Notes 6 - Modelling

Jillian Morrison

November 1, 2019

What will we cover

Supervised Learning techniques

- Numerical Response
  - Simple Linear Regression
  - Multiple Linear Regression
- Categorical Response (Classification)
  - Logistic Regression
  - Multiple Logistic Regression
- Cross Validation
- Model Selection

We will not go into the Math of these models in detail, as these are covered in future classes required for the data science major/minor.

The goal is that you are able to identify when these models can be used and implement them in R

This covers various sections of chapters of "An Introduction to Statistical Learning" by James, et. al. Also read "Doing Data Science: Straight Talk from the Frontline" by O'Neil and Schutt for more information.

Simple Linear Regression

Important Notes:

- QUANTITATIVE response variable (i.e. the thing you want to predict) eg. of quantitative predictors include: height, weight, price, length, etc.
- Can have both quantitative and qualitative (e.g. color) predictor variables (i.e. the things you want to use to predict your response)
- very useful, simple and wisely used
- this is the basis for many more complicated prediction methods (you will want to learn this before moving to more complex methods)

Example: Simple Linear Regression

Let's take a look at `Advertising` dataset at http://faculty.marshall.usc.edu/gareth-james/ISL/data.html:

```
> library(dplyr)
> Advertising <- read.csv("Advertising.csv")
> glimpse(Advertising)
Observations: 200
Variables: 5
$ X         <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 1...
```

```
$ TV        <dbl> 230.1, 44.5, 17.2, 151.5, 180.8, 8.7, 57...
$ radio     <dbl> 37.8, 39.3, 45.9, 41.3, 10.8, 48.9, 32.8...
$ newspaper <dbl> 69.2, 45.1, 69.3, 58.5, 58.4, 75.0, 23.5...
$ sales     <dbl> 22.1, 10.4, 9.3, 18.5, 12.9, 7.2, 11.8, ...
```

Information about variables:

`sales`- sales of the product in 200 markets

`TV, radio, newspaper` - budgets for the product in each of the markets in these mediums

Some more info: It is not possible for the client to directly increase the sales of the product, but they can control the expenditure in each of the 3 media.

What question might you ask?

Example: Simple Linear Regression

- Is there an association/relationship between the sales and the advertising budgets?

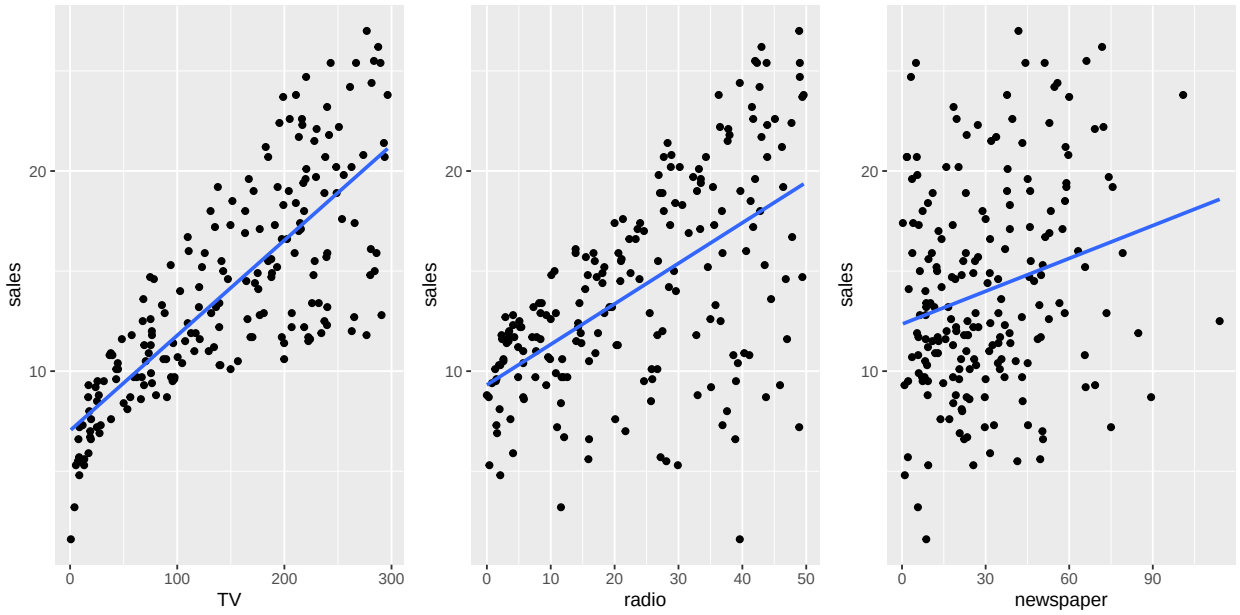
If you knew this, you would be able to advise your client on how to adjust advertising budgets, which may indirectly increase sales.

