STAT 380 – Introduction to Regression Part 1 (Lecture 4)

Definition: Statistical learning refers to a vast set of tools for understanding data. These tools can be classified as supervised or unsupervised.

Definition: Broadly speaking, supervised learning involves building a statistical model for predicting, or estimating, an output based on one or more inputs.

NOTE: Input variables go by many names including: x, explanatory, predictor, regressor, independent variable, feature, and attribute.

NOTE: The output variable also has many names including: y, response, dependent variable, target

Definition: In unsupervised learning, there are inputs but no supervising output; nevertheless we can learn relationships and structure from such data.

NOTE: Supervised learning problems fall into two broad categories: regression and classification. Regression techniques are used when the output/response variable is quantitative. Classification techniques are used when the response is categorical.

NOTE: In this class, we will primarily focus on supervised learning problems. Suppose that we observe a quantitative response, Y, and p predictors, $X_1, X_2, \dots, X_p$. Assuming there is a relationship between Y and $X = (X_1, X_2, \dots, X_p)$, we describe the general form of the relationship as:

General Form of Regression: $Y = f(X) + \epsilon$

EXAMPLE 1: In simple linear regression: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

EXAMPLE 2: In multiple linear regression: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$

NOTE: There are two main reasons for estimating f: Prediction and Inference

Prediction:

- Setting: X's are readily available and we want to predict Y as accurately as possible
- Implementation: Predict Y using $\hat{Y} = \hat{f}(X)$ where $\hat{f}$ is an estimate of f and will not be perfect

Inference:

- Setting: We have a collection of X's and Y's and we want to understand/explain the association between X and Y
- Implementation: Understand the nature of $\hat{f}$
- Types of Questions:
    - Which predictors are associated with the response?
    - Is a linear relationship appropriate?
    - How does each predictor affect the response?

EXAMPLE 3: The file L4_Vehicles.csv is a simulated dataset that contains information regarding the mileage (in thousands of miles), the price (in thousands of dollars), manufacturer, and model for a random sample of used cars currently for sale. All vehicles in the dataset were manufactured in 2016.

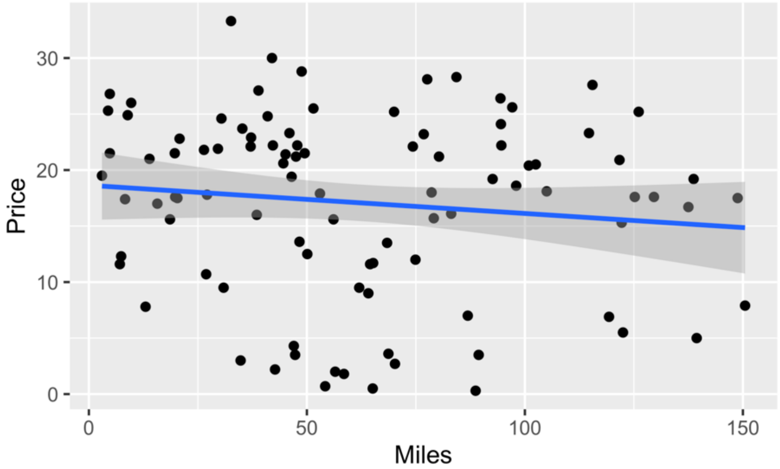
a. The column names are not provided. What is the order of the columns? Explain your reasoning.

First Second Third Fourth

b. Read the dataset into R and name the result Vehicles. Create descriptive variable names (i.e., do not use V1, V2, V3, and V4).

