HW11

106022113

5/4/2021

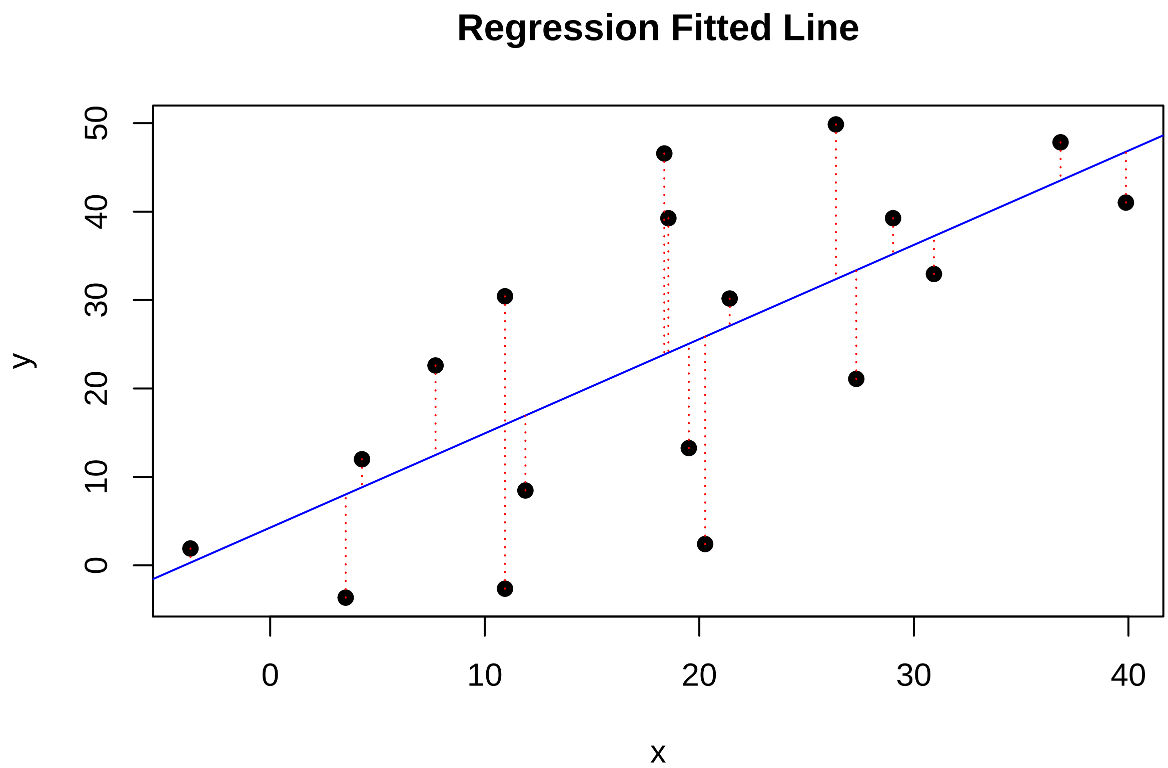
Question 1 : Answer Questions by simulating the four scenarios below

a. What regression is doing to compute model fit

```
pts <- data.frame(x = c(-3.719323,3.516220,10.942172,4.277856,11.894217,20.272214,  7.705219, 19.510578
regr <- lm(y ~ x, data=pts)
summary(regr)
```

```
##
## Call:
## lm(formula = y ~ x, data = pts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.460 -10.867   2.341   8.669  22.735
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.270      5.969   0.715  0.48462
## x              1.065      0.273   3.903  0.00126 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.39 on 16 degrees of freedom
## Multiple R-squared:  0.4878, Adjusted R-squared:  0.4558
## F-statistic: 15.24 on 1 and 16 DF,  p-value: 0.001264
```

