Short HW 1

Bhavika Sewpal - 300089940

Part 1

1)

H_0 = IQ has no effect on diligence
H_1 = IQ has an effect on diligence

2)

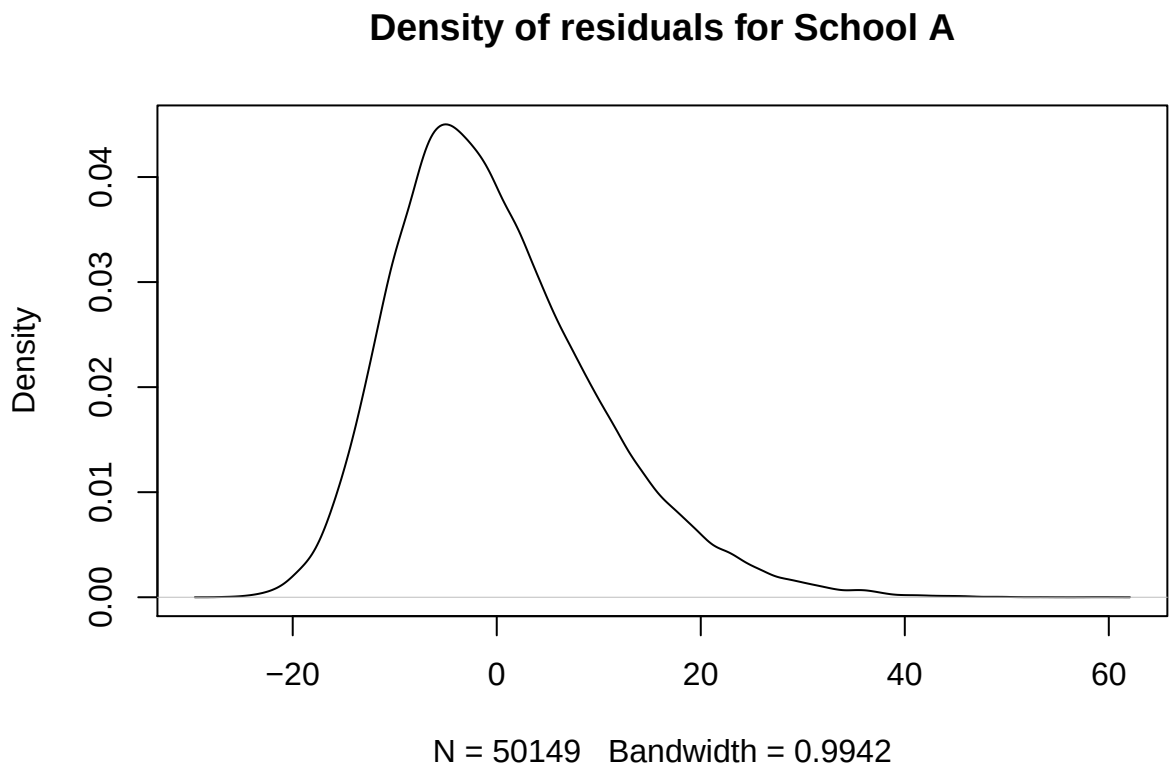
```
schoolA <- read.csv("SchoolA.csv")
y <- schoolA$V2
x <- schoolA$V1
fit <- lm(y~x)
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.613  -7.104  -1.450   5.790  59.090
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 173.427803   0.415271   417.6   <2e-16 ***
## x            -0.554576   0.003701  -149.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.938 on 50147 degrees of freedom
## Multiple R-squared:  0.3092, Adjusted R-squared:  0.3092
## F-statistic: 2.245e+04 on 1 and 50147 DF,  p-value: < 2.2e-16
```

Since the p-value is less than 0.05, we reject the null hypothesis. (i.e. we conclude that IQ has an effect on diligence)

3)

