 14.1621  0.2114
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                   Npar Df Npar F   Pr(F)
## (Intercept)
## s(radiation, 3)         2 0.9000 0.41141
## s(temperature, 3)       2 4.5925 0.01351 *
## s(wind, 3)              2 3.1107 0.05106 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```