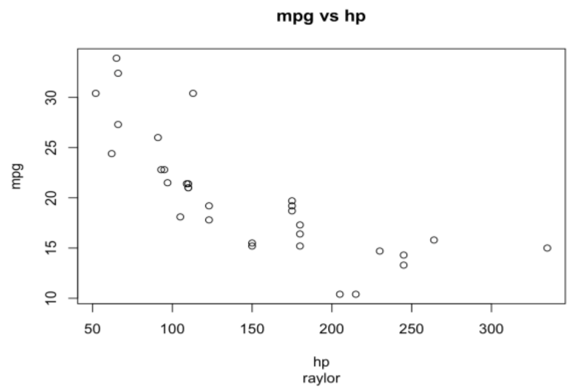
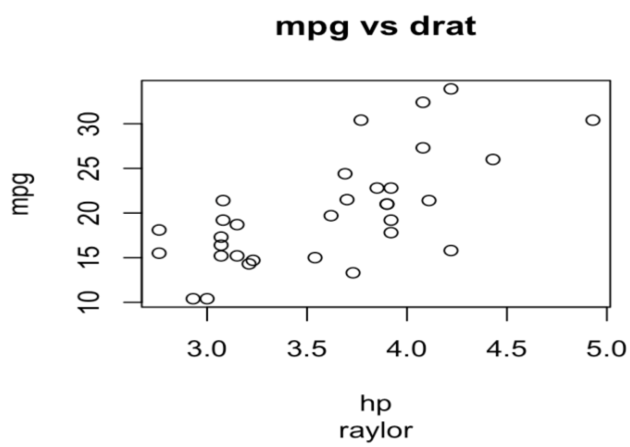
Problem 2:

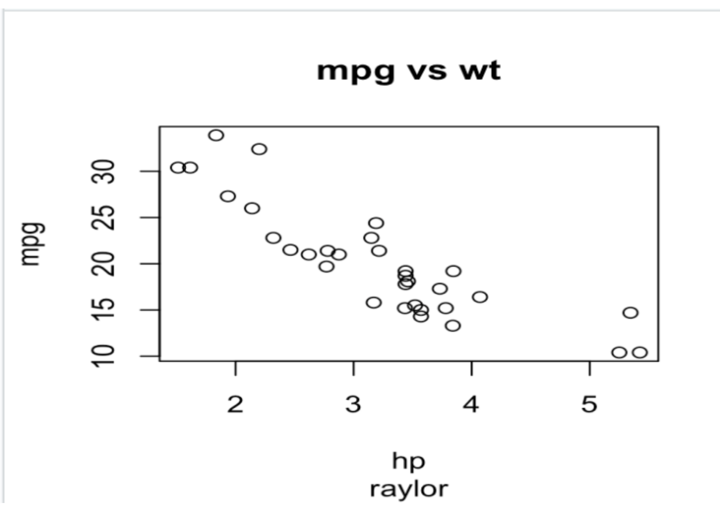
A. Give scatter plots of **mpg** vs. **horsepower, mpg** vs. **rear axle ratio mpg** vs. **weight** and **mpg** vs. **time for ¼ mile**. Comment about linearity on each plot. Is there any relationship you want to explore more?



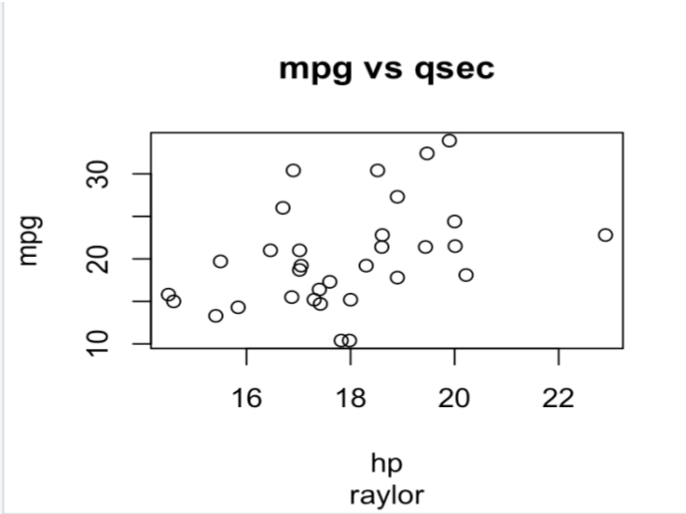
It has some negative linear association between mpg and hp.



