Advanced Applied Statistics Homework 5 Solutions

Joshua D. Ingram

2022-10-14

1.

Read the updated document on smoothing, which is Chapter 11 in the handout on "Review of the simple linear regression model"

2.

I have found some wonderful reading of the material we covered in class this week, in the freely available book Introduction to Statistical Learning, 2nd edition, which you can access here for free:

statlearning

a. Smoothing

Please read Chapter 7, Sections 7.1 to 7.5, all of which should sound familiar.

Then complete Exercise 9, part (e) only, using at least three smooth fits, one of which has to be a regular cubic regression spline, and one of which has to be a natural cubic regression spline. You can also try a smoothing spline. (see the handout mentioned in 1) above)

Exercise 9 (e)

This question uses the variable `dis` (the weighted mean of distances to five Boston employment centers) and `nox` (nitrogen oxides concentration in parts per 10 million) from the `Boston` data. We will treat `dis` as the predictor and `nox` as the response.

Now fit a regression spline for a range of degrees of freedom, and plot the resulting fits and report the resulting RSS. Describe the results obtained.

Answer:

```
library(ISLR)
library(MASS)
library(splines)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
```

```
attach(Boston)
```

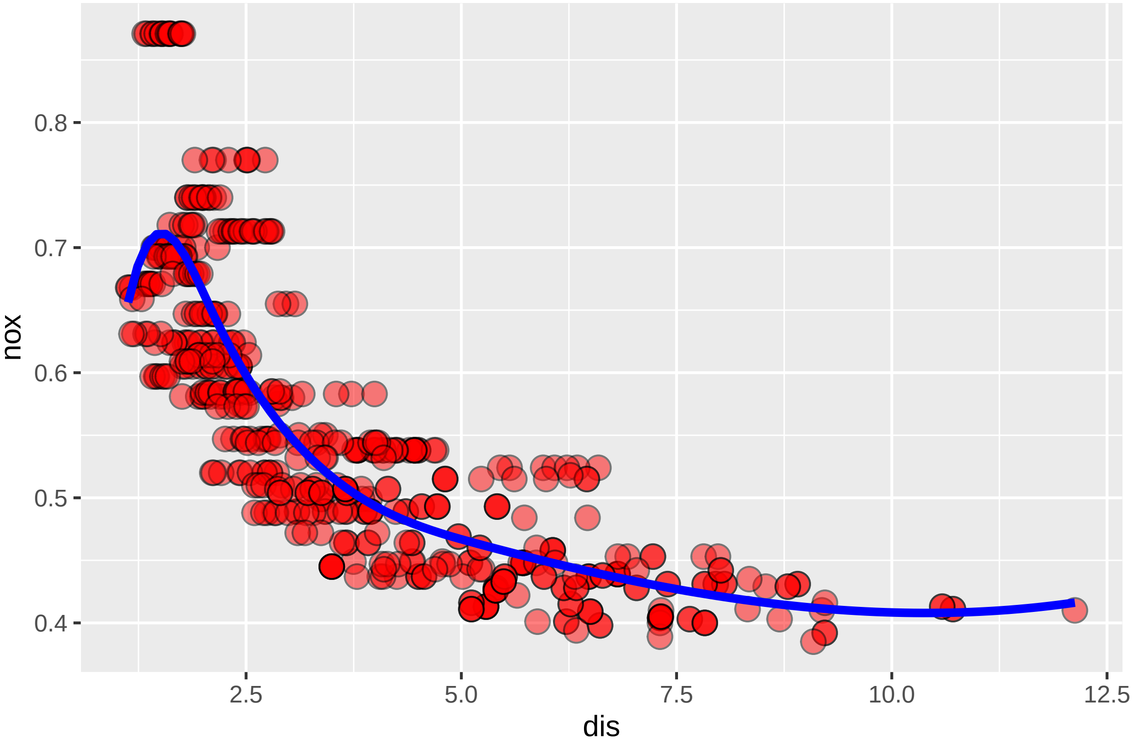
Regular Cubic Regression Spline:

```
x <- Boston$dis

# knots, dividing into 3 regions
eta1 <- quantile(x, 0.25)
eta2 <- quantile(x, 0.5)
eta3 <- quantile(x, 0.75)

# the bases functions
h1 <- rep(1,length(x))
h2 <- x
h3 <- x^2
h4 <- x^3
h5 <- (x - eta1)^3*(x > eta1)
h6 <- (x - eta2)^3*(x > eta2)
h7 <- (x - eta3)^3*(x > eta3)

# design matrtix
X = cbind(h1, h2, h3, h4, h5, h6, h7)

# fit the regression model
fit_cubic_spline <- lm(nox ~ -1 + X, data = Boston)
summary(fit_cubic_spline)
```

```
##
## Call:
## lm(formula = nox ~ -1 + X, data = Boston)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.128538 -0.037813 -0.009987  0.022644  0.195494
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## Xh1 -0.647049   0.406641  -1.591 0.112197
## Xh2  2.240035   0.648001   3.457 0.000593 ***
## Xh3 -1.181649   0.335520  -3.522 0.000468 ***
## Xh4  0.194751   0.056623   3.439 0.000632 ***
## Xh5 -0.201051   0.064695  -3.108 0.001993 **
## Xh6  0.002351   0.011321   0.208 0.835580
## Xh7  0.004049   0.002345   1.727 0.084817 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06062 on 499 degrees of freedom
## Multiple R-squared:  0.9887, Adjusted R-squared:  0.9886
## F-statistic:  6244 on 7 and 499 DF,  p-value: < 2.2e-16
```

```
betas <- coefficients(fit_cubic_spline)
f_x <- function(x) betas[1]*1 + betas[2]*x + betas[3]*x^2 + betas[4]*x^3 +
betas[5]*(x - eta1)^3*(x > eta1) + betas[6]*(x - eta2)^3*(x > eta2) + betas[7]*(x - eta3)^3*(x > eta3)
```

```
scatplot <- ggplot(data=Boston, aes(x=dis, y=nox)) +
  geom_point(pch=21, fill="red", size=4, alpha=0.5)

scatplot +
  geom_function(fun = f_x, col="blue", size=1.5, alpha=1)
```



```
# RSS
sum(resid(fit_cubic_spline)^2)
```

```
## [1] 1.833966
```

