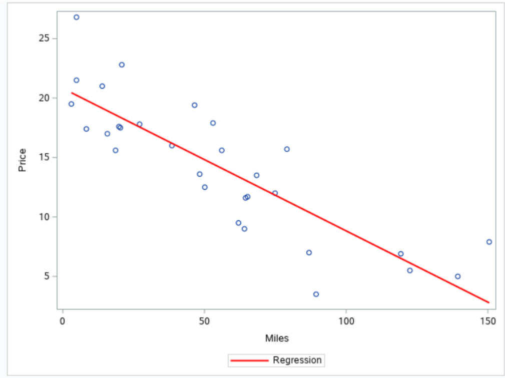
STAT 380 – Introduction to Regression Part 2 (Lecture 5)

Simple Linear Regression

IDEA: In simple linear regression, we construct a model for predicting a quantitative response y_i for a given value of the predictor variable (or input), x_i.

RECALL: The simple linear regression (SLR) model given by:

$$\text{True/Population Model: } y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$



where

$$\beta_0 = \text{the intercept which represents the } expected\ value \text{ (or mean value) } of\ y\ when\ x = 0$$

$$\beta_1 = \text{the slope which represents the } average\ change\ in\ y\ for\ a\ 1\ unit\ increase\ in\ x$$

$$\epsilon = \text{the error term which is a catch all for what we missed with the simple model}$$

NOTE: What are the assumptions/conditions?

1. Linearity - The mean response, $\mu_{y_i|x_i}$, is a linear function of the predictor x_i.

2. The errors, ϵ_i's, are independent.

3. The errors, ϵ_i's are normally distributed at each value of the predictor x_i.

4. The errors, ϵ_i's, have equal (constant) variance at any given x_i (e.g., $Var[\epsilon_i \mid x_i] = \sigma^2$). This is called homoscedasticity of the errors.

5. The errors, ϵ_i's, are on average 0 (i.e., $E[\epsilon_i \mid x_i] = 0$).

NOTE: Assumptions 2-5 can be written using:

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

which means that the error terms are independent and identically distributed (iid) according to a Normal distribution with mean 0 and variance σ^2.

NOTE: In the simple linear regression model, we have to estimate the unknown parameters: β_0, β_1, and σ^2. In terms of the β's, we can estimate the β's using the least-squares condition which tells us to minimize the sum of the square residuals:

$$\text{Objective Function: } Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

NOTE: We will denote the sum of the squared residuals as RSS (Residual Sum of Squares).

NOTE: The estimated model is given by: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

NOTE: In order to assess how well the simple linear regression model fits the data, we often use:

- R^2 which is the proportion of the total variation in the response (y) explained by the model with x.
- Residual Standard Error (RSE) which is an estimate of the standard deviation of ϵ. You can think of RSE as the average amount the response will deviate from the line. RSE is calculated using

$$RSE = \sqrt{\frac{RSS}{n-p}}$$

where n is the number of observations and p represents the number of β's in your model.

IDEA: In simple linear regression, the "simple" tells us that there is a single predictor (x) variable; however, there is nothing that limits us to using only one predictor variable in our regression models.

Definition: Multiple linear regression (MLR) extends single predictor variable regression (simple linear regression) to the case that still has one response but many predictors.

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k} + \epsilon_i$$

NOTE: The x terms now have 2 subscripts:

$$x_{i,2}$$

EXAMPLE 1: Our goal is to build a regression model for predicting Price based on miles and model for the dataset L4_Vehicles.csv.

a. Since our second variable is categorical, we must convert it to an indicator (dummy) variable consisting of 0's and 1's. Although the software will do this automatically, it is good to do it yourself so that you have control over the way in which the 0's and 1's are assigned. Create two new variables. One should be called Accord and takes on a value of 1 if the model is Accord and a 0 otherwise. The other should be called Tacoma and takes on a value of 1 if the model is Tacoma and a 0 otherwise. Write the mathematical expressions below and write R code to create the new variables.

