HW07_Sampathirao_A

Anvita Sampathirao

7/7/2019

R Markdown

#1.

```
data("mtcars")
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

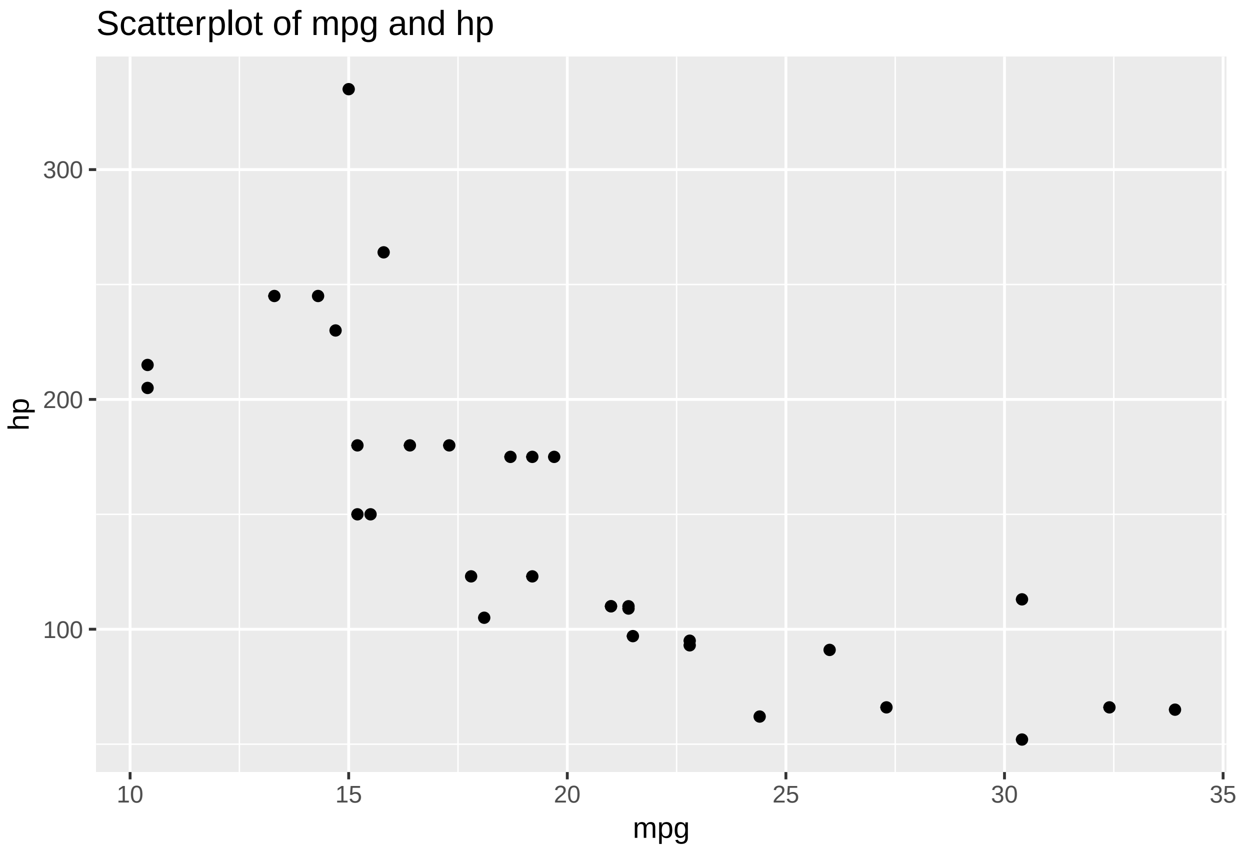
```
df<- data.frame(mtcars$mpg,mtcars$hp,
                row.names = row.names(mtcars))
colnames(df)<- c("mpg","hp")
head(df)
```

```
##                    mpg  hp
## Mazda RX4         21.0 110
## Mazda RX4 Wag     21.0 110
## Datsun 710        22.8  93
## Hornet 4 Drive    21.4 110
## Hornet Sportabout 18.7 175
## Valiant           18.1 105
```

```
library(ggplot2)
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
ggplot(df, aes(x= mpg, y= hp)) + geom_point() + ggtitle("Scatterplot of mpg and hp")
```



From the plot, it looks like hp and mpg are negatively related. There is no reason to believe that relation is non linear. A relation can be said as non linear if the points are scattered all over and do not coincide which doesn't seem like in our case.

#2.