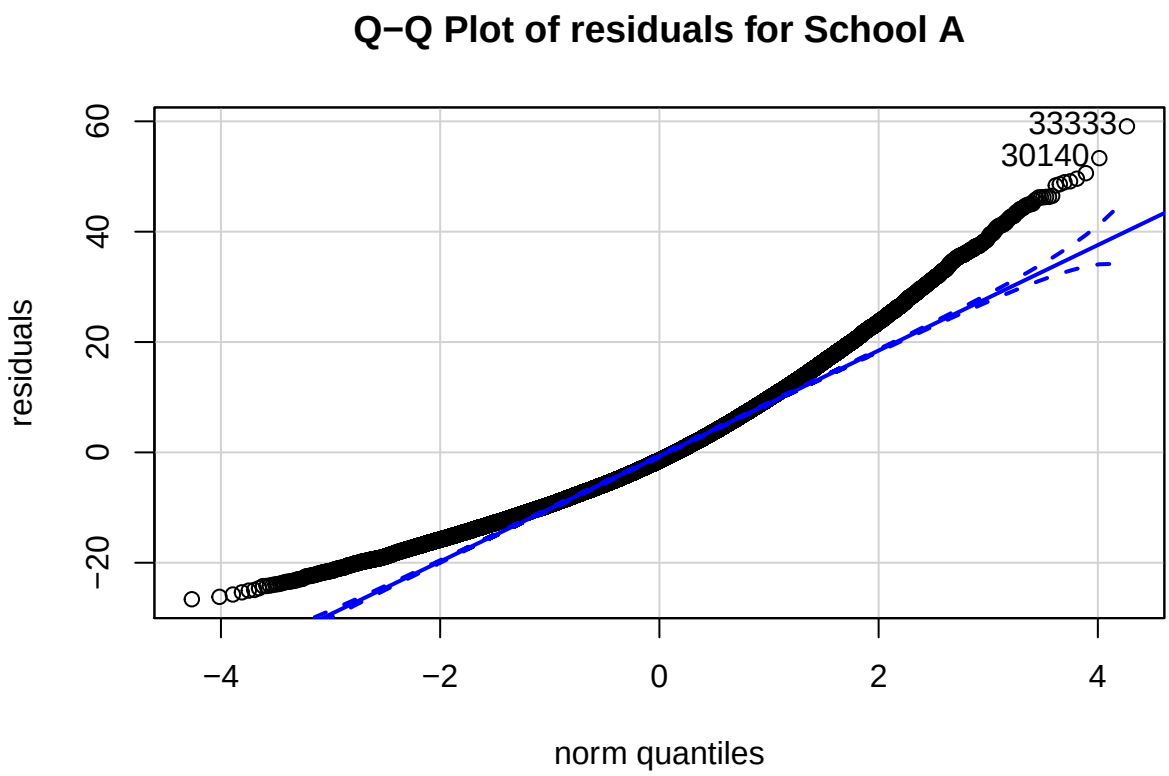
```
residuals = fit$residuals
plot(density(residuals) , main = "Density of residuals for School A")
```




```
library("car")
```

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

```
qqPlot(residuals , main = "Q-Q Plot of residuals for School A")
```




```
## [1] 33333 30140
```

It seems that the residuals are skewed to the left according to the density plot.
On the qqplot, the residuals don't fall on a straight line.
All these observations indicate that the residuals are not Gaussian

Part 2

1)

