ETC3250 2018 - Lab 4 solutions

Souhaib Ben Taieb

15 March 2018

Exercise 1

Understand all the steps in the proof of the bias-variance decomposition (see https://github.com/bsouhaib/BA2018/raw/master/slides/week2/proof-bv.pdf).

Let $y = f(x) + \varepsilon$ where ε is iid noise with zero mean and variance σ^2. Using the bias-variance decomposition, show that $E[(y - \hat{f}(x_0))^2]$ is minimum when $\hat{f}(x_0) = E[y|x = x_0]$. What is this minimum value?

Replacing $\hat{f}(x_0)$ by $E[y|x = x_0]$ in the bias-variance decomposition, shows that $E[(y - \hat{f}(x_0))^2] = \sigma^2$ which is the irreducible error (and the minimum value).

Exercise 2

Do the exercise 1 in Section 7.9 of ISLR.

1. (a) $a_1 = \beta_0$, $b_1 = \beta_1$, $c_1 = \beta_2$, $d_1 = \beta_3$
2. (b) $a_2 = \beta_0 - \beta_4\xi^3$, $b2 = \beta_1 + 3\beta_4\xi^2$, $c_2 = \beta_2 - 3\beta_4\xi$, $d_2 = \beta_3 + \beta_4$
3. (c), (d) and (e) Just develop the different terms for each function.

Exercise 3

Do some exploratory data analysis on the `Wage` data set (available in the ISLR package).

- Tabulate education and marital status
- Tabulate education and race
- Tabulate marital status race
- Plot marital status as a function of age
- Try other combinations

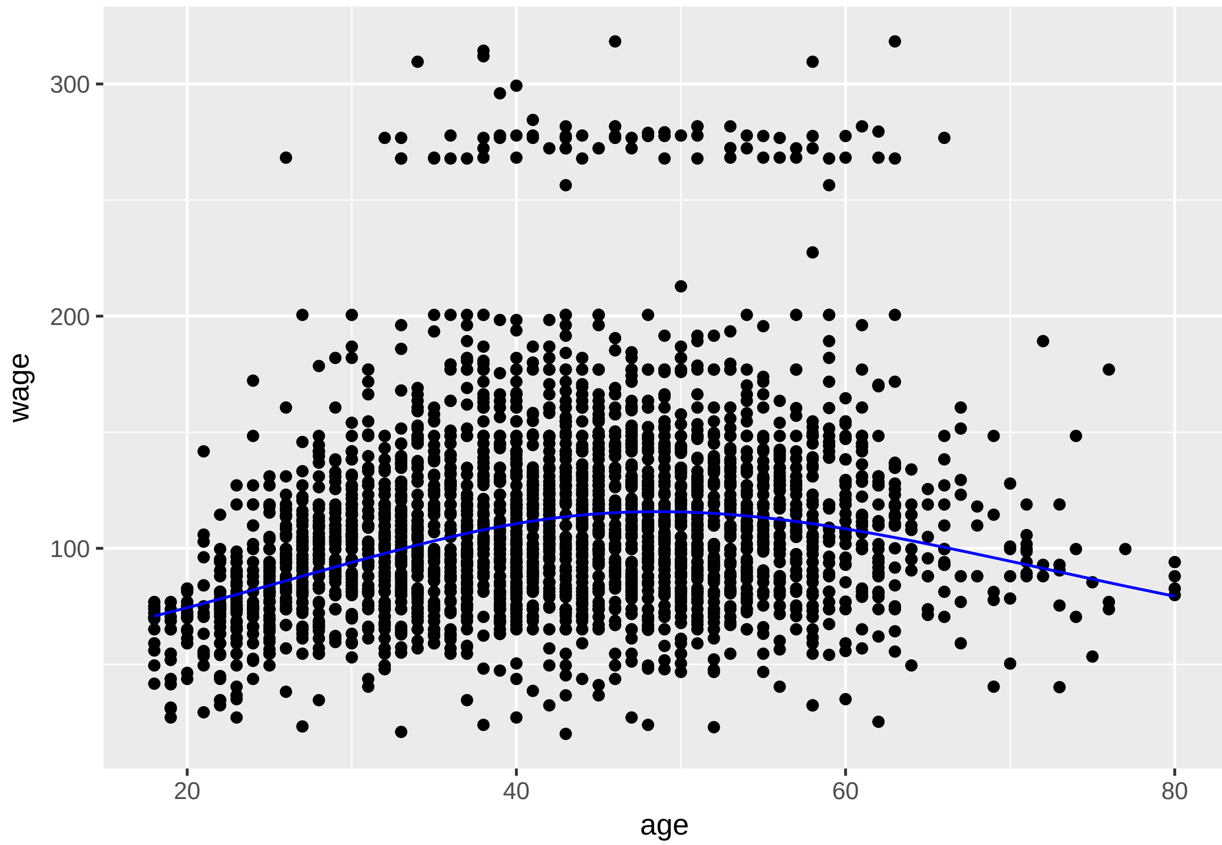
Exercise 4

- Fit a spline curve to the relationship between wage and age using two degrees of freedom (`df=2`).
- Experiment with different values of `df` (degrees of freedom)
- Select one that you think is about right.

```r
library(ISLR)
library(splines)
library(ggplot2)
p <- qplot(age, wage, data=Wage)

fit <- lm(log(wage) ~ ns(age, df=2), data=Wage)
Wage$fc <- exp(fitted(fit))

p + geom_line(aes(age, fc), data=Wage, col='blue')
```



Exercise 5

Now we will test which value of `df` minimizes the MSE on some test data.