So your **GOAL** should be to develop an accurate model that can be used to **predict sales** on the **basis of the media budgets**

Example: Simple Linear Regression

Let's Visualize the data

```
> library(ggplot2)
> library(gridExtra)
> TV=ggplot(Advertising,aes(x=TV, y=sales))+geom_point()+geom_smooth(method=lm, se=FALSE)
> radio=ggplot(Advertising,aes(x=radio, y=sales))+geom_point()+geom_smooth(method=lm, se=FALSE)
> sales=ggplot(Advertising,aes(x=newspaper, y=sales))+geom_point()+geom_smooth(method=lm, se=FALSE)
> grid.arrange(TV,radio,sales, nrow=1)
```



Example: Simple Linear Regression

So, we have a few questions here:

- Is there a relationship between the advertising budgets and sales?
- How strong is the a relationship between the advertising budgets and sales?
- Which media contribute to sales?
- How accurately can we estimate the effect of each medium on sales?
- How accurately can we predict future sales?
- Is the relationship linear?
- Is there a synergy among the advertising media?

Example: Simple Linear Regression

So, let's think answers

Is there a relationship between the advertising budgets and sales? - Is there evidence of an association? If it is weak, then we probably should not spend money on advertising using these media!

How strong is the a relationship between the advertising budgets and sales? -If there is a relationship, we want to know how strong! If the relationship is strong, predicting sales will be soooo much better and accurate!

Which media contribute to sales? - Whiich one of these budgets should we increase or decrease or leave the same? We have to figure out individual effects of advertising in these media.

How accurately can we estimate the effect of each medium on sales? - If we increase the TV or radio or newspaper budget by $1, by how much will the sales increase? How accurately can you predict this increase?

How accurately can we predict future sales? - For any budget level for these mediums, how good is the prediction of sales?

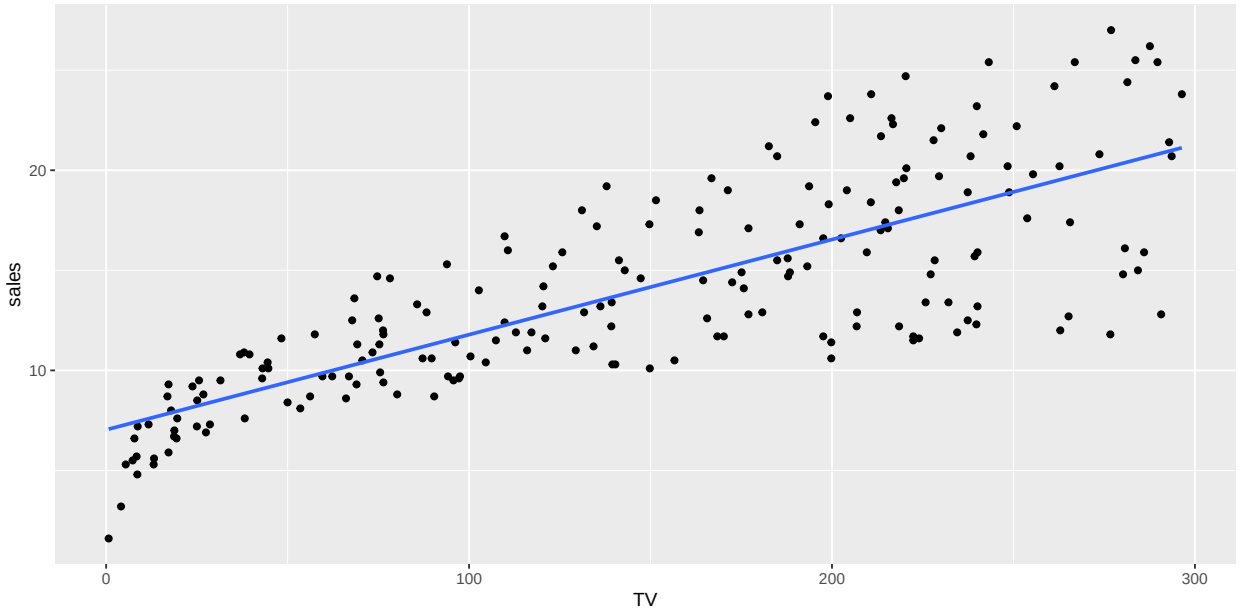
Is the relationship linear? - If the relationship is not linear, DO NOT USE THIS METHOD! Maybe try nonlinear or polynomial regression or something else! But we will save these for another course.