Mpg and drat: It is positive but the linearity is weak.



Mpg and weight have a strong negative linear relationship.



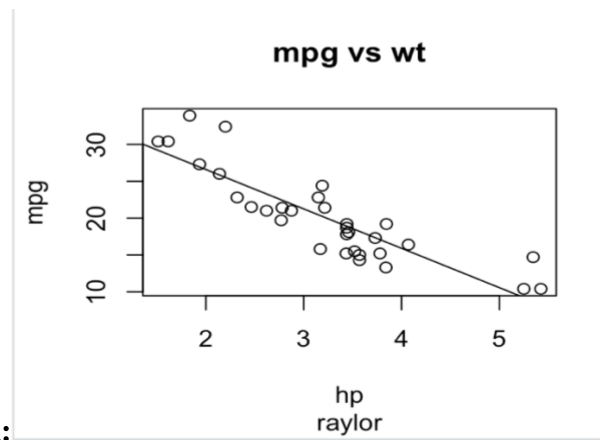
Mpg and qsec is seems to be positive but weak linearity.
I want to explore mpg and weight relationship more because it is the one which have the strongest linear relationship. It is more meaningful.

B. Fit a simple linear regression model for **mpg** vs. **weight.**

i) Give the details for the regression model, including:

β_0=37.2851 β_1 =-5.3445 σ^2=3.046^2=9.278116
the equation of the estimated regression line: MPG=37.2851-5.3445 weight
R^2=0.7528

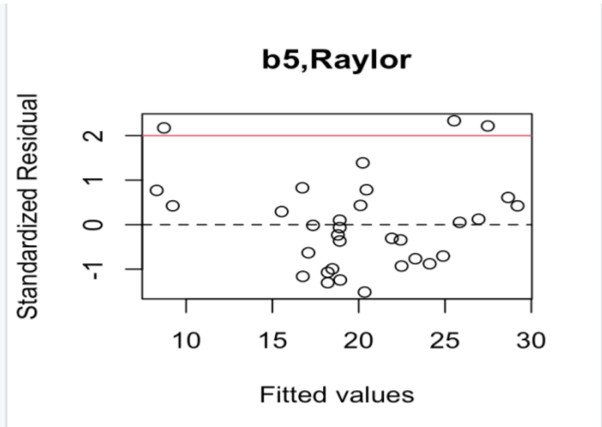
Regression plot:



ii) **Confidence interval is between -6.486308 and -4.202635. We can conclude with 95% of confidence level that the mpg increases between -6.486308 and -4.202635 for each additional unit of weight. The difference is E(-1.5b1)=-1.5*(-6.486308 , -4.202635)=(9.7294, 6.3039)**

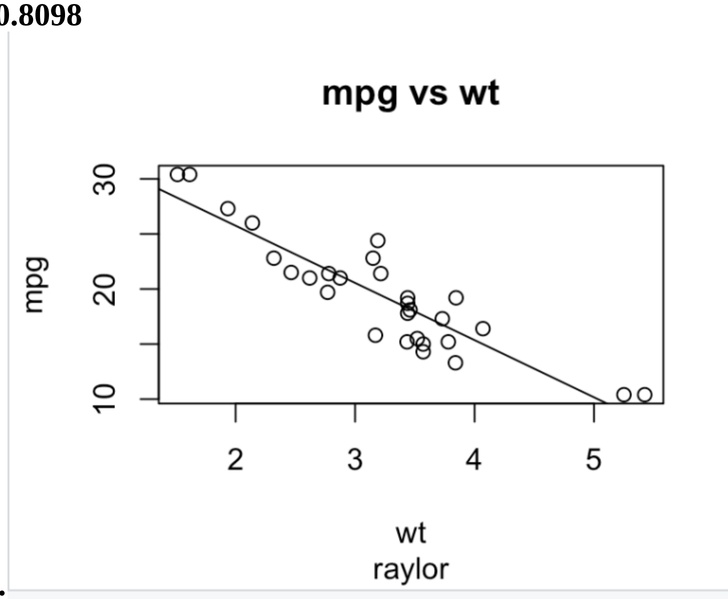
iii) **The intercept is 37.2851. Assume x=o, intercept is the value of y. In other words, if the weight is 0, the mpg is 37.2851 even though the weight couldn't be zero.**

