Matters of Discussion

1) Simple Linear Regression computation

2) ANOVA in R

3) Autocorrelation

1. Simple Linear Regression

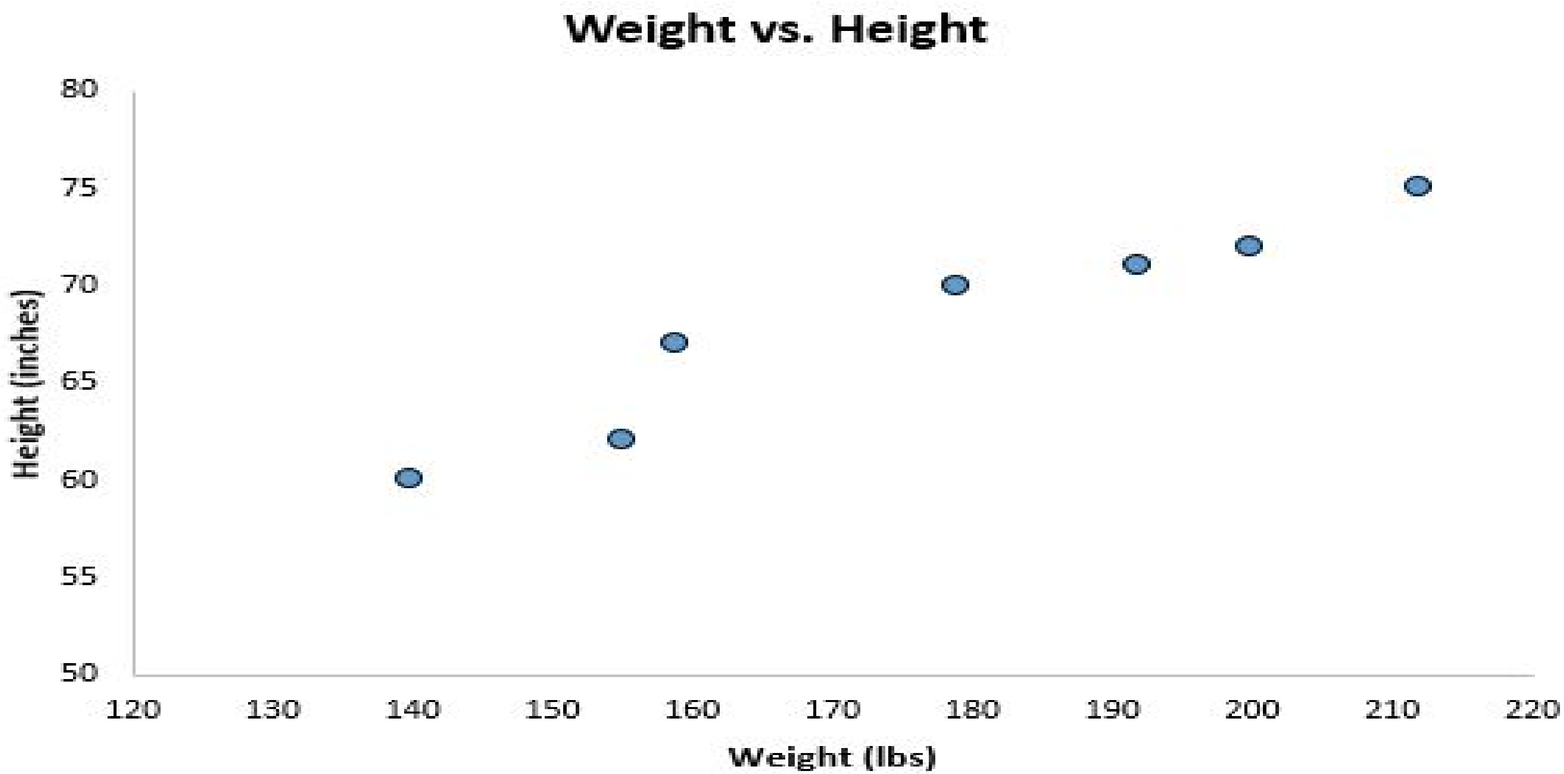
- Simple linear regression is a statistical method use to understand the relationship between two variables, x and y.
- One variable, x, is known as the predictor variable.
- The other variable, y, is known as the response variable.
- For example, suppose we have the following dataset with the weight and height.