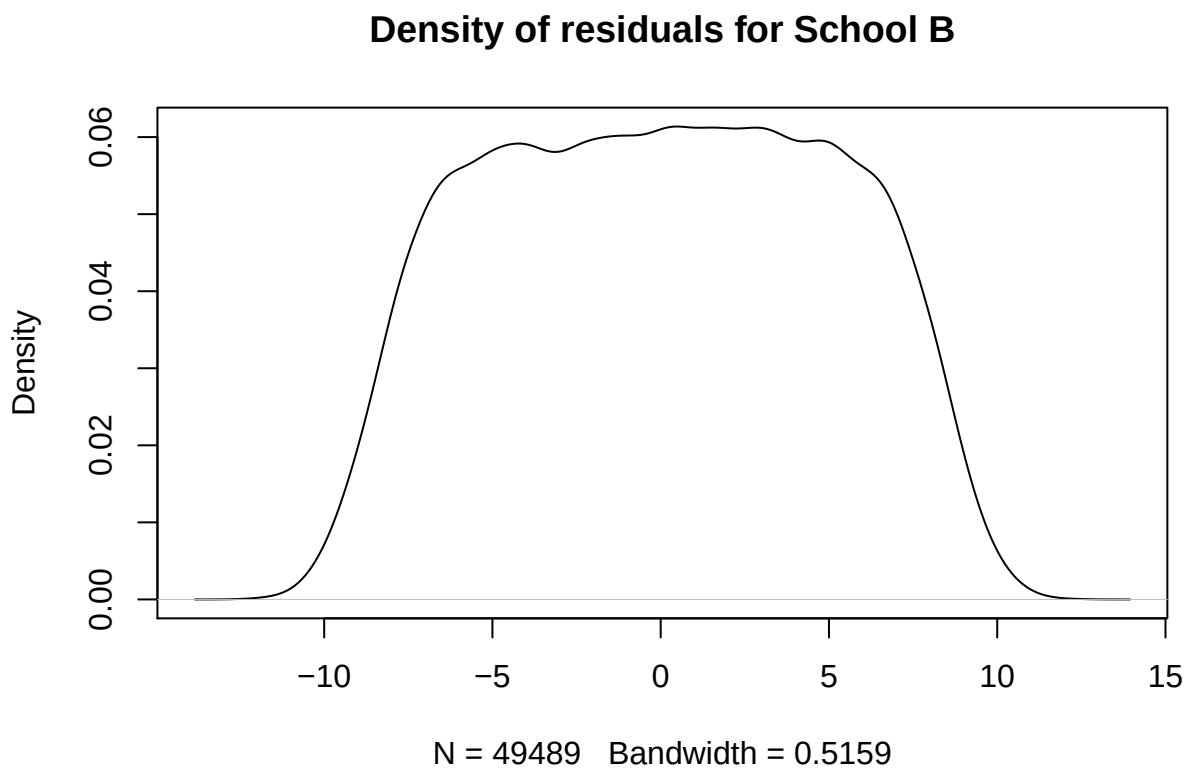
```
schoolB <- read.csv("SchoolB.csv")
y2 <- schoolB$V2
x2 <- schoolB$V1
fit2 <- lm(y2 ~ x2)
summary(fit2)
```

```
##
## Call:
## lm(formula = y2 ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.2952  -4.1631   0.0484   4.1402  12.3960
##
## Coefficients:
```

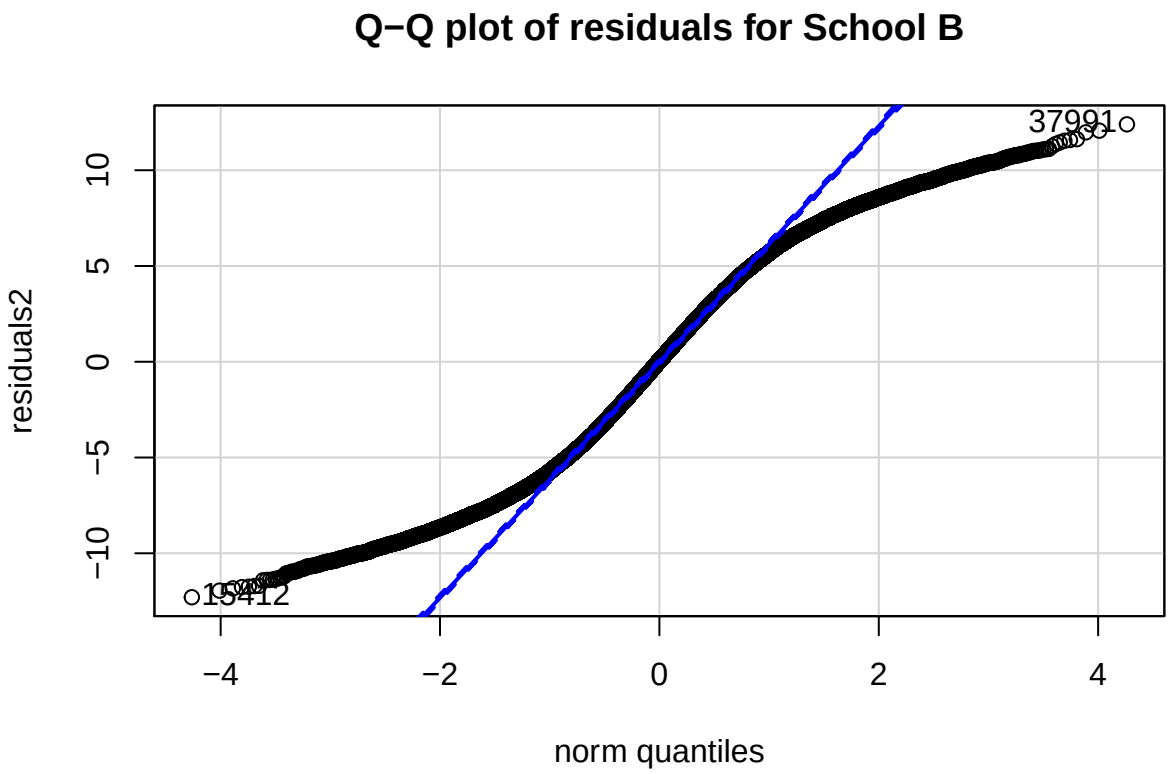
```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 189.056290   0.206066   917.5   <2e-16 ***
## x2           -0.890102   0.002047  -434.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.98 on 49487 degrees of freedom
## Multiple R-squared:  0.7925, Adjusted R-squared:  0.7925
## F-statistic: 1.89e+05 on 1 and 49487 DF,  p-value: < 2.2e-16
```

