Assignment-EDA-2

vinoth Aryan Nagabooshanam

September 12, 2017

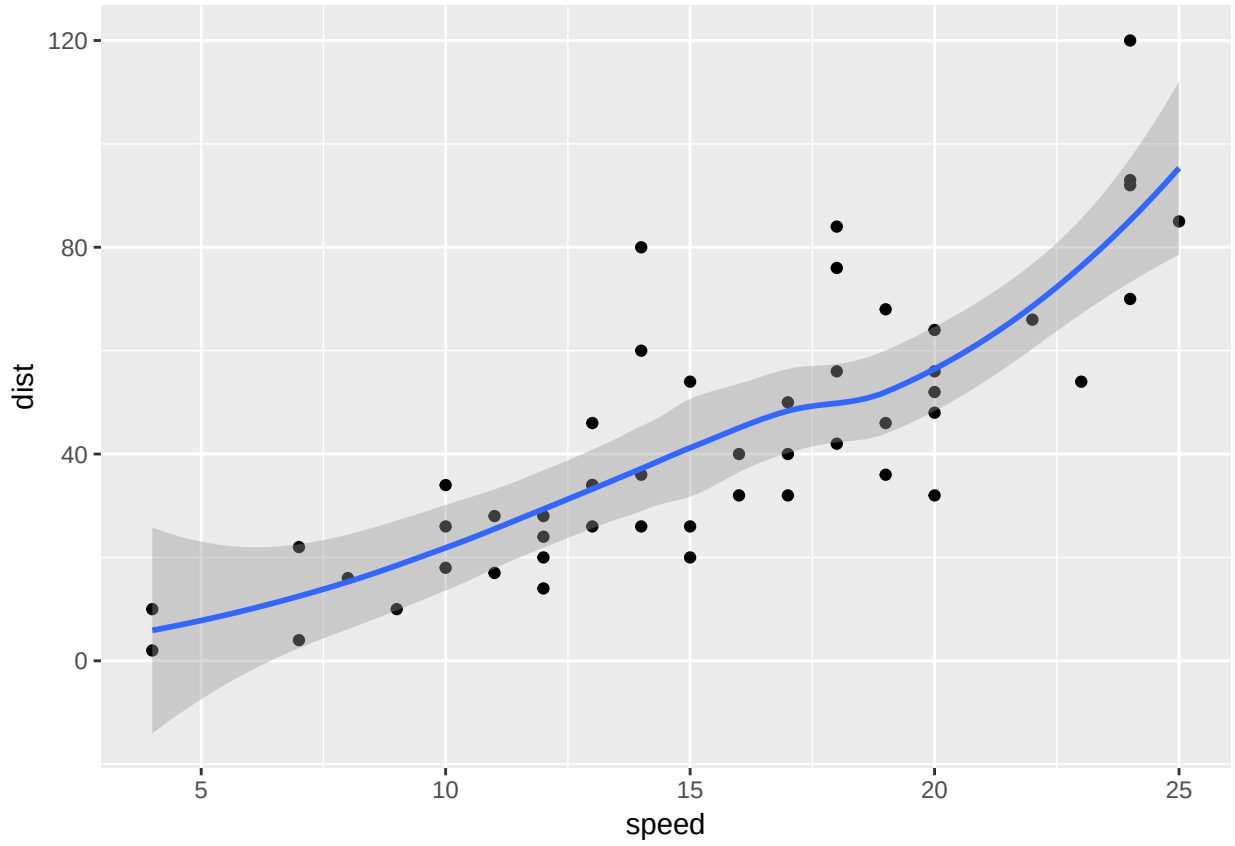
Solution 1 First we are drawing the scatterplot and adding smoth curve to plot

```
library("ggplot2")
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

```
cars.gg=ggplot(cars,aes(x=speed,y=dist))+geom_point()+geom_smooth()
cars.gg
```

```
## `geom_smooth()` using method = 'loess'
```



here the code which is used to model simple linear model

```
library(ggplot2)
library(broom)
```

```
## Warning: package 'broom' was built under R version 3.3.3
```

```
cars.lm=lm(dist~speed,data=cars)
#install.packages('broom')
car.lm.df=augment(cars.lm)
```

