503 HW #3

Diana Liang

2/27/2020

Problem 1: g1 versus g2

a. λ going towards infinity, training error

g2 would have the smaller training error. Both curves would be fully defined by minimizing the third or fourth derivative. Setting the third derivative to zero would allow for less overfitting than setting the fourth derivative to zero, so g2 would allow for more overfitting and have a smaller training error as a result.

b. λ going towards infinity, test error

The test error measures how well a model fits new data. Since we don't know anything about this new data, such as how complex it is, it is impossible to tell whether g1 or g2 will fit the new data better or whether g1 or g2 has the smaller test error. If the new data was not very complex, g1 would fit the new data better and have a smaller test error; but if the new data was very complex, g2 would fit the new data better and have a smaller test error.

c. λ going towards zero

Both g1 and g2 would mainly be defined by the same first term, so they would have similar training and test errors.

Problem 2: Ozone Data

```
# Set up training and test data
train_idx = sample(nrow(oz_df), floor(nrow(oz_df)*0.7), replace = FALSE)
train_oz = oz_df[train_idx, ]
test_oz = oz_df[-train_idx, ]
```

a. Linear Model

```
oz_lm = lm((ozone)^(1/3)~., data = train_oz)
summary(oz_lm)

##
## Call:
## lm(formula = (ozone)^(1/3) ~ ., data = train_oz)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.10324 -0.36010 -0.04046  0.34087  1.51026
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0804374  0.6751425   0.119 0.905491
## radiation    0.0022583  0.0006951   3.249 0.001753 **
## temperature  0.0450046  0.0076053   5.918 9.79e-08 ***
## wind        -0.0752063  0.0188491  -3.990 0.000156 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5221 on 73 degrees of freedom
## Multiple R-squared:  0.6847, Adjusted R-squared:  0.6717
## F-statistic: 52.83 on 3 and 73 DF,  p-value: < 2.2e-16
```

```
lm_train_err = mean(((train_oz$ozone^(1/3)) - predict(oz_lm, train_oz[, -
1]))^2)
lm_train_err
```

```
## [1] 0.2583903
```

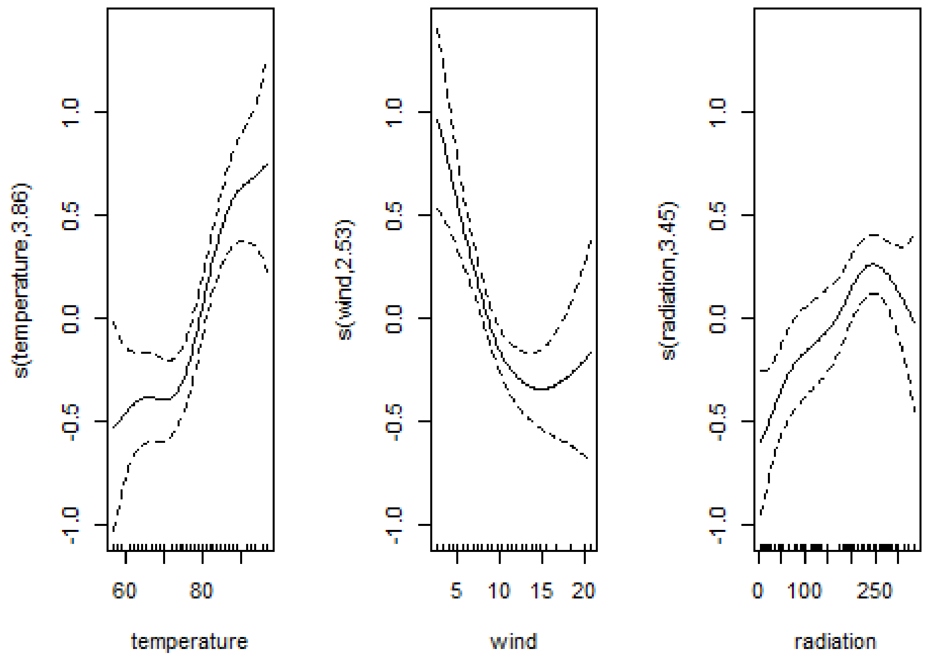
The linear model shows that all 3 variables are important in predicting ozone although the R^2 value of 0.68 suggests that a linear model may not be the best fit for this data. The MSE shown above will be used as a measure of how well this model fits the training data to compare with other models.

b. Smoothing GAM Model

```
oz_gam = gam((ozone^(1/3))~s(temperature)+s(wind)+s(radiation),
              data = train_oz, method = "GCV.Cp")
summary(oz_gam)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## (ozone^(1/3)) ~ s(temperature) + s(wind) + s(radiation)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.20402    0.05057   63.36   <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) 3.857  4.758 7.847 9.86e-06 ***
## s(wind)        2.527  3.177 8.769 3.99e-05 ***
## s(radiation)   3.450  4.247 4.961  0.00148 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.763   Deviance explained = 79.3%
## GCV = 0.22917  Scale est. = 0.19692   n = 77

par(mfrow = c(1,3))
plot(oz_gam)
```



```
gam_train_err = mean(((train_oz$ozone^(1/3)) - predict(oz_gam, train_oz[, -
1]))^2)
gam_train_err

## [1] 0.1692172
```

The CV chose degrees of freedom for each predictor variable based on the smoothing parameter that minimizes the training error. The parameters chosen for each is shown in the plots above. The lowest training error for this GAM model also shown above is much smaller than that of the linear model. Since the GAM model does not force the relationship between the response and predictor variables to be linear, the model can fit more closely to the training data. So a lower training error for the GAM model compared to that of the linear model is expected.

c. Test Error

```
lm_test_err = mean(((test_oz$ozone^(1/3)) - predict(oz_lm, test_oz[, -1]))^2)
lm_test_err

## [1] 0.2414873
```

```r
gam_test_err = mean(((test_oz$ozone^(1/3)) - predict(oz_gam, test_oz[, -1]))^2)
gam_test_err

## [1] 0.2212909
```

The linear model test error is similar to the GAM model test error, which suggests that the linear model is just as good at predicting ozone compared to the GAM model.

d. Non-linear Relationships

Since the linear model has a similar test error to the GAM model, this suggests that assuming all the predictor variables have a linear relationship with the response variable led to a model just as good as that from allowing for non-linear relationships. From this, there is no evidence that any of the variables have a non-linear relationship with the cube root of ozone.