STAT/MATH 495: Problem Set 03

Brenna Sullivan

17-09-26

Contents

Question

For both `data1` and `data2` tibbles (a tibble is a data frame with some metadata attached:

- Find the splines model with the best out-of-sample predictive ability.
- Create a visualizaztion arguing why you chose this particular model.
- Create a visualizaztion of this model plotted over the given (x_i, y_i) points for $i = 1, \ldots, n = 3000$.
- Give your estimate $\widehat{\sigma}$ of σ where the noise component ϵ_i is distributed with mean 0 and standard deviation σ.

Cross Validation Function (for Data1 and Data2)

```r
#test to delete
crossval <- function(train, test){
  dfRMSE <- data.frame()
  output <- NULL
  i <- 5
  while(i <= 50){
    model <- smooth.spline(x = train$x,y = train$y, df=i)
    new_x <- test$x
    output <- predict(model, new_x)
    output <- as.data.frame(output)
    output <- cbind(output, y_obs = test$y)
    RMSE <- sqrt(mean(output$y_obs - output$y)^2)
    new_row <- c(i, RMSE)
    dfRMSE <- rbind(dfRMSE, new_row)
    i = i + 1
  }
  names(dfRMSE) <- c("df","RMSE")
  return(dfRMSE)
}
```

Data1

Five Folds

```
#create the folds
set.seed(1)
splitdata <- sample(1:5, size = nrow(data1), replace=T, prob=c(0.2,0.2,0.2,0.2,0.2))
data11 <- data1[splitdata==1,]
data12 <- data1[splitdata==2,]
data13 <- data1[splitdata==3,]
data14 <- data1[splitdata==4,]
data15 <- data1[splitdata==5,]

#run each fold agains the other folds
fold1_data1 <- crossval(data1[splitdata != 1,], data11)
fold2_data1 <- crossval(data1[splitdata != 2,], data12)
fold3_data1 <- crossval(data1[splitdata != 3,], data13)
fold4_data1 <- crossval(data1[splitdata != 4,], data14)
fold5_data1 <- crossval(data1[splitdata != 5,], data15)
```

meanRMSE

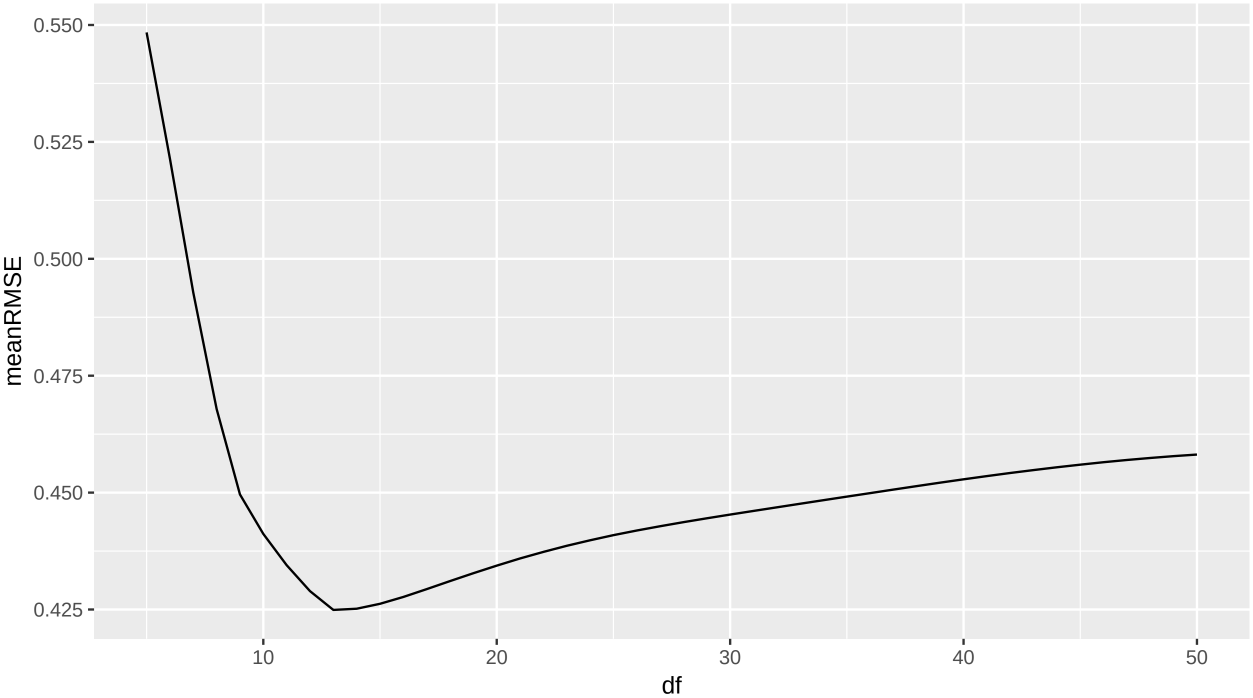
```
#calculate the RMSE for each fold and take the average of the 5 to hey meanRMSE
data1_folds <- join_all(list(fold1_data1, fold2_data1, fold3_data1, fold4_data1, fold5_data1), by = "df
names(data1_folds) <- c("df","fold1","fold2","fold3","fold4","fold5")
data1_folds$meanRMSE <- (data1_folds$fold1 + data1_folds$fold2 + data1_folds$fold3 + data1_folds$fold4
```