```
a<- cov(df$mpg,df$hp)
a
```

```
## [1] -320.7321
```

a) There is a statistical association between mpg and hp because the variation of mpg coincides with the variation in hp on an average.
b) The sign of the relation is negative, indicating that there is a negative association between mpg and hp
c) The magnitude is relatively high, thus it is a strong association. However, covariance is not an apt measure to determine the strength of the relationship

#3.

```
b<- cor(df$mpg,df$hp)
b
```

```
## [1] -0.7761684
```

a) There is a statistical relation between mpg and hp because the variation of mpg coincides with the variation in hp on an average

b) The sign of the relation is negative, indicating that there is a negative association between mpg and hp
c) The strength is considerably strong as the magnitude of the correlation coefficient is closer to the bound of -1.

#4. No we cannot conclude that hp causes mpg. From 2., we can infer that there is a negative relation between hp and mpg. And From 3., we can observe that the correlation coefficient is closer to the -1 bound. Hence, there is a negative correlation between hp and mpg. However, we cannot deduce from this that hp causes mpg. "Correlation does not imply causation."

#5.

$$\beta_1 = \frac{\sigma_{(\mathrm{x,y})}}{sd_x^2}$$

$$\beta_0 = \bar{y} - \beta_1 * \bar{x}$$

```
b1<- a/(sd(mtcars$mpg)^2)
b1
```

```
## [1] -8.829731
```

```
b0<- mean(mtcars$hp)- (b1*mean(mtcars$mpg))
b0
```

```
## [1] 324.0823
```

Beta_0 is the y intercept of the fitted line in our linear model. It is the average value of hp when mpg is 0.

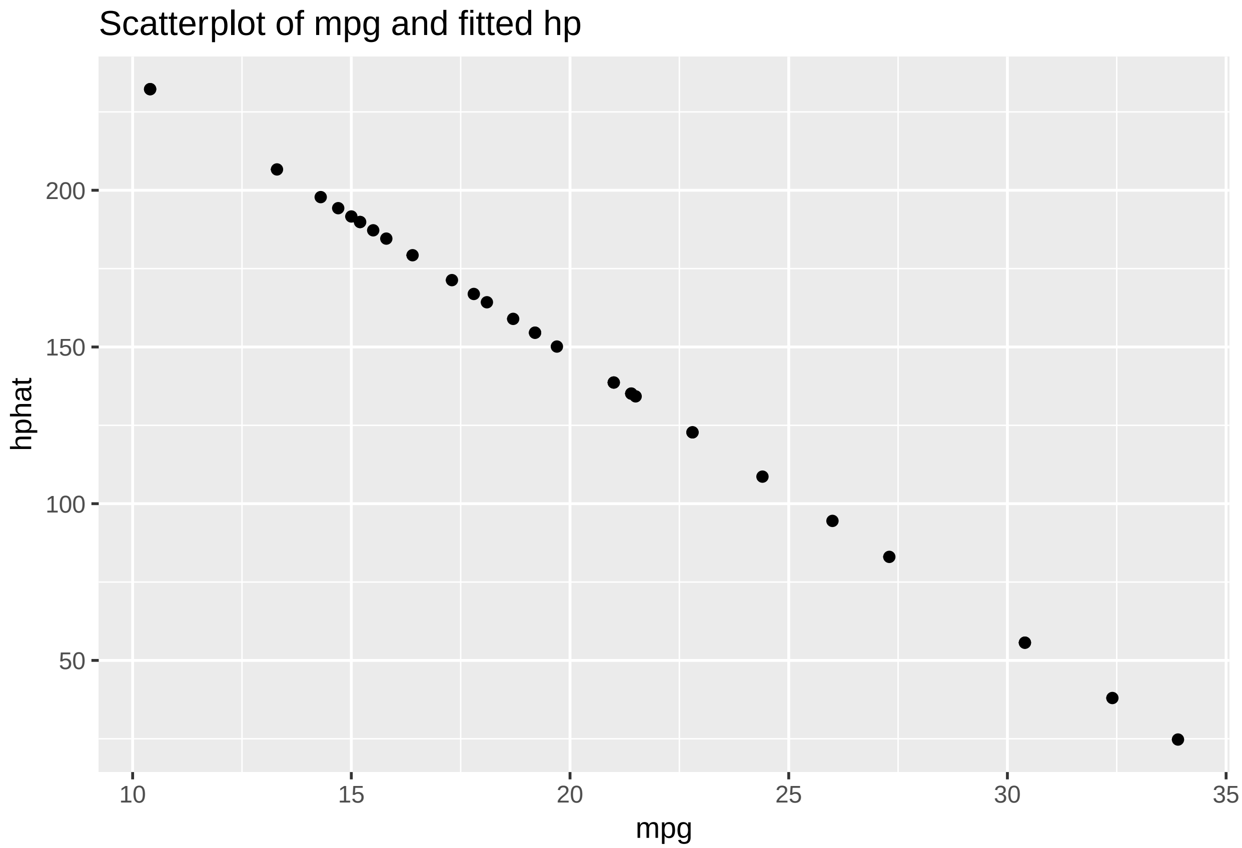
Beta_1 is the slope of the fitted line in our linear model. It means, hp will change Beta_1 times with an incremental change in mpg, i.e., hp drops by 8.83 units when there is a unit increase in mpg.

#6.

