Business Analytics - ETC3250 2017 - Lab 4

Souhaib Ben Taieb

16 August 2017

Regression

Exercise 1

Understand all the steps in the proof of the bias-variance decomposition (see https://github.com/bsouhaib/BA2017/blob/master/slides/2/2-biasvardecomp.pdf)

Assignment - Question 1

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?

Exercise 2

Do some exploratory data analysis on the `Wage` data set.

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

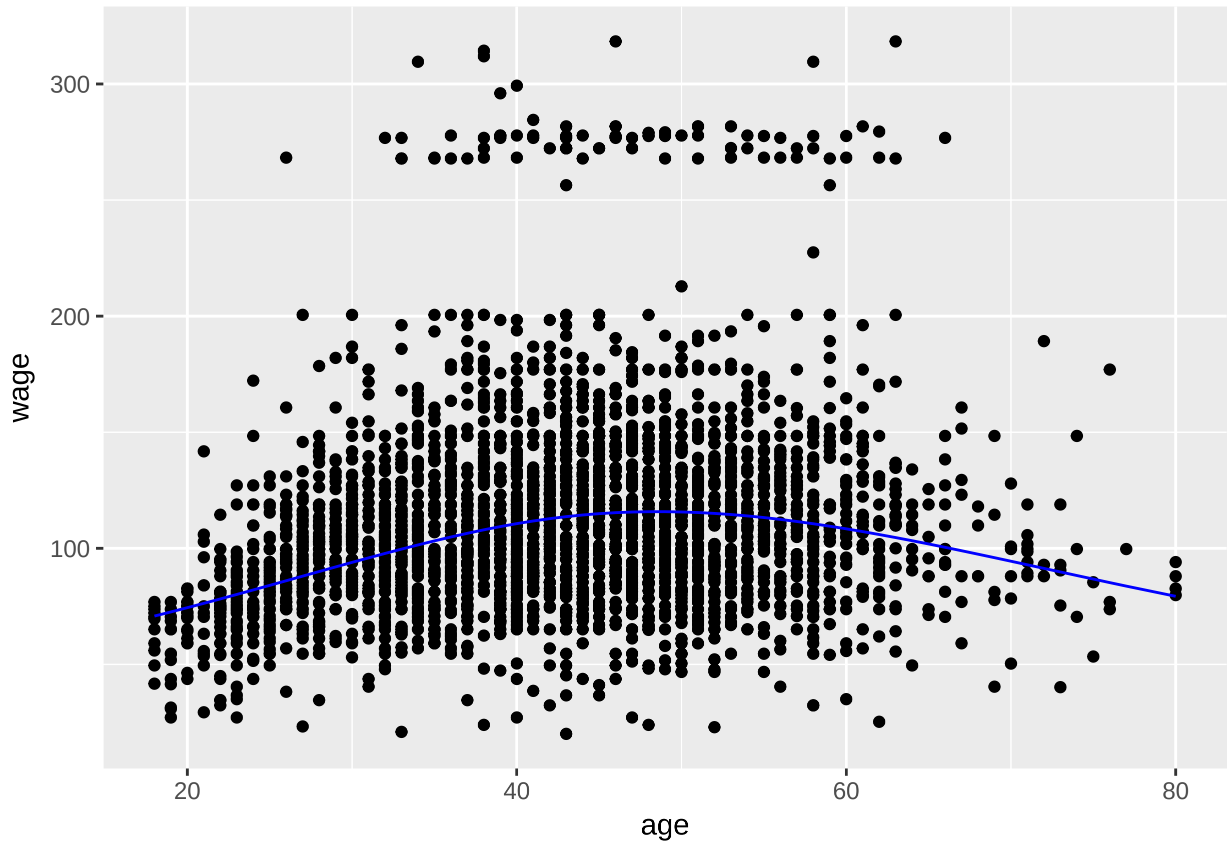
Exercise 3

The following code fits a spline curve to the relationship between wage and age.

```
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')
```



- Experiment with different values of `df` (degrees of freedom)
- Select one that you think is about right.

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 1000 observations in the test data.