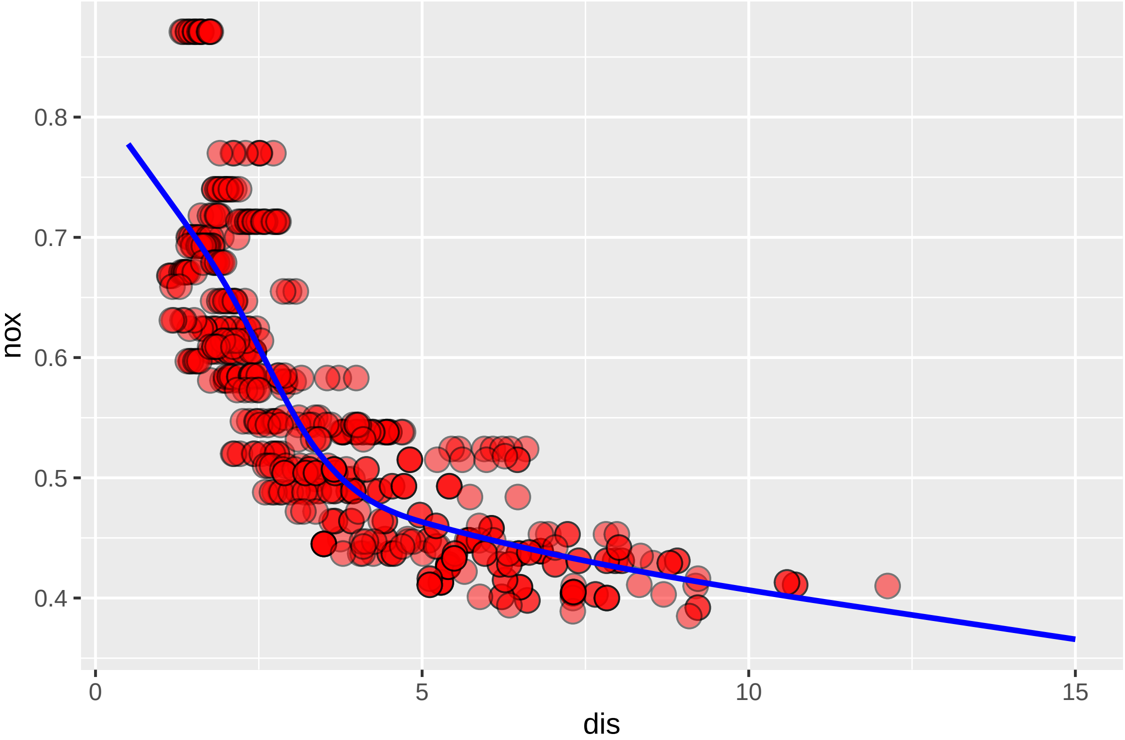
Natural Cubic Regression Spline:

```
# fit natural cubic regression
fit_natural <- lm(nox ~ ns(dis, knots = c(eta1, eta2, eta3)), data = Boston)
summary(fit_natural)
```

```
##
## Call:
## lm(formula = nox ~ ns(dis, knots = c(eta1, eta2, eta3)), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12940 -0.04073 -0.00805  0.02494  0.19059
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           0.73032    0.01276   57.23   <2e-16 ***
## ns(dis, knots = c(eta1, eta2, eta3))1 -0.24312    0.01373  -17.70   <2e-16 ***
## ns(dis, knots = c(eta1, eta2, eta3))2 -0.27001    0.01724  -15.67   <2e-16 ***
## ns(dis, knots = c(eta1, eta2, eta3))3 -0.38799    0.03179  -12.21   <2e-16 ***
## ns(dis, knots = c(eta1, eta2, eta3))4 -0.30464    0.03105   -9.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06135 on 501 degrees of freedom
## Multiple R-squared:  0.7219, Adjusted R-squared:  0.7197
## F-statistic: 325.1 on 4 and 501 DF,  p-value: < 2.2e-16
```

```
# Compute predictions
newDis = seq(0.5, 15, by=0.1)
natural <- predict(fit_natural,newdata=data.frame(dis=newDis))
```

```
scatplot +
  geom_line(data = data.frame(x = newDis, y = natural), aes(x=x,y=y),color = "blue", size = 1 )
```