```
yhatfun<- function(x){
  yhat<- b0 + (b1*x)
  return(yhat)
}
hpfit<- yhatfun(df$mpg)
df1<- data.frame(df$mpg, df$hp, hpfit, row.names = row.names(df))
colnames(df1)<- c("mpg","hp","hphat")
head(df1)
```

```
##                    mpg  hp    hphat
## Mazda RX4         21.0 110 138.6580
## Mazda RX4 Wag     21.0 110 138.6580
## Datsun 710        22.8  93 122.7644
## Hornet 4 Drive    21.4 110 135.1261
## Hornet Sportabout 18.7 175 158.9663
## Valiant           18.1 105 164.2642
```

```
ggplot(df1, aes(x= mpg, y= hphat)) + geom_point() +
  ggtitle("Scatterplot of mpg and fitted hp")
```



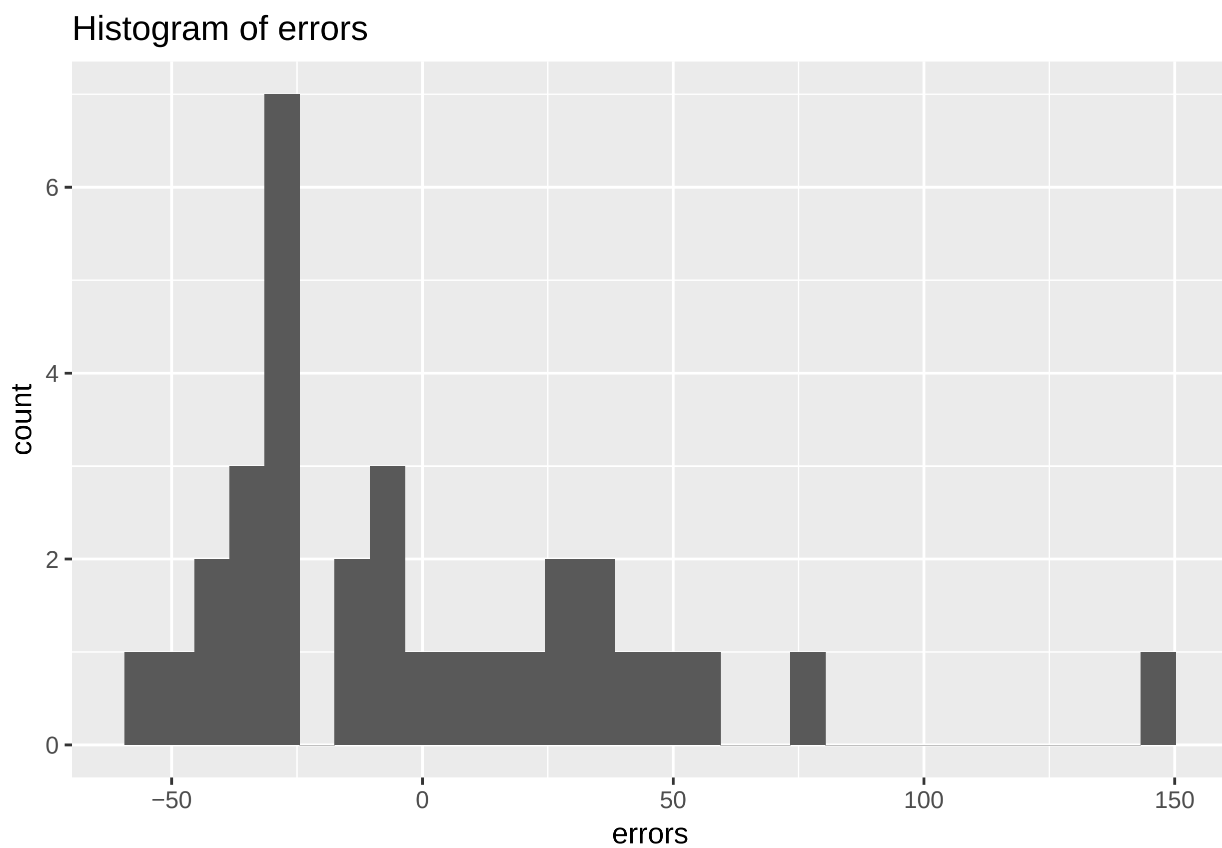
#7.

