Problem Set 1

Seth Harrison

1/17/2020

Statistical and Machine Learning

```
# Load Required Packages

library(tidyverse)
library(readr)

# Load Data

wage_data <- read_csv("wage_data.csv")
```

- Describe the difference between supervised and unsupervised learning.

What is the relationship between the Xs and Y?

In supervised learning, the inputs (X) are given and the outputs (Y) are given by 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 models requires training data to establish the relationship between X and Y. Unsupervised learning, by contrast, does not use output data to cluster or otherwise identify structure in the data.

What are our goals in approaching data?

In supervised learning, the approach is to identify relationships between variables in the data. While in unsupervised learning, the goal is to identify structures in the data. As a result, supervised learning requires preprocessing to label, classify, or categorize the data.

How is learning conceptualized?

In supervised learning, the algorithm infers a function from from analyzing the training data. In unsupervised learning, the algorithm identifies patterns in the data set.

Linear Regression

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

```
Model1 <- lm(mtcars$mpg~mtcars$cyl,
             data = mtcars)

summary(Model1)
```

```
##
## 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
```

- 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.

```
Model2 <- lm(mtcars$mpg~(mtcars$cyl+mtcars$wt))

summary(Model2)
```

```
##
## 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
```

The relationship between number of cylinders and miles per gallon remained negative and significant, meaning every increase in the number of cylinders was associated with a decrease in miles per gallon. The coefficient value became closer to 0 for the car cylinder variable from the previous model, signifying a difference in the scales that measure cylinders and car weight. As a result, it is necessary to look to a standaradized measure of effect size, the R^2 value. Here, the adjusted R^2 jumped from ~.72 to ~.82, indicating that cylinders and car weight can predict miles per gallon better than cylinders alone.

- Interact weight and cylinders and report the results. What is the same or different? What are we theoretically asserting by including a multiplicative interaction term in the function?

```
Model3 <- lm(mtcars$mpg~(mtcars$cyl*mtcars$wt))

summary(Model3)
```

```
##
## Call:
## lm(formula = mtcars$mpg ~ (mtcars$cyl * mtcars$wt))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2288 -1.3495 -0.5042  1.4647  5.2344
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           54.3068     6.1275   8.863 1.29e-09 ***
## mtcars$cyl            -3.8032     1.0050  -3.784 0.000747 ***
## mtcars$wt             -8.6556     2.3201  -3.731 0.000861 ***
## mtcars$cyl:mtcars$wt   0.8084     0.3273   2.470 0.019882 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.368 on 28 degrees of freedom
## Multiple R-squared:  0.8606, Adjusted R-squared:  0.8457
## F-statistic: 57.62 on 3 and 28 DF,  p-value: 4.231e-12
```

Here the R-squared increased to ~.85, suggesting that a multiplicative relationship between weight and cylinders best predicts miles per gallon of the three. Theoretically, interacting the two predictor variables means that the value of one variable depends on the value of the second variable. We can see in this case that this accurately models the real world: cars that are heavier tend to have more cylinders in the engine.

Non-linear Regression

- Fit a polynomial regression, predicting wage as a function of a second order polynomial for age.

