# Project 1 Part A

```
getwd()
```

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

```
wdir <- "C:\\Users\\Darian\\Documents\\Project1Files"
setwd(wdir)
PartA_IV <- read.csv('Project1_IV_Values.csv', header = TRUE)
#Get IV values for PartA
PartA_DV <- read.csv('Project1_DV_Values.csv', header = TRUE)
#Get DV Values for PartA
PartA <- merge(PartA_IV, PartA_DV, by = 'ID')
#Merge the IV and DV values by our identifier, the ID of the observations.
str(PartA)
```

```
## ’data.frame’:    438 obs. of  3 variables:
##  $ ID: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ IV: num  16.5 14 13.3 20 16.7 ...
##  $ DV: num  184 151 146 218 162 ...
```

```
View(PartA)
#View the merged data using srt() and View()
any(is.na(PartA[,2]) == TRUE)
```

```
## [1] TRUE
```

```
any(is.nan(PartA[,2]) == TRUE)
```

```
## [1] FALSE
```

```
any(is.null(PartA[,2]) == TRUE)
```

```
## [1] FALSE
```

```
# From the above, any(is.na(PartA[,2]) == TRUE) we can see that we have missing
# values that are labeled as na.
any(is.na(PartA[,3]) == TRUE)
```

```
## [1] TRUE
```

```
any(is.nan(PartA[,3]) == TRUE)
```

```
## [1] FALSE
```

```
any(is.null(PartA[,3]) == TRUE)
```

```
## [1] FALSE
```

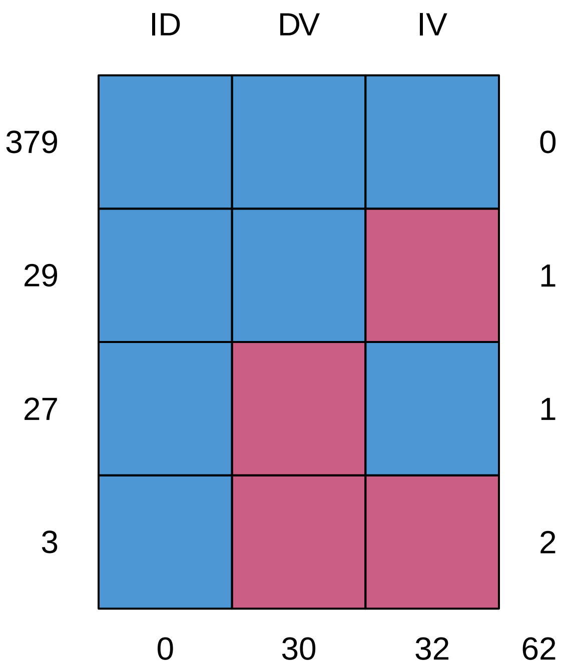
```
# From the above, any(is.na(PartA[,3]) == TRUE) we can see that we have missing
# values that are labeled as Na.
PartA_incomplete <- PartA
library(mice)
```

```
##
## Attaching package: 'mice'

## The following object is masked from 'package:stats':
##
##     filter

## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

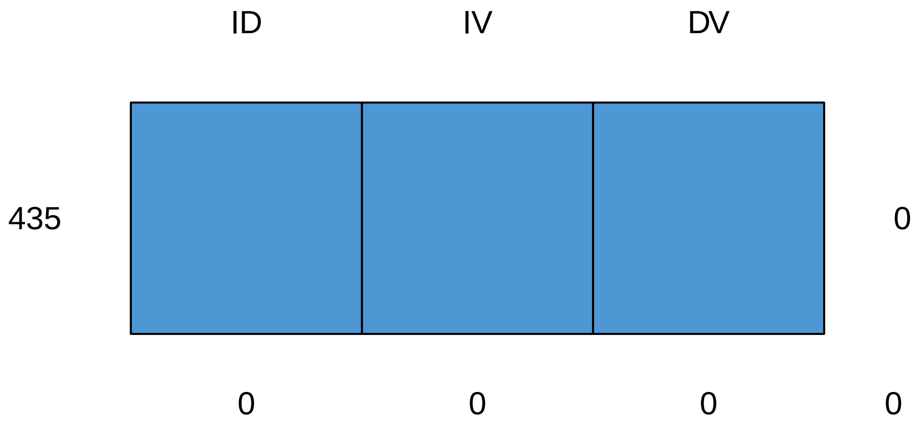
```
md.pattern(PartA_incomplete)
```