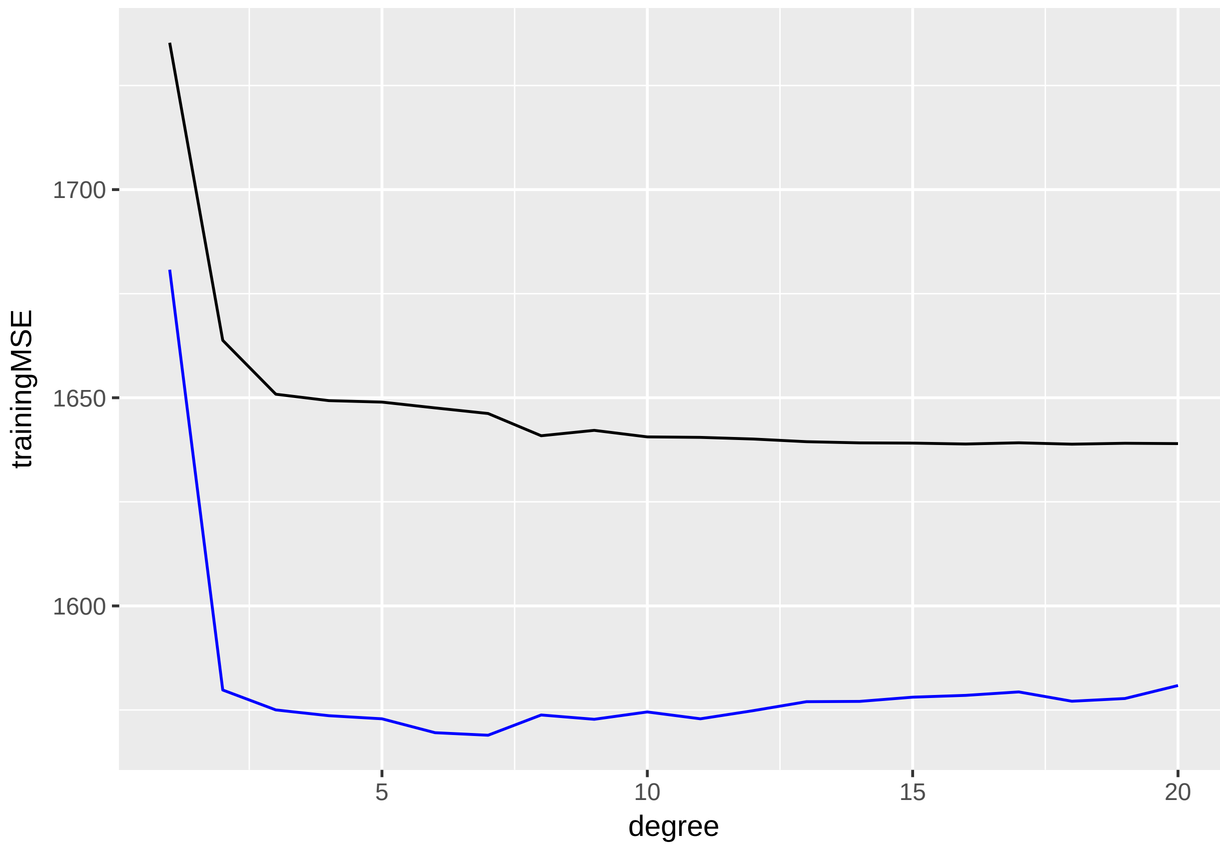
```
idx <- sample(1:nrow(Wage), size=2000)
train <- Wage[idx,]
test <- Wage[-idx,]
```

Exercise 4

Next try different values of `df`:

```
# 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)
}
```

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



- 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?

A model for wages

Repeat this analysis, but use the full linear model including the other variables in the data set. That is, fit models like this (but choose the optimal `df`):

