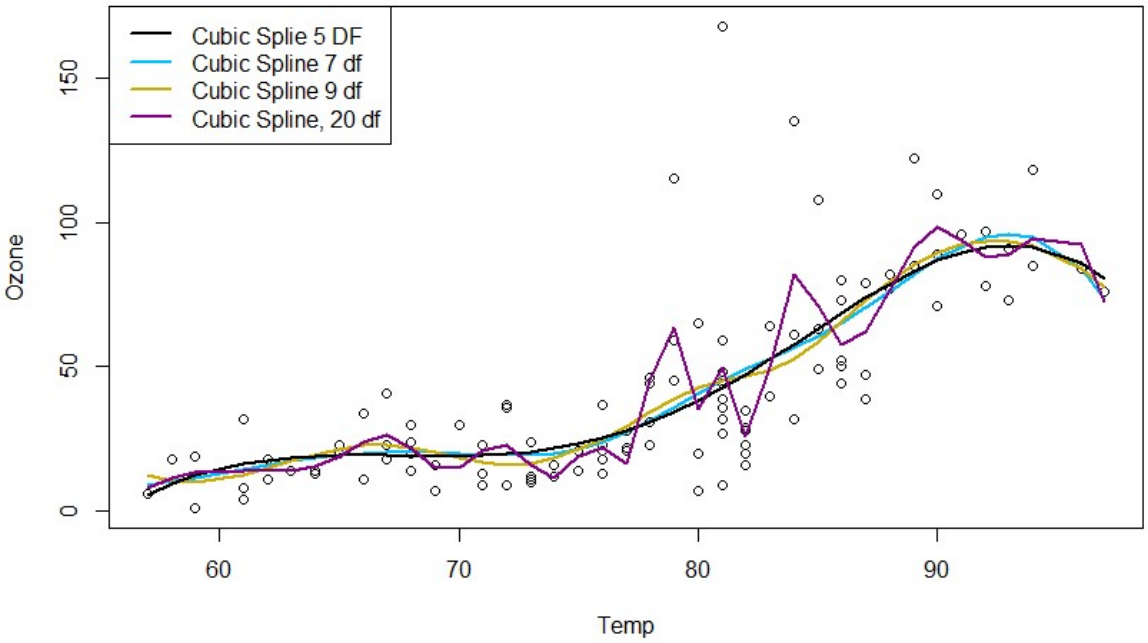
5 Exercises

Application

Refer to the Air Quality data described previously, and the analyses we have done with Ozone as the response variable, and the five explanatory variables (including the two engineered features).

1. Use cubic splines to model the relationship between Ozone and Temp:
(a) On one graph, plot the data along with fits of
i. Cubic regression
ii. Cubic splines with 5, 7, 9, and 20 DF.
Present the plot. Be sure to add a legend and use different colours for the different functions

```r
library(dplyr)
library(MASS)   # For ridge regression
library(glmnet) # For LASSO
library(splines)
source("Helper Functions.R")
data = na.omit(airquality[, 1:4])
data$TWcp = data$Temp*data$Wind
data$TWrat = data$Temp/data$Wind
data = data[-c(2, 3, 5, 6)]
#Ordered Data
ordered.data = data[order(data$Temp),]

#1. Use cubic splines to model the relationship between Ozone and Temp:
# (a) On one graph, plot the data along with fits of
# i. Cubic regression
# ii. Cubic splines with 5, 7, 9, and 20 DF.
# Present the plot. Be sure to add a legend and use different colours for
# the different functions
x11(h=7, w=10)
with(data, plot(Temp, Ozone))
legend("topleft", legend=c("Cubic Splie 5 DF", "Cubic Spline 7 df",
                           "Cubic Spline 9 df", "Cubic Spline, 20 df"),
       lty="solid", col=colors()[c(24,121,145,84)], lwd=2)

poly3 <- lm(data=data, Ozone ~ poly(x=Temp, degree=3))


# 5 DF spline
cub.spl.5 <- lm(data=ordered.data, Ozone ~ bs(Temp,df=5))
lines(x=ordered.data$Temp, y=predict(cub.spl.5, newdata=ordered.data), col=colors()[24], lwd=2)

# 7 DF spline
cub.spl.7 <- lm(data=ordered.data, Ozone ~ bs(Temp,df=7))
lines(x=ordered.data$Temp, y=predict(cub.spl.7, newdata=ordered.data), col=colors()[121], lwd=2)

# 9 DF spline
cub.spl.9 <- lm(data=ordered.data, Ozone ~ bs(Temp,df=9))
lines(x=ordered.data$Temp, y=predict(cub.spl.9, newdata=ordered.data), col=colors()[145], lwd=2)

# 20 DF spline
cub.spl.20 <- lm(data=ordered.data, Ozone ~ bs(Temp,df=20))
lines(x=ordered.data$Temp, y=predict(cub.spl.20, newdata=ordered.data), col=colors()[84], lwd=2)
```



b) Which model seems to have the most bias? (**Just report the name**)

```
#(b) Which model seems to have the most bias? (Just report the name)
# Cubic spline with 5 DF seems to be the most biased.
```

(c) Do any functions have a tendency to overfit? If so, **which one(s), and what do you see that causes you to think it/they overfit?**

```
# (c) Do any functions have a tendency to overfit? If so, which one(s), and what
# do you see that causes you to think it/they overfit?
# Cubic spline with 20 DF seems to have a tendency to overfit. Too many DF may cause overfit
```

(d) If you had to choose one model, **which would it be? Why?**

```
#(d) If you had to choose one model, which would it be? Why?
# Cubic spline with 7, because it is not biased and it shows the trend of data naturally.
```