```
##     ID DV IV
## 379  1  1  1  0
## 29   1  1  0  1
## 27   1  0  1  1
## 3    1  0  0  2
##      0 30 32 62
```

```r
# The above tells us that there are 379 complete observations.
# The IV is missing in 32 observations (29 have the DV, 3 have neither).
# The DV is missing in 30 observations (27 have the IV, 3 have neither).
# Both the IV and DV are missing in 3 observations (as stated earlier).
PartA_imp <- PartA[!is.na(PartA$IV)==TRUE|!is.na(PartA$DV)==TRUE,]
#Get rid of observations that have both variables missing (3 as stated earlier)
#So we have 435 observations (379 complete, 29 with the DV, 27 with the IV)
imp <- mice(PartA_imp, method = "norm.boot", printFlag = FALSE)
#linear regression using bootstrap method is used here to approximate missing
#IV and DV values
PartA_complete <- complete(imp)
View(PartA_complete)
md.pattern(PartA_complete)
```

```
##  /\     /\
## {  ‘---’  }
## {  O   O  }
## ==>  V <==  No need for mice. This data set is completely observed.
##  \  \|/  /
##   ‘-----’
```



```
##     ID IV DV  
## 435  1  1  1 0
##      0  0  0 0
```

```
#The above tells us that after imputation, the data set is complete with
#435 Observations.
#Recall that there was no data (no DV and IV) for 3 observations!
M <- lm(DV ~ IV, data=PartA_complete)
#Make a linear regression model using the complete data of 435 observations
#after imputation and save it to the object 'M' (M for Model)
summary(M)
```

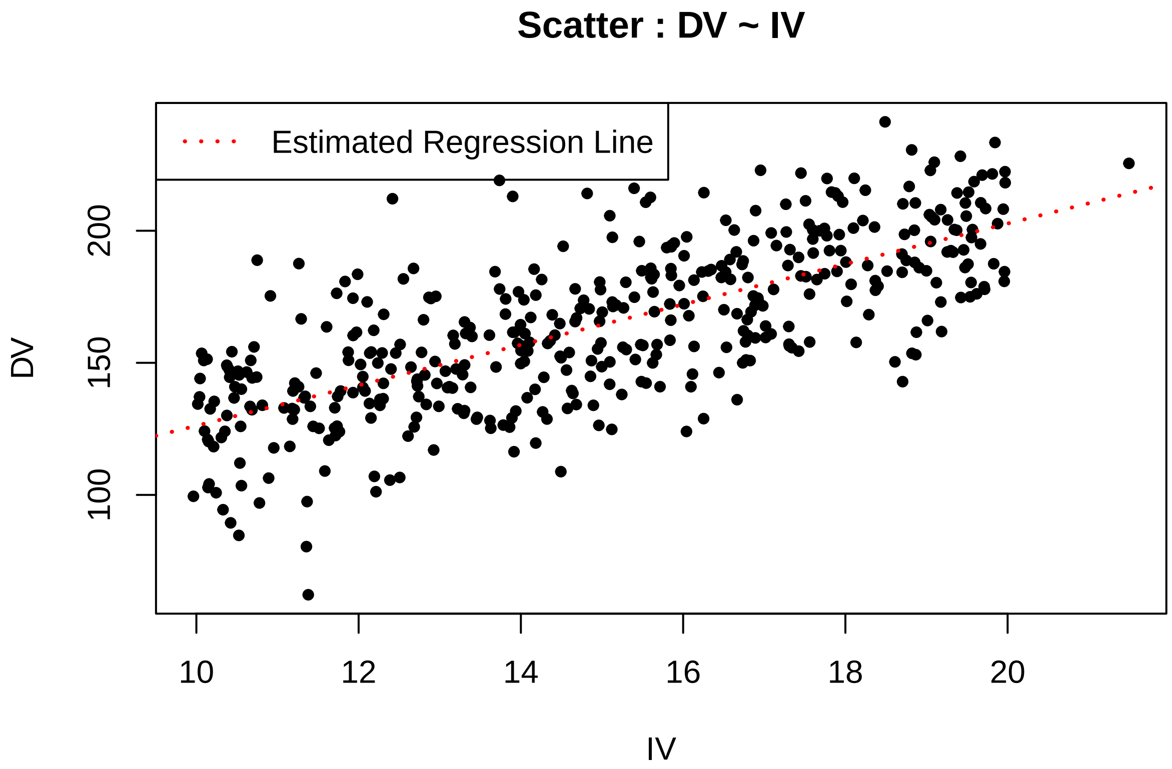
```
## 
## Call:
## lm(formula = DV ~ IV, data = PartA_complete)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -74.534 -14.360   0.462  13.428  67.435 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  49.6392     5.3881   9.213   <2e-16 ***
## IV            7.6546     0.3536  21.650   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.39 on 433 degrees of freedom
## Multiple R-squared:  0.5198, Adjusted R-squared:  0.5187
## F-statistic: 468.7 on 1 and 433 DF,  p-value: < 2.2e-16
```