$x_{i,1}$ = miles for the i^{th} vehicle

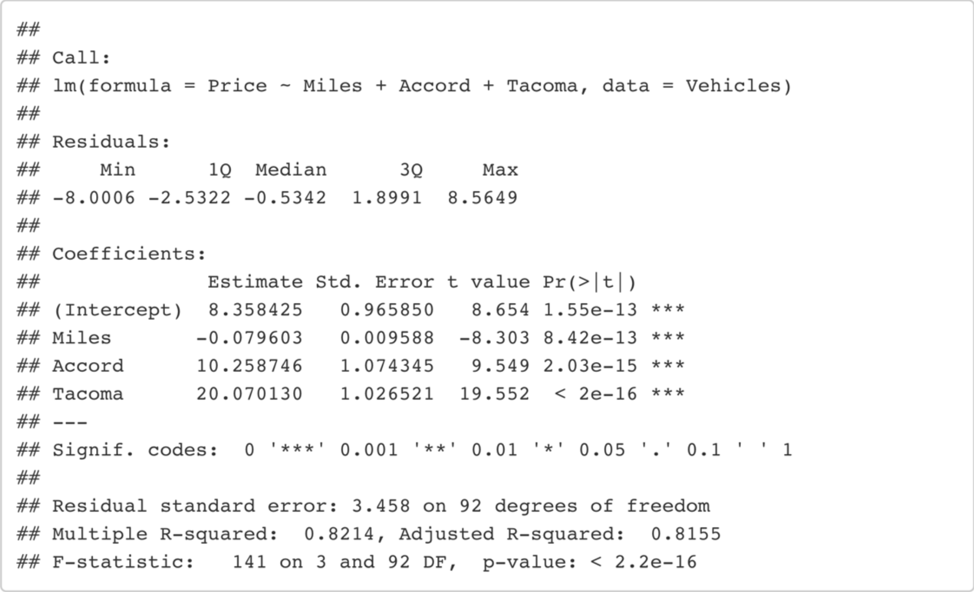
b. If both Accord and Tacoma are 0, what is the model of the vehicle?

Accord ($x_{i,2}$)	Tacoma ($x_{i,3}$)	Implied Model of Car
1	0	
0	1	
0	0	

NOTE: In general, for a categorical variable with k levels, we must define $k - 1$ indicator variables.

c. Find the estimated regression equation for predicting Price based on miles and model. (Use the indicator variables from Part a.) In R, name the results model1c.

R CODE:



```
##
## Call:
## lm(formula = Price ~ Miles + Accord + Tacoma, data = Vehicles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0006 -2.5322 -0.5342  1.8991  8.5649
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.358425   0.965850   8.654 1.55e-13 ***
## Miles       -0.079603   0.009588  -8.303 8.42e-13 ***
## Accord      10.258746   1.074345   9.549 2.03e-15 ***
## Tacoma      20.070130   1.026521  19.552  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.458 on 92 degrees of freedom
## Multiple R-squared:  0.8214, Adjusted R-squared:  0.8155
## F-statistic:   141 on 3 and 92 DF,  p-value: < 2.2e-16
```

d. Interpret the coefficients.

y-intercept: $\hat{\beta}_0 = 8.36$.

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

NOTE: Here, are talking about Miles = 0, Accord = 0, and Tacoma = 0

Context specific: The average (mean) price is 8.36 (thousands of dollars) for a Volt with 0 miles.

Partial Slope for Miles: $\hat{\beta}_1 = -0.08$.

General interpretation: The slope represents the average change in y as x increases by 1 unit, assuming all other x's are held constant.

Context specific: As the miles increase by 1 (thousands of miles), we expect the price to decrease by 0.08 (thousands of dollars), on average, assuming the model does not change.

Context specific (change units): For each additional 1000 miles, we expect the price to decrease by $80, on average, assuming the model does not change.

Context specific (change increment): For each additional 10,000 miles, we expect the price to decrease by $800, on average, assuming the model does not change.

