Project 1 Part B

```r
getwd()
```

```
## [1] "C:/Users/Darian/Documents/Project1Files"
```

```r
wdir <- "C:\\Users\\Darian\\Documents\\Project1Files"
setwd(wdir)
getwd()
```

```
## [1] "C:/Users/Darian/Documents/Project1Files"
```

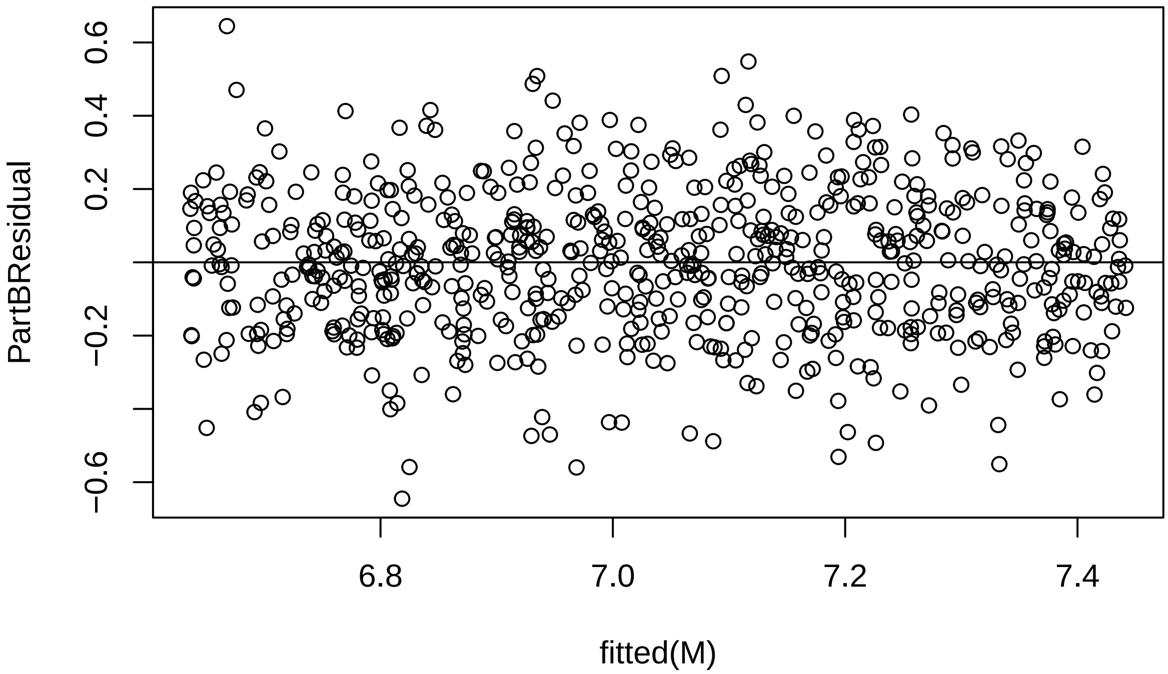
```r
PartB <- read.csv('Project1_PartB.csv', header = TRUE)
str(PartB)
```

```
## 'data.frame':    657 obs. of  3 variables:
##  $ ID: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ IV: num  3.98 4.46 3.07 4.67 2.05 ...
##  $ DV: num  7.06 7.31 7.01 7.46 7.21 ...
```

```r
View(PartB)
#So we have no missing values in Part B! We have 657 complete observations!
M <- lm(DV ~ IV, data=PartB)
summary(M)
```

```
##
## Call:
## lm(formula = DV ~ IV, data = PartB)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.64499 -0.13137 -0.00267  0.13210  0.64453
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.433531   0.022053  291.73   <2e-16 ***
## IV          0.201973   0.006909   29.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1973 on 655 degrees of freedom
## Multiple R-squared:  0.5661, Adjusted R-squared:  0.5655
## F-statistic: 854.7 on 1 and 655 DF,  p-value: < 2.2e-16
```

```r
#r^2 is approximately .5661 so 56.61% of the variance in y can be explained by
#changes in x, the other 43.49% is presumably due to random variability or other
# unknown variables
PartBResidual <- resid(M)
plot(fitted(M), PartBResidual)
abline(0,0)
```