Is there a synergy among the advertising media? - Maybe, your answer is not just increasing one or the other, but increasing these mediums in combination. For example, you may have to increase the TV budget by \$1000 and the radio budget by \$500 at the same time (and not just one) in order to have the best sales. If this is so, we will have to consider what we call an `interaction` or `synergy` effect. We will leave this for another class as well since this will require way more math and statistics to understand.

Simple Linear Regression - What is it?

You thought right! Fit a line to the data!

Do we remember what the equation of a line is?



Simple Linear Regression - What is it?

Equation of a line

$$y = b + mx$$

Simple Linear Regression equation

$$Y \approx \beta_0 + \beta_1 X$$

Where:

- Y is the response variable - the thing you want to predict (on the y axis)
- X is the predictor or explanatory variable - the thing you want to use to predict Y (on the x axis)
- β_0 is the Y intercept
- β_1 is the slope of the line OR the effect of X on Y. Read - for a unit change in X, there is a change in Y by β_1

The reason why we use $\approx$ instead of $=$ is because all the data points are not on the line that you fit. You are using this line to approximate the data!

Simple Linear Regression - Estimating the Coefficients

How do we estimate β_0 and β_1?

This is actually a pretty simple process (probably the simplest process of all other data science and statistics methods!). Now, the computer can do this for you, but it's worth having an idea of how it works!

ANSWER:

1. Calculate the error. This error is the distance each point is from the line.
2. Square this error. If you do not square it, they all add to zero (recall the standard deviation issue)!
3. Minimize the sum of the squared error. You can actually do this just by using what you learned in Calculus 111. Basically, you are finding the minimum point of this squared error function (i.e. find the values of β_0 and β_1 that makes the derivative 0 and see which of these values produces the minimum of the function!)

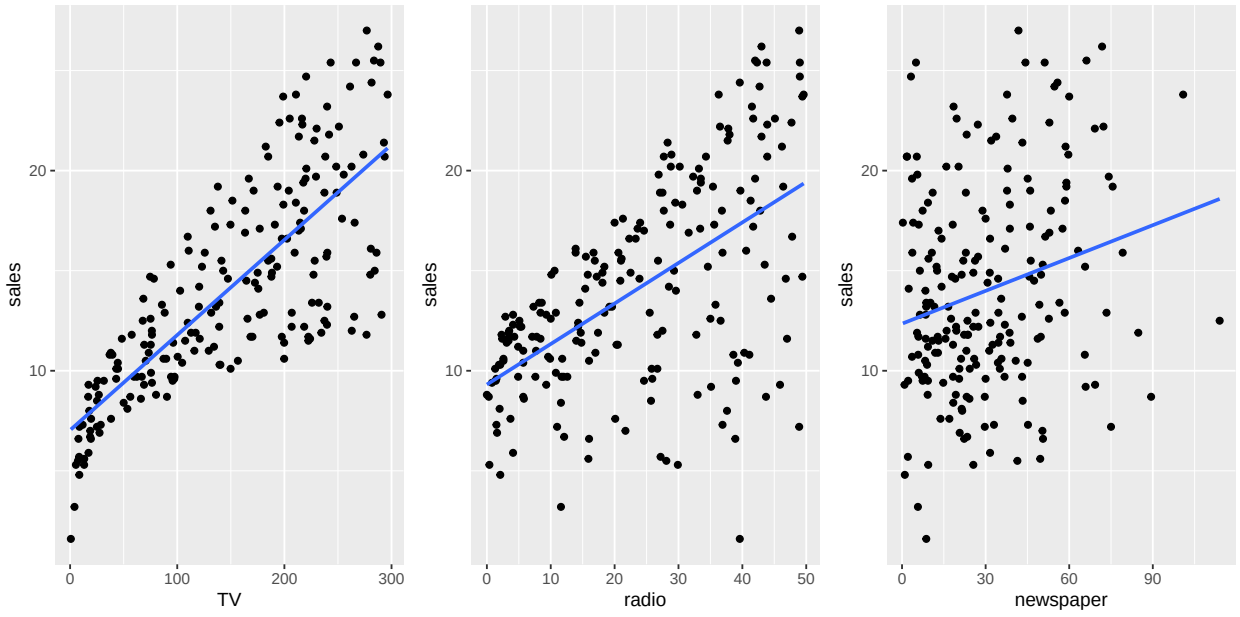
NOTE: This is why CALCULUS is so very important to do Statistics and Data Science!