iv) **Cl of mean (3.21725) is between 18.99098 and 21.19027. and predict interval is between 13.77366 and 26.40759.**

v) 

We can see there are three outliers.

New model: β_0=36.1376 β_1 =-5.1948 σ^2=2.283^2=5.212
the equation of the estimated regression line: MPG=36.1376-5.1948 weight
R^2=0.8098

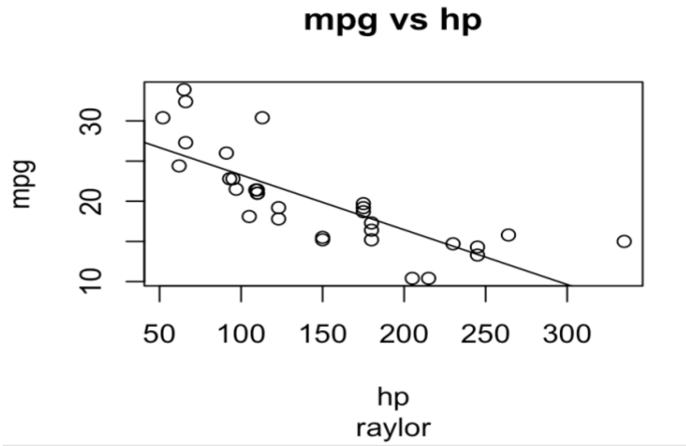
Plot: 

I found the good regression model should avoid outliers and the larger r^2 means more linear. Like the new model is better.

C. Fit a simple linear regression model **mpg** vs. **horsepower**

i) **β_0=30.09886 β_1 =-0.06823 σ^2=3.863^2=14.8996**

the equation of the estimated regression line: MPG= 30.09886-0.06823hp
R^2=0.6014



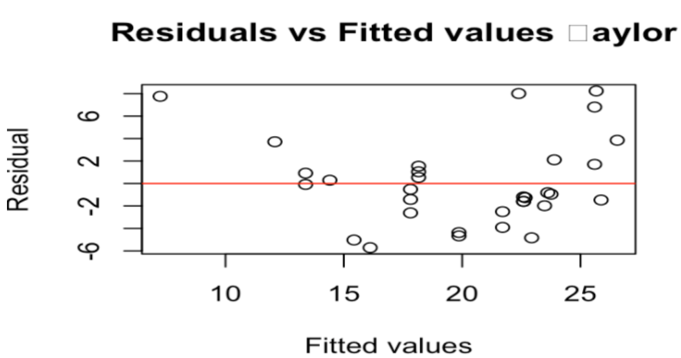
Plot

I prefer model from part b. That is because part b has larger r^2, which is more linear and more useful. Moreover, σ^2 of part b model is smaller thanσ^2 of this model.

Yes , it is necessarily indicated that the variable used in the preferred model is much more informative to our common response variable.

ii) **Cl of mean (3.21725) is between 18.69599 and 21.48526 and predict interval is between 12.07908 and 28.10217.**

Compared with cl of mean, I found 95% confidence interval for the mean response when the horsepower of motors reaches at an average value is similar with part b vi, but the 95% confidence interval of the miles a motor with average horsepower can run by 1 gallon of oil is not similar.



iii)