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

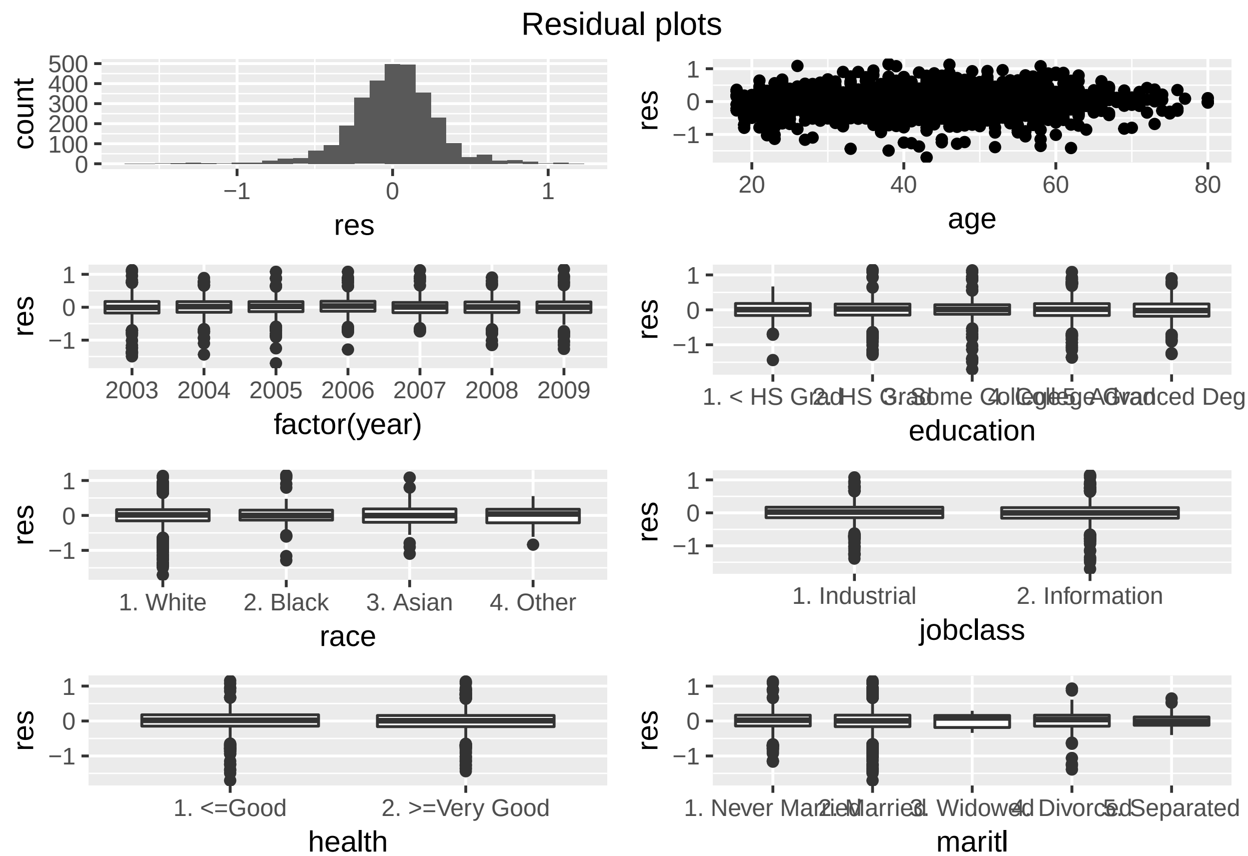
How much better is the test MSE once you include the other predictor variables?

Finally, we will check the residuals, assuming your best model is stored as `fit`.

```r
library(gridExtra)
res <- residuals(fit)
resplots <- list()
```

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

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



Do you see anything unusual in the residual plots?

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

What makes the outlier unusual?

Assignment - Question 2

Write code that includes:

1. fitting your final model above;
2. summary statistics for the residuals;
3. a plot of the residuals against the fitted values.

Classification

Assignment - Question 3

Do the exercise 7 in Section 2.4 of ISLR.

Assignment - Question 4

We want to predict whether a given car gets high or low gas mileage based on a set of features describing the car. We will use the *Auto* dataset in library ISLR.

1. Create a binary variable, *mpg01*, that contains a 1 if *mpg* contains a value above its median, and a 0 if *mpg* contains a value below its median. Add the variable *mpg01* to the data.frame *Auto*
2. Explore the data graphically in order to investigate the association between *mpg01* and the other features. Which of the other features seem most likely to be useful in predicting *mpg01*? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.
3. Split the data into a training set and a test set.
4. Perform KNN on the training data, with several values of K, in order to predict *mpg01*. Use only the variables that seemed most associated with *mpg01*. Plot the training and testing errors a function of $1/k$. Compare which value of K seems to perform the best when using training or testing errors?

TURN IN

- Your `.Rmd` file (which should knit without errors and without assuming any packages have been pre-loaded)
- Your Word (or pdf) file that results from knitting the Rmd.
- DUE: 20 August, 11:55pm (late submissions not allowed), loaded into moodle