```
## Warning: Deprecated: please use `purrr::possibly()` instead

## Warning: Deprecated: please use `purrr::possibly()` instead

## Warning: Deprecated: please use `purrr::possibly()` instead

## Warning: Deprecated: please use `purrr::possibly()` instead

## Warning: Deprecated: please use `purrr::possibly()` instead
```

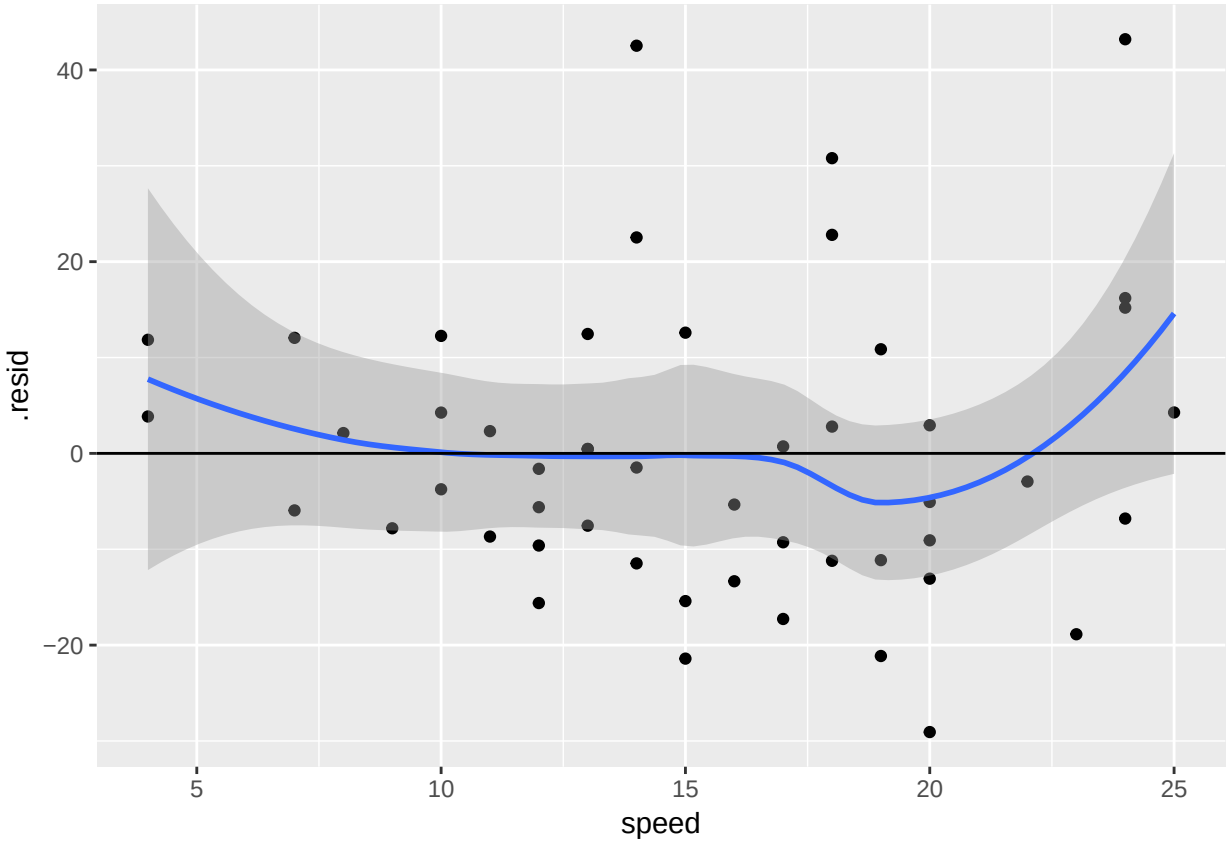
```
summary(car.lm.df)
```