Example: Simple Linear Regression

We should probaby check to see if it is valid to fit a linear model. i.e. anwering th question: **Is there a relationship between the advertising budgets and sales?**

Let's look at the plots again:

```
> grid.arrange(TV,radio,sales, nrow=1)
```



Looks like we may be able to fit a line for TV and radio. Maybe not so much newspaper.

Example: Simple Linear Regression

Let's think of one media: We want to see the effect of TV advertisements on sales, i.e. we want to estimate the equation

$$Sales \approx \beta_0 + \beta_1 * TV$$

So, this is how we do it:

```
> model1=lm(sales~TV, data = Advertising)
```

`lm` is a function that estimates linear models and it comes preloaded in R.

Example: Simple Linear Regression

Let's see what is inside `model1`

```
> model1

Call:
lm(formula = sales ~ TV, data = Advertising)

Coefficients:
(Intercept)           TV
    7.03259      0.04754
```

We see that $\beta_0 = 7.03259$ and $\beta_1 = 0.0474$, so our estimated equation is:

$$Sales \approx 7.03259 + 0.04754 * TV$$

Meaning: for a \$1 increase in TV advertising budget, we expect sales to increase by \$0.0474,

Example: Simple Linear Regression

Now, Let's answer some questions:

How strong is this relationship/association? Here we can look at the R^2 value. This is the **Coefficient of Determination** and it is simply $Correlation^2$. Recall that the correlation measures the strength and direction of a relationship and can be $-1 \leq R \leq 1$.