```
err<- df1$hp - df1$hphat
df2<- data.frame(df1$mpg, df1$hp, df1$hphat, err,
                 row.names = row.names(df1))
colnames(df2)<- c("mpg","hp","hphat","errors")
head(df2)
```

```
##                    mpg  hp    hphat    errors
## Mazda RX4         21.0 110 138.6580 -28.65796
## Mazda RX4 Wag     21.0 110 138.6580 -28.65796
## Datsun 710        22.8  93 122.7644 -29.76445
## Hornet 4 Drive    21.4 110 135.1261 -25.12607
## Hornet Sportabout 18.7 175 158.9663  16.03366
## Valiant           18.1 105 164.2642 -59.26418
```

```
ggplot(df2, aes(x= errors)) + geom_histogram() + ggtitle("Histogram of errors")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Yes, they look normally distributed but the distribution looks skewed to the right.

```
SSE<- sum((df2$hp-df2$hphat)^2)
SSE #Sum of Standard Errors
```

```
## [1] 57935.56
```

#8.

$$se_{\beta_1} = se_{\hat{y}} \frac{1}{\sqrt{\sum(x_i - \bar{x})^2}}$$

where,

$$se_{\hat{y}} = \sqrt{\frac{\sum(y_i - \hat{y})^2}{n-2}}$$

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

```
n<-length(df2$hp)
k<-1 #1 variable in linear model
dof<- n-k-1 #Degree of freedom
stderr_y <- sqrt(sum((df2$hp-df2$hphat)^2)/dof)
stderr_b1 <- stderr_y* 1/(sqrt(sum((df2$mpg - mean(df2$mpg))^2)))
stderr_b1 #Standard Error of Beta1
```

```
## [1] 1.309585
```

```r
t_val<- (b1-0)/stderr_b1
t_val
```

```
## [1] -6.742389
```

```r
t_crit<-qt(c(0.975,0.025),dof)
t_crit
```

```
## [1]  2.042272 -2.042272
```

```r
CI<- b1+(t_crit*stderr_b1)
CI #95% Confidence Interval
```

```
## [1]  -6.155202 -11.504260
```

```r
2*pt(t_val,dof)
```

```
## [1] 1.787835e-07
```

We can see that t_value lies outside of our acceptance region of t distribution, therefore we reject the null hypothesis. Also, we note that p value is less than 0.05, it confirms that we can reject the null hypothesis, i.e., Beta1 is not equal to 0 and this implies there exists a linear relationship between mpg and hp.