Context specific (change increment): For each additional 500 miles, we expect the price to decrease by $40, on average, assuming the model does not change.

Partial Slope for Accord: $\hat{\beta}_2 = 10.26$.

NOTE: Since $x_{i,2}$ can only take the values 0 and 1, a 1 unit increase in $x_{i,2}$ can only happen if we go from 0 to 1. $x_{i,3}$ cannot simultaneously change because we make the assumption that "all other x's are held constant". So, a 1 unit increase in $x_{i,2}$ means we are going from

$$x_{i,2} = 0 \text{ and } x_{i,3} = 0 \quad \text{to} \quad x_{i,2} = 1 \text{ and } x_{i,3} = 0$$

Context specific: As we go from a Volt to an Accord, we expect the price to increase by 10.26 thousands of dollars, on average, assuming the miles do not change.

Alternative phrasing: We expect the price of an Accord to be $10,260 higher than the price of a Volt, on average, assuming the miles are the same.

Partial Slope for Accord: $\hat{\beta}_3 = 20.07$.

Context specific:

e. What proportion of the total variation in the Price is explained by the model containing both miles and model?

Problem 1: Every time you add a predictor to a model, the R^2 increases, even if due to chance alone. It never decreases. Consequently, a model with more terms may appear to have a better fit simply because it has more terms.

Problem 2: If a model has too many predictors and/or higher order polynomials, it begins to model the random noise in the data. This condition is known as overfitting the model and it produces misleadingly high R^2 values and a lessened ability to make predictions for new observations.

Definition: The adjusted R^2, denoted by R^2_{adj}, is incorporates a penalty for the number of predictors in your model. In MLR, it is common to use R^2_{adj} for comparing models.

NOTE: R^2_{adj} does not address how well the model generalizes to new data. (More on this in the future.)

f. Report R^2_{adj} for the model based on miles and model.

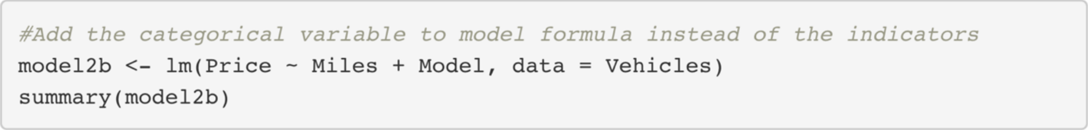
g. Consider 3 models: 1) the model containing miles only, 2) the model containing model only, and 3) the model containing miles and price. Which works the best based on R^2_{adj}? Based on RSE?

Terms	R^2	R^2_{adj}	RSE
Miles			
Model (use Accord and Tacoma)			
Miles and Model (use Miles, Accord, & Tacoma)			

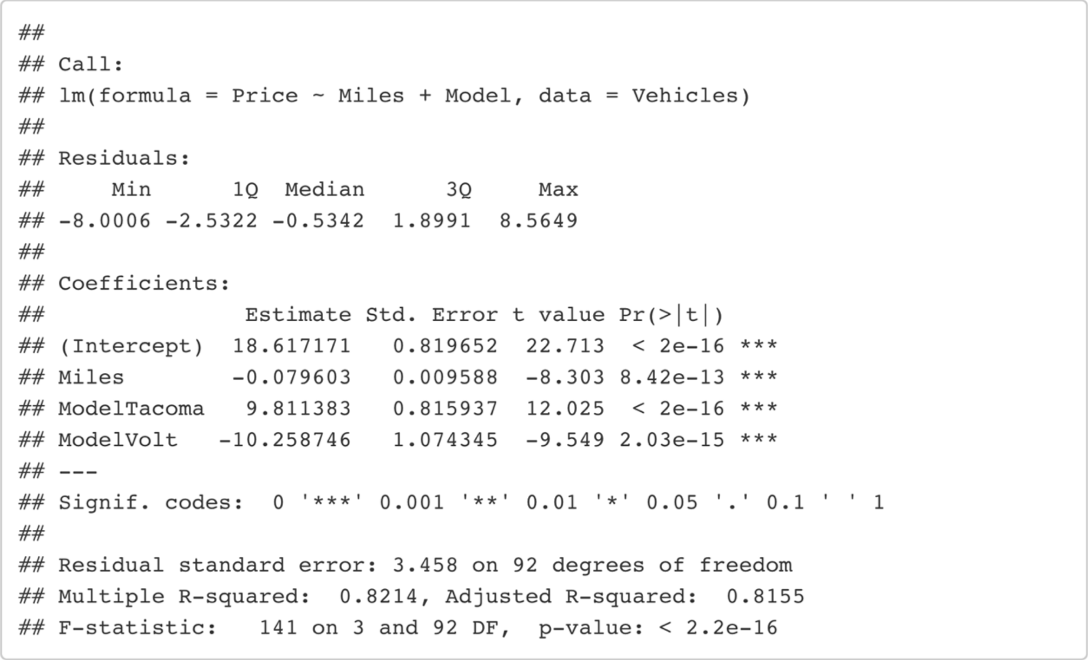
EXAMPLE 2: (Additional Considerations)