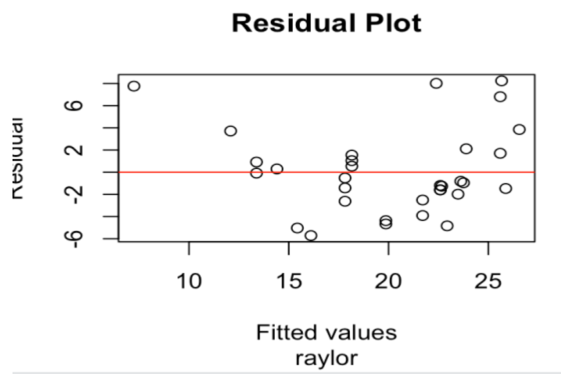
$$H_0: Equal\ Variance \sigma^2(\varepsilon_i) = \sigma^2 \ \ vs\ H_a: Unequal\ Variance$$

BP = 0.047689, df = 1, p-value = 0.8271> 0.05

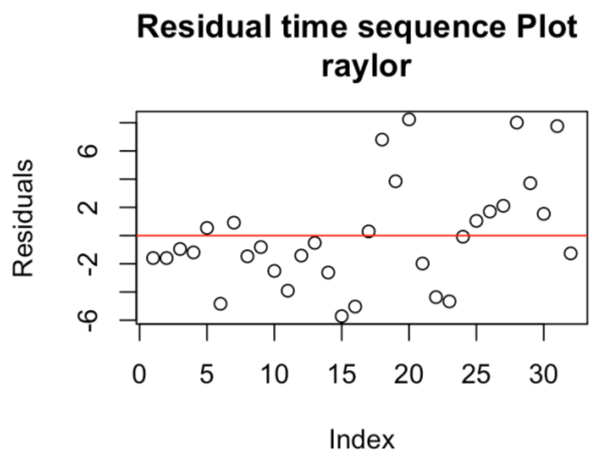
We cannot reject null hypothesis. There is no significant evidence that the variances are unequal.

iv) Check other three basic assumptions.

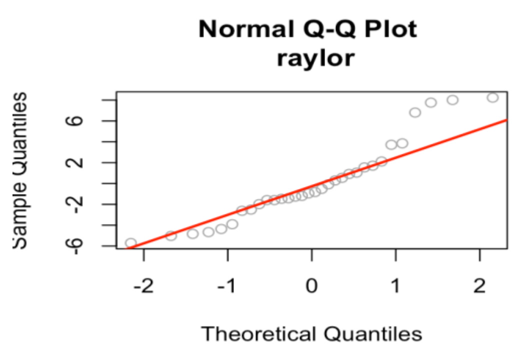
Firstly, use proper plots to test these three assumptions and make comments on them. Secondly, check the independence and normality with proper hypothesis testings and give your conclusions.



They are not linear, and the errors do not seem to have constant variances.



Based on the index vs residuals plot, it suggests that the errors are not independent.



Based on the Q-Q plot, it seems that the points are not well fitted with the QQ line. We may conclude that the errors are not very normally distributed.

Dw Test:

H_0: *Errors are uncorrelated over time* H_a: *Errors are positively correlated*

DW = 1.1338, p-value = 0.00411<0.05, We reject null hypothesis. There is significant evidence that the errors are positively correlated.

Sh:

H_0: *Errors are from normal distibution* H_a: *Errors are not from normal distribution*

W = 0.92337, p-value = 0.02568>0.05, We reject null hypothesis. There is significant evidence that the errors are not from normal distribution.

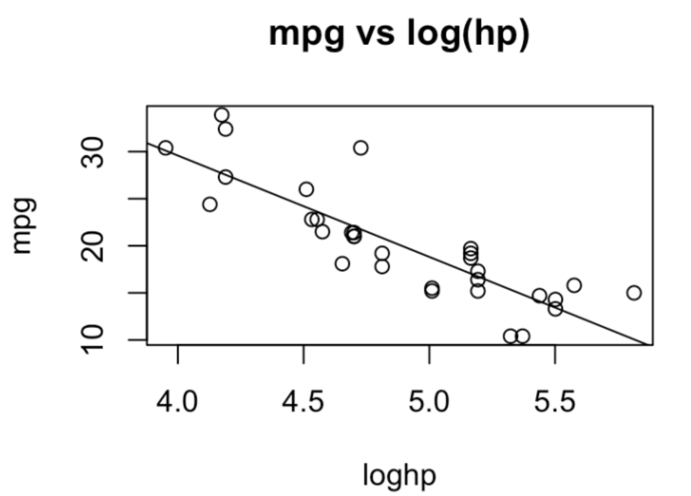
v) Assume that you regard the model as non-linearity, what would you do to improve your model? Analysis with the residual plot. Give a possible way and show as much details as you can. Notice, the method is expect doing log transformation on horsepower, since we will do it later in Part D.
Does your method work?

