stats503hw3

Ningyuan Wang

2/27/2020

Question 1

a. As lambda goes to infinity, in g1, the integration of squares of third derivatiation goes to a straight line. Similarly, in g2, the integration of squares of fourth derivation goes to a straight line as well. In that case, g2 will have a smaller training error since it has higher degree and therefore more flexibility.

b. As lambda goes to infinity, since we do not know the structure of testing data, we cannot tell which model will have smaller testing error.

c. For lambda = 0, g1 and g2 are the same, so they will have same training and testing error.

Question 2

a.

In the linear regression model, all predictors are significant. R squared is 0.7174, which means around 71.74% variation of cubic root of ozone can be caught with the model. Overall, this is a good fit. The testing error of the model is 0.388.

```
dat = read.table("ozone_data.txt", header = T)
dat$ozne_cubr = dat$ozone^(1/3)
summary(dat$ozne_cubr) # check missing values
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.621   3.141   3.248   3.958   5.518
```

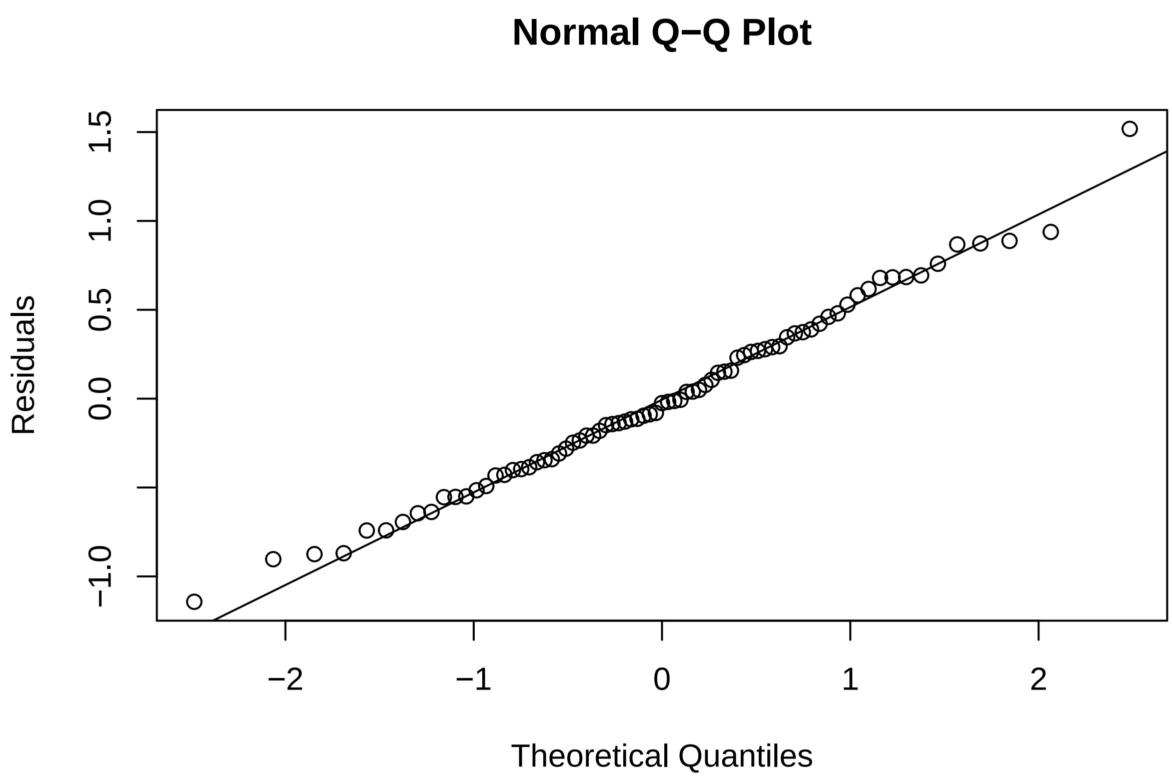
```
# split dataset
train_id = sample(1:nrow(dat), floor(0.7*nrow(dat)))
train = dat[train_id, ]
test = dat[-train_id,]

# normalize the variables
# mean_lm_train = colMeans(train)
# sd_lm_train = apply(train,2,sd)
# lm_train = scale(train,center = mean_lm_train,scale = sd_lm_train)
# lm_train = as.data.frame(lm_train)
# lm_test = scale(test,center = mean_lm_train,scale = sd_lm_train)
# lm_test = as.data.frame(lm_test)
# fit linear regression
lm.fit = lm(ozne_cubr ~ temperature + wind + radiation, data = train )
summary(lm.fit)
```