```
##       dist            speed         .fitted          .se.fit
##  Min.   :  2.00   Min.   : 4.0   Min.   :-1.849   Min.   :2.181
##  1st Qu.: 26.00   1st Qu.:12.0   1st Qu.:29.610   1st Qu.:2.393
##  Median : 36.00   Median :15.0   Median :41.407   Median :2.640
##  Mean   : 42.98   Mean   :15.4   Mean   :42.980   Mean   :2.967
##  3rd Qu.: 56.00   3rd Qu.:19.0   3rd Qu.:57.137   3rd Qu.:3.358
##  Max.   :120.00   Max.   :25.0   Max.   :80.731   Max.   :5.212
##      .resid             .hat             .sigma         .cooksd
##  Min.   :-29.069   Min.   :0.02012   Min.   :14.10   Min.   :0.0000113
##  1st Qu.: -9.525   1st Qu.:0.02420   1st Qu.:15.37   1st Qu.:0.0017903
##  Median : -2.272   Median :0.02946   Median :15.47   Median :0.0069914
##  Mean   :  0.000   Mean   :0.04000   Mean   :15.38   Mean   :0.0210046
##  3rd Qu.:  9.215   3rd Qu.:0.04774   3rd Qu.:15.53   3rd Qu.:0.0191012
##  Max.   : 43.201   Max.   :0.11486   Max.   :15.54   Max.   :0.3403959
##    .std.resid
##  Min.   :-1.924523
##  1st Qu.:-0.627833
##  Median :-0.151050
##  Mean   : 0.002765
##  3rd Qu.: 0.610374
##  Max.   : 2.919060
```

now we are plot stopping distance(response variable) the speed (explantory variable) then add a loess curve.if the confidence band contains the line y = 0, then maybe the model is fitting well.

```
ggplot(car.lm.df,aes(x=speed,y=.resid))+geom_point()+geom_smooth()+geom_abline(slope=0,intercept = 0)
```

```
## `geom_smooth()` using method = 'loess'
```

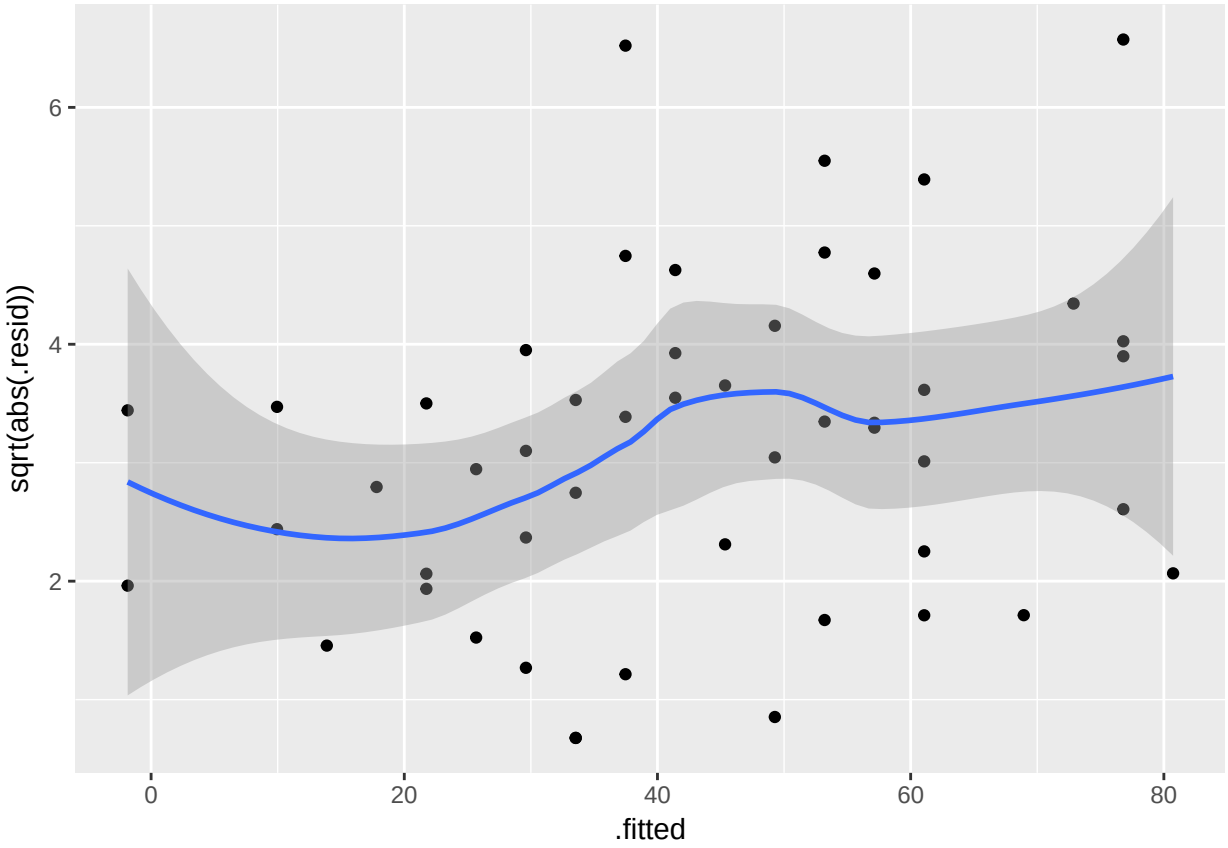


From the above figure, we can clearly see the residuals is wiggling around zero, and given the distribution of residual with respect to speed, we can say that the simple linear model fits.

more option we will check homosedasticity