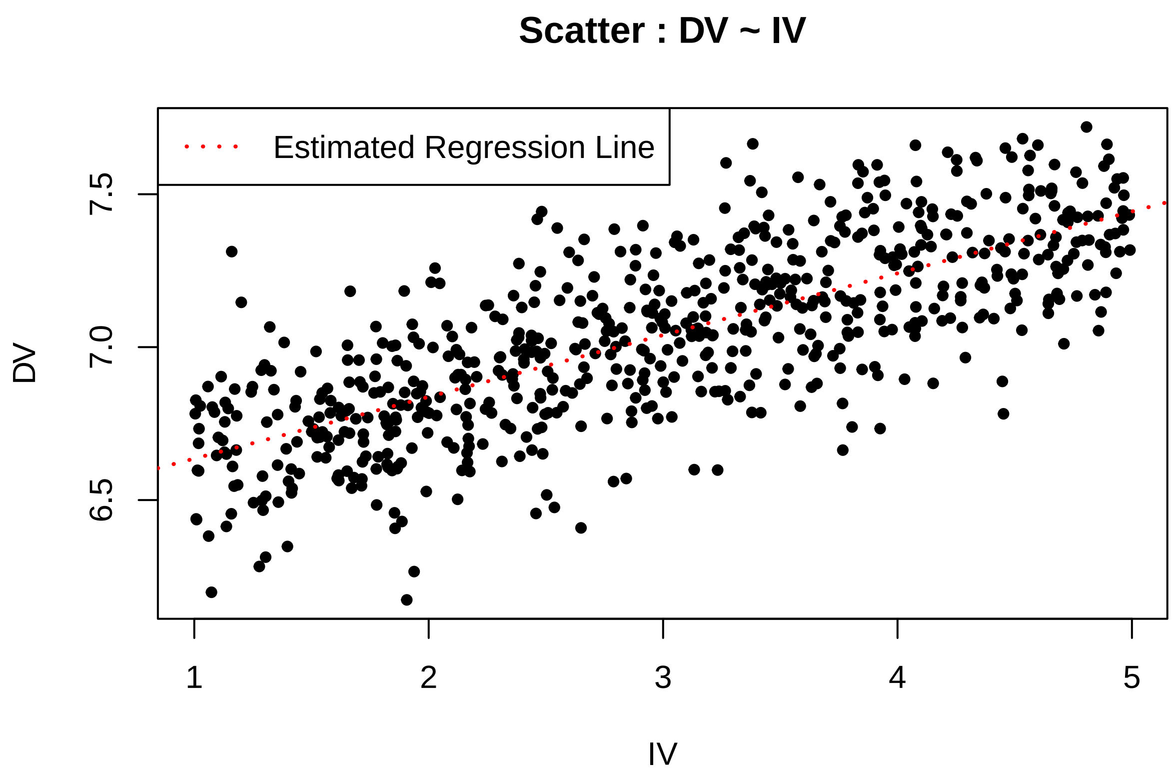

```r
# M is the object that represents the linear regression model. M for Model.
#residual plot seems to indicate linearity and no lack of fit so no
#transformation seems necessary
library(knitr)
kable(anova(M), caption='ANOVA Table')
```

Table 1: ANOVA Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
IV	1	33.28580	33.2858037	854.6925	0
Residuals	655	25.50882	0.0389448	NA	NA

```r
#F-value associated with Regression(x) is 854.6925 with df1=1, df2=655
# and p-value = 0
#So reject null hypothesis that slope is 0. So x has some value in predicting y.
plot(PartB$DV ~ PartB$IV, main='Scatter : DV ~ IV', xlab='IV', ylab='DV', pch=20)
```

```r
abline(M, col='red', lty=3, lwd=2)
legend('topleft', legend='Estimated Regression Line', lty=3, lwd=2, col='red')
```




```r
groups <- cut(PartB$IV,breaks=c(-Inf,seq(min(PartB$IV)+0.3, max(PartB$IV)-0.3,by=0.3),Inf))
table(groups)
```

```
## groups
## (-Inf,1.3]  (1.3,1.6]  (1.6,1.9]  (1.9,2.2]  (2.2,2.5]  (2.5,2.8]  (2.8,3.1]
##         42         39         64         51         55         45         51
##  (3.1,3.4]  (3.4,3.7]    (3.7,4]    (4,4.3]  (4.3,4.6] (4.6, Inf]
##         57         52         52         47         41         61
```

```r
x <- ave(PartB$IV, groups)
data_bin <- data.frame(x=x, y=PartB$DV)
library(remotes)
library(alr3)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```r
fit_b <- lm(y ~ x, data = data_bin)
pureErrorAnova(fit_b)
```

```
## Analysis of Variance Table
##
## Response: y
##                Df Sum Sq Mean Sq  F value Pr(>F)
## x               1 33.014  33.014 833.7511 <2e-16 ***
## Residuals     655 25.781   0.039
##  Lack of fit   11  0.281   0.026   0.6441 0.7913
##  Pure Error   644 25.500   0.040
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#After binning and grouping data the adjusted F-value for Regression(x) is
#833.7511 with df1=1 and df2=655, and p = 0, it's still huge and p is still 0
#so there is no real change or consequences from this adjusted value
#The LOF F value is .6441 with df1=11 and df2=644 with a p-value of .7913
#showing that there is no significant lack of fit
#99% C.I. for slope and intercept
confint(M, level = 0.99)
```

```
##                 0.5 %    99.5 %
## (Intercept) 6.3765603 6.4905017
## IV          0.1841255 0.2198201
```