Weight (lbs)	Height (inches)
140	60
155	62
159	67
179	70
192	71
200	72
212	75

Let weight be the predictor variable(I/P) and let height be the response variable (O/P).

Cont..

- If we graph these two variables using a scatterplot, with weight on the x-axis and height on the y-axis, here's what it would look like:



Cont..

- Suppose we're interested in understanding the relationship between weight and height.
- From the scatterplot we can clearly see that as weight increases, height tends to increase as well,
- but to actually quantify this relationship between weight and height, we need to use **linear regression**.

Cont..

- Using linear regression, we can find the line that best "fits" our data.
- This line is known as the least squares regression line and it can be used to help us understand the relationships between weight and height.
- The formula for the line of best fit is written as:

$$\hat{y} = b0 + b1x$$

- where $\hat{y}$ is the predicted value of the response variable,
- b0 is the y-intercept, b1 is the regression coefficient, and x is the value of the predictor variable.

Quantify the relationship through Linear Regression

- Simple linear regression is a statistical method you can use to quantify the relationship between a predictor variable and a response variable.
- Example:

Weight (lbs)	Height (inches)
140	60
155	62
159	67
179	70
192	71
200	72
212	75

Refer ---Linear Regression in Weka

Cont..

- Use the following steps to fit a linear regression model to this dataset, using weight as the predictor variable(I/P) and height as the response variable(O/P).
- Step 1: Calculate X*Y, X2, and Y2

Weight (lbs)	Height (inches)	X*Y	X^2	Y^2
140	60	8400	19600	3600
155	62	9610	24025	3844
159	67	10653	25281	4489
179	70	12530	32041	4900
192	71	13632	36864	5041
200	72	14400	40000	5184
212	75	15900	44944	5625

Refer ---Linear Regression in Weka

Cont..

❖ Step 2: Calculate ΣX, ΣY, ΣX*Y, ΣX2, and ΣY2

	Weight (lbs)	Height (inches)	X*Y	X^2	Y^2
	140	60	8400	19600	3600
	155	62	9610	24025	3844
	159	67	10653	25281	4489
	179	70	12530	32041	4900
	192	71	13632	36864	5041
	200	72	14400	40000	5184
	212	75	15900	44944	5625
Σ	**1237**	**477**	**85125**	**222755**	**32683**

Refer ---Linear Regression in Weka

Cont.. $\hat{y} = b0 + b1x$

❖ **Step 3: Calculate b0**

The formula to calculate b0 is:

$[(\Sigma Y)(\Sigma X^2) – (\Sigma X)(\Sigma XY)] \;/\; [n(\Sigma X^2) – (\Sigma X)^2]$

In this example,

b0 =

[(477)(222755) – (1237)(85125)] / [7(222755) – (1237)2]

= 32.783

NB- n is the sample size= 7

Cont.. $\hat{y}$ = b0 + b1x

❖ **Step 4: Calculate b1**

The formula to calculate b1 is:

$$[n(\Sigma XY) - (\Sigma X)(\Sigma Y)] \;/\; [n(\Sigma X^2) - (\Sigma X)^2]$$

In this example,

b1 =

$$[7(85125) - (1237)(477)] \;/\; [7(222755) - (1237)^2]$$

= 0.2001

Cont.. $\hat{y} = b0 + b1x$

- Step 5: Place b0 and b1 in the estimated linear regression equation.

The estimated linear regression equation is:

$\hat{y} = b_0 + b_1 * x$

In our example,

it is

$$\hat{y} = 32.783 + (0.2001) * x$$

b0 = 32.7830.

When weight is zero pounds, the predicted height is 32.783 inches. Sometimes the value for b0 can be useful to know, but in this example it doesn’t actually make sense to interpret b0 since a person can’t weigh zero pounds.

b1 = 0.2001. A one pound increase in weight is associated with a 0.2001 inch increase in height.

2. ANOVA in R

- ANOVA also known as Analysis of variance
- used to investigate relations between categorical variables and continuous variable in R Programming.
- It is a type of hypothesis testing for population variance.
- **R – ANOVA Test**

ANOVA test involves setting up:

- **Null Hypothesis:** All population means are equal.
- **Alternate Hypothesis:** At least one population mean is different from other.

