Fitting of regression model and Test of significance of regression coefficient of simple linear regression model

**INTRODUCTION:**

***OBJECTIVE:*** In this problem we want to observe the relationship between variables i.e speed and distance of cars from scatter plot and we are also interested in fitting the regression model to our dataset and further we are intersted in testing the significance of regression coefficient. Here we are also interested in checking if the correlation between observed value and fitted value is same as the correlation between Y and X and interpret the result. Further, in this practical we also focus on the describing the steps that can be taken to measure the quality of the regression fit.

***DATA DESCRIPTION:*** The cars dataset give the speed of cars and the distances taken to stop.The data were recorded in the 1920s. We have two variables in our dataset i.e. 1. Distance - the distances taken to stop(Dependent variable). 2. Speed - the speed of car(Independent variable).

**ANALYSIS:**

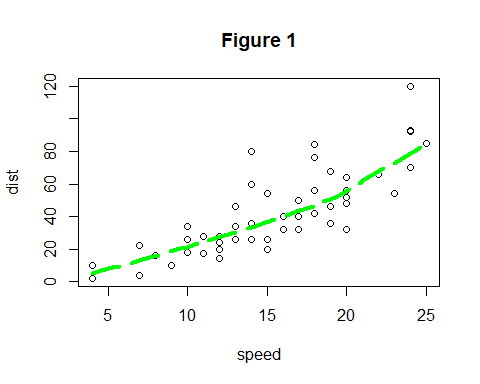
*#Obtaining the first few records of our dataset 'cars'.*  
**head**(cars)

## speed dist  
## 1 4 2  
## 2 4 10  
## 3 7 4  
## 4 7 22  
## 5 8 16  
## 6 9 10

*#attach() function helps us to access the variables present in our dataset without calling the dataset.*   
**attach**(cars)

Plotting the scatter diagram for the data and finding coefficient of correlation.

*#Obtaining the scatter plot between our two variables speed and distance of cars.*  
**scatter.smooth**(speed,dist,lpars = **list**(col="green", lwd=4, lty=5), main="Figure 1")



*#Now we obtain the coefficient of correlation between speed and distance of cars.*  
**cor**(speed,dist)

## [1] 0.8068949

***Interpretation:*** From the figure 1 we observe there exist a linear relation between the variables speed and distance.Therefor, we try to fit a simple linear regression model to the datset.

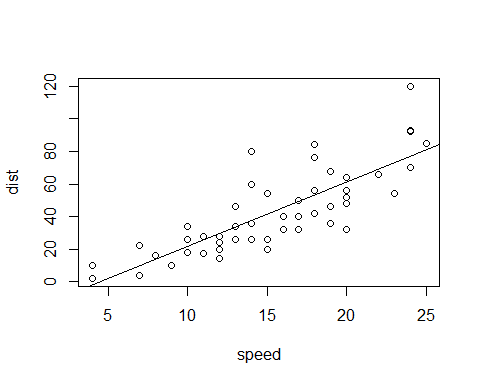
We also observe that speed of a car is our independent variable and stopping distance car is dependent variable.

Estimating the parameters of a simple linear regression model and fitting a regression line.

*#Fitting a simple linear regression model to our cars data set.*  
reg=**lm**(dist**~**speed)  
reg

##   
## Call:  
## lm(formula = dist ~ speed)  
##   
## Coefficients:  
## (Intercept) speed   
## -17.579 3.932

*#Plotting the values of speed and distance of cars.*  
**plot**(speed,dist)  
  
*#Fitting a regression line to our dataset.*  
**abline**(reg)



***Interpretation:*** Hence, our estimated values of parameters are obtained i.e. intercept which is Beta0\_hat = -17.579 and regression coefficient which is Beta1\_hat = 3.932. Hence the fitted model is Yhat=-17.579+3.932X. We also obtained the fitted regression line to our datset above.

Testing the significance of the regression coefficient.