a. Create an indicator for Volt.

b. In Example 1c, we built the model using the indicator variables that we created Accord and Tacoma. R will actually create indicators for you. i) Based on the following output, which indicators did R use in the model? ii) Does the fit of the model change? iii) What does change?

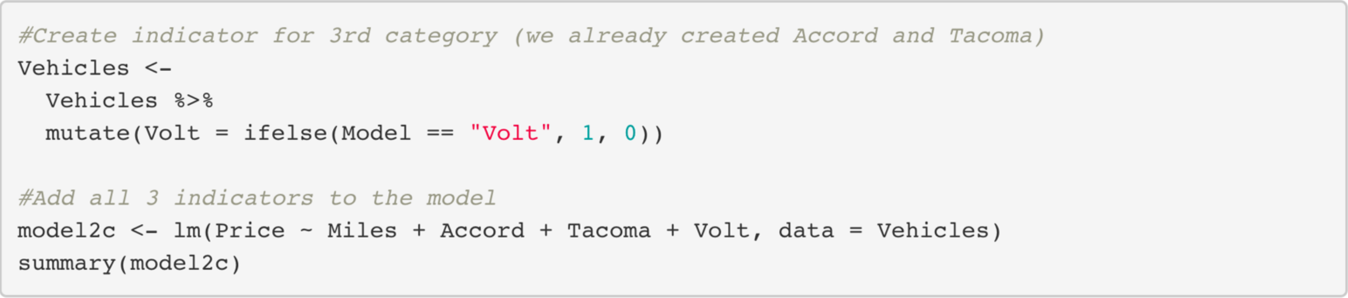


```
#Add the categorical variable to model formula instead of the indicators
model2b <- lm(Price ~ Miles + Model, data = Vehicles)
summary(model2b)
```