First, we randomly split the `Wage` data set into training and test sets, with 2000 observations in the training data and the remaining 1000 observations in the test data.

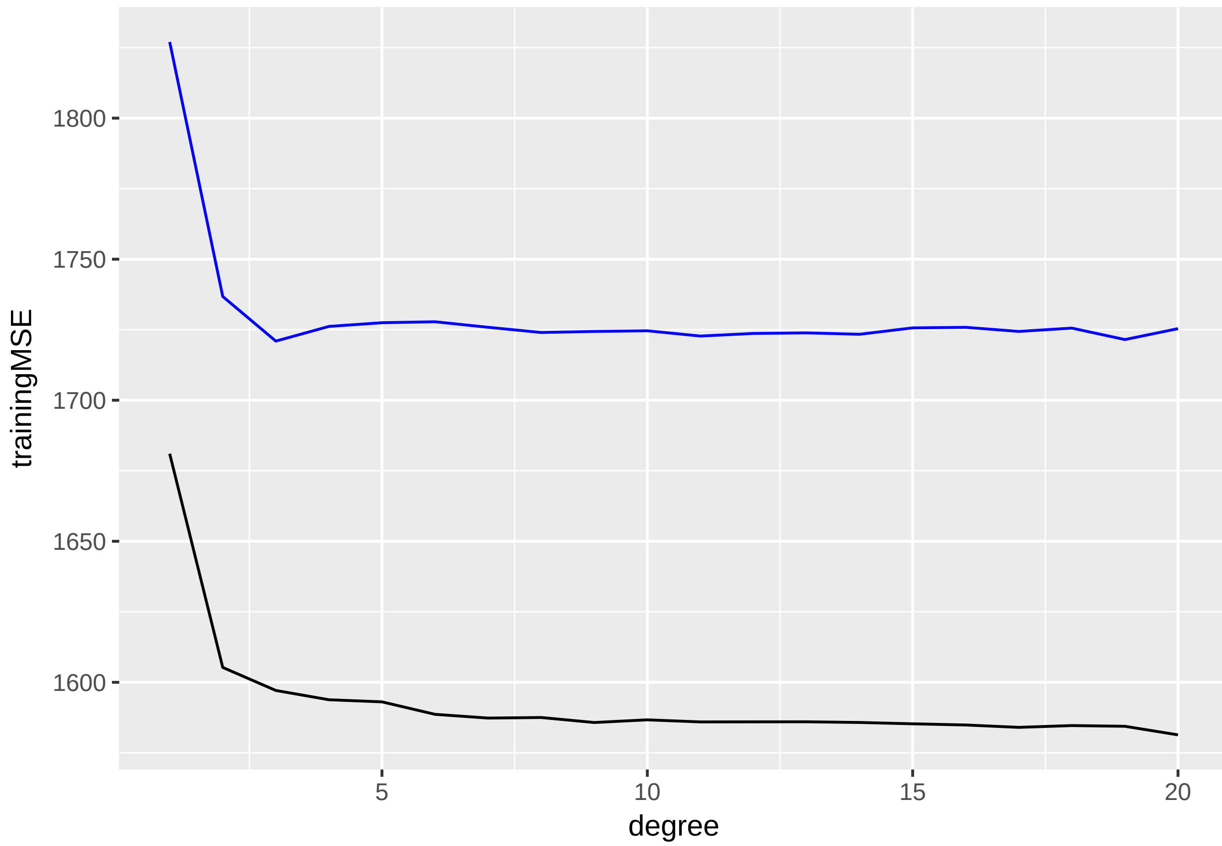
```
library(ISLR)
idx <- sample(1:nrow(Wage), size=2000)
train <- Wage[idx,]
test <- Wage[-idx,]
```

- Using a loop, compute the training and test MSE for `df = 1, 2, ..., 20`, and store it in two vectors `trainingMSE` and `testMSE`.
- Plot both `trainingMSE` and `testMSE` as a function of `df`.
- Which value of `df` gives the minimum training MSE?
- Which value of `df` gives the minimum test MSE?
- Plot a vertical line at your "guessed" value of `df`. How close is it to the optimal?
- Do you get the same results if you repeat the exercise on different splits of training and test data? Why?

```
# MSE on training and test sets
trainingMSE <- testMSE <- numeric(20)
for(i in 1:20)
{
  fit <- lm(log(wage) ~ ns(age, df=i), data=train)
  trainingMSE[i] <- mean((train$wage - exp(fitted(fit)))^2)
  testMSE[i] <- mean((test$wage - exp(predict(fit,newdata=test)))^2)
```

```
}

qplot(degree, trainingMSE, geom="line",
      data=data.frame(degree=1:20, trainingMSE, testMSE)) +
  geom_line(aes(degree, testMSE), col='blue')
```



