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?

Supervised learning techniques, including linear regressions require that certain assumptions be made about the data. For example,

What are our goals in approaching data?

In unsupervised learning, the goal is to identify structures in the 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?

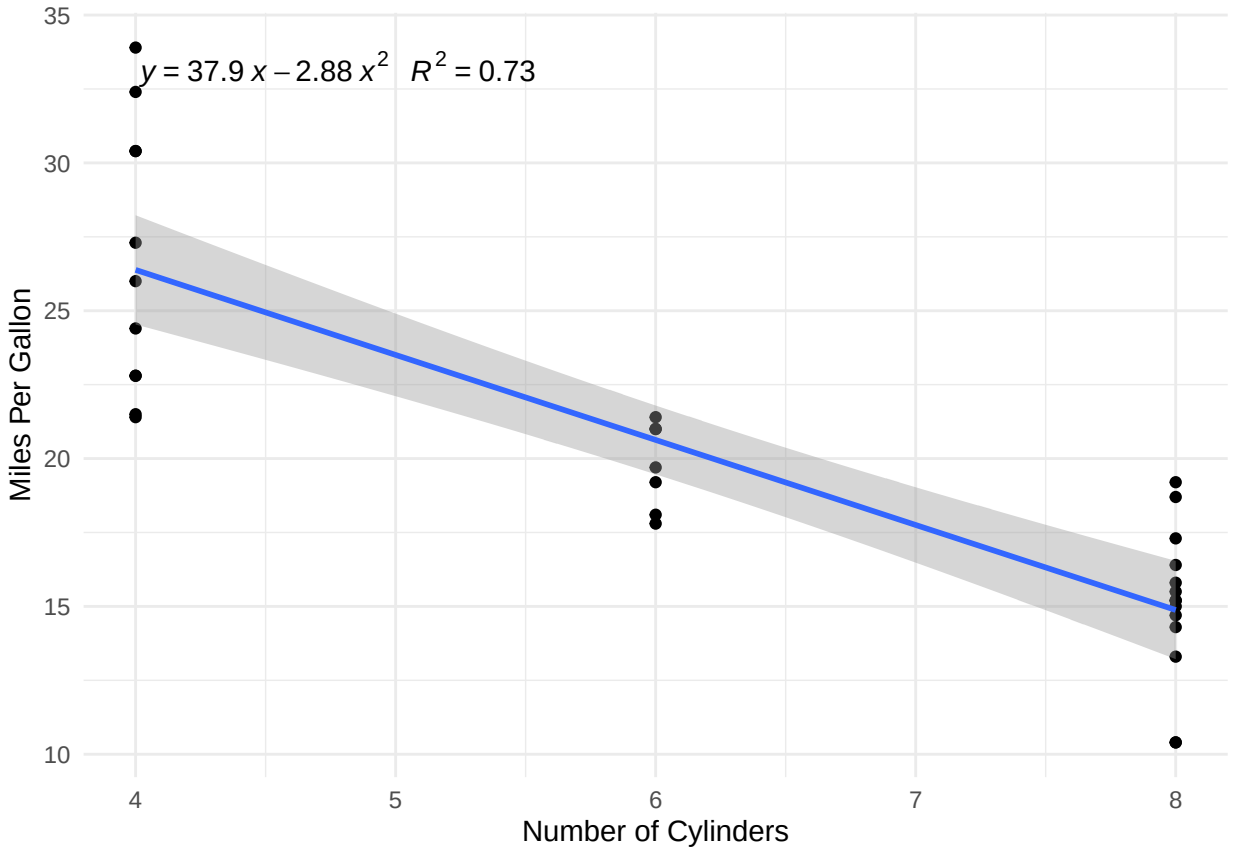
```
# Load Required Packages

library(tidyverse)
library(ggpmisc)
```

```
# Linear Model Function

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

summary(LM1)
```

```
##
## Call:
## lm(formula = mtcars$mpg ~ mtcars$cyl, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9814 -2.1185  0.2217  1.0717  7.5186
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.8846     2.0738   18.27  < 2e-16 ***
## mtcars$cyl   -2.8758     0.3224   -8.92 6.11e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.206 on 30 degrees of freedom
## Multiple R-squared:  0.7262, Adjusted R-squared:  0.7171
## F-statistic: 79.56 on 1 and 30 DF,  p-value: 6.113e-10
```

```r
# 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

- Add vehicle weight (wt) to the specification. Report the results and talk about differences in coefficient size, effects, etc.

```
LM2 <- lm(mtcars$mpg~(mtcars$cyl+mtcars$wt))

summary(LM2)
```

```
##
## Call:
## lm(formula = mtcars$mpg ~ (mtcars$cyl + mtcars$wt))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2893 -1.5512 -0.4684  1.5743  6.1004
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.6863     1.7150  23.141  < 2e-16 ***
## mtcars$cyl   -1.5078     0.4147  -3.636 0.001064 **
## mtcars$wt    -3.1910     0.7569  -4.216 0.000222 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.568 on 29 degrees of freedom
## Multiple R-squared:  0.8302, Adjusted R-squared:  0.8185
## F-statistic: 70.91 on 2 and 29 DF,  p-value: 6.809e-12
```