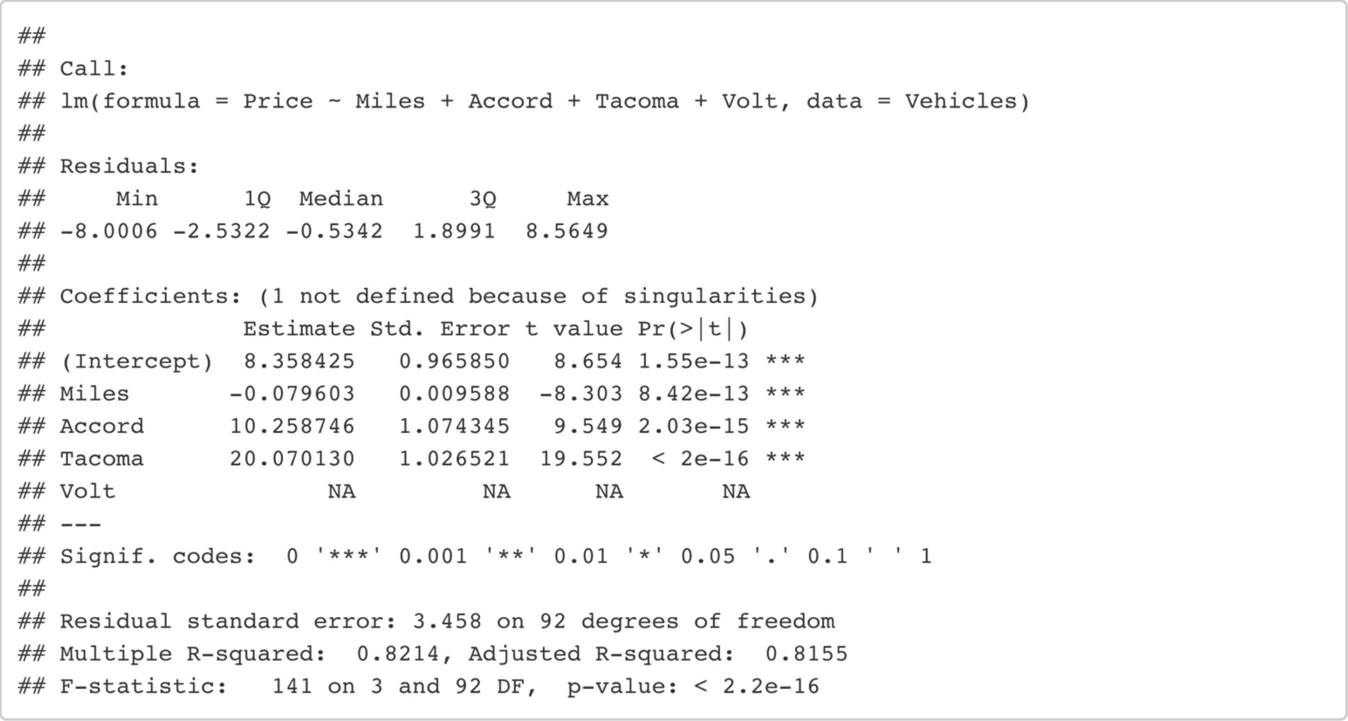
```
##
## Call:
## lm(formula = Price ~ Miles + Model, data = Vehicles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0006 -2.5322 -0.5342  1.8991  8.5649
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.617171   0.819652  22.713  < 2e-16 ***
## Miles        -0.079603   0.009588  -8.303 8.42e-13 ***
## ModelTacoma   9.811383   0.815937  12.025  < 2e-16 ***
## ModelVolt   -10.258746   1.074345  -9.549 2.03e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.458 on 92 degrees of freedom
## Multiple R-squared:  0.8214, Adjusted R-squared:  0.8155
## F-statistic:   141 on 3 and 92 DF,  p-value: < 2.2e-16
```

c. Earlier, we mentioned that for a categorical variable with k levels, we must define $k - 1$ indicator variables. Suppose we used all k indicators instead. Explain what the output.



```r
#Create indicator for 3rd category (we already created Accord and Tacoma)
Vehicles <-
  Vehicles %>%
  mutate(Volt = ifelse(Model == "Volt", 1, 0))

#Add all 3 indicators to the model
model2c <- lm(Price ~ Miles + Accord + Tacoma + Volt, data = Vehicles)
summary(model2c)
```



```
##
## Call:
## lm(formula = Price ~ Miles + Accord + Tacoma + Volt, data = Vehicles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0006 -2.5322 -0.5342  1.8991  8.5649
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.358425   0.965850   8.654 1.55e-13 ***
## Miles       -0.079603   0.009588  -8.303 8.42e-13 ***
## Accord      10.258746   1.074345   9.549 2.03e-15 ***
## Tacoma      20.070130   1.026521  19.552  < 2e-16 ***
## Volt               NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.458 on 92 degrees of freedom
## Multiple R-squared:  0.8214, Adjusted R-squared:  0.8155
## F-statistic:   141 on 3 and 92 DF,  p-value: < 2.2e-16
```