So, R^2 measures this relationship, but $0 \leq R^2 \leq 1$

- Closer to 1 means the relationship is stonger
- Closer to 0 means relationship is weaker

```
> summary(model1)

Call:
lm(formula = sales ~ TV, data = Advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3860 -1.9545 -0.1913  2.0671  7.2124
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119,    Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

Here, it' called Multiple R-squared and it is 0.6119 for the relationship between TV and sales.

Example: Simple Linear Regression

Let's try to answer: **Which media contribute to sales?** and we will start out by asking: *Does TV advertising contribute to sales?*

We do this by doing a **Hypothesis test** as follows:

1. Set your Research Hypothesis: **TV contributes to sales**
2. Create a hypothesis that is the opposite of your Research hypothesis: **TV does not contribute to sales**
    - Let's call this your Null Hypothesis

THE GOAL of this test is to **REJECT your Null Hypothesis** to show that you have evidence based on your data to **SUPPORT your Research Hypothesis**

Example: Simple Linear Regression

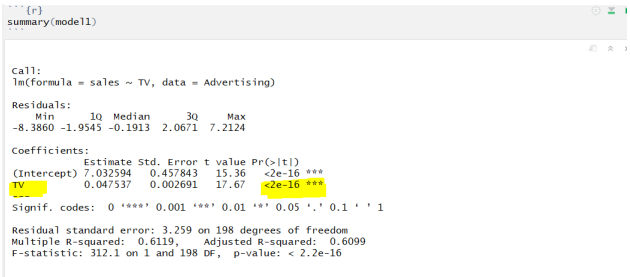
Still answering: **Does TV contribute to sales?**

So we have:

Null Hypothesis: slope or $\beta_1 = 0$ (if the slope is zero, there is no relationship between tv and sales)

Research Hypothesis: slope or $\beta_1 \neq 0$ (if the slope is not zero, there is some relationship between tv and sales)

Well, without all the math and statistics details, we can look at results from model1 to decide which hypothesis we should go with.



Example: Simple Linear Regression

Still answering: **Does TV contribute to sales?**

- If Pr(>|t|) for TV is less than or equal to 0.05. You can reject your null hypothesis. This means that you believe that TV contributes to predicting Sales
- If Pr(>|t|) for TV is more than 0.05. You cannot reject your null hypothesis. This means that you DO NOT believe that TV contributes to predicting Sales

OUR ANSWER: Since Pr(>|t|) is $< 2e-16 = 2$ x 10^{-16} , this is way way way less than 0.05 so we believe that TV contributes to sales

NOTE: $< 2e-16$ is R's way of writing in scientific notation.

Example: Simple Linear Regression

What have we done so far?

1. We have looked at the plots to determine if there is an association between the media budgets and Sales
2. We have also looked at these plots to see if the relationship may be linear.
3. We fit a linear model and estimated the linear equation that can be used to predict sales based on the TV budget.
4. We looked at the R^2 value to see the strength of the association between the points and the fitted line.
5. We did a hypothesis test to determine if TV budget was important to predicing sales. It was.

What else do we want to do?

1. We want to know how accurate the predictions are when we increase or decrease the TV budget. So, Let's so this!

Example: Simple Linear Regression

How accurately can we predict future sales?

Well, first we have to find out how bad we did with our predictions! So, We already have the actual data, but we do not have the predicted values for our data based on the estimated equation. The residual would be:

$$Residual = ActualSales - PredictedSales$$

OR

$$Residuals = Y_{Actual} - Y_{Predicted}$$

The predicted sales can be calculated using the `predict()` function.

```
> ##Using the TV Budget to calulate predicted sales using model1
> pred= predict(model1, data=Advertising)
```

Example: Simple Linear Regression

How accurately can we predict future sales?

```
> library(dplyr)
> #Adding the predicted sales to the original dataset
> Advertising2=Advertising%>%mutate(Predicted_Sales=pred)
> head(Advertising2)
  X    TV radio newspaper sales Predicted_Sales
1 1 230.1  37.8      69.2  22.1       17.970775
2 2  44.5  39.3      45.1  10.4        9.147974
3 3  17.2  45.9      69.3   9.3        7.850224
4 4 151.5  41.3      58.5  18.5       14.234395
5 5 180.8  10.8      58.4  12.9       15.627218
6 6   8.7  48.9      75.0   7.2        7.446162
```

Notice that Your Predicted Sales based on the TV Budget is not the same as the actual sales... AS EXPECTED!

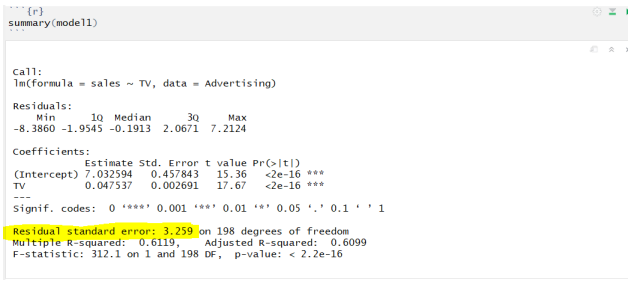
- It would only be the same if all your data points fit exactly on the line!

Example: Simple Linear Regression

Still answering: **How accurately can we predict future sales?**

So how do we use this to summarize how good we did?

- Well we can basically use the residual to calculate a 'standard deviation' to see how far away (on average) the predicted value is away from the actual value.
- This is called the RSE (Residual Standard Error) - which can be found in `summary(model1)`



Example: Simple Linear Regression

Answering: **How accurately can we predict future sales?**

Well, we can use RSE or RMSE (which is Root Mean Squared Error which is a slight variation of RSE) to determine how accurate a model is at making predictions. we will come back to this over and over when we want to select the best model.

Think about this: How can we use this RSE or RMSE to select the best model?

Model Selection using RSE

We can select the best model by choosing those models that smallest RSE

- Remember, smallest RSE means that the models have the smallest error when it comes to making predictions!

So, we know the RSE for the model to predict sales using TV is 3.259.

Let's find the RSE when we use Newspaper and then wehen we use Radio to predict Sales.

RSE with Radio

RSE is 4.275