If let me improve it, I will transfer the value of horsepower by log horsepower.

```
Residuals:
    Min      1Q  Median      3Q     Max 
-4.9427 -1.7053 -0.4931  1.7194  8.6460 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   72.640      6.004  12.098 4.55e-13 ***
loghp        -10.764      1.224  -8.792 8.39e-10 ***
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.239 on 30 degrees of freedom
Multiple R-squared:  0.7204,	Adjusted R-squared:  0.7111 
F-statistic:  77.3 on 1 and 30 DF,  p-value: 8.387e-10
```



raylor

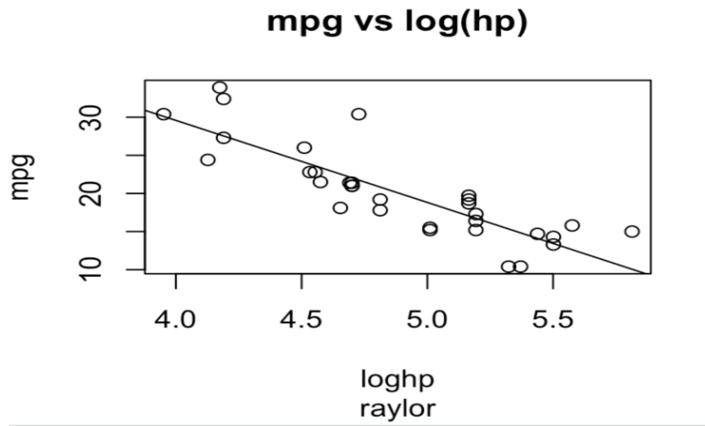
It works, we can see it is more linear and normal distribution. And we have larger r^2.

D. Fit a simple linear regression model **mpg** vs. log(**horsepower)**

i) Give the details for the regression model, including:

β_0=72.64 β_1 =-10.764 σ^2=3.239^2=10.4911

the equation: MPG=72.64-10.764 log(**horsepower) R^2=0.7204**

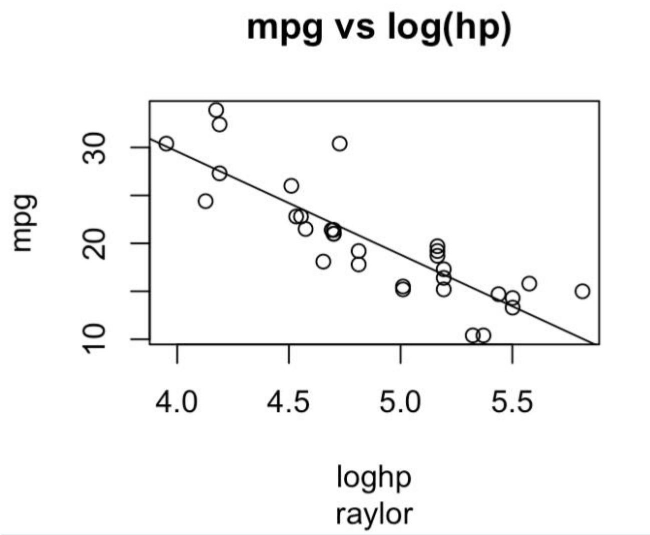


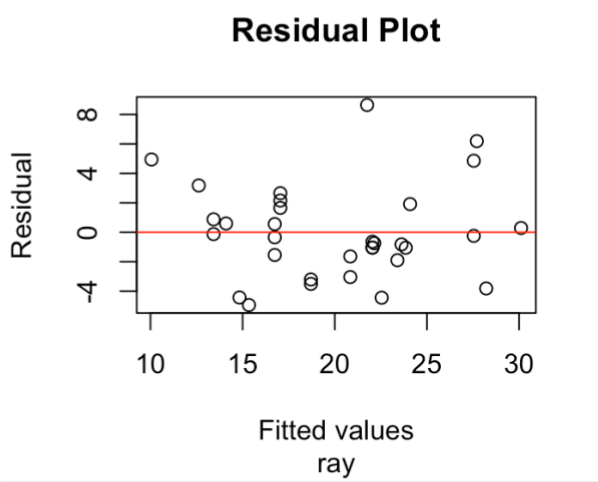
We estimate that the mean mpg decreases by 10.764 for each additional unit in log horsepower. 72.04 percent of the variation in mpg is explained by the variation in log horsepower.

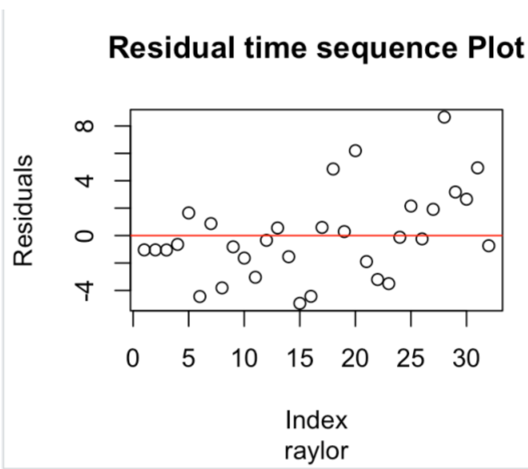
ii) **I prefer this model and my own improved model. That is because they are more linear and has large r^2 and smaller sigma.**

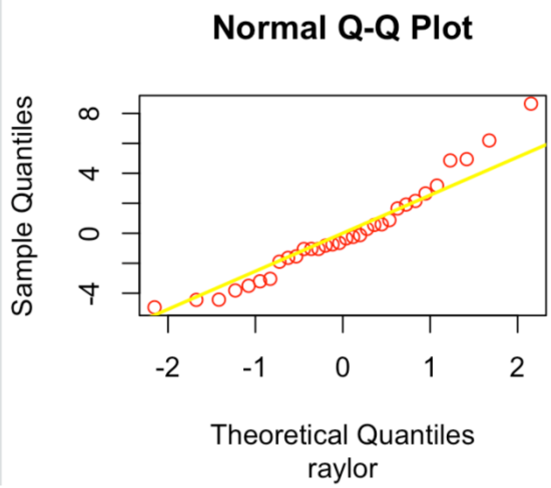
iii) **By r, 80: (23.31,27.63) 160: (16.41, 19.606) 240: (11.25, 16.04)**
We can conclude with 95% of confidence level for mean response at hp=80 is between 23.31 and 27.63.
We can conclude with 95% of confidence level for mean response at hp=160 is between 16.41and 19.606.
We can conclude with 95% of confidence level for mean response at hp=240 is between 11.25 and 16.04.
Hp=240, interval is widest because it is not stable if hp is very large.

iv) Check the basic assumption for this simple linear regression. Give the proper plots for testing each assumption. Show complete hypothesis testing procedures for checking equal variances, independence, and normality.


mpg vs log(hp)
mpg
10
20
30
4.0
4.5
5.0
5.5
loghp
raylor


Residual Plot
Residual
-4
0
4
8
10
15
20
25
30
Fitted values
ray


Residual time sequence Plot
Residuals
-4
0
4
8
0
5
10
15
20
25
30
Index
raylor



Bp: $\boldsymbol{H_0: Equal\ Variance\ \sigma^2(\varepsilon_i) = \sigma^2\ \ vs\ H_a: Unequal\ Variance}$

BP = 0.19869, df = 1, p-value = 0.6558 $\boldsymbol{> 0.05}$

We cannot reject null hypothesis. There is no significant evidence that the variances are unequal.

Dw Test:

$\boldsymbol{H_0: Errors\ are\ uncorrelated\ over\ time\quad H_a: Errors\ are\ positively\ correlated}$

DW = 1.3826, p-value = 0.03109<0.05, We reject null hypothesis. There is significant evidence that the errors are positively correlated.