Since the p-value is less than 0.05, we reject the null hypothesis. (i.e. we conclude that IQ has an effect on diligence)

```
residuals2 = fit2$residuals
plot(density(residuals2), main = "Density of residuals for School B")
```




```
library("car")
qqPlot(residuals2, main = "Q-Q plot of residuals for School B")
```




```
## [1] 37991 15412
```

The density plot for the residuals is not bell-shaped. Instead, it is very wide.
The residuals don't fall on a straight line on the qqplot.
From these observations, we can conclude that the residuals are not Gaussian.

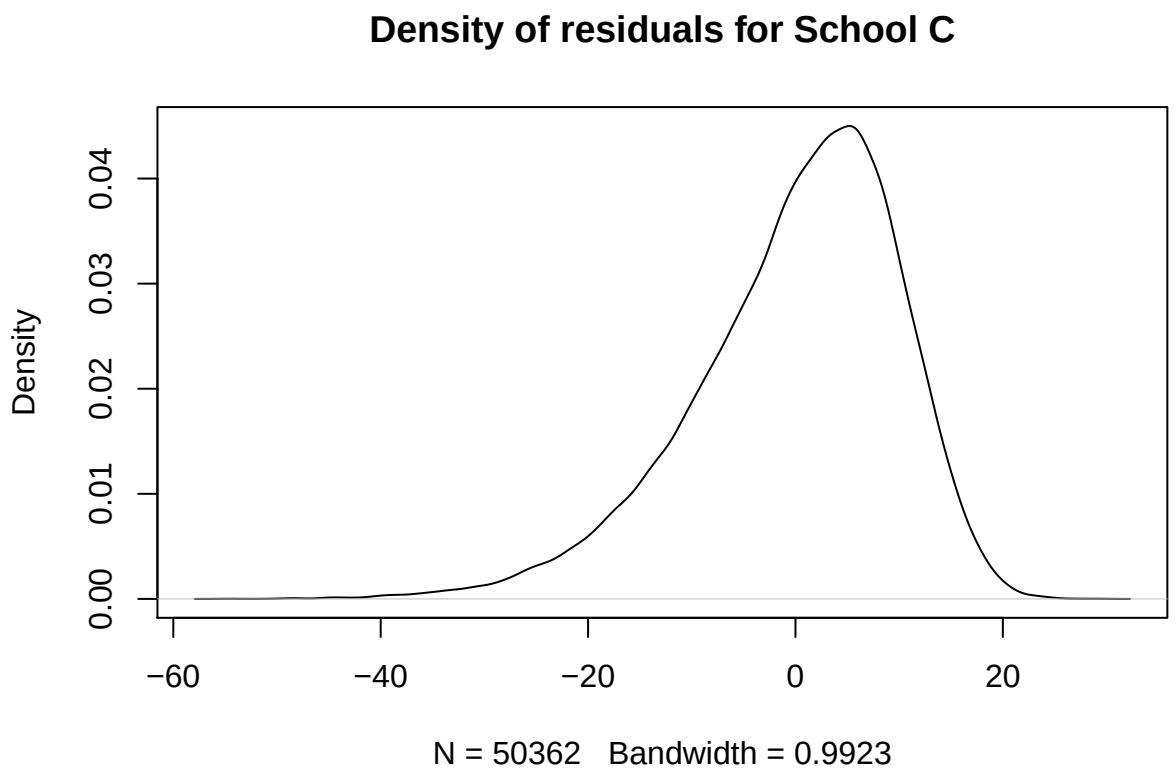
```
schoolC <- read.csv("SchoolC.csv")
y3 <- schoolC$V2
x3 <- schoolC$V1
fit3 <- lm(y3 ~ x3)
summary(fit3)
```

```
##
## Call:
## lm(formula = y3 ~ x3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.945  -5.753   1.483   7.128  29.280
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 138.226799   0.330068   418.8   <2e-16 ***
## x3           -0.561711   0.003697  -152.0   <2e-16 ***
## ---
```