Best Out-Of-Sample Predictive Ability

```
#minimum point and graph of df
min_point <- data1_folds[which(data1_folds$meanRMSE == min(data1_folds$meanRMSE)),]
min_point
```

```
##   df       fold1     fold2     fold3     fold4   fold5  meanRMSE
## 9 13 0.002326137 0.3348121 0.4184415 0.5839789 0.78499 0.4249097
```

```
ggplot() +
  geom_line(data = data1_folds, aes(x = df, y = meanRMSE))
```

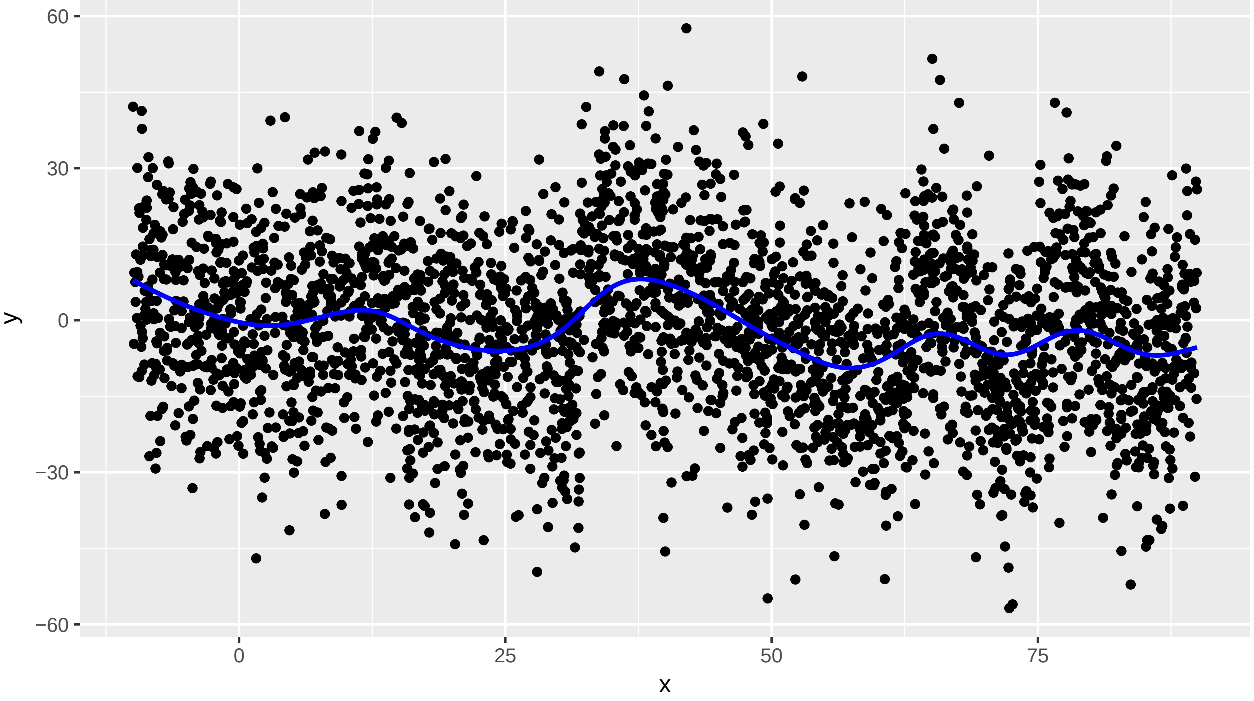


Here, from the calculation of the minimum of the RMSE, as well as the graph, you can see that RMSE is the smallest when df=13. There, the RMSE is .42491.

