Problem Set 1

Seth Harrison

1/17/2020

Statistical and Machine Learning

1. Describe in 500-800 words the difference between supervised and unsupervised learning.

What is the relationship between the X's and Y?

In supervised learning, the outputs (Y) are based on expected values based on prior knowledge. By contrast, in unsupervised learning, the outputs (Y) are not based on any external knowledge and are based only on the structure of the data.

What is the target we are interested in?

For supervised learning, because the outputs are based off of expected values, common targets include both classification and regression. For unsupervised learning, the targets are not specified beforehand and, as a result, common targets include clustering and principal component analysis. A broader distinction for targets can be stated as supervised targets have meaning external to the data, while unsupervised targets do not.

How do we think about data generating processes?

What are our goals in approaching data?

How is learning conceptualized?

Linear Regression

- Predict miles per gallon (mpg) as a function of cylinders (cyl). What is the output and parameter values for your model?

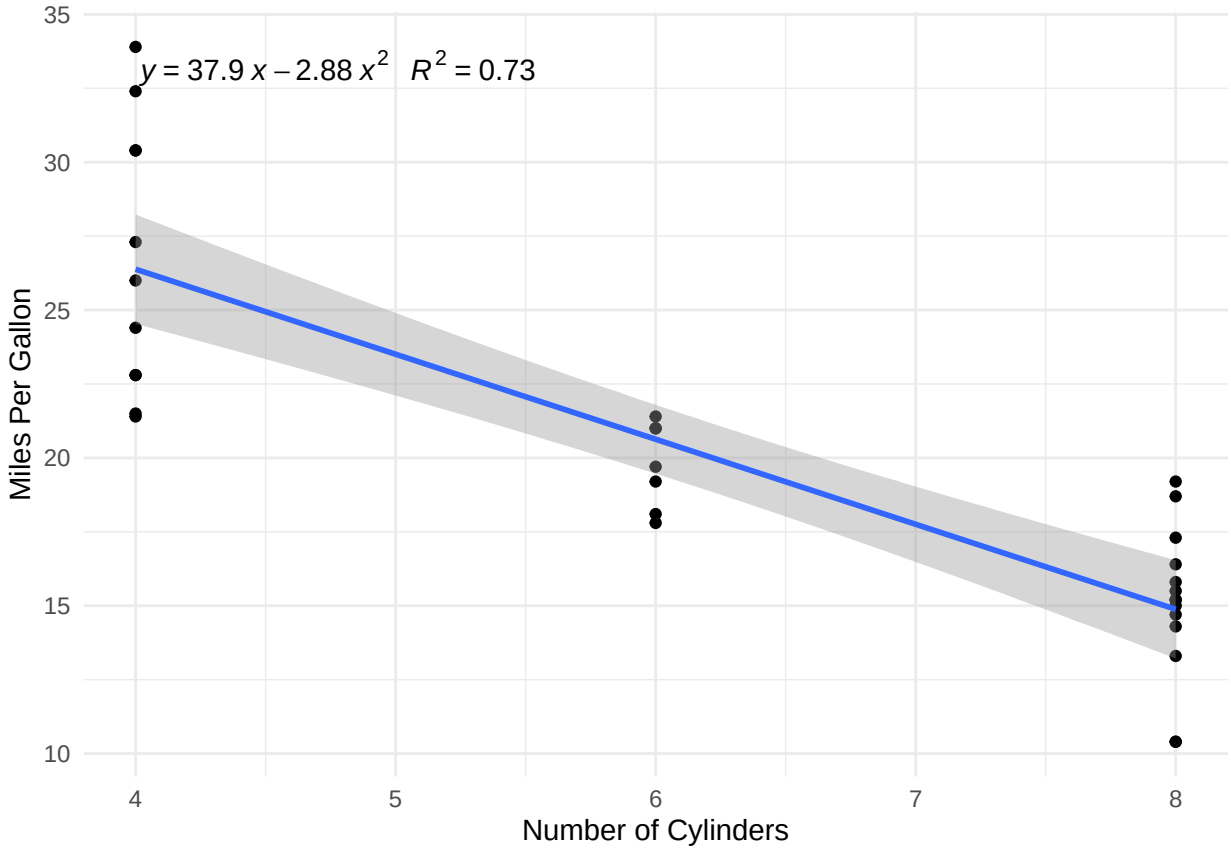
```
# Load Required Packages

library(tidyverse)
library(ggpmisc)
```

```
# Linear Model Function

LM1 <- lm(mtcars$mpg~mtcars$cyl,
            data = mtcars)

LM1
```

```
##
## Call:
## lm(formula = mtcars$mpg ~ mtcars$cyl, data = mtcars)
##
## Coefficients:
## (Intercept)   mtcars$cyl
##      37.885       -2.876
```

```
# Visualize Relationship

LM1 %>%
  ggplot(
    aes(x=mtcars$cyl,
```

```
    y=mtcars$mpg))+
  geom_point()+
  geom_smooth(method = "lm")+
  labs(x="Number of Cylinders", y="Miles Per Gallon")+
  theme_minimal()+
  stat_poly_eq(formula = LM1,
               aes(label = paste(..eq.label.., ..rr.label..,
                  sep = "~~~")),
                  parse = TRUE)
```



- Write the statistical form of the simple model in the previous question.

  y = -2.876x+37.885