```
ggplot(car.lm.df,aes(x=.fitted,y=sqrt(abs(.resid))))+geom_point()+geom_smooth()
```

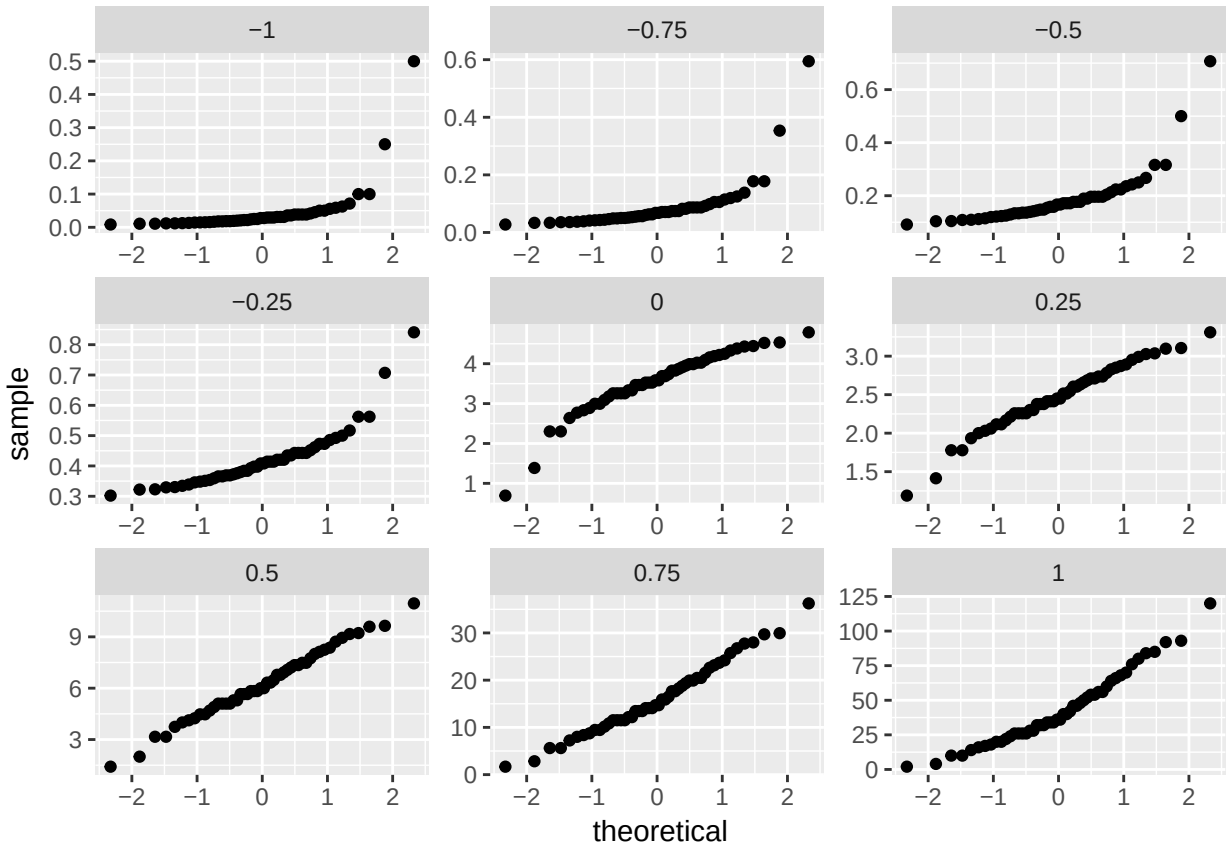
```
## `geom_smooth()` using method = 'loess'
```



if we see above given figure its clear its fitted the simple linear model even clear by plot using homosedasticity check processor.

Solution 2 : Using ggplot() to reproduce the normal Q-Q plots on the next page for different power transformations of stopping distance.