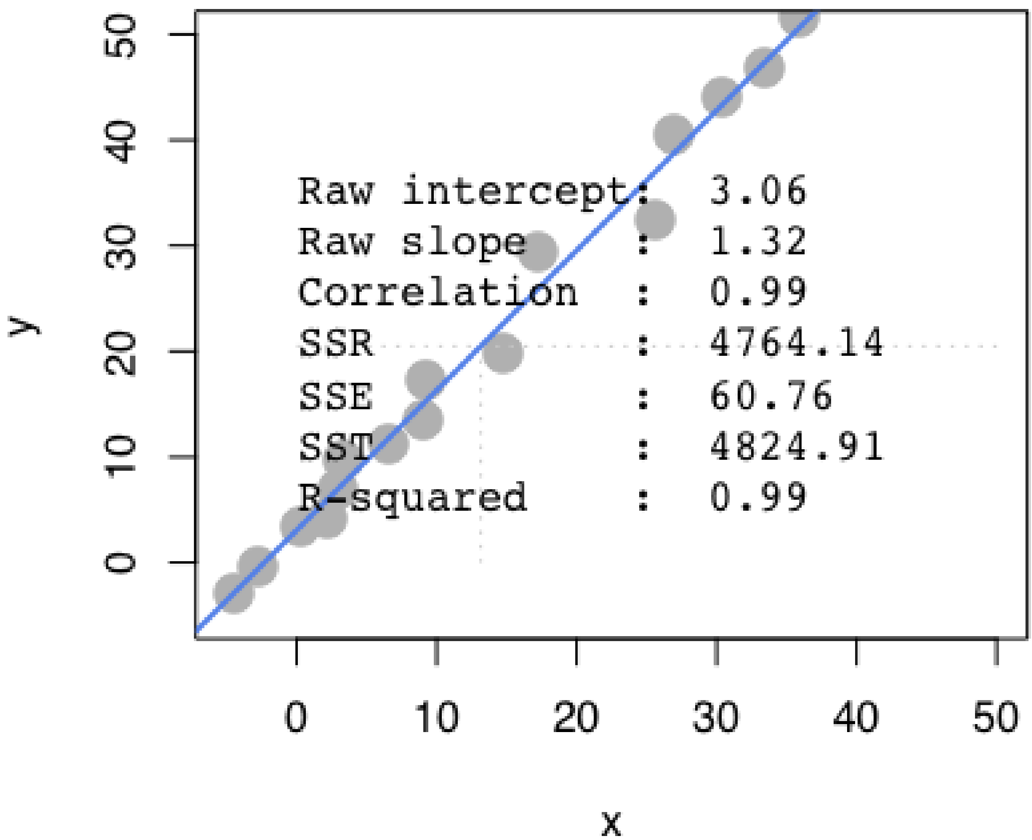
```
y_hat <- regr$fitted.values
plot(pts,main = "Regression Fitted Line",pch = 19)
abline(regr, col = 'blue')
segments(pts$x, pts$y, pts$x, y_hat, col="red", lty="dotted")
```




```
regression_fit <- function(pts){
  regr <- lm(y~x,  data = pts)
  y_hat <- regr$fitted.values
  SSE <- sum((pts$y-y_hat)^2)
  SSR <- sum((mean(pts$y)-y_hat)^2)
  SST <- sum((pts$y-mean(pts$y))^2)
  Rsq <- SSR/SST
  return(data.frame(SSE = SSE, SSR = SSR, SST= SST, Rsq = Rsq))
}
regression_fit(pts)
```