Exercise 5

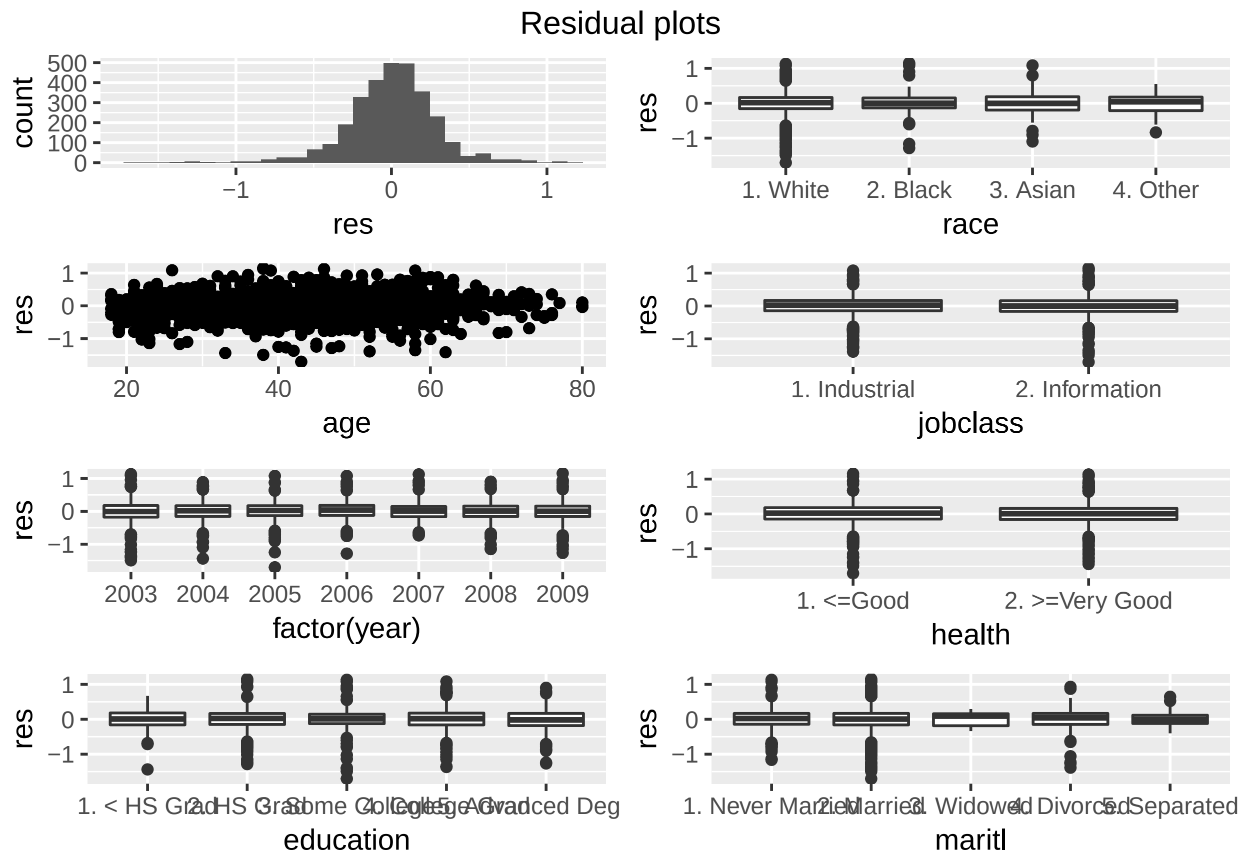
- Repeat the previous analysis, but use the full linear model including the other variables in the data set.
- How much better is the test MSE once you include the other predictor variables?
- Check your model by plotting the residuals as a function of each predictor variable. Do you see anything unusual in the residual plots?

```
fit <- lm(log(wage) ~ year + ns(age, df=5) + education + race + jobclass + health + maritl, data=Wage)
```

```
library(gridExtra)
res <- residuals(fit)
resplots <- list()
resplots[[1]] <- qplot(res)
resplots[[2]] <- qplot(age,res, data=Wage)
resplots[[3]] <- qplot(factor(year),res, data=Wage, geom="boxplot")
resplots[[4]] <- qplot(education,res, data=Wage, geom="boxplot")
resplots[[5]] <- qplot(race,res, data=Wage, geom="boxplot")
resplots[[6]] <- qplot(jobclass,res, data=Wage, geom="boxplot")
resplots[[7]] <- qplot(health,res, data=Wage, geom="boxplot")
```

```r
resplots[[8]] <- qplot(maritl,res, data=Wage, geom="boxplot")

marrangeGrob(resplots, ncol=2, nrow=4, top="Residual plots")
```




```r
res <- residuals(fit)
outliers <- subset(Wage, abs(res) > 1.5)
```