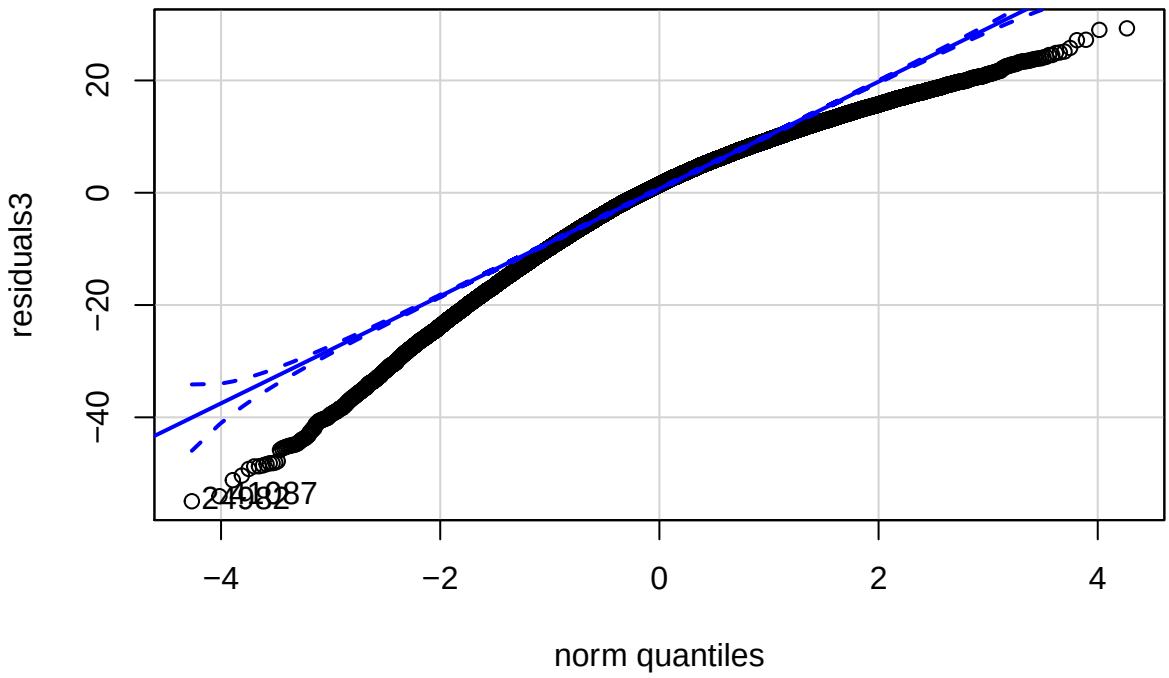
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.947 on 50360 degrees of freedom
## Multiple R-squared:  0.3144, Adjusted R-squared:  0.3143
## F-statistic: 2.309e+04 on 1 and 50360 DF,  p-value: < 2.2e-16
```

Since the p-value is less than 0.05, we reject the null hypothesis. (i.e. we conclude that IQ has an effect on diligence)