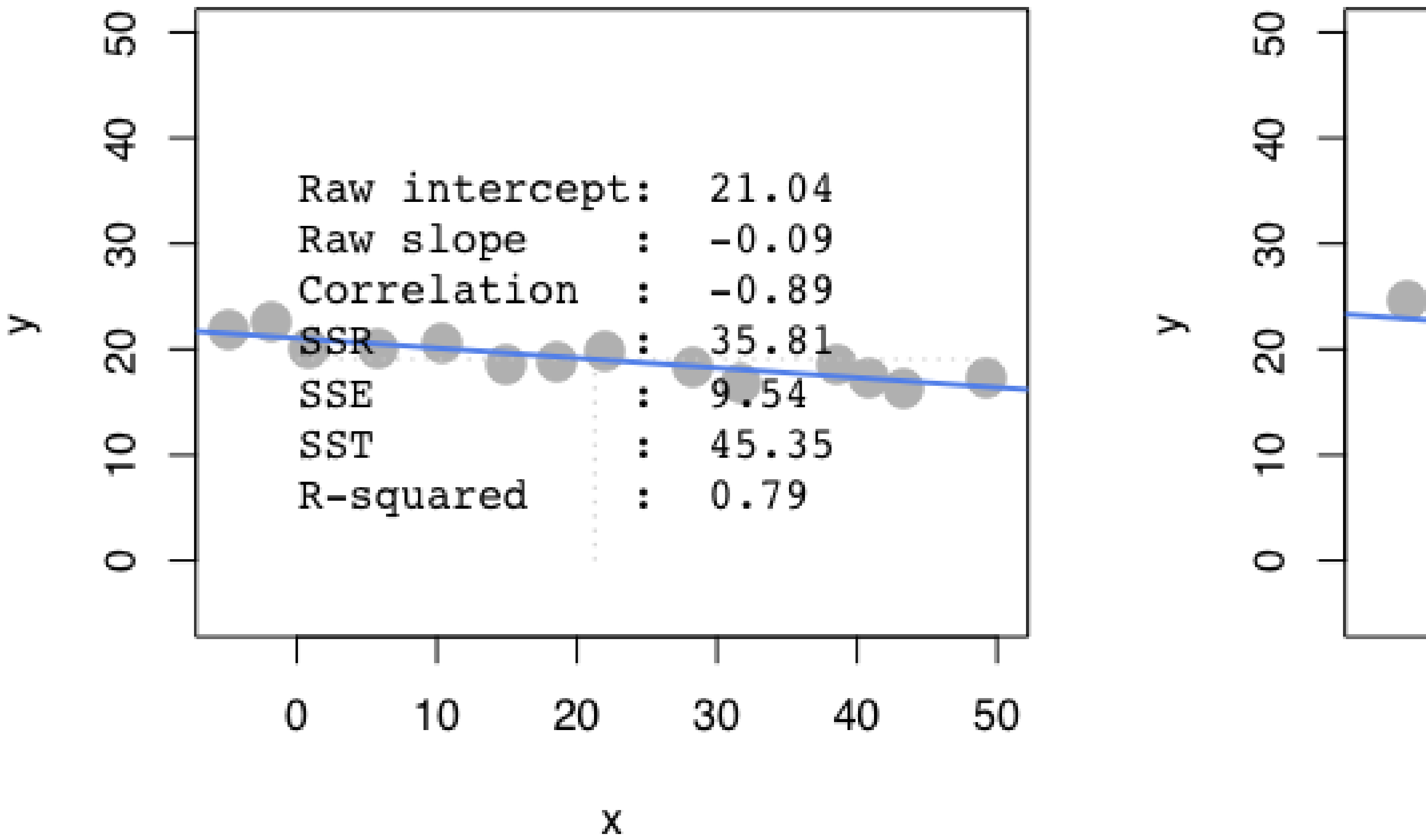
```
##        SSE      SSR      SST       Rsq
## 1 2868.667 2731.911 5600.578 0.4877909
```

b. Compare Scenario 1, 2 which to expect to have stronger R square



ANSWER: Since scenario 1 is obviously more closer to a fitted line because it's more dense, which also indicates it obtains a more stronger linear characteristic. Hence, we should expect Scenario 1 obtain higher R sqaure than Scenario 2

c. Compare Scenario 3, 4 for larger R sqaure



ANSWER: Since scenario 3 is more denser and obtain more linear characteristic, it will have higher r square compared to the more widespread distribution of scenario 4.

d. Comparing scenarios 1 and 2, which do we expect has bigger/smaller SSE, SSR, and SST?

ANSWER: SST and SSE of scenario 1 will be smaller than scenario 2, while SSR(depends on slope) for scenario 2 will be smaller than scenario 1. Because scenario 1 has a better fit than scenario 2.

e. Comparing scenarios 3 and 4, which do we expect has bigger/smaller SSE, SSR, and SST?

ANSWER: SST and SSE of scenario 3 will be smaller than scenario 4, while SSR(depends on slope) for scenario 4 will be smaller than scenario 3. Because scenario 3 has a better fit than scenario 2.

Question 2.