Sw:

$\boldsymbol{H_0: Errors\ are\ from\ normal\ distibution\ H_a: Errors\ are\ not\ from\ normal\ distribution}$

W = 0.9533, p-value = 0.1788>0.05, We reject null hypothesis. There is significant evidence that the errors are not from normal distribution.

```
> data("mtcars")
> mtcars
                  mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
Merc 280C         17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
```

```
Merc 450SE          16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
Merc 450SL          17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
Merc 450SLC         15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
Fiat 128            32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
Honda Civic         30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
Toyota Corolla      33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
Toyota Corona       21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
Dodge Challenger    15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
AMC Javelin         15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
Camaro Z28          13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
Pontiac Firebird    19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
Fiat X1-9           27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
Porsche 914-2       26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
Lotus Europa        30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
Ford Pantera L      15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
Ferrari Dino        19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
Maserati Bora       15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
Volvo 142E          21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
> plot(mtcars$mpg~mtcars$hp, xlab="hp",ylab = "mpg", main="mpg vs hp",sub="raylor")
> plot(mtcars$mpg~mtcars$drat, xlab="hp",ylab = "mpg", main="mpg vs drat",sub="raylor")
> plot(mtcars$mpg~mtcars$wt, xlab="hp",ylab = "mpg", main="mpg vs wt",sub="raylor")
> plot(mtcars$mpg~mtcars$qsec, xlab="hp",ylab = "mpg", main="mpg vs qsec",sub="raylor")
> regmodel=lm(mtcars$mpg~mtcars$wt)
> summary(r)

Call:
lm(formula = mtcars$mpg ~ mtcars$wt)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5432 -2.3647 -0.1252  1.4096  6.8727

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
mtcars$wt    -5.3445     0.5591  -9.559 1.29e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,  Adjusted R-squared:  0.7446
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
> plot(mtcars$mpg~mtcars$wt, xlab="hp",ylab = "mpg", main="mpg vs wt",sub="raylor")
```

```
> abline(lm(mtcars$mpg~mtcars$wt))
> confint(regmodel)
                2.5 %    97.5 %
(Intercept) 33.450500 41.119753
mtcars$wt   -6.486308 -4.202635
>mpg= mtcars$mpg
>  wt= mtcars$wt
.newdata = data.frame(wt=1.5)
> predict(regmodel, newdata, interval="confidence")
       fit     lwr      upr
1 29.26842 27.0203 31.51653
>mean(wt)
> newdataa = data.frame(wt=3.21725)
> predict(regmodel, newdataa, interval="confidence")
> predict(regmodel, newdataa, interval="prediction")
> standard_res <- rstandard(regmodel)
> plot(regmodel$fitted.values,standard_res,main="b5,Raylor",xlab="Fitted values",ylab="Standardized Residual")
> abline(h=c(-2,0,2),col=c(2,1,2),lty=c(1,2,1))
> data<- data.frame(mpg, wt)
> new.data<- data[abs(rstandard(regmodel))<2 , ]
> reg=lm(new.data
+ )
> summary(reg)

Call:
lm(formula = new.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.8700 -1.8324 -0.0635  1.5353  4.8339 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  36.1376     1.6200   22.31  < 2e-16 ***
wt           -5.1948     0.4846  -10.72 3.13e-11 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.283 on 27 degrees of freedom
Multiple R-squared:  0.8098,  Adjusted R-squared:  0.8027 