Spline Model for Visualization

```
#spline model with the optimal degrees of freedom
data1_spline <- smooth.spline(data1$x, data1$y, df=min_point$df)


model_spline_data_frame <- data1_spline %>%
  broom::augment()
ggplot(model_spline_data_frame, aes(x=x)) +
  geom_point(aes(y=y)) +
  geom_line(aes(y=.fitted), col="blue", size=1)
```



Data 2

Five Folds

```
#create the folds
set.seed(1)
splitdata <- sample(1:5, size = nrow(data2), replace=T, prob=c(0.2,0.2,0.2,0.2,0.2))
data21 <- data2[splitdata==1,]
data22 <- data2[splitdata==2,]
data23 <- data2[splitdata==3,]
data24 <- data2[splitdata==4,]
data25 <- data2[splitdata==5,]

#run each fold agains the other folds
fold1_data2 <- crossval(data2[splitdata != 1,], data21)
fold2_data2 <- crossval(data2[splitdata != 2,], data22)
fold3_data2 <- crossval(data2[splitdata != 3,], data23)
fold4_data2 <- crossval(data2[splitdata != 4,], data24)
fold5_data2 <- crossval(data2[splitdata != 5,], data25)
```

meanRMSE

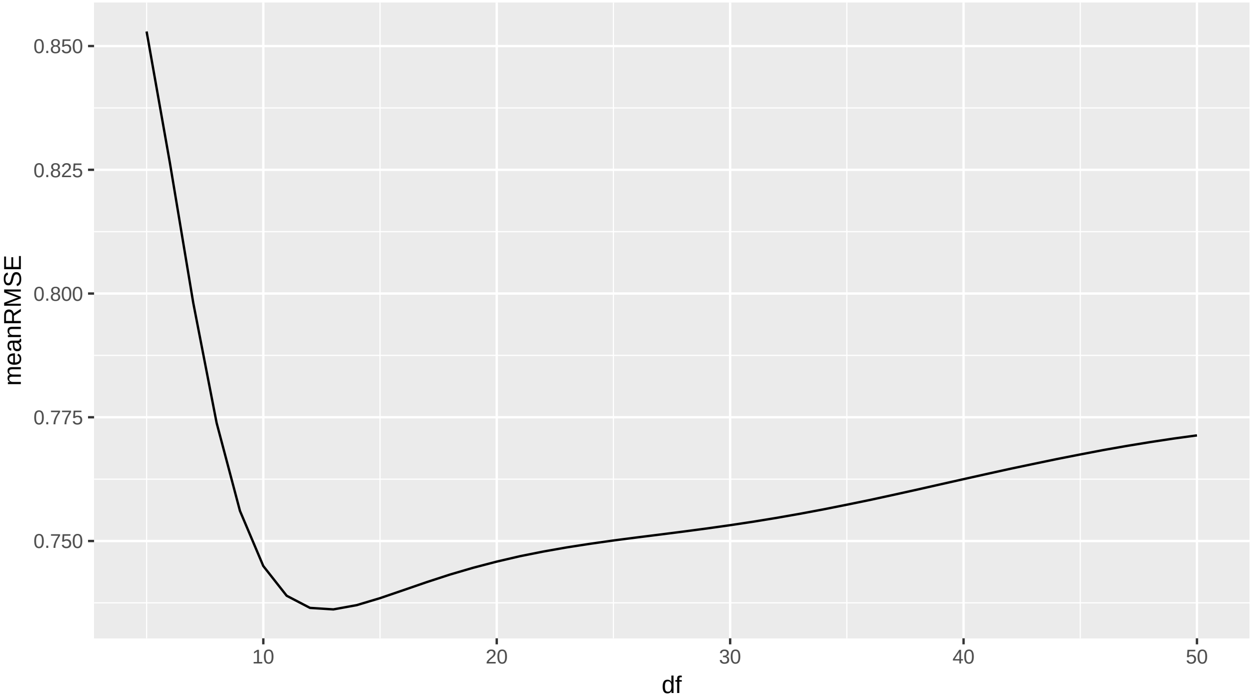
```
#calculate the RMSE for each fold and take the average of the 5 to get meanRMSE
data2_folds <- join_all(list(fold1_data2, fold2_data2, fold3_data2, fold4_data2, fold5_data2), by = "df
names(data2_folds) <- c("df","fold1","fold2","fold3","fold4","fold5")
data2_folds$meanRMSE <- (data2_folds$fold1 + data2_folds$fold2 + data2_folds$fold3 + data2_folds$fold4 +
```