```
dissort=sort(cars$dist)
n=length(dissort)
# here the length is equal
power=rep(seq(-1,1,0.25),each=n)
dsort=c(dissort^-1,dissort^-0.75,dissort^-0.5,dissort^-0.25,log(dissort),dissort^0.25,dissort^0.5,disso
ggplot(data.frame(power,dsort ), aes(sample = dsort)) + stat_qq() + facet_wrap(~power,scales = "free")
```



if we look above given qqplot and power transformation of Cars$dist with vlaue 1,0.75,0.5,0.25 look a normal distrution.

more the 0.5 poewr transformation look so prefect as normal distribution.

q3 Based on the normal Q-Q plots of 0.5 power transformation that looks the best. and i am Use the transformed stopping distance to fit the linear model again and comparing on whether the model fits better after transformation or not.

```
library(ggplot2)
library(grid)
library(gridExtra)
cars.lm=lm(cars$dist~cars$speed)
car.lm.df=augment(cars.lm)
```

```
## Warning: Deprecated: please use `purrr::possibly()` instead
```

```
## Warning: Deprecated: please use `purrr::possibly()` instead
```

```
## Warning: Deprecated: please use `purrr::possibly()` instead
```

```
## Warning: Deprecated: please use `purrr::possibly()` instead
```

```
## Warning: Deprecated: please use `purrr::possibly()` instead
```

```
a=ggplot(car.lm.df, aes(x = cars$speed, y = .resid)) + geom_point() + geom_smooth() +
geom_abline(slope = 0, intercept = 0)+ggtitle(" this plot show before pow transformatin")
car.p05.lm = lm((dissort)^0.5 ~ cars$speed)
car.p05.lm.df = augment(car.p05.lm)
```

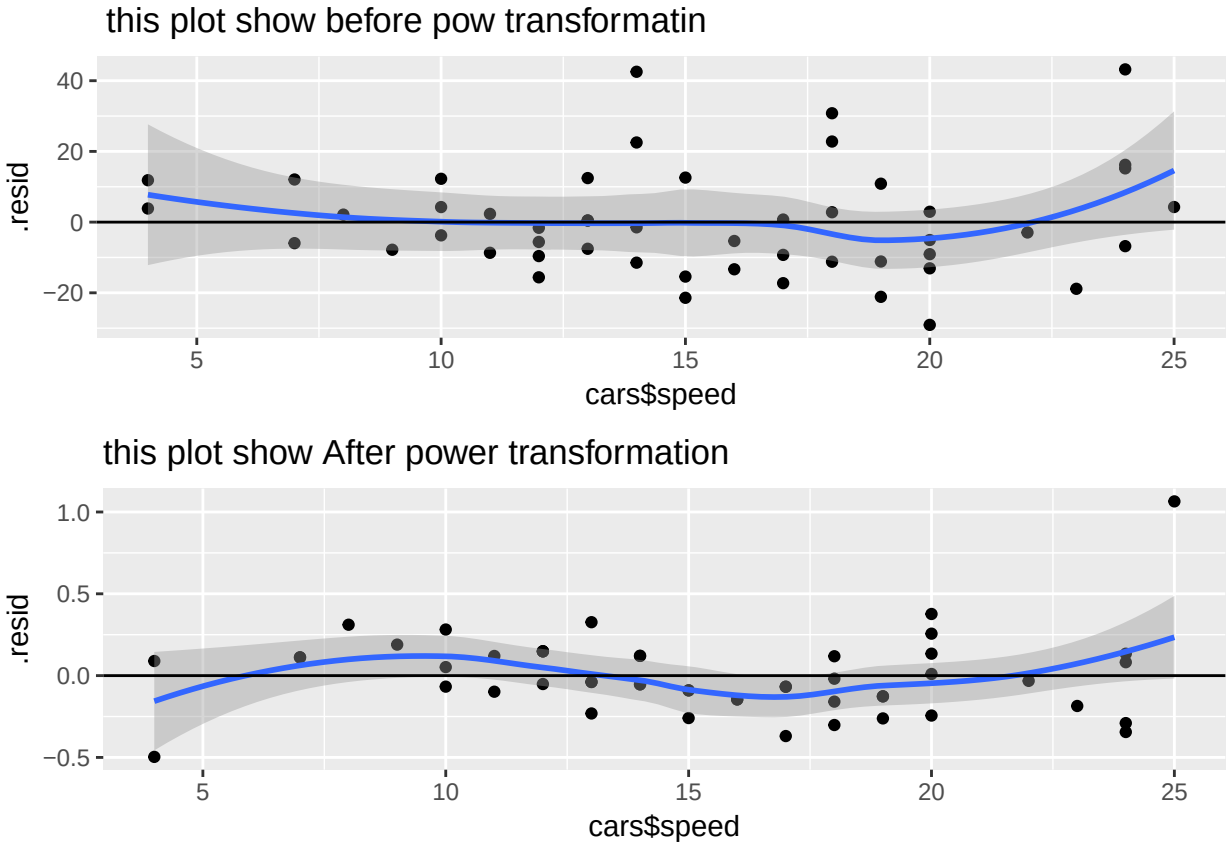
```
## Warning: Deprecated: please use `purrr::possibly()` instead
```

```
## Warning: Deprecated: please use `purrr::possibly()` instead
## Warning: Deprecated: please use `purrr::possibly()` instead
## Warning: Deprecated: please use `purrr::possibly()` instead
## Warning: Deprecated: please use `purrr::possibly()` instead
```

```
b=ggplot(car.p05.lm.df, aes(x = cars$speed, y = .resid)) + geom_point() + geom_smooth() +
geom_abline(slope = 0, intercept = 0)+ggtitle("this plot show After power transformation")