*#Obtaining the summary of our fitted regression model.*  
**summary**(reg)

##   
## Call:  
## lm(formula = dist ~ speed)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -29.069 -9.525 -2.272 9.215 43.201   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -17.5791 6.7584 -2.601 0.0123 \*   
## speed 3.9324 0.4155 9.464 1.49e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.38 on 48 degrees of freedom  
## Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438   
## F-statistic: 89.57 on 1 and 48 DF, p-value: 1.49e-12

***Interpretation:*** From the above calculation we observe that calculated t value=|-2.601|=2.601 is greater than t value with 48 dof and 0.025 level of significance is 2.021 therefore we reject null hypothesis and conclude that regression coefficient is having a significant effect on data

To check if the correlation between observed value and fitted value is same as the correlation between Y and X.

*#Obtaining the fitted values of the model.*  
fit\_values=**fitted.values**(reg)  
fit\_values

## 1 2 3 4 5 6 7 8   
## -1.849460 -1.849460 9.947766 9.947766 13.880175 17.812584 21.744993 21.744993   
## 9 10 11 12 13 14 15 16   
## 21.744993 25.677401 25.677401 29.609810 29.609810 29.609810 29.609810 33.542219   
## 17 18 19 20 21 22 23 24   
## 33.542219 33.542219 33.542219 37.474628 37.474628 37.474628 37.474628 41.407036   
## 25 26 27 28 29 30 31 32   
## 41.407036 41.407036 45.339445 45.339445 49.271854 49.271854 49.271854 53.204263   
## 33 34 35 36 37 38 39 40   
## 53.204263 53.204263 53.204263 57.136672 57.136672 57.136672 61.069080 61.069080   
## 41 42 43 44 45 46 47 48   
## 61.069080 61.069080 61.069080 68.933898 72.866307 76.798715 76.798715 76.798715   
## 49 50   
## 76.798715 80.731124

*#Obtaining the coefficient of correlation between the observed values and fitted values.*  
**cor**(dist,fit\_values)

## [1] 0.8068949

***Interpretation:*** Hence from the above calculations it can be observed that the the correlation between observed value and fitted value is same as the correlation between Y and X which is equal to 0.8068949 which implies that X and Y are linearly related.

The steps that can be taken to measure the quality of the regression fit are

1. In test of significance of regression coefficient the t value or p value helps us measure the strength of linear relationship between X and Y i.e. larger is the t value(or smaller the p value) stronger is the linear relationship between X and Y.
2. We can also directly assess the strength of the relationship from scatter plot.
3. Coefficient of determination also helps us assess the relationship i.e. larger the coefficient of determination better is the fit of model.

**CONCLUSION:**

From the above analysis,

We observe from the scatter plot i.e. figure1 that there is a linear relationship between the speed and stopping distance of the cars and also obtain the coefficient of correlation between the variable i.e. 0.8068949 which is positive which means the speed and stopping distance are positively correlated which means that more is the speed more distance will be taken by car to stop.

Since the relationship between variables is linear we proceed to fit the simple linear regression model to our dataset. The fitted model is Yhat=-17.579+3.932X.

We obtained the regression coefficient as 3.932 which is positive which indicates that for increase in one unit change in predictor variable i.e. speed there will be a increase mean of response variable i.e. stopping distance.

The equation shows that the coefficient for the speed of a car is 3.932 therefore the distance taken to stop the car increases by 3.932 as the speed increases.

Since our intercept is negative the expected mean stopping distance will be negative when the speed of the car is 0 which means the car is not moving only.

From the test of significance we observe that the calculated t = 2.601 is larger than its critical value 2.021 therefore we will reject the null hypothesis, which means that the predictor variable i.e. speed of a car is a statistically significant predictor of the response variable i.e. stopping distance of a car.

We also observe that the correlation between observed value and fitted value is same as the correlation between stopping distance of a car(Y) and speed of a car (X) which implies that there is a linear relationship netween speed and stopping distance of a car.