Model Comparison Example Wrap-Up

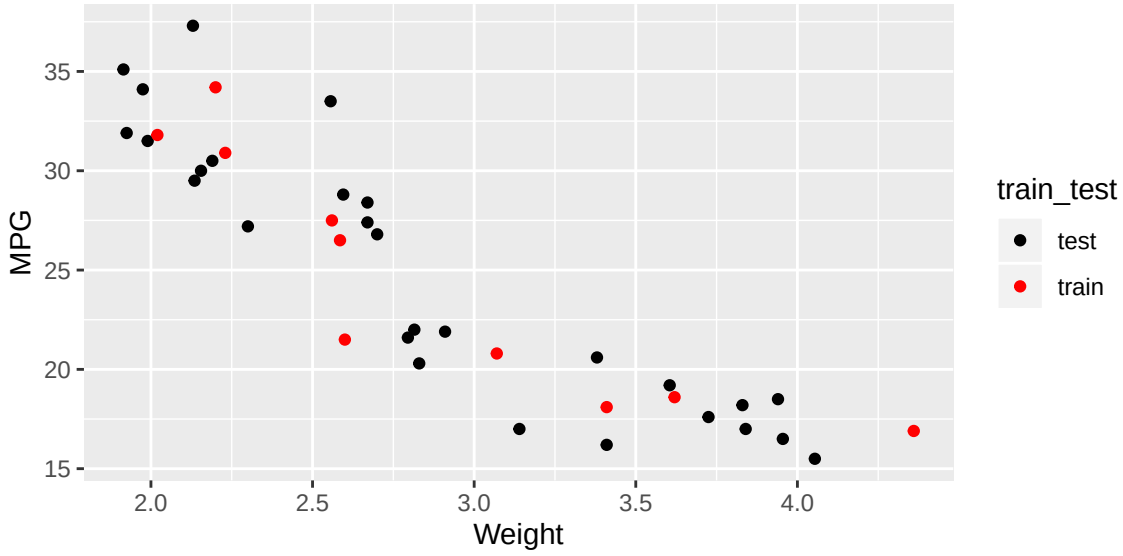
Recall we fit a line, a parabola, and a degree 9 polynomial to model the relationship between a car's weight (`Weight`) and its fuel efficiency (`MPG`).

We fit these models to 10 cars that were selected from a larger data set of 38 cars. These 10 cars are the **training set**: they were used to **train** the model, or estimate the model parameters.