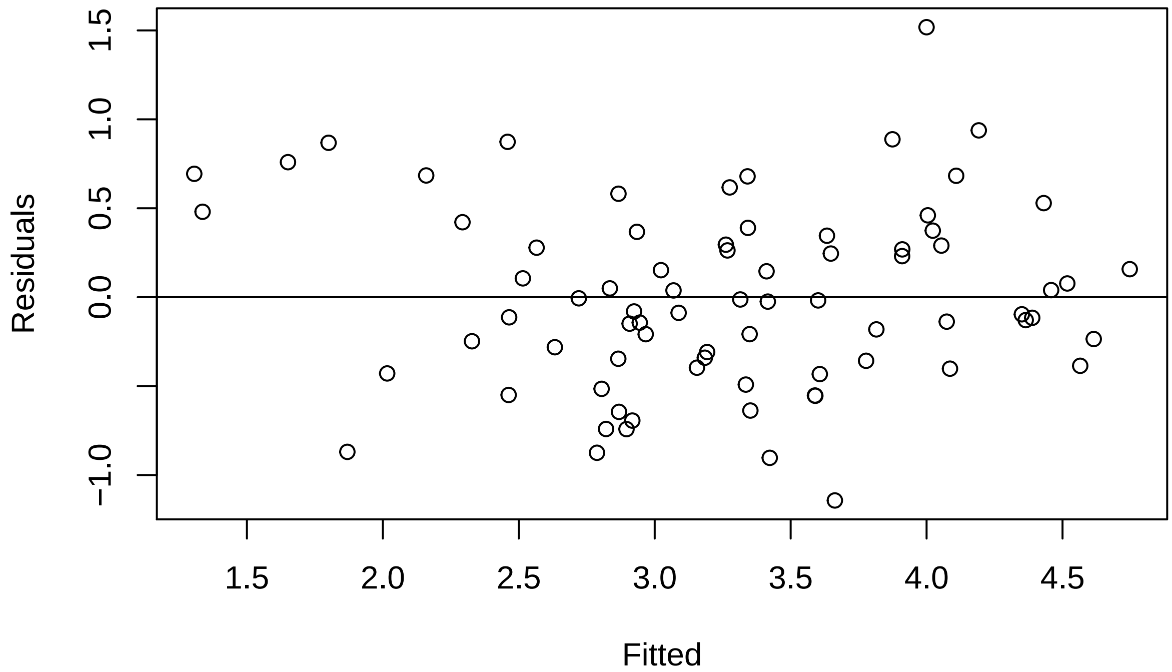
```
##
## Call:
## lm(formula = ozne_cubr ~ temperature + wind + radiation, data = train)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.14281 -0.35754 -0.02483  0.34558  1.51792
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.6901796  0.6778639  -1.018 0.311961
## temperature  0.0540850  0.0076927   7.031 9.14e-10 ***
## wind        -0.0669379  0.0177700  -3.767 0.000332 ***
## radiation    0.0022555  0.0007117   3.169 0.002235 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5222 on 73 degrees of freedom
## Multiple R-squared:  0.6999, Adjusted R-squared:  0.6876
## F-statistic: 56.75 on 3 and 73 DF,  p-value: < 2.2e-16
```

```r
# check model assumpiton
qqnorm(lm.fit$residuals, ylab = "Residuals")
qqline(lm.fit$residuals) # normality is good
```




```r
plot(lm.fit$fitted.values, lm.fit$residuals, xlab = "Fitted", ylab = "Residuals")
abline(h=0) # constant variance
```




```
# calculate testing error
prd_test = predict(lm.fit, test)
lm_test_err = mean((test$ozne_cubr - prd_test)^2) #0.186826
```

b.

Above linear model has significant advantanges in terms of interpretation and inference. However, it also has significant limitations in terms of prediction power. Therefore, we try to fit a generalized additive model (GAM) to relax from linear limitation and achieve a non-linear fit.