```
> model_radio=lm(sales~radio, data = Advertising)
> summary(model_radio)

Call:
lm(formula = sales ~ radio, data = Advertising)

Residuals:
     Min       1Q   Median       3Q      Max
-15.7305  -2.1324   0.7707   2.7775   8.1810

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.31164    0.56290  16.542   <2e-16 ***
radio        0.20250    0.02041   9.921   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.275 on 198 degrees of freedom
Multiple R-squared:  0.332, Adjusted R-squared:  0.3287
F-statistic: 98.42 on 1 and 198 DF,  p-value: < 2.2e-16
```

RSE with newspaper

RSE is 5.092

```
> model_newspaper=lm(sales~newspaper, data = Advertising)
> summary(model_newspaper)

Call:
lm(formula = sales ~ newspaper, data = Advertising)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2272  -3.3873  -0.8392   3.5059  12.7751

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 12.35141    0.62142   19.88  < 2e-16 ***
newspaper    0.05469    0.01658    3.30  0.00115 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.092 on 198 degrees of freedom
Multiple R-squared:  0.05212,   Adjusted R-squared:  0.04733
F-statistic: 10.89 on 1 and 198 DF,  p-value: 0.001148
```

Which model is better?

RSE with TV as predictor of Sales is 3.259

RSE with Radio as predictor of sales is 4.275

RSE with Newspaper as predictor of Sales is 5.092

Which is better?

Which Model is better?

Ofcourse, the model with the smaller RSE!

Remember that RSE measures how far (on average) the predicted sales/responses are away from the actual sales/responses. So, the smaller the RSE, the better!

Using RSE, we see that of the 3 variables, TV predicts sales better than Newspaper and Radio. You can also make the othercomparisons.

Model Selection: Is using RSE the only way?

NOOO! There are many things/criteria we can use to compare how good models are. For example:

- Akaike information criterion (AIC)
- Bayes Information factor (BIC)
- Likelihood ratio test
- False Discovery Rate (FDR)
- Deviance Information Criterion (DIC)
- R - Pearson Correlation Coefficient (which is just correlation as we have been calling it)
- R^2 or Coefficient of Determination which is the square of the Pearson Correlation coefficient
- Deviance
- Kendall's τ
- Somer's D
- Kruskal's γ

And the list goes on...

However, note that some of these are used for numerical data while some are used for categorical data.

CROSS-VALIDATION - A better way to determine RSE for Model Selection

We will use Cross-Validation to help us check how well a model does at making predictions.

IDEA: You want to split your data into a Training dataset and a Testing dataset. Use the training dataset to fit your model and test your model using your testing dataset.

Biggest Question: How do you determine which observations are in your testing versus training dataset? We will consider:

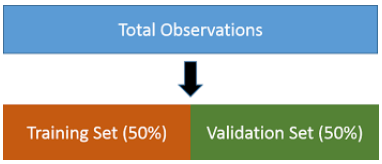
```
- Validation Set Approach
- Leave-One-Out Cross-Validation
- K-Fold Cross Validation (will only discuss)
```

These can be use REGARDLESS of the Model used to fit the data. This will give you a better estimation of the criteria you used for model selection.

Validation Set Approach

ALgorithm:

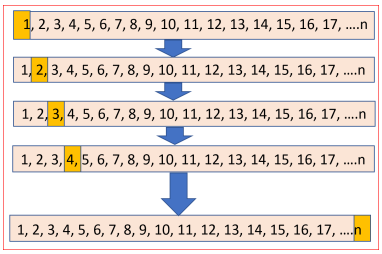
1. Randomly divide observations into 2 parts - training and testing
2. Use the training set to fit model.
3. Use model from (2) to predict using the testing set.
4. You already know the real response value, so you use this to see to see how well you predicted. So, calculate RMSE using
   - $MSE = mean((actualresponse - Predicted)^2)$ and then
   - $RMSE = sqrt(MSE)$.



Leave-One-Out Cross Validation

Algorithm:

1. In this method, you split your dataset into a training and testing datast, n times, where your training set has n-1 observations and your testing set has the last observation.
2. You do this until every observation has a chance to be the testing set.
3. Each time you will fit the model using the training set and test the model with the observation left out for the testing.
4. Each time, calculate RMSE and summarize all the RMSE's (for example using mean and standard deviation) to come up with an overall RMSE.



Implementing Leave-One-Out Cross Validation

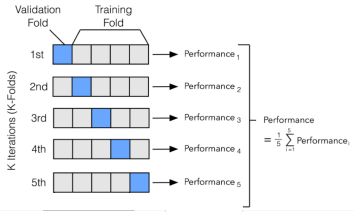
```
> #install.packages("caret")
> library(caret)
> data_ctrl <- trainControl(method = "LOOCV")
> model_caret <- train(sales~TV,   # model to fit
+                      data = Advertising,
+                      trControl = data_ctrl,              # folds
+                      method = "lm",                      # specifying regression model
+                      na.action = na.pass)                # pass missing data to model - some models w
```