grid.arrange(a,b, nrow=2)
```

```
## `geom_smooth()` using method = 'loess'
## `geom_smooth()` using method = 'loess'
```

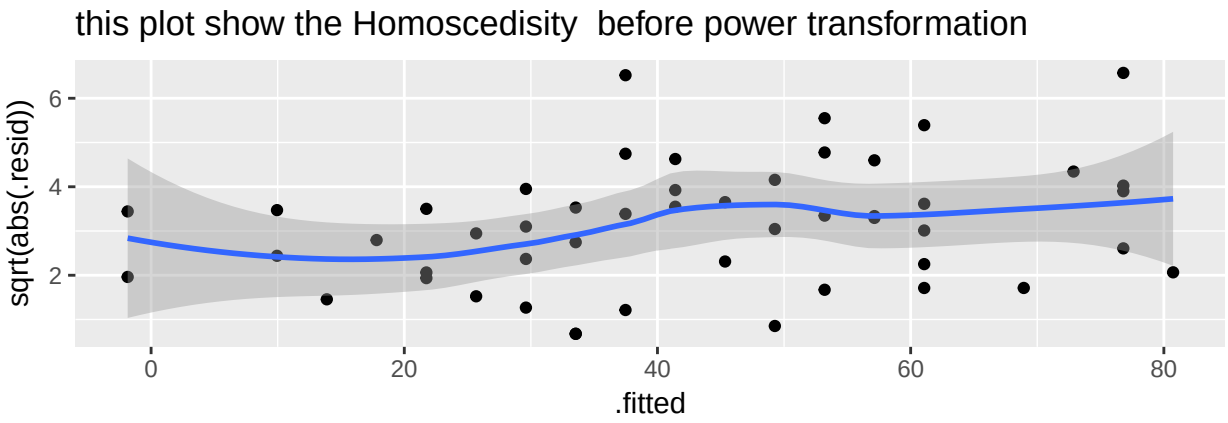


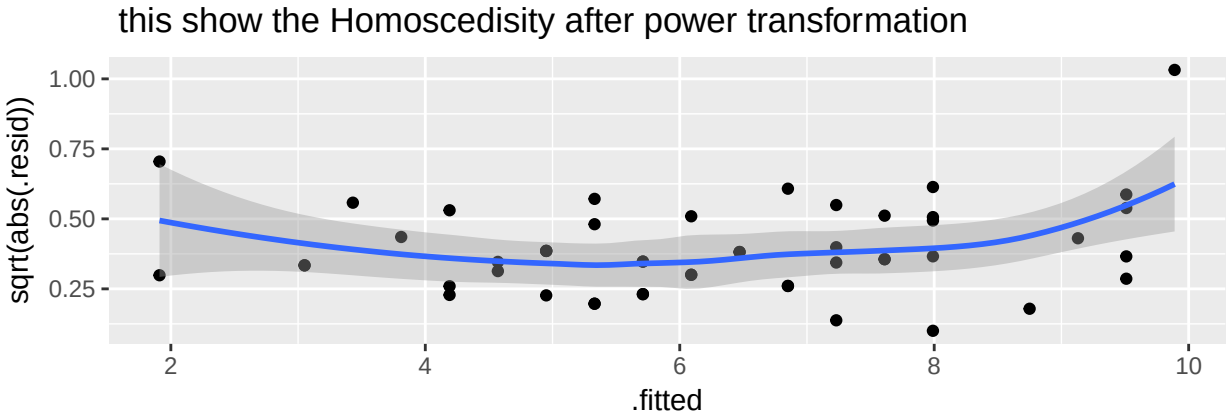
if we see the above given plot the residual wiggle around 0 so we can say it fitts linear model for both cases.

now we are going to check with homoscedisity check

```
c <- ggplot(car.lm.df, aes(x = .fitted, y = sqrt(abs(.resid)))) + geom_point() +
geom_smooth()+ggtitle("this plot show the Homoscedisity  before power transformation")
d<- ggplot(car.p05.lm.df, aes(x = .fitted, y = sqrt(abs(.resid)))) + geom_point() +
geom_smooth()+ggtitle(" this show the Homoscedisity after power transformation")
grid.arrange(c,d, nrow=2)
```