Cont..

❖ ANOVA tests are of two types:

•**One way ANOVA:** It takes one categorical group into consideration.

•**Two way ANOVA:** It takes two categorical group into consideration.

❖ **The Dataset [Motor Trend Car Road Tests]**

✓ The mtcars (motor trend car road test) dataset is used which consist of 32 car brands and 11 attributes.

✓ The dataset comes preinstalled in **dplyr** package in R.

✓ To get started with ANOVA, we need to install and load the **dplyr** package.

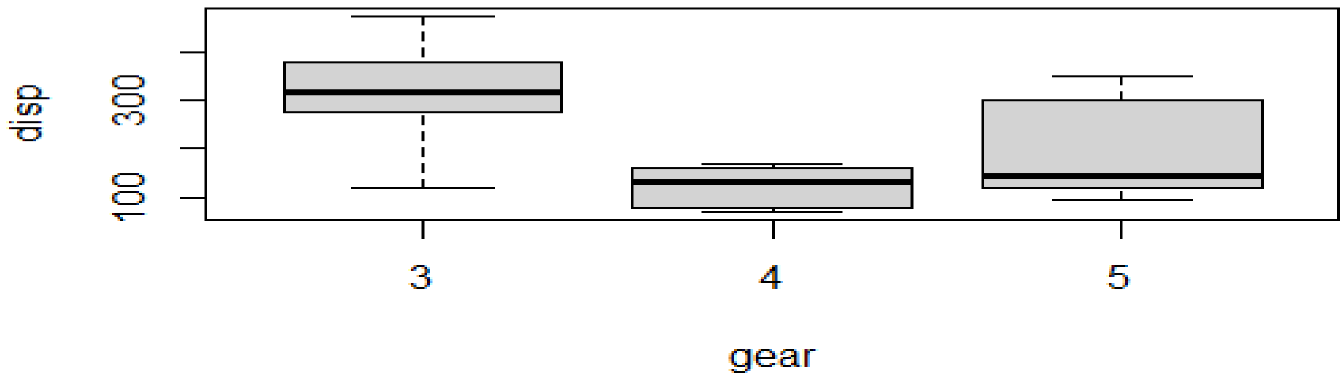
Performing One Way ANOVA test in R

- One way ANOVA test is performed using mtcars dataset which comes preinstalled with dplyr package between --disp attribute, a continuous attribute and gear attribute, a categorical attribute.

[, 1]	mpg	Miles/(US) gallon
[, 2]	cyl	Number of cylinders
[, 3]	**disp**	Displacement (cu.in.)
[, 4]	hp	Gross horsepower
[, 5]	drat	Rear axle ratio
[, 6]	wt	Weight (1000 lbs)
[, 7]	qsec	1/4 mile time
[, 8]	vs	Engine (0 = V-shaped, 1 = straight)
[, 9]	**am**	Transmission (0 = automatic, 1 = manual)
[,10]	**gear**	Number of forward gears
[,11]	carb	Number of carburetors

```r
# Installing the package
install.packages(dplyr)

# Loading the package
library(dplyr)

# Variance in mean within group and between group
boxplot(mtcars$disp~factor(mtcars$gear),
        xlab = "gear", ylab = "disp")

# Step 1: Setup Null Hypothesis and Alternate Hypothesis
# H0 = mu = mu01 = mu02(There is no difference
# between average displacement for different gear)
# H1 = Not all means are equal

# Step 2: Calculate test statistics using aov function
mtcars_aov <- aov(mtcars$disp~factor(mtcars$gear))
summary(mtcars_aov)

# Step 3: Calculate F-Critical Value
# For 0.05 Significant value, critical value = alpha = 0.05

# Step 4: Compare test statistics with F-Critical value
# and conclude test p < alpha, Reject Null Hypothesis
```

Result Analysis



mean values of gear with respect of displacement.

categorical variable is gear on which factor function is used and continuous variable is disp.

