Model Comparison – Example 2 MSE, Bias-Variance Motivation

September 10, 2018

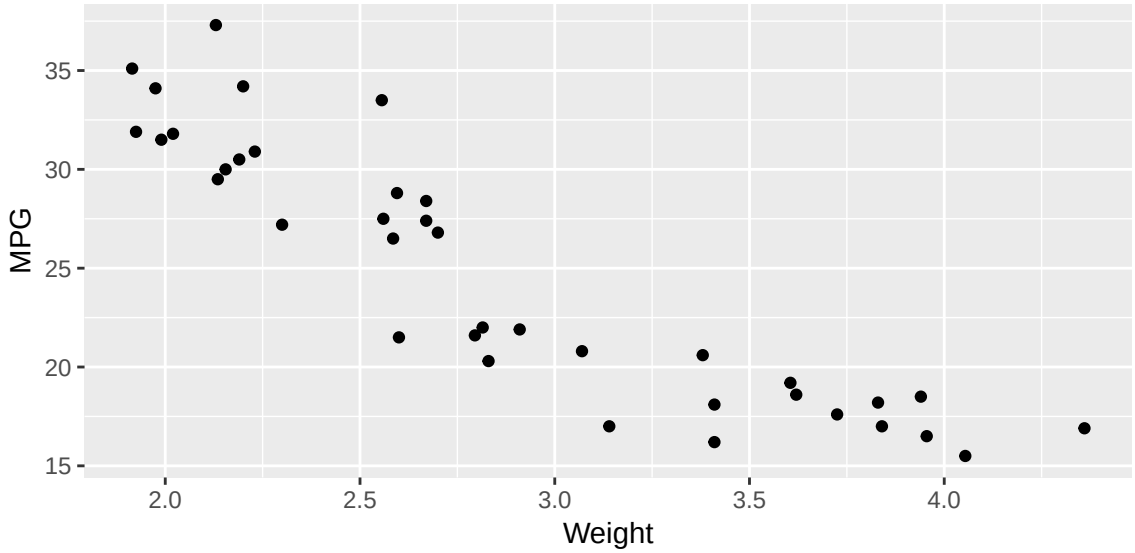
Let's consider one more time the example of polynomial models for the relationship between a car's weight (`Weight`) and its fuel efficiency (`MPG`).

We will compare degree 1 (i.e., linear) and degree 5 polynomials fit to 1000 different randomly selected subsets of our data of size 10.

```
library(dplyr) # for data manipulation functions
library(tidyr) # for data manipulation functions
library(readr) # for read_csv, which can read csv files from the internet
library(ggplot2) # for making plots
library(gridExtra) # for grid.arrange, which arranges the plots next to each other
library(polynom) # for obtaining the third polynomial fit below

cars <- read_csv("http://www.evanlray.com/data/sdm4/Cars.csv")

ggplot() +
  geom_point(data = cars, mapping = aes(x = Weight, y = MPG))
```




```
set.seed(773850)
prediction_grid <- seq(from = min(cars$Weight), to = max(cars$Weight), length = 1001)
prediction_df <- data.frame(
  Weight = prediction_grid
)

preds <- bind_rows(lapply(1:100, function(i) {
  obs_inds <- sample(seq_len(nrow(cars)), size = 10, replace = FALSE)
  cars_subset <- cars[obs_inds, ]
  fit1 <- lm(MPG ~ Weight, data = cars_subset)
  fit5 <- lm(MPG ~ poly(Weight, 5), data = cars_subset)
```

```
  return(data.frame(
    sample_ind = i,
    model = c(rep("degree 1", length(prediction_grid)), rep("degree 5", length(prediction_grid))),
    Weight = rep(prediction_grid, 2),
    MPG = c(predict(fit1, prediction_df), predict(fit5, prediction_df))
  ))
}))

preds <- preds %>%
  mutate(sample_model = paste(sample_ind, model))
```

```
ggplot() +
  geom_line(data = preds, mapping = aes(x = Weight, y = MPG, color = model, group = sample_model), alpha
  geom_point(data = cars, mapping = aes(x = Weight, y = MPG)) +
  ylim(range(cars$MPG))
```

```
## Warning: Removed 29486 rows containing missing values (geom_path).
```