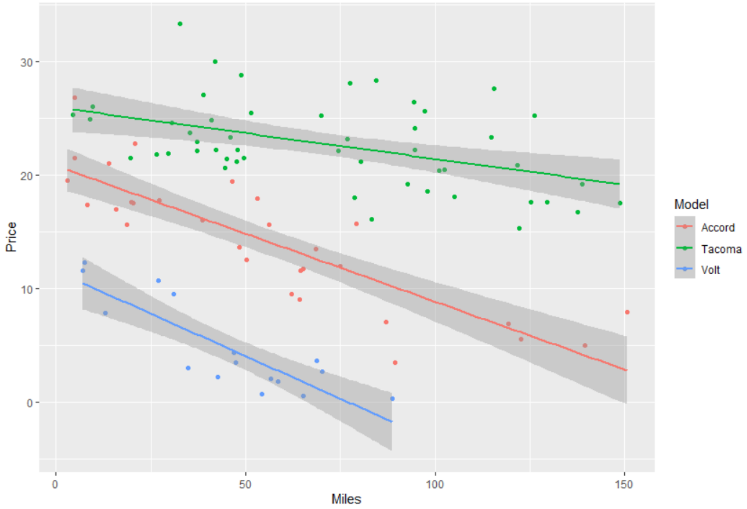
c. Using the quantitative variables, which variable do you think is most likely to explain the other?

d. Create a scatterplot showing the relationship between mileage and price and include the linear smoother. Do you think there is an association between price and miles? Find the correlation.



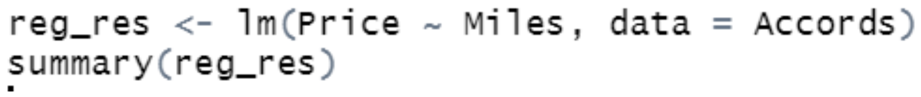
NOTE: If we believe one variable explains the other, the convention is to plot the explanatory variable on the horizontal (x) axis.

e. Create a scatterplot showing the relationship between mileage and price. Use a different color for the points (observations) associated with each model and include the linear model smoother. What did you learn by examining the plot? Do you think that there is a linear relationship between mileage and price when accounting for the model?

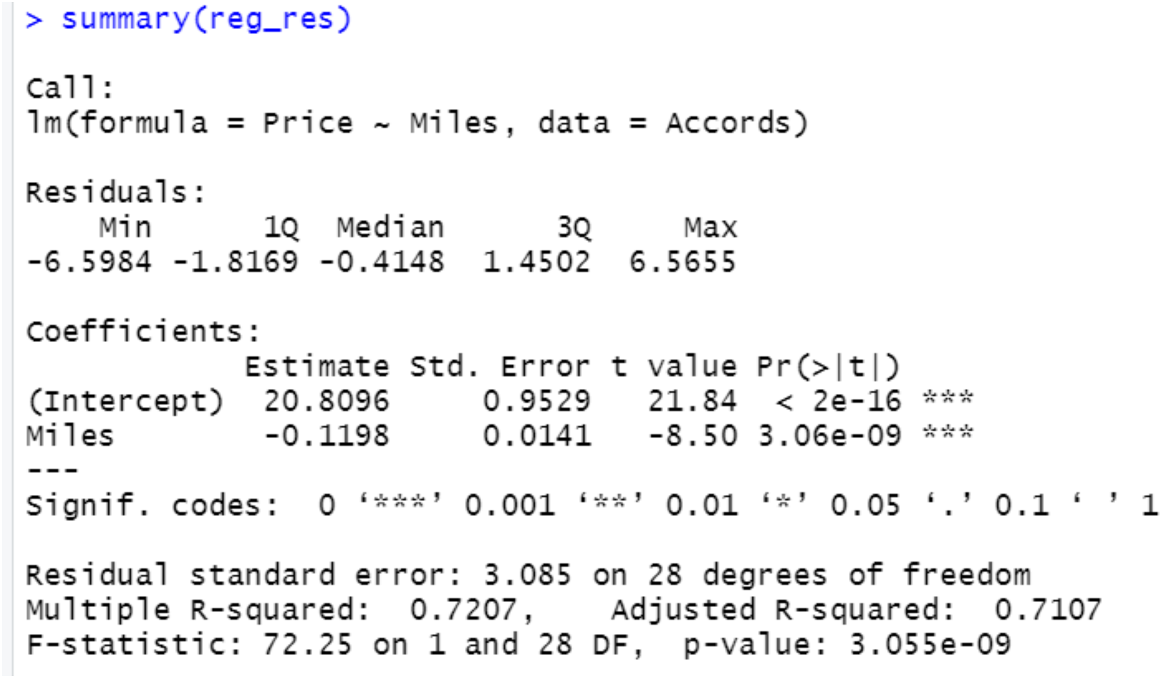


f. Create a new dataset called Accords that only includes the Honda Accords.