Best Out-Of-Sample Predictive Ability

```
#minimum point and graph of df
min_point2 <- data2_folds[which(data2_folds$meanRMSE == min(data2_folds$meanRMSE)),]
min_point2
```

```
##   df    fold1     fold2     fold3     fold4    fold5  meanRMSE
## 9 13 0.191195 0.5116611 0.7059952 0.8416529 1.430397 0.7361802
```

```
ggplot() +
  geom_line(data = data2_folds, aes(x = df, y = meanRMSE))
```

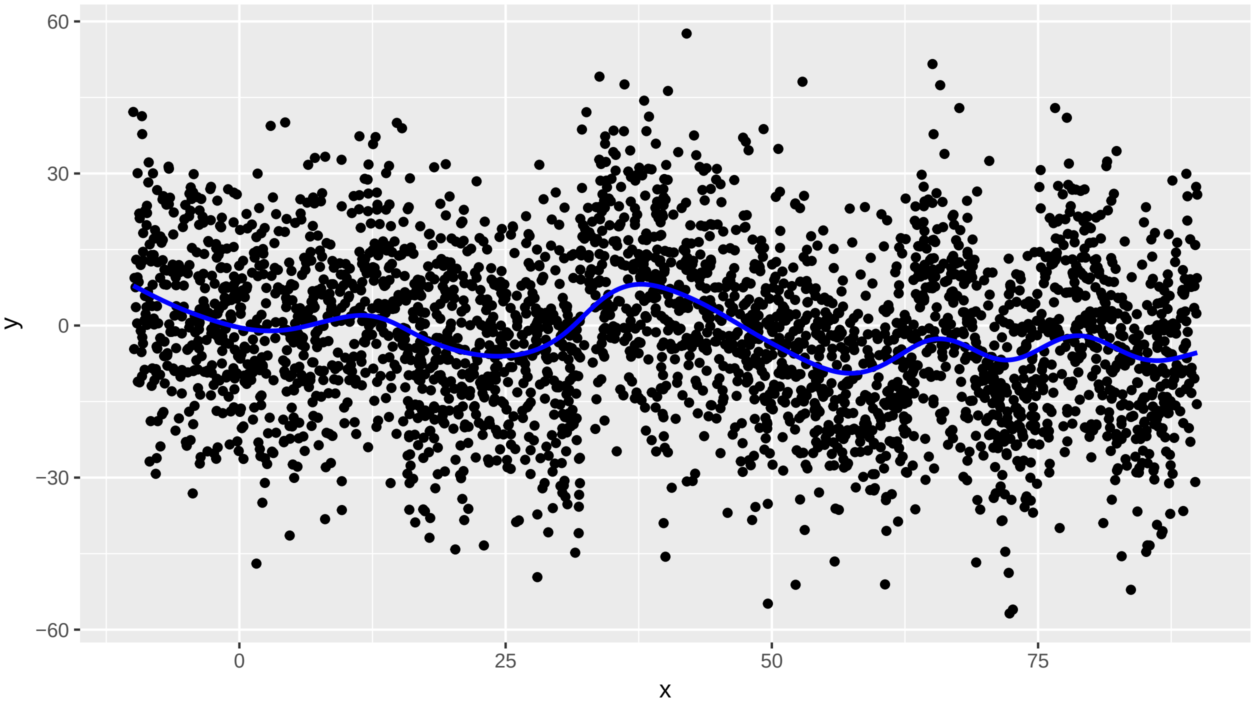


Here, from the calculation of the minimum of the RMSE, as well as the graph, you can see that RMSE is the smallest when df=13. There, the RMSE is .7361.

Spline Model for Visualization

```
#spline model with the optimal degrees of freedom
data2_spline <- smooth.spline(data2$x, data2$y, df=min_point2$df)


model_spline_data_frame <- data1_spline %>%
  broom::augment()
ggplot(model_spline_data_frame, aes(x=x)) +
  geom_point(aes(y=y)) +
  geom_line(aes(y=.fitted), col="blue", size=1)
```


60
30
0
−30
−60
y
0
25
50
75
x