```
> model_caret
Linear Regression

200 samples
  1 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 199, 199, 199, 199, 199, 199, ...
Resampling results:

  RMSE     Rsquared   MAE
  3.27736  0.6034685  2.576222

Tuning parameter 'intercept' was held constant at a value
 of TRUE
```

K- fold Cross Validation

Algorithm:

1. Decide how many folds you want. That is, how many times you want to train and test. This is k.
2. Split the dataset into k-parts.
3. Train the model on the dataset made up of k-1 of these parts/folds and use the last part/fold as the testing set.
4. Do this until each part/fold has the opportunity to be the testing dataset.
5. Each time, calculate RMSE and summarize (for example using mean and standard deviation) all the RMSE's to come up with an overall RMSE.



Implementing K- fold Cross Validation

```
> #install.packages("caret")
> library(caret)
> data_ctrl <- trainControl(method = "cv", number=2)
> model_caret <- train(sales~TV,   # model to fit
+                      data = Advertising,
+                      trControl = data_ctrl,              # folds
+                      method = "lm",                      # specifying regression model
+                      na.action = na.pass)                # pass missing data to model - some models w
```

```
> model_caret
Linear Regression

200 samples
  1 predictor

No pre-processing
Resampling: Cross-Validated (2 fold)
Summary of sample sizes: 101, 99
Resampling results:

  RMSE      Rsquared  MAE
  3.274747  0.617468  2.579016

Tuning parameter 'intercept' was held constant at a value
 of TRUE
```

Appendix

Appendix - Implementing the Validation Set approach

```
> #install.packages("ISLR")
> library(ISLR)
> ##randomly split your dataset
> set.seed(200)
> train = sample(x=1:200,size=100)
> # Use train dataset to fit model
> model1_train=lm(sales~TV, data=Advertising, subset=train)
```

```
> summary(model1_train)

Call:
lm(formula = sales ~ TV, data = Advertising, subset = train)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5104 -1.6220  0.1345  1.9769  5.5839
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.965068   0.598906   11.63   <2e-16 ***
TV          0.048230   0.003596   13.41   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.096 on 98 degrees of freedom
Multiple R-squared:  0.6473,    Adjusted R-squared:  0.6437
F-statistic: 179.9 on 1 and 98 DF,  p-value: < 2.2e-16
```

NOTE: `sample()` takes a sample of the specified `size` from the elements of `x` using either with or without replacement

Appendix - Implementing the Validation Set approach

Now, Let's use the testing dataset to see how well our model fits this 'new' datset.

```
> #Predicted sales for all observations based on model
> ##model_1 train is the model fitted using the training data
> #Advertising is the dataset we are looking at
> predicted = predict(model1_train, Advertising)
```

```
> ##But you only want the predicted sales for your testing dataset which you can access using [-train]
> ## This removes all the predicted values that matches the indexes of those in the trainign dataset
> predicted_testing=predicted[-train]
```

```
> ##You also want your actual sales from your testing datase (so you can compare to the predicted)
> ## you can also access this using [-train]
> actual_testing= Advertising$sales[-train]
```

```
> ##Now Calculate RMSE
> MSE = mean((predicted_testing-actual_testing)^2)
> RMSE = sqrt(MSE)
> RMSE
[1] 3.411773
```