```
residuals3 = fit3$residuals
plot(density(residuals3), main = "Density of residuals for School C")
```




```
library("car")
qqPlot(residuals3)
```




```
## [1] 24982 41087
```

The density plot for the residuals is skewed to the right.
The residuals don't fall on a straight line on the qqplot.
From these observations, we can conclude that the residuals are not Gaussian.

2)

Since the p-values for all the the three studies is smaller than 0.05, we can reject the null hypothesis for all of them. Hence, it seems that IQ has an effect on diligence for all three studies.

Part 3

1)

A 100 (1−) percent confidence interval on β_1 is obtained as follows:

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} se(\hat{\beta}_1)$$

where the standard error of $\hat{\beta}_1$ is given by:

$$se(\hat{\beta}_1) = \frac{\sqrt{(y_i - \hat{y}_i)^2}}{\sqrt{(x_i - \bar{x})^2}(n-2)}$$

When we combine independent datasets, the number of observation,n, increases.
$\frac{\sqrt{(y_i-\hat{y}_i)^2}}{n-2}$ converges to σ^2 as n gets bigger.

The denominator $\sqrt{(x_i - \bar{x})^2}$ gets bigger as n increases.
Hence, the standard error of $\hat{\beta}_1$ decreases as n gets bigger, leading to smaller confidence intervals.

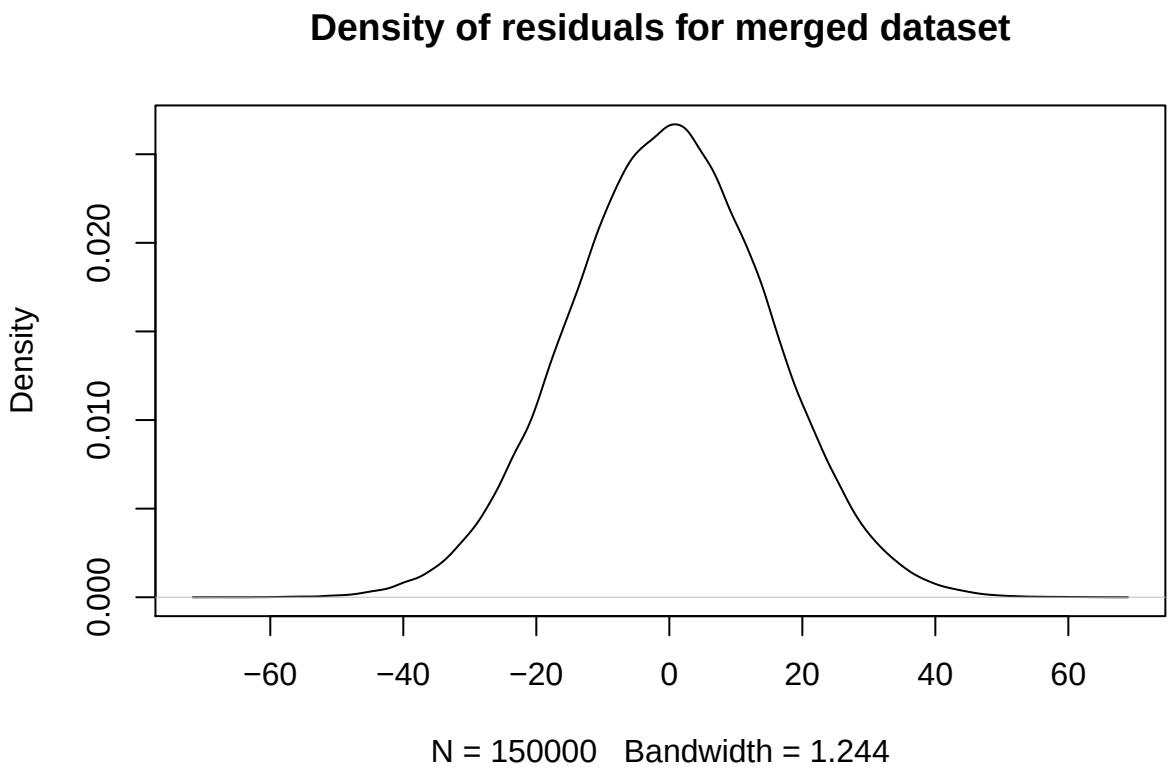
2)