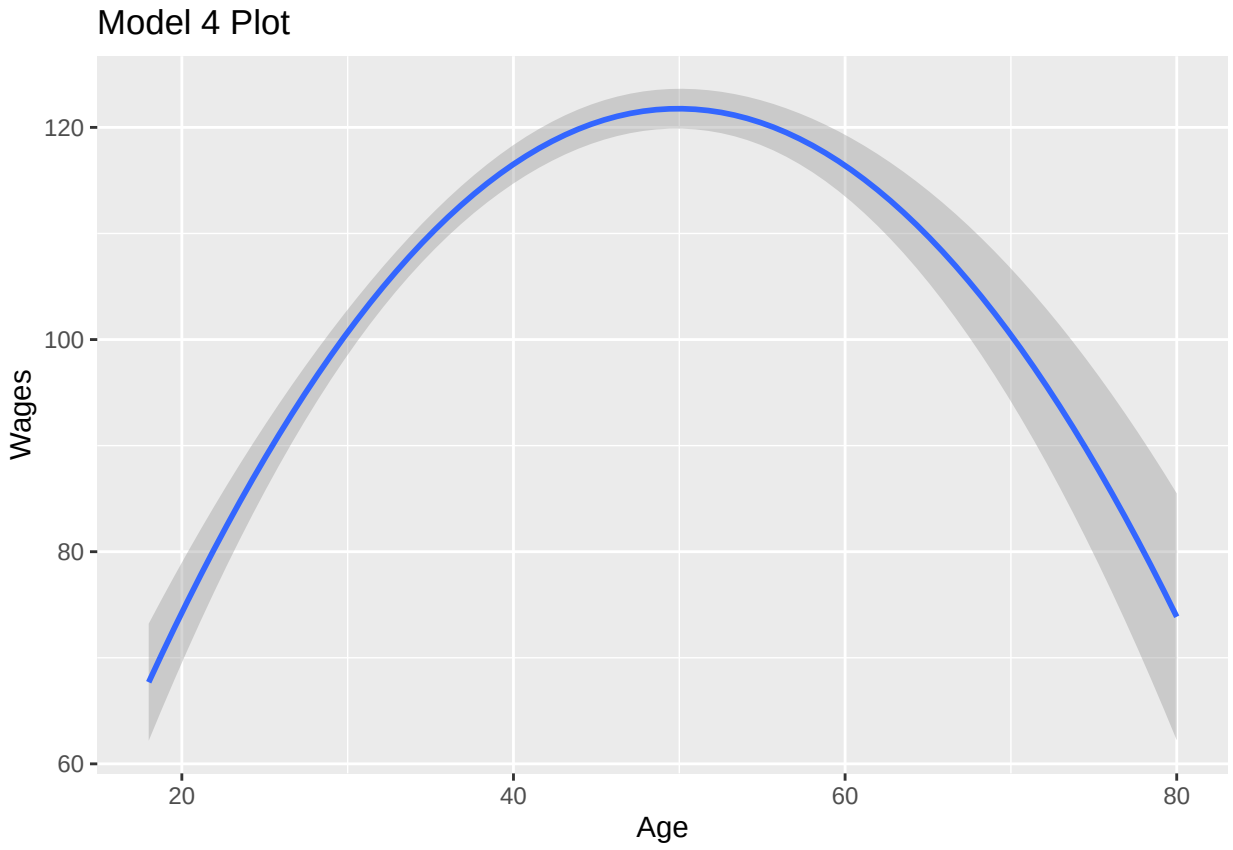
```
Age2 <- wage_data$age^2

Model4 <- lm(wage_data$wage~(wage_data$age+Age2))

summary(Model4)
```

```
##
## Call:
## lm(formula = wage_data$wage ~ (wage_data$age + Age2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -99.126 -24.309  -5.017  15.494 205.621
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -10.425224   8.189780  -1.273    0.203
## wage_data$age   5.294030   0.388689  13.620   <2e-16 ***
## Age2           -0.053005   0.004432 -11.960   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.99 on 2997 degrees of freedom
## Multiple R-squared:  0.08209,    Adjusted R-squared:  0.08147
## F-statistic:   134 on 2 and 2997 DF,  p-value: < 2.2e-16
```

- Plot the function with 95% confidence interval bounds.

```
wage_data %>%
    ggplot(
      aes(wage_data$age,wage_data$wage)) +
      stat_smooth(method = "lm",
                  formula = y ~ x + I(x^2), size = 1) +
      labs(x= "Age", y= "Wages", title ="Model 4 Plot")
```



- Describe the output. What do you see substantively? What are we asserting by fitting a polynomial regression?
- How does a polynomial regression differ both statistically and substantively from a linear regression (feel free to also generalize to discuss broad differences between non-linear and linear regression)?