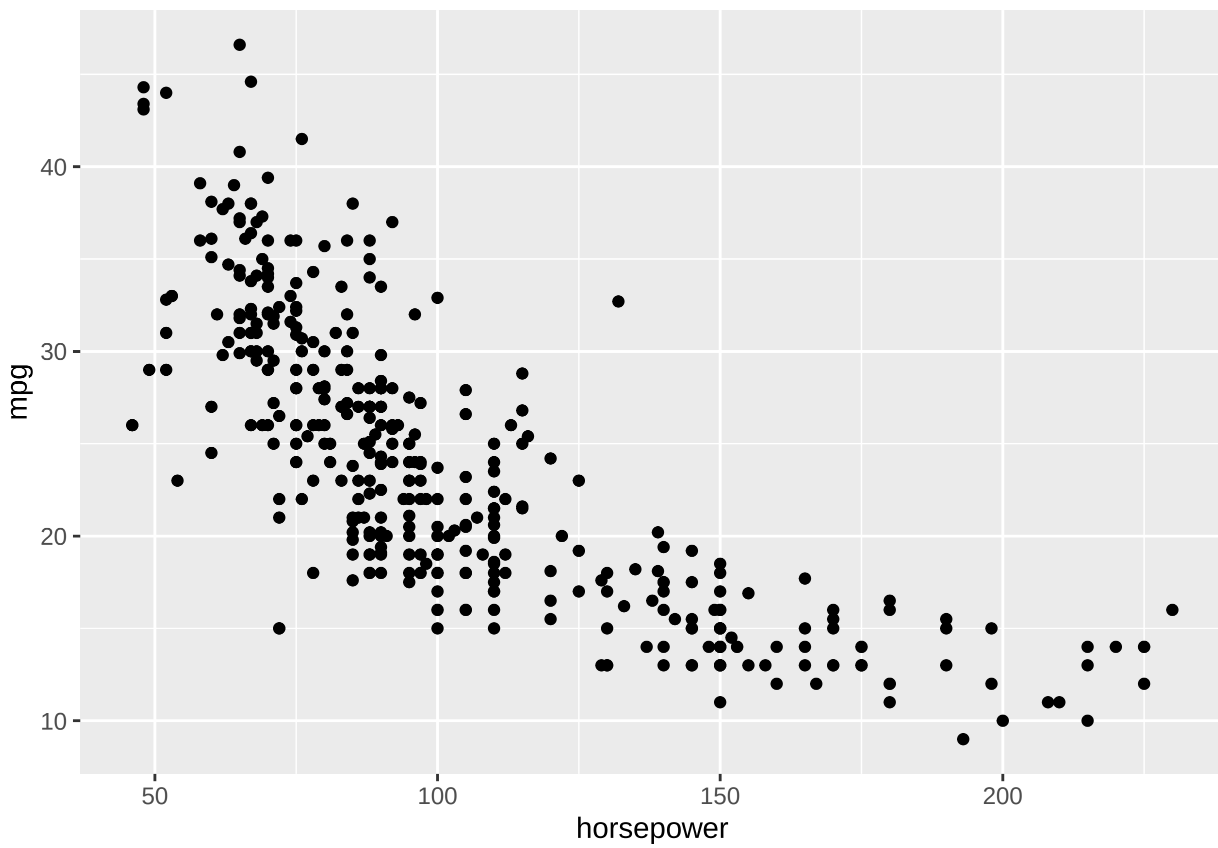
```
auto <- read.table("auto-data.txt", header=FALSE, na.strings = "?")
names(auto) <- c("mpg", "cylinders", "displacement", "horsepower", "weight",
                 "acceleration", "model_year", "origin", "car_name")
```

a. Explore Data

```
library(ggplot2)
ggplot(auto, aes(x=horsepower, y=mpg)) + geom_point()
```

i. Visualize

```
## Warning: Removed 6 rows containing missing values (geom_point).
```



This is a plot indicating that with bigger horsepower, the cars cannot achieve good fuel efficiency. Hence, there is a negative correlation here between horsepower and mpg.

```r
cor_table <- cor(auto[,1:8], use = "pairwise.complete.obs")
cor_table
```

ii. Correlation table

```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7753963   -0.8042028 -0.7784268 -0.8317409
## cylinders    -0.7753963  1.0000000    0.9507214  0.8429834  0.8960168
## displacement -0.8042028  0.9507214    1.0000000  0.8972570  0.9328241
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8317409  0.8960168    0.9328241  0.8645377  1.0000000
## acceleration  0.4202889 -0.5054195   -0.5436841 -0.6891955 -0.4174573
## model_year    0.5792671 -0.3487458   -0.3701642 -0.4163615 -0.3065643
## origin        0.5634504 -0.5625433   -0.6094094 -0.4551715 -0.5810239
##              acceleration model_year     origin
## mpg             0.4202889  0.5792671  0.5634504
## cylinders      -0.5054195 -0.3487458 -0.5625433
## displacement   -0.5436841 -0.3701642 -0.6094094
## horsepower     -0.6891955 -0.4163615 -0.4551715
## weight         -0.4174573 -0.3065643 -0.5810239
## acceleration    1.0000000  0.2881370  0.2058730
## model_year      0.2881370  1.0000000  0.1806622
## origin          0.2058730  0.1806622  1.0000000
```