Console Terminal Jobs

~/

```
>
> # Step 2: Calculate test statistics using aov function
> mtcars_aov <- aov(mtcars$disp~factor(mtcars$gear))
> summary(mtcars_aov)
                    Df Sum Sq Mean Sq F value   Pr(>F)    
factor(mtcars$gear)  2 280221  140110   20.73 2.56e-06 ***
Residuals           29 195964    6757                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # Step 3: Calculate F-Critical Value
> # For 0.05 Significant value, critical value = alpha = 0.05
>
> # Step 4: Compare test statistics with F-Critical value
> # and conclude test p < alpha, Reject Null Hypothesis
>
```

The degrees of freedom (DF) are the number of independent pieces of information.

Cont..

- ❖ The summary shows that the gear attribute is very significant to displacement (Three stars denoting it).
- ❖ Also, the P value is less than 0.05, so proves that gear is significant to displacement i.e related to each other and we reject the Null Hypothesis.

✓ **Obtained significant result……….**

✓ **Displacement is strongly related to Gears in cars i.e. displacement is dependent on gears with $p < 0.05$.**

Key Insights

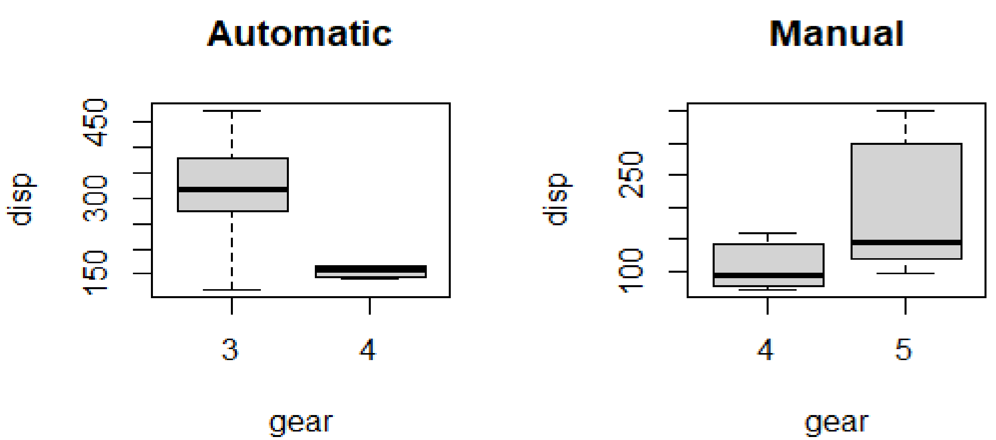
- ❖ The F-value is simply **a ratio of two variances**.
- ❖ The F value in one way ANOVA is a tool to help you answer the question "Is the variance between the means of two populations significantly different?"
- ❖ The F value in the ANOVA test also determines the P value;
- ❖ The P value is the probability of getting a result at least as extreme as the one that was actually observed.
- ❖ **The higher the F-value, the lower the corresponding p-value**.
- ❖ If the p-value is below a certain threshold (e.g. $\alpha = .05$), we can reject the null hypothesis of the ANOVA and conclude that there is a statistically significant difference between group means.

Two Way ANOVA test in R

- Two-way ANOVA test is performed using mtcars dataset which comes preinstalled with dplyr package between
- disp attribute, a continuous attribute and gear attribute, a categorical attribute, **am** attribute, a categorical attribute.

am----Transmission (0 = automatic, 1 = manual)

Disp—displacement ; **gear -**Number of forward gears



```r
# Installing the package
install.packages(dplyr)

# Loading the package
library(dplyr)

# Variance in mean within group and between group
boxplot(mtcars$disp~mtcars$gear, subset = (mtcars$am == 0),
        xlab = "gear", ylab = "disp", main = "Automatic")
boxplot(mtcars$disp~mtcars$gear, subset = (mtcars$am == 1),
            xlab = "gear", ylab = "disp", main = "Manual")

# Step 1: Setup Null Hypothesis and Alternate Hypothesis
# H0 = mu0 = mu01 = mu02(There is no difference between
# average displacement for different gear)
# H1 = Not all means are equal

# Step 2: Calculate test statistics using aov function
mtcars_aov2 <- aov(mtcars$disp~factor(mtcars$gear) *
                            factor(mtcars$am))
summary(mtcars_aov2)

# Step 3: Calculate F-Critical Value
# For 0.05 Significant value, critical value = alpha = 0.05

# Step 4: Compare test statistics with F-Critical value
# and conclude test p < alpha, Reject Null Hypothesis
```