```
## `geom_smooth()` using method = 'loess'
## `geom_smooth()` using method = 'loess'
```





if we see the above given plot the plot without power transformation which is like horizontal line so we can this homoscedasicity is not correct method for this cases. for with power transformaton its do not like horizontial line so the is ok for homoscadesity is corret for this. so we are say model fit linear and homescadesity for power taransformation.

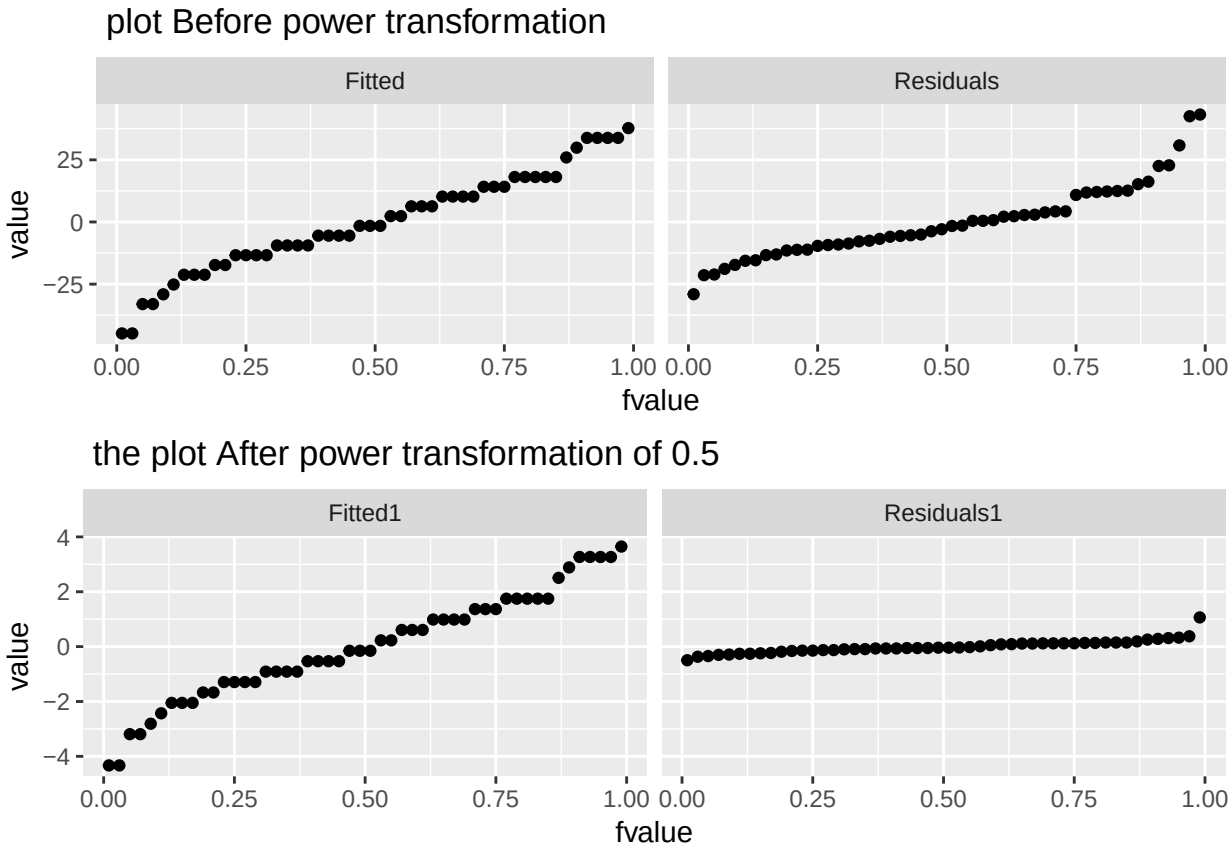
we are going to check residual fit plot for two cases

```
require(gridExtra)
l = nrow(car.lm.df)
fvalue = (0.5:(l - 0.5))/l
car.ft = data.frame(fvalue, Fitted = sort(car.lm.df$.fitted) - mean(car.lm.df$.fitted),
Residuals = sort(car.lm.df$.resid))
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.3.3
```

```
car.ft.lo = car.ft %>% gather(type, value, Fitted:Residuals)
p1 <- ggplot(car.ft.lo, aes(x = fvalue, y = value)) +
geom_point() + facet_wrap(~type) + ggtitle(" plot Before power transformation")
l1 = nrow(car.p05.lm.df)
f1value = (0.5:(l - 0.5))/l1
car.pt.ft = data.frame(fvalue, Fitted1 = sort(car.p05.lm.df$.fitted) - mean(car.p05.lm.df$.fitted),
Residuals1 = sort(car.p05.lm.df$.resid))
library(tidyr)
car.pt.ft.lo = car.pt.ft %>% gather(type, value, Fitted1:Residuals1)
p2 <- ggplot(car.pt.ft.lo, aes(x = fvalue, y = value)) +
geom_point() + facet_wrap(~type) +
ggtitle(" the plot After power transformation of 0.5")
```

```
grid.arrange(p1,p2,nrow=2)
```



in with out the transformation both fitted and residual look normal in after transormation fitted and residual is different so linear model fit in befor transformation.