Problem Set 1

Rei Bertoldi

1/15/2020

```
wage_df <- read_csv("wage_data.csv")
data(mtcars)
```

Statistical and Machine Learning

Supervised learning analyzes 'training data' and uses it to generate a function to be used to generalize reasonable future predictions. In supervised learning, we have the input (X) and output (Y) variables, and are capturing their relationship by approximating a mapping function (f) from X to Y. There are two different types of supervised learning: classification and regression. Classification approximates the mapping function from X to discrete Y variables. Discrete variables can be thought of a categorical variables or dummy variables belonging to one of two possibilities. For instance - male or female, enrolled in a program or not enrolled, control or treatment, ect. Regression modeling, similar to classification, but instead of dealing with discrete variables, the Y variables are real-values, i.e. integers. Often, these values are amount of sizes. In the `mtcars` data, for instance, miles per gallon are real-values. Regression Linear regression will capture the best fit line, which minimized the residual sum of squares, i.e. minimizing `sum((y - (alpha + beta*x))^2)`.

Unsupervised learning, on the other hand, is where we only have the input (X) data and no corresponding output (Y) variables. It is using the underlying structure or distribution of the data to learn. Algorithms find previously unknown patterns in the data without the pre-existing labels we see when using supervised learning analysis. There are two kinds of unsupervised learning: clustering and association. Clustering seeks to find inherent grouping in the data while association seeks to find rules that describe large portions of the data. An example of an unsupervised learning algorithm is K-means clustering, where the algorithm looks for a fixed number (k) of clusters in the data. 'Means' in K-means refers to the averaging of the data to find the center or centroid of the clusters in the data. The algorithm will iterate and optimize the position of the centroid for each cluster.

Linear Regression Regression

```
reg_model_1 <- lm(mpg ~ cyl, data = mtcars)
reg_model_2 <- lm(mpg ~ cyl + wt, data = mtcars)
reg_model_3 <- lm(mpg ~ cyl + wt + cyl*wt, data = mtcars)

mod_stargazer <- function(...){
  output <- capture.output(stargazer(...))
  output <- output[4:length(output)]
  cat(paste(output, collapse = "\n"), "\n")
}

mod_stargazer(reg_model_1,
          reg_model_2,
          reg_model_3,
          title = "Linear Regression Results")
```

1. Regression results are reported in the regression output table below.

a. Our α or intercept estimate is 37.8846. Our β or coefficient estimate on `cyl` is -2.8758. This is telling us that an additional behicle cylinder is associated with a -2.8758 decrease in miles per gallon.

Table 1: Linear Regression Results

	Dependent variable:		
	mpg		
	(1)	(2)	(3)
cyl	-2.876^{***}	-1.508^{***}	-3.803^{***}
	(0.322)	(0.415)	(1.005)
wt		-3.191^{***}	-8.656^{***}
		(0.757)	(2.320)
cyl:wt			0.808^{**}
			(0.327)
Constant	37.885^{***}	39.686^{***}	54.307^{***}
	(2.074)	(1.715)	(6.128)
Observations	32	32	32
R^2	0.726	0.830	0.861
Adjusted R^2	0.717	0.819	0.846
Residual Std. Error	3.206 (df = 30)	2.568 (df = 29)	2.368 (df = 28)
F Statistic	79.561^{***} (df = 1; 30)	70.908^{***} (df = 2; 29)	57.618^{***} (df = 3; 28)
Note:			*p<0.1; **p<0.05; ***p<0.01

b.

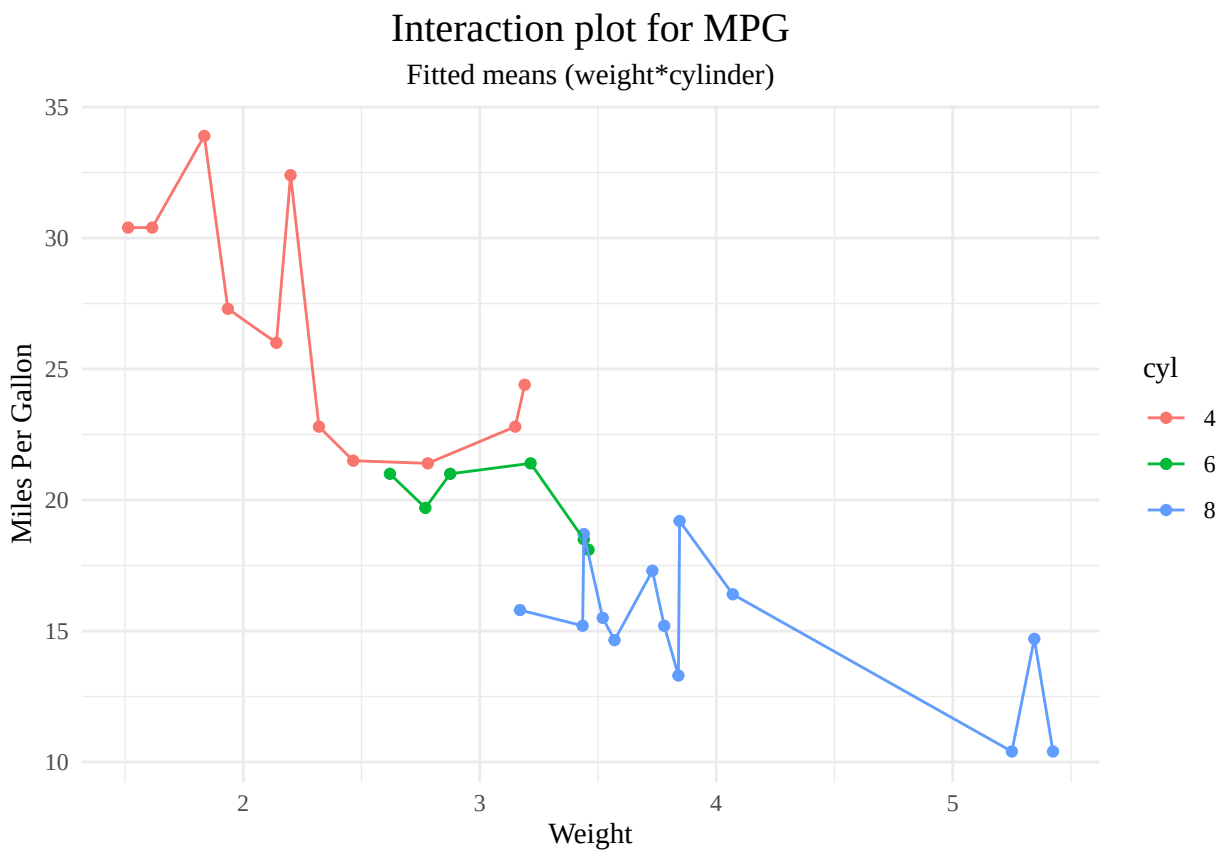
$$Y_{mpg} = \alpha - \beta * X_{cyl} + \epsilon$$
$$Y_{mpg} = 37.8846 - 2.8758 * X_{cyl} + \epsilon$$

