Rashika Pramod Singh
IST 687

Lecture Notes- 8

An object in R that is used for understanding data and prediction.

Linear Model:
It is used for prediction but it not perfect and simply tries to minimize distance from points to the line.
Y=MX+B
Example- predict the cost of car repairs using past repairs, miles driven and number of oil changes during the past three years data.

Which are independent and which are dependent variables?
Oil changes, repairs, miles are independent variables in the data which are not a function of other variables and cost of repair is a dependent variable as it depends on independent variables for the prediction. It depends on the problem a variable in one case might be independent in other case dependent.

Example:
14 observations and 3 attributes
Plot(oil$oilChanges, oil$repairs) #exploring the data
Model1<-lm(formula=repairs~oilChanges, data=oil)
#build a linear model function, tilde says it is function of oil changes
Summary(model1) #summarize linear model

R-squared error:
R squared is errors in model and is called coefficient of determination represents proportion of variation for the dependent variable by whole set of independent variables. The closer to 1, the greater the influence the independent variable has on predicting dependent variable.

Abline(model1)- shows the error.

How "model" the cost? What might be some ranges of the cost?
The cost can be modelled by analyzing the predictors that affect the price of the oil. Changes can be due to the market pressure where other vendors might charge more so the oil prices might increase. Demand and supply also affect the price.

Code:

```
Oil$oilChangeCost<-Oil$oilChanges*350  #multiplying the values by 350
Oil$totalCost<- oil$oilChangeCost+oil$repairs #add the cost and repair to daily costs
M<-lm(formula=totalCost~oilChanges, data =oil) #apply the linear model
Plot (oil$oilChanges, oil$totalCost) #plot the predictors with the predicted value
Abline(m) #helps to see the linear trend
```

Test=data.frame(oilChanges=0)
Predict(m,test,type=”response”)

How accurate is the model? Did we have all the facts? Did we have all the data?
You can check the accuracy of the model by the adjusted R-squared values. The p-values can be used to check if the predictors are significant. There might be biases if we do not have all the data.

Code:

```
X<-c(1:10) #sample data
Y<-c(1:10)
Df<-data.frame(x,y) #build a dataframe
Plot (df$x, df$y) #plot the data
M<-lm(formula=y~x, data=df) #build the model
Summary(m) #summarize the model
Abline(m) #shows the linear trend

G<-ggplot(df, aes(x=x, y=y)) + geom_plot()
G
G+ stat_smooth(method=”lm”, col= “red”) #scatter plot
```

To see evaluation metrics:

```
#too see adjusted r-square
Sum.model<-summary(mpg.lm)
Paste(“p-values: “)
Sum.model$coef[,4] #pvalues
Paste(“adjusted r squared:’, sum.model$adj.r.sq) #R^2
```