F-statistic: 114.9 on 1 and 27 DF,  p-value: 3.126e-11
> plot(new.data$mpg~new.data$wt, xlab="hp",ylab = "mpg", main="mpg vs wt",sub="raylor")
> abline(lm(new.data$mpg~new.data$wt))
> regmodel1= lm(mpg~hp)
```

```
> summary(regmodel1)

Call:
lm(formula = mpg ~ hp)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.7121 -2.1122 -0.8854  1.5819  8.2360 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
hp          -0.06823    0.01012  -6.742 1.79e-07 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.863 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
> mean(hp)
[1] 146.6875
>  newdata1= data.frame(hp=146.6875)
>  predict (regmodel1, newdata1, interval="confidence")
       fit      lwr      upr
1 20.09062 18.69599 21.48526
>  predict (regmodel1, newdata1, interval="prediction")
       fit      lwr      upr
1 20.09062 12.07908 28.10217
> 
> plot(regmodel1$fitted.values,regmodel1$residuals,main=" Residuals vs Fitted values \raylor",
+      xlab="Fitted values",ylab="Residual")
> abline(h=0,col="red")
> bptest(regmodel1, studentize=FALSE)

	Breusch-Pagan test

data:  regmodel1
BP = 0.047689, df = 1, p-value = 0.8271

> plot (regmodel1$fitted.values, regmodel1$residuals, main="Residual Plot", sub=
"raylor",xlab="Fitted values",ylab="Residual")
> abline(h=0, col="red")
> abline(h=0,col="red")
> plot(regmodel1$residuals, ylab="Residuals",main="Residual time sequence Plot")
> abline(h=0,col="red")
```

```
> plot(regmodel1$residuals, ylab="Residuals",main="Residual time sequence Plot \n raylor")
> abline(h=0,col="red")
> qqline(resid(regmodel1), col = "red", lwd = 2)
>
> qqnorm(resid(regmodel1), main = "Normal Q-Q Plot \n raylor", col = "darkgrey")
> qqline(resid(regmodel1), col = "red", lwd = 2)
> dwtest(mpg~hp)

        Durbin-Watson test

data:  mpg ~ hp
DW = 1.1338, p-value = 0.00411
alternative hypothesis: true autocorrelation is greater than 0

> shapiro.test(resid(regmodel1))

        Shapiro-Wilk normality test

data:  resid(regmodel1)
W = 0.92337, p-value = 0.02568
> loghp=log(hp)
> re = lm(mpg~loghp)
> summary(re)

Call:
lm(formula = mpg ~ loghp)

Residuals:
    Min      1Q  Median      3Q     Max
-4.9427 -1.7053 -0.4931  1.7194  8.6460

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   72.640      6.004  12.098 4.55e-13 ***
loghp        -10.764      1.224  -8.792 8.39e-10 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.239 on 30 degrees of freedom
Multiple R-squared:  0.7204,  Adjusted R-squared:  0.7111
F-statistic:  77.3 on 1 and 30 DF,  p-value: 8.387e-10
> plot (mpg~loghp, main="mpg vs log(hp)", sub="raylor")
> abline(lm(mpg~loghp))
.>scheffe(re,data.frame(hp=80))
>scheffe(re,data.frame(hp=160))
```

```
>scheffe(re,data.frame(hp=240))
> plot(mpg~loghp, main=" mpg vs log(hp)", sub="raylor")
> abline(lm(mpg~loghp))
> plot (re$fitted.values,re$residuals, main="Residual Plot", sub= "ray",xlab="Fitted values",ylab="Residual")
> abline(h=0, col="red")
> plot(re$residuals, ylab="Residuals",main="Residual time sequence Plot", sub= "raylor")
> abline(h=0, col="red")
> qqnorm(resid(re), main = "Normal Q-Q Plot", col = "red", sub="raylor")
> qqline(resid(re), col = "yellow", lwd = 2)
> bptest (re, studentize = FALSE)

        Breusch-Pagan test

data:  re
BP = 0.19869, df = 1, p-value = 0.6558

> dwtest(mpg~loghp)

        Durbin-Watson test

data:  mpg ~ loghp
DW = 1.3826, p-value = 0.03109
alternative hypothesis: true autocorrelation is greater than 0

> bptest(re, studentize=FALSE)

        Breusch-Pagan test

data:  re
BP = 0.19869, df = 1, p-value = 0.6558

> dwtest(mpg~loghp)

        Durbin-Watson test

data:  mpg ~ loghp
DW = 1.3826, p-value = 0.03109
alternative hypothesis: true autocorrelation is greater than 0

> shapiro.test(resid(re))

        Shapiro-Wilk normality test

data:  resid(re)
W = 0.9533, p-value = 0.1788
```