```
# r^2 is .5195 on my run (Note: Results/values might be slightly different
# because bootstrap will give different approximation values each run,
# but results will be similar at least) so that's the proportion of variation
# so 51.95% of the variance in y can be explained by the changes in x.
# the other 48.05% is presumably due to random variability or unknown variables
library(knitr)
kable(anova(M), caption='ANOVA Table')
```

Table 1: ANOVA Table

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---:|---:|---:|---:|---:|
| IV | 1 | 214363.8 | 214363.7660 | 468.7267 | 0 |
| Residuals | 433 | 198024.8 | 457.3321 | NA | NA |

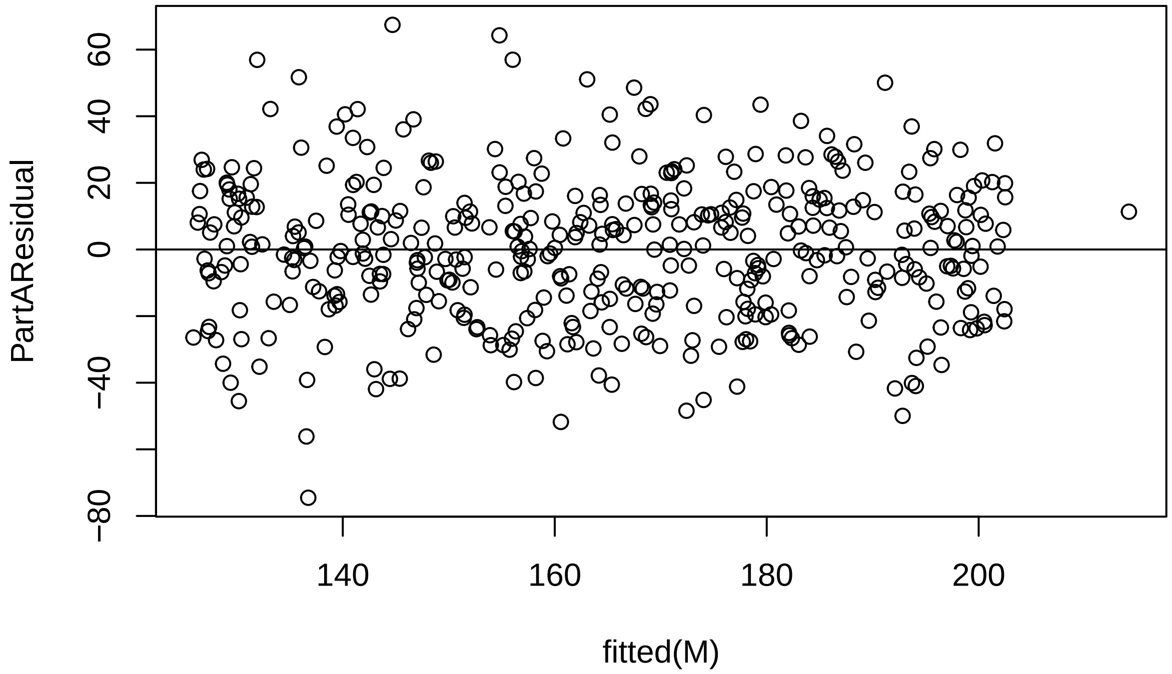
```
#The F-value associated with Regression(x) is extremely large! With p=0! so
#reject the null hypothesis that the slope of the model is zero.
plot(PartA_complete$DV ~ PartA_complete$IV, main='Scatter : DV ~ IV', xlab='IV',
     ylab='DV', pch=20)
abline(M, col='red', lty=3, lwd=2)
legend('topleft', legend='Estimated Regression Line', lty=3, lwd=2, col='red')
```



```
#99% CI of Slope and Intercept
confint(M, level = 0.99)
```

```
##                  0.5 %    99.5 %
## (Intercept) 35.698827 63.579519
## IV           6.739879  8.569366
```

```
#analyzing the residual plot for lack of fit and linearity, there appears
#to be linearity and no lack of fit from the random scatter of values with no
#pattern!
PartAResidual <- resid(M)
plot(fitted(M), PartAResidual)
abline(0,0)
```



```
#Note: M is the object that represents the linear regression model. M for Model.
```