c. Our α or intercept estimate is 39.6863. Our β or coefficient estimate on `cyl` is -1.5078. Our β or coefficient estimate on `wt` is -3.1910. We see that the absolute value of the coefficient estimate on `cyl` decreased after adding `wt` as a control. We might attribute this decrease in effect to bias in our previous model. In other words, our previous model was over-estimating the effect of `cyl` on `mpg`, and adding `wt` as a control in our second model allowed `wt` to eat up some of the effect estimate.

d. It means that the slope of one continuous variable on the response variable changes as the values on a second continuous change. In other words, the estimated effect of weight on average miles per gallon, changes as the the number of cylinders in the vehicle changes. When we are interacting `wt` and `cyl` in our model, we are theoretically asserting this interaction between weight and cylinders. In the interaction plot below, we can see that as the number of cylinders increases, so does the weight of the vehicle, which is associated with lower mean miles per gallon.

```
mtcars %>%
  mutate(cyl = as.character(cyl)) %>%
  ggplot() +
  aes(x = wt, color = cyl, group = cyl, y = mpg) +
  stat_summary(fun.y = mean, geom = "point") +
  stat_summary(fun.y = mean, geom = "line") +
  labs(x = "Weight",
       y = "Miles Per Gallon",
       title = "Interaction plot for MPG",
       subtitle = "Fitted means (weight*cylinder)") +
```

```
  theme_minimal(base_family = "serif") +
  theme(plot.title = element_text(size=15, hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))
```



Non-linear Regression

```
df <- data.frame(wage=wage_df$wage, age=wage_df$age)
reg_results <- lm(wage ~ age + I(age^2), data = df)
predict <- as.data.frame(
  predict(reg_results, x=wage, interval = 'confidence', level=0.95))
predict_df <- cbind(df, predict)
```

```
ggplot(predict_df, aes(x = age)) +
  geom_point(aes(y = wage),
             color ="purple",
             alpha = 0.2) +
  geom_line(aes(y = fit)) +
  geom_ribbon(aes(ymin = lwr, ymax = upr),
              fill = "blue", alpha = 0.2) +
  labs(
    x = "Age",
    y = "Wage",
    title =
      "Predicting wage as a function of a second order polynomial for age",
    subtitle = "With 95% confidence interval bounds") +
```

```
theme_minimal(base_family = "serif") +
theme(plot.title = element_text(size=15, hjust = 0.5),
      plot.subtitle = element_text(hjust = 0.5))
```

c. Here we are using polynomial regression to fit a quadratic curve to our data. In our plot we see the results of fitting a 2-degree polynomial using least squares (the solid blue lines). When we fit a quadratic curve, we are relaxing the linearity assumption in our data and allowing for non-linearity. We are making assumptions of best fit in the sense that we must make the decision, as researchers, on the degree of polynomial to use. The shaded lines on our graph are our standard error, which correspond to an approximate 95% confidence interval. There also appears to be two groups within the population, earners above $250,000$ and income earners below that threshold. We also see that the standard error coverage gets larger at the tails of our fitted curve, which may be due to smaller n. Furthermore, the concave fit of our curve is tell us that wage increases with age until around 50, after which is begins to decrease.

d. Linear regression can have limited predictive power and sometimes, using polynomial regression can improve least squares. Polynomial regression, statistically, is achieved by raising predictors to a power, which allows for a non-linear fit to our data. Where, a standard linear regression, i.e. $Y = \alpha + \beta * x + \epsilon$, might predict that as age increases, so does wage, while adding a power, i.e. $Y = \alpha + \beta * x + \beta * x^2 + \epsilon$, allows for the fit we see in the graph below where wage increases until 50, then decreases. This may be closer to the true model of our data.