O/P

```
>
> # Step 1: Setup Null Hypothesis and Alternate Hypothesis
> # H0 = mu0 = mu01 = mu02(There is no difference between
> # average displacement for different gear)
> # H1 = Not all means are equal
>
> # Step 2: Calculate test statistics using aov function
> mtcars_aov2 <- aov(mtcars$disp~factor(mtcars$gear) *
+                     factor(mtcars$am))
> summary(mtcars_aov2)
                    Df Sum Sq Mean Sq F value   Pr(>F)    
factor(mtcars$gear)  2 280221  140110  20.695 3.03e-06 ***
factor(mtcars$am)    1   6399    6399   0.945    0.339    
Residuals           28 189565    6770                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> # Step 3: Calculate F-Critical Value
> # For 0.05 Significant value, critical value = alpha = 0.05
>
> # Step 4: Compare test statistics with F-Critical value
> # and conclude test p < alpha, Reject Null Hypothesis
> 
```

O/P analysis

1) The summary shows that gear attribute is very significant to displacement(Three stars denoting it)
2) and am attribute is not much significant to displacement.
3) P-value of gear is less than 0.05, so it proves that gear is significant to displacement i.e related to each other.
4) P-value of am is greater than 0.05, am is not significant to displacement i.e not related to each other.

Final result on mtcat

1) Displacement is strongly related to Gears in cars i.e displacement is dependent on gears with $p < 0.05$.

2) Displacement is strongly related to Gears but not related to transmission mode in cars with p 0.05 with am.

3. Autocorrelation

- Already we have discussed the Time-series data to identify the trend, sessional, and cyclic patterns.
- Autocorrelation, also known as serial correlation, refers to the degree of correlation of the same variables between two successive time intervals.
- It is mainly used to measure the relationship between the actual values and the previous values.
- The value of autocorrelation ranges from -1 to 1.
- A value between -1 and 0 represents negative autocorrelation. A value between 0 and 1 represents positive autocorrelation.
- Autocorrelation gives information about the trend of a set of historical data, so it can be useful in the technical analysis for the equity market.

Cont..

❖ In R, we can calculate the autocorrelation in a vector by using the module tseries. Within this module, we have to use acf() method to calculate autocorrelation.

Syntax:

acf(vector, lag, pl)

Parameter:

•*vector is the input vector*

•*lag represents the number of lags*

•*pl is to plot the auto correlation*

❖ A "lag" is **a fixed amount of passing time**; One set of observations in a time series is plotted (lagged) against a second, later set of data. The k^{th} lag is the time period that happened "k" time points before time i.

auto correlation in a vector with different lags

```
# load tseries module
library(tseries)

# create vector1 with 8 time periods
vector1=c(34,56,23,45,21,64,78,90)

# calculate auto correlation with no lag
print(acf(vector1,pl=FALSE))

# calculate auto correlation with lag 0
print(acf(vector1,lag=0,pl=FALSE))

# calculate auto correlation with lag 2
print(acf(vector1,lag=2,pl=FALSE))

# calculate auto correlation with lag 6
print(acf(vector1,lag=6,pl=FALSE))
```

lag" is a fixed amount of passing time

auto correlation in a vector with different lags

```
Autocorrelations of series 'vector1', by lag

     0      1      2      3      4      5      6      7 
 1.000  0.257  0.208 -0.389 -0.093 -0.268 -0.064 -0.151 

Autocorrelations of series 'vector1', by lag

0 
1 

Autocorrelations of series 'vector1', by lag

    0     1     2 
1.000 0.257 0.208 

Autocorrelations of series 'vector1', by lag

     0      1      2      3      4      5      6 
 1.000  0.257  0.208 -0.389 -0.093 -0.268 -0.064 
```

ACTIVITY-09(Lab-04)

Formulate a null Hypothesis by considering any scenario and Investigate the computational analysis of one way and two way ANOVA to estimate the P-value to take a decision.

Cheers For the Great Patience!

Query Please?