g. Find the estimated regression equation for predicting the price of an Accord based on its mileage.



```
reg_res <- lm(Price ~ Miles, data = Accords)
summary(reg_res)
```



```
> summary(reg_res)

Call:
lm(formula = Price ~ Miles, data = Accords)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.5984 -1.8169 -0.4148  1.4502  6.5655 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  20.8096     0.9529   21.84  < 2e-16 ***
Miles        -0.1198     0.0141   -8.50 3.06e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.085 on 28 degrees of freedom
Multiple R-squared:  0.7207,	Adjusted R-squared:  0.7107 
F-statistic: 72.25 on 1 and 28 DF,  p-value: 3.055e-09
```

h. Predict the price for an Accord with 62.0 (really 62,000) miles.

NOTE: The predicted (or estimated) value of the response is denoted as

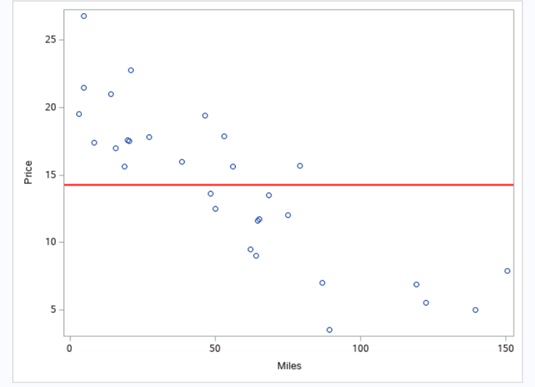
i. Within the dataset, there is one observation with 62000 miles. How does the actual price (also called observed price) of that car compare to the predicted price?

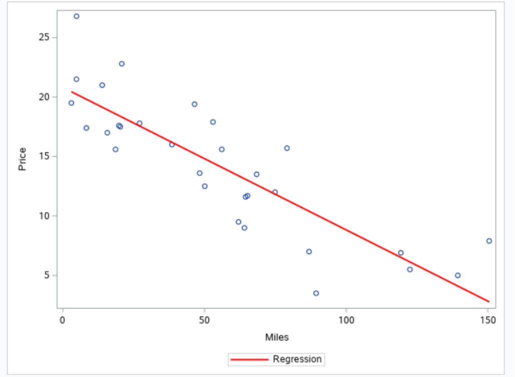
Definition: The residual for the i^{th} observation, denoted as e_i, is the difference between the observed value, y_i, and the predicted value, $\hat{y}_i$. In symbols,

IMPORTANT QUESTION: How do we determine the optimal line?

Least-squares Condition – Minimize the sum of the squared residuals. NOTE: R denotes the sum of the squared residuals as RSS.

j. In each picture, label the residual associated with car that has the most miles.





k. (Inference) Based on the model we are using, explain the relationship between prices and miles. (To answer this, you should interpret the slope in the context of the problem.)

General: The slope represents the average change in y as x increases by 1 unit.

l. Interpret the y-intercept in the context of the problem. Do you trust the model to make this prediction? Explain.

General: The y-intercept represents the expected value (mean value) of y when x is 0.

CAUTION: Beware predictions made outside the range of the observed data (X's) –

m. What proportion of the total variation in price in explained by the model including miles for the Accords dataset?

n. What other variables could. be used to help explain the variation in price?

o. What is the "average" amount the observed responses (price) will deviate from the line?

Facts about R^2

1. R^2 measures how well the regression line explains the response. In particular, R^2 is the proportion of total variation in y explained by the least-squares regression line.
2. R^2 is a number between 0 and 1.
3. The closer R^2 is to 1, the better the model explains the relationship between x and y.