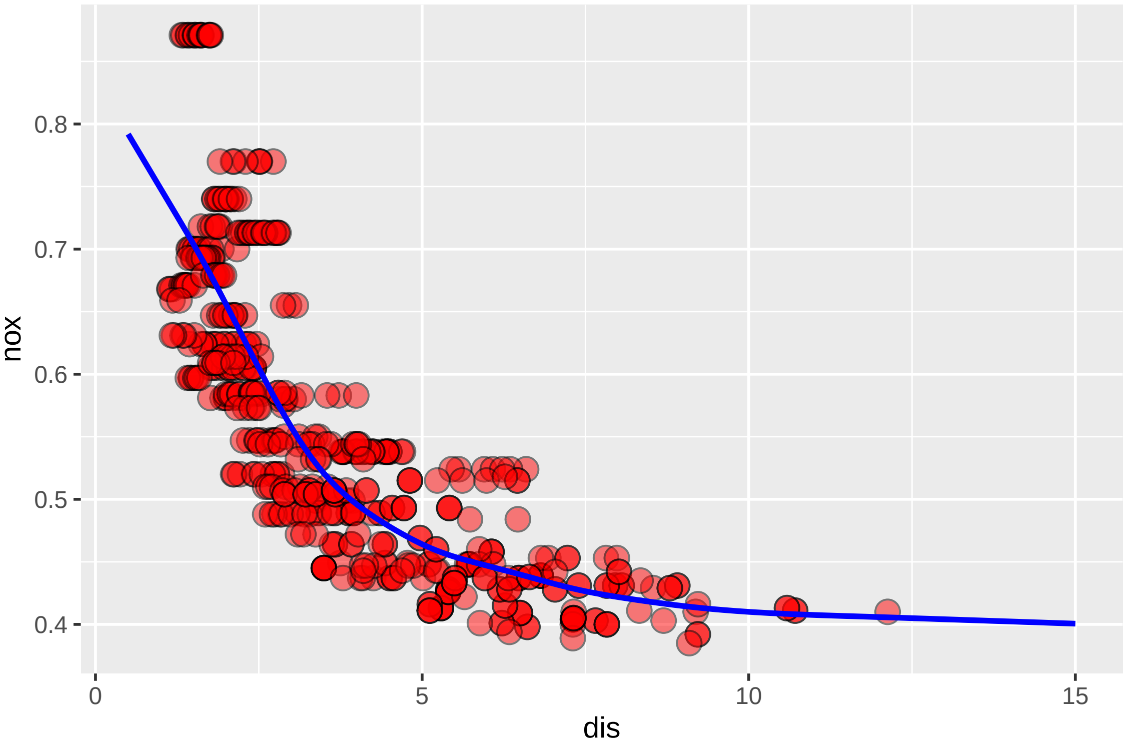


```
# RSS
sum(resid(fit_natural)^2)
```

```
## [1] 1.885805
```

Smoothing Spline:

```
fit_smoothing = smooth.spline(Boston$dis, Boston$nox, df=6)

estimated_f = predict(fit_smoothing, newDis)

scatplot +
  geom_line(data = data.frame(x = newDis, y = estimated_f$y), aes(x=x,y=y),color = "blue", size = 1 )
```



```
sum(resid(fit_smoothing)^2)
```

```
## [1] 1.885953
```

We obtain a RSS of 1.833966 for the regular cubic regression spline, a RSS of 1.885805 for the natural cubic regression spline, and a RSS of 1.885953 for the smoothing spline where df = 6.

b. Multiple Testing and the Bonferroni Correction

Read Chapter 13, up to the Bonferroni Method in Section 13.3.2.

Then, complete Exercise 4, parts (a) and (b) only!

Exercise 4 (a)

Suppose we test m = 10 hypotheses, and obtain the p-values shown in Table 13.4.

Suppose that we wish to control the Type I error for each null hypothesis at level $\alpha = 0.05$. Which null hypotheses will we reject?

Answer:

We reject the null hypotheses where the p-value is less than our confidence level, $\alpha = 0.05$. We reject $H_{01}, H_{02}, H_{03}, H_{08}, H_{09}, H_{10}$.

Exervise 4 (b)

Now suppose that we wish to control the FWER at level $\alpha = 0.05$. Which null hypotheses will we reject? Justify your answer.

Answer:

We reject the null hypotheses where the p-value is less than $\frac{a}{m} = \frac{0.05}{10} = 0.005$. We reject H_{01}, H_{09}, H_{10}.