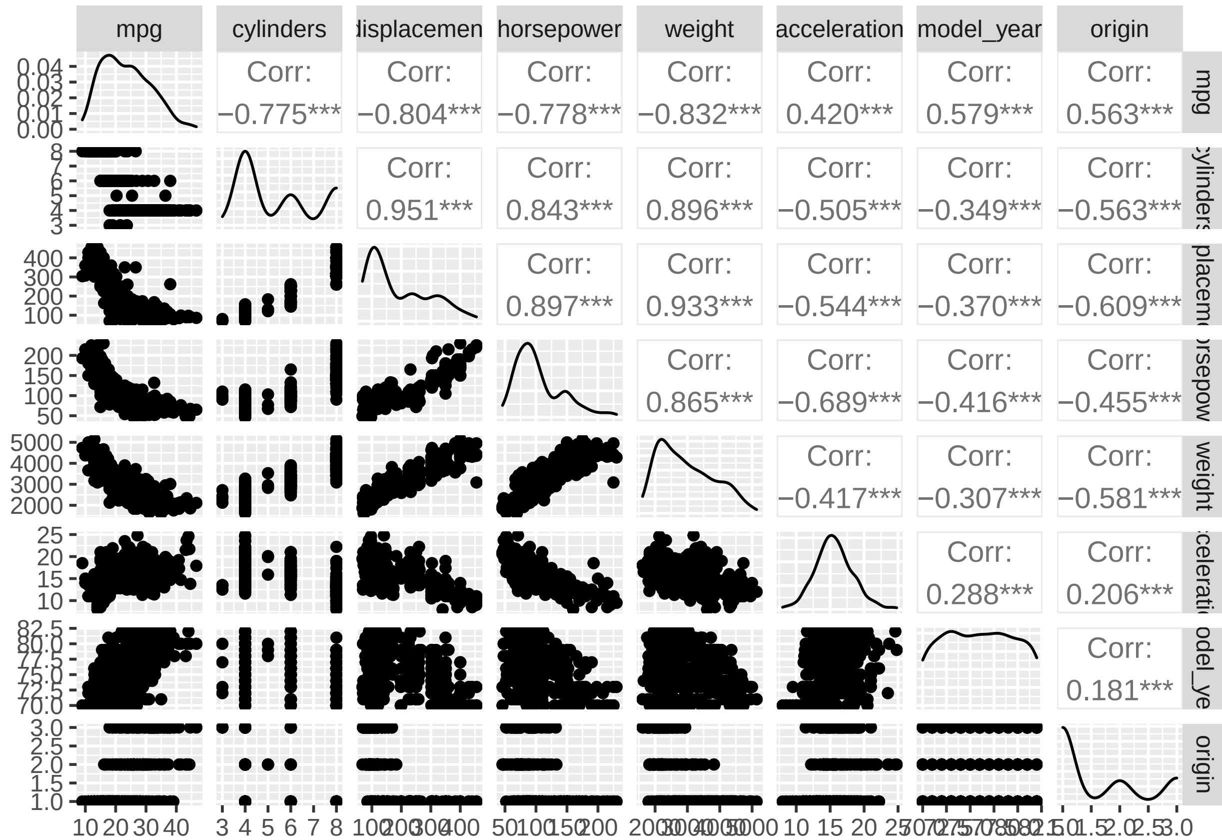
iii. Which variables seems to related to mpg **ANSWER:** According to the table above, seems like *cylinders, displacement, horsepower,weight,model_year* are negetively correlated to mpg. The other factors doesn't have a very strong positive correlation to mpg.

```r
library(GGally)
```

iv. Which relations might not be linear?

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
ggpairs(auto[1:8])
```



From the plot above, relationships between *model_year, origin, cylinders* doesn't seem to show linear characteristics.