We fit a GAM to predict the cubic root of ozone concentration using natural spline of predicotrs temprature, radiation and wind speed. We apply packae "mgcv" to help us decide the optimal degree of freedom based on cross validation. We apply smoothing spline as our function for each predictor variables.

```
library(gam)
library(mgcv)
```

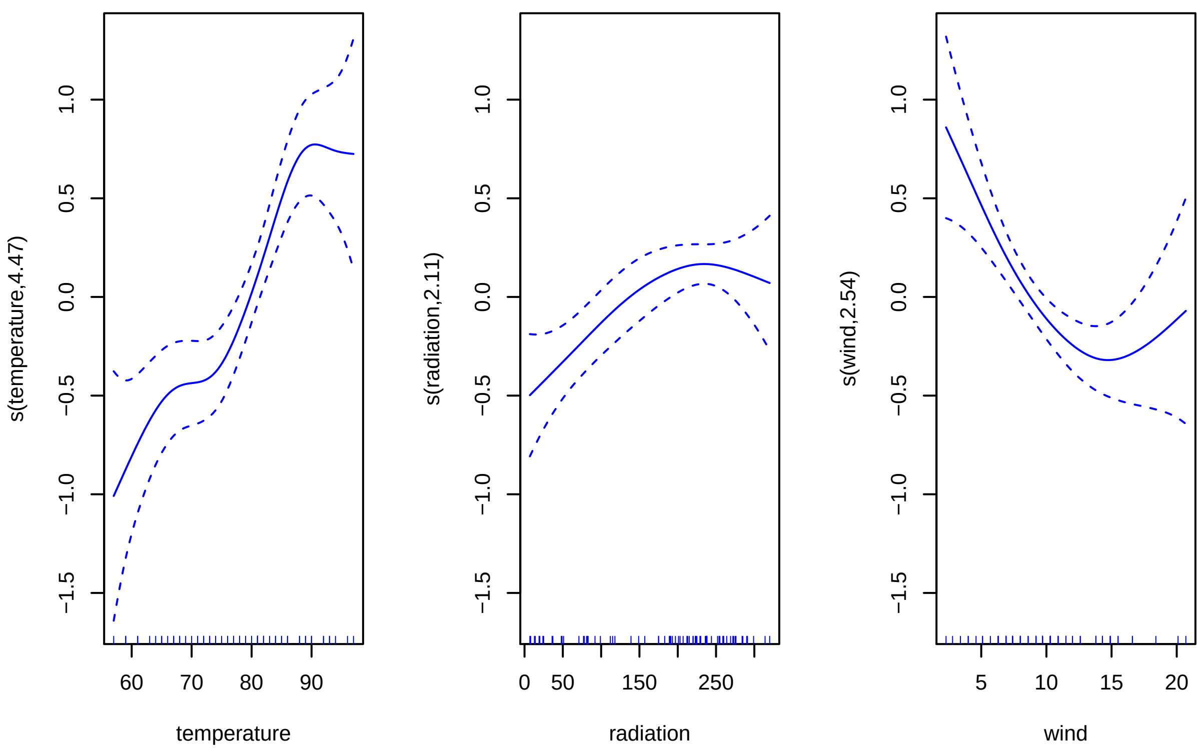
The plots below show the relationships of response and each variable. Based on the plot, we think the model is a good fit of training data set since all predictions are in the confidence intervals for every variable.

```
# smoothing spline approach
gam2=gam(ozne_cubr~s(temperature)+s(radiation)+s(wind), method = "GCV.Cp", data= train)
summary(gam2)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## ozne_cubr ~ s(temperature) + s(radiation) + s(wind)
##
```

```
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.26400    0.05325    61.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                   edf Ref.df      F  p-value
## s(temperature) 4.467  5.474 10.281 5.69e-08 ***
## s(radiation)   2.113  2.648  5.145 0.004095 **
## s(wind)        2.536  3.197  6.844 0.000309 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.75   Deviance explained =   78%
## GCV = 0.25137  Scale est. = 0.21835   n = 77
```

```r
par(mfrow=c(1,3))
plot(gam2 ,col="blue")
```




```r
# compare testing errors of two models
lm_test_err #0.186826
```

```
## [1] 0.2393778
```

```r
gam_test_err = mean((test$ozne_cubr-predict(gam2,test))^2)
gam_test_err #0.1964542
```

```
## [1] 0.1886778
```

Based on the previous 3-plot set, we conclude that variables temperature and radiation have non-linear relationship with response variable.