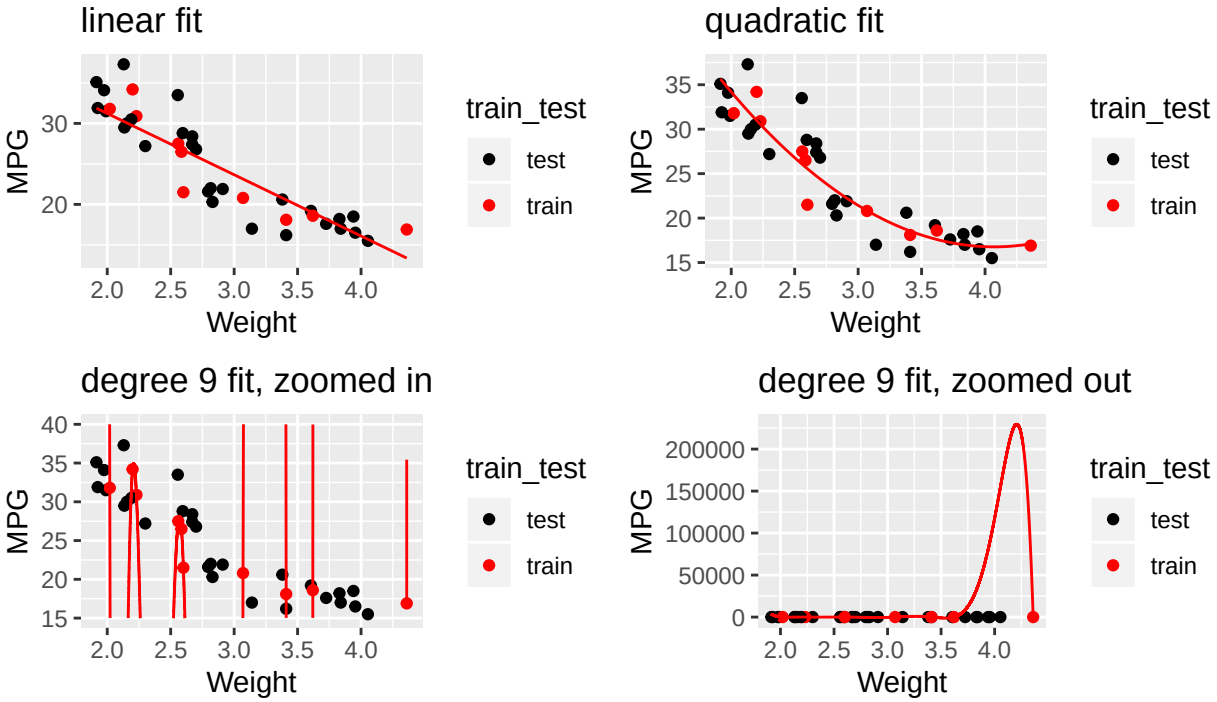
The remaining 28 cars can be used as a **test set**: a set of observations that were *not* used in model estimation, and can therefore be used to independently check the quality of the model fit.

The train and test set are labeled as such in the plot below.

```
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")
train_inds <- c(1, 6, 8, 14, 15, 16, 21, 32, 33, 37)
train_cars <- cars %>% slice(train_inds) # 10 observations to use in getting fits below.

cars$train_test <- "test"
cars$train_test[train_inds] <- "train"

ggplot() +
  geom_point(data = cars, mapping = aes(x = Weight, y = MPG, color = train_test)) +
  scale_color_manual(breaks = c("test", "train"), values = c("black", "red"))
```



Here are the plots again, over both the training and test sets. The estimated curves are shown in red, to indicate that they were fit to the training data set.



Below is R code for calculating and plotting the Mean Squared Error (MSE), separately for the train and test sets for each of our three candidate models:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2 = \frac{1}{n}SSR$$

The MSE contains the exact same information as SSR but doesn't grow with the sample size. This means it's more comparable across data sets of different sizes.

```
cars_residuals <- cars %>%
  mutate(
    residual_linear = MPG - predict_1(Weight),
    residual_quadratic = MPG - predict_2(Weight),
    residual_degree9 = MPG - predict_9(Weight)
  )

model_residual_summaries <-
  rbind(
    cars_residuals %>%
      dplyr::group_by(train_test) %>%
      dplyr::summarize(
        MSE = mean(residual_linear^2),
      ) %>%
      dplyr::mutate(
        degree = "1"
      ),
    cars_residuals %>%
      dplyr::group_by(train_test) %>%
      dplyr::summarize(
```

```
      MSE = mean(residual_quadratic^2),
    ) %>%
    dplyr::mutate(
      degree = "2"
    ),
  cars_residuals %>%
    dplyr::group_by(train_test) %>%
    dplyr::summarize(
      MSE = mean(residual_degree9^2),
    ) %>%
    dplyr::mutate(
      degree = "9"
    )
)

model_residual_summaries
```

```
## # A tibble: 6 x 3
##   train_test     MSE degree
##   <chr>        <dbl> <chr>
## 1 test       8.23e+0 1
## 2 train      7.40e+0 1
## 3 test       6.97e+0 2
## 4 train      3.51e+0 2
## 5 test       1.46e+9 9
## 6 train      1.10e-6 9
```

```
ggplot(data = model_residual_summaries) +
  geom_point(mapping = aes(x = degree, y = MSE)) +
  facet_wrap( ~ train_test, scales = "free")
```