```
library(reshape2)
diag(cor_table) <- 0
cor_melt <- melt(cor_table)
new_cor <- cor_melt[abs(cor_melt$value)>0.7,]
new_cor[!duplicated(new_cor[1:2]),]
```

v. Are there any pairs of independent variables that are highly correlated

```
##            Var1         Var2      value
## 2     cylinders          mpg -0.7753963
## 3  displacement          mpg -0.8042028
## 4    horsepower          mpg -0.7784268
## 5        weight          mpg -0.8317409
## 9           mpg    cylinders -0.7753963
## 11 displacement    cylinders  0.9507214
## 12   horsepower    cylinders  0.8429834
## 13       weight    cylinders  0.8960168
## 17          mpg displacement -0.8042028
## 18    cylinders displacement  0.9507214
## 20   horsepower displacement  0.8972570
```

```
## 21       weight displacement  0.9328241
## 25          mpg   horsepower -0.7784268
## 26    cylinders   horsepower  0.8429834
## 27 displacement   horsepower  0.8972570
## 29       weight   horsepower  0.8645377
## 33          mpg       weight -0.8317409
## 34    cylinders       weight  0.8960168
## 35 displacement       weight  0.9328241
## 36   horsepower       weight  0.8645377
```

' ### b. Create a linear regression model where mpg is dependent upon all other suitable variables

```r
regr <- lm(mpg ~ cylinders+displacement+horsepower+weight+acceleration+model_year+factor(origin)
           ,data = auto)
summary(regr)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     acceleration + model_year + factor(origin), data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0095 -2.0785 -0.0982  1.9856 13.3608
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.795e+01  4.677e+00  -3.839 0.000145 ***
## cylinders       -4.897e-01  3.212e-01  -1.524 0.128215
## displacement     2.398e-02  7.653e-03   3.133 0.001863 **
## horsepower      -1.818e-02  1.371e-02  -1.326 0.185488
## weight          -6.710e-03  6.551e-04 -10.243  < 2e-16 ***
## acceleration     7.910e-02  9.822e-02   0.805 0.421101
## model_year       7.770e-01  5.178e-02  15.005  < 2e-16 ***
## factor(origin)2  2.630e+00  5.664e-01   4.643 4.72e-06 ***
## factor(origin)3  2.853e+00  5.527e-01   5.162 3.93e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.307 on 383 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.8242, Adjusted R-squared:  0.8205
## F-statistic: 224.5 on 8 and 383 DF,  p-value: < 2.2e-16
```

i. which factors have significant on mpg at 1% significance? ANSWER: By the summary upon, the *intercept, displacement, weight, model_year, and origin have significant on mpg at 1% significance.

ii. Is it possible to determine which independent variables are most effective at increasing mpg? ANSWER: Not possible, since the variables aren't standardized, the scales for the factors are different. Hence we can not merely observe the coefficients and give out answers for this question.

###. c. Create standardized regression results

```r
sd_data <- cbind(scale(auto[1:7]),auto$origin)
colnames(sd_data) <- colnames(auto[1:8])
sd_df <- as.data.frame(sd_data)
new_regr <- lm(mpg~ cylinders+displacement+horsepower+weight+acceleration+model_year+factor(origin),data
summary(new_regr)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     acceleration + model_year + factor(origin), data = sd_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.15270 -0.26593 -0.01257  0.25404  1.70942
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.13323    0.03174  -4.198 3.35e-05 ***
## cylinders       -0.10658    0.06991  -1.524  0.12821
## displacement     0.31989    0.10210   3.133  0.00186 **
## horsepower      -0.08955    0.06751  -1.326  0.18549
## weight          -0.72705    0.07098 -10.243  < 2e-16 ***
## acceleration     0.02791    0.03465   0.805  0.42110
## model_year       0.36760    0.02450  15.005  < 2e-16 ***
## factor(origin)2  0.33649    0.07247   4.643 4.72e-06 ***
## factor(origin)3  0.36505    0.07072   5.162 3.93e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.423 on 383 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.8242, Adjusted R-squared:  0.8205
## F-statistic: 224.5 on 8 and 383 DF,  p-value: < 2.2e-16
```

i. Are these figures easier to interpret? Yes, it will be easier to interpret since we can see that weight is most effective at increasing mpg. Which is quite reasonable.

```r
fit1 <- lm(mpg~cylinders,data = sd_df)
fit2 <- lm(mpg~displacement,data = sd_df)
fit3 <- lm(mpg~horsepower,data = sd_df)
fit4 <- lm(mpg~weight,data = sd_df)
fit5 <- lm(mpg~acceleration,data = sd_df)
fit6 <- lm(mpg~model_year,data = sd_df)
fit7 <- lm(mpg~origin,data = sd_df)
signifi<- function(fit){
  return (signif(summary(fit)$coef[2,4],2))
}
paste('cylinders:',signifi(fit1))
```

ii. Regress mpg over each nonsignificant independent variable. Which one will become significant over mpg?