#9.

$$R^2 = \frac{TSS - SSE}{TSS}$$

```r
TSS <- sum((df2$hp - mean(df2$hp))^2)
Rsq <- (TSS -SSE)/TSS
Rsq
```

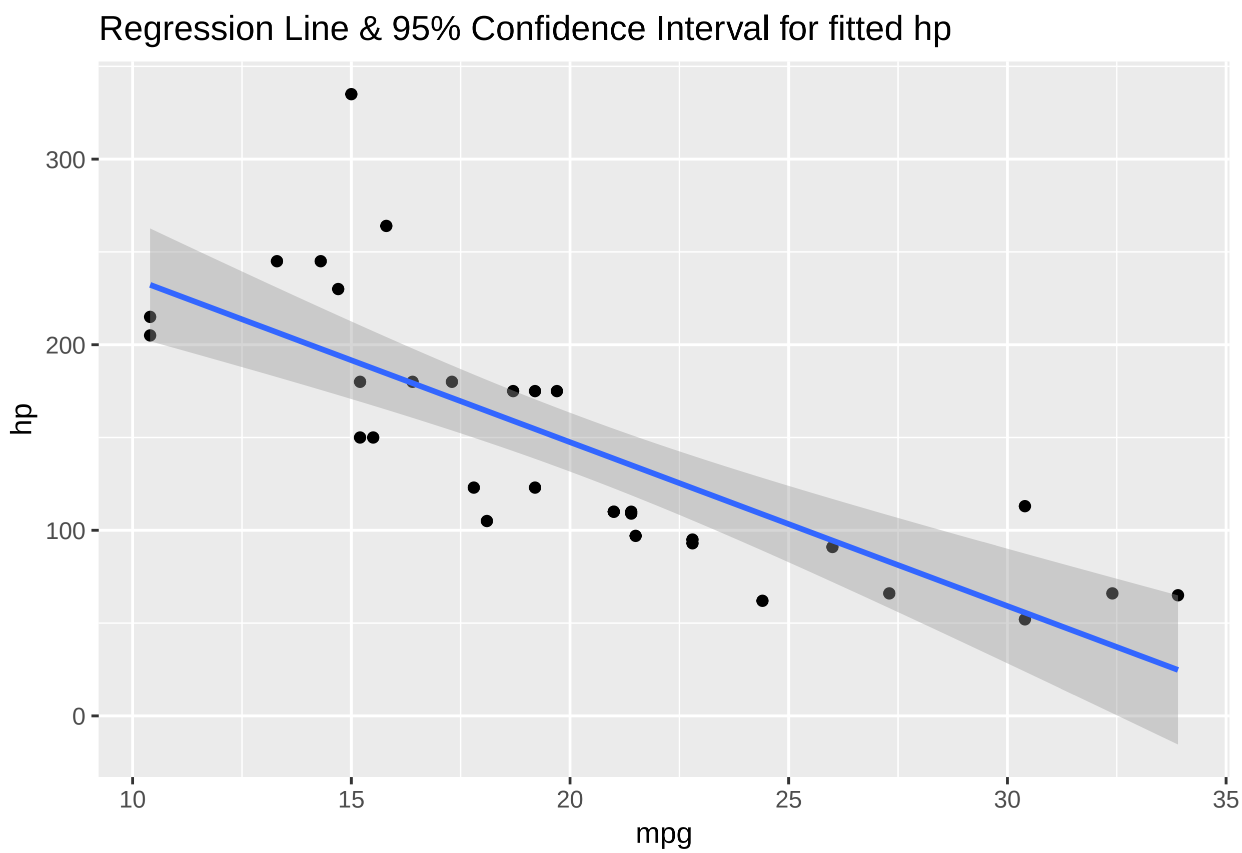
```
## [1] 0.6024373
```

R^2= 0.6024 implies that 60.24% of the variation in hp is defined by our linear model, i.e.,

$$hp = -8.83 * mpg + 324.08$$

#10.

```r
ggplot(df2, aes(x=mpg, y=hp)) + geom_point() + geom_smooth(method=lm) + ggtitle("Regression Line & 95% (
```



#11.

```r
summary(lm(hp ~ mpg, data = df2))
```

```
##
## Call:
## lm(formula = hp ~ mpg, data = df2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -59.26 -28.93 -13.45  25.65 143.36
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   324.08      27.43  11.813 8.25e-13 ***
## mpg            -8.83       1.31  -6.742 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.95 on 30 degrees of freedom
## Multiple R-squared:  0.6024, Adjusted R-squared:  0.5892
## F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

Thus, we see that Beta0= 324.08 Beta1= -8.83 Standard Error of Beta 1= 1.31 t-test for Beta 1 (t-value)= -6.742 and p-value= 1.79e-07 match our values in 5. and 8. Hence, proved!