```
## [1] "cylinders: 4.5e-81"
```

```
paste('displacement:',signifi(fit2))
```

```
## [1] "displacement: 1.7e-91"
```

```
paste('horsepower:',signifi(fit3))
```

```
## [1] "horsepower: 7e-81"
```

```
paste('weight:',signifi(fit4))
```

```
## [1] "weight: 3e-103"
```

```
paste('accerleration:',signifi(fit5))
```

```
## [1] "accerleration: 1.8e-18"
```

```
paste('mdoel_year:',signifi(fit6))
```

```
## [1] "mdoel_year: 4.8e-37"
```

```
paste('origin:',signifi(fit7))
```

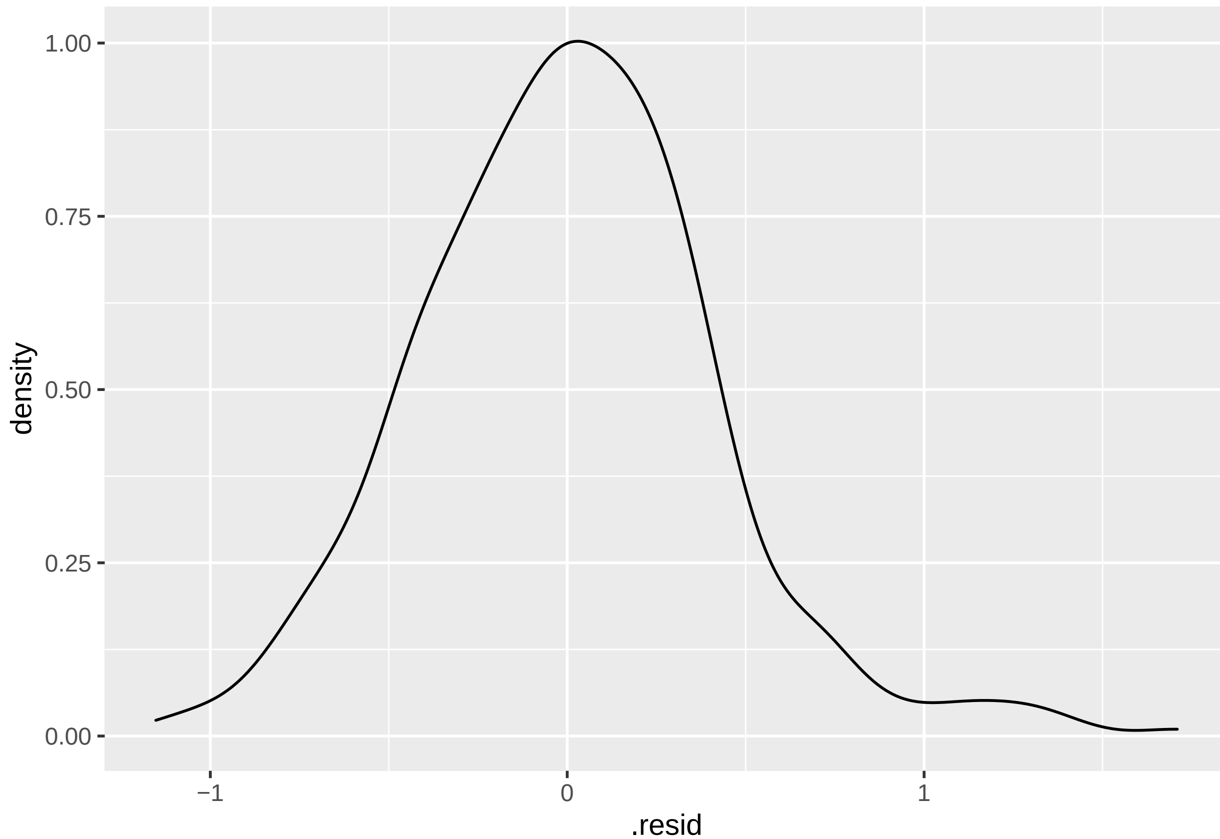
```
## [1] "origin: 1e-34"
```

ANSWER: After fitted with every independent variable, they are all significant since the p-values are very low.

```


library(ggplot2)
regr_plt <- fortify(new_regr)
ggplot(new_regr,aes(.resid))+ geom_density()
```

iii. Plot the density of the residuals, are they normally distributed and centered around zero?



ANSWER: It's near a normal distribution with mean near 0. Can be verified by QQ plot.