```
merged_data <- rbind(schoolA, schoolB, schoolC)
y4 <- merged_data$V2
x4 <- merged_data$V1
fit4 <- lm(y4~x4)
summary(fit4)
```

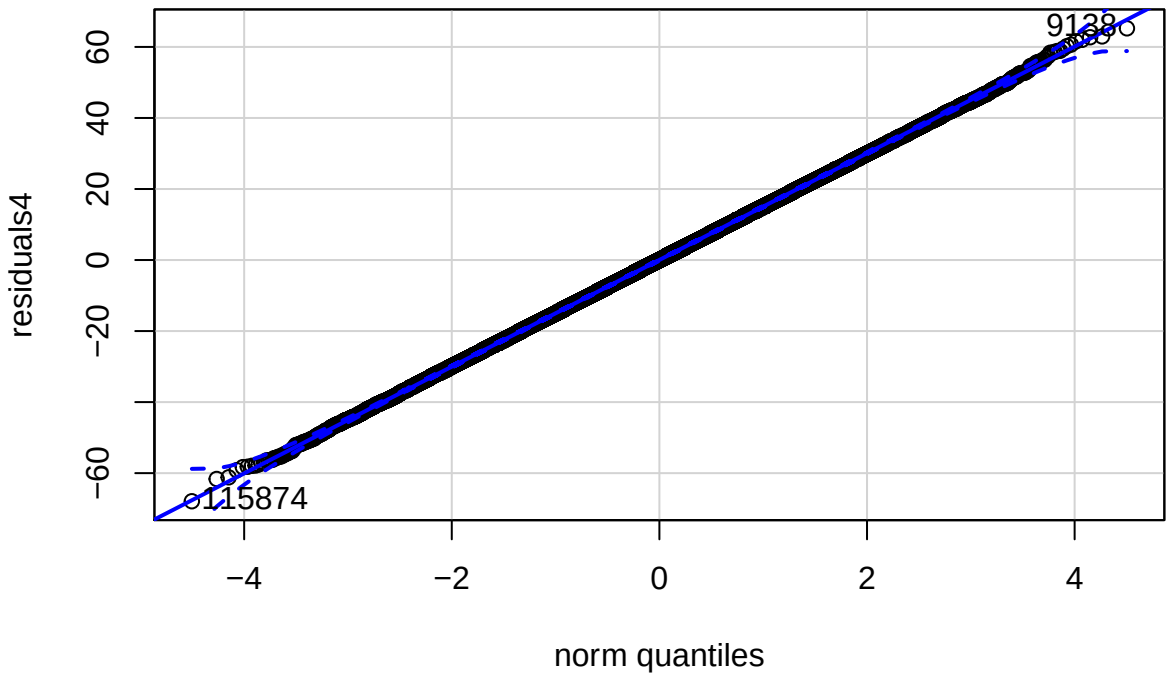
```
##
## Call:
## lm(formula = y4 ~ x4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.912 -10.112   0.035  10.153  65.230
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.995e+01  2.609e-01 383.033   <2e-16 ***
## x4          6.607e-04  2.580e-03   0.256    0.798
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.99 on 149998 degrees of freedom
## Multiple R-squared:  4.372e-07,  Adjusted R-squared:  -6.23e-06
## F-statistic: 0.06558 on 1 and 149998 DF,  p-value: 0.7979
```

For the merged dataset, the p-value is greater than 0.05. The null hypothesis is not rejected. i.e., we conclude that IQ has no effect on diligence.

```
residuals4 = fit4$residuals
plot(density(residuals4), main = "Density of residuals for merged dataset")
```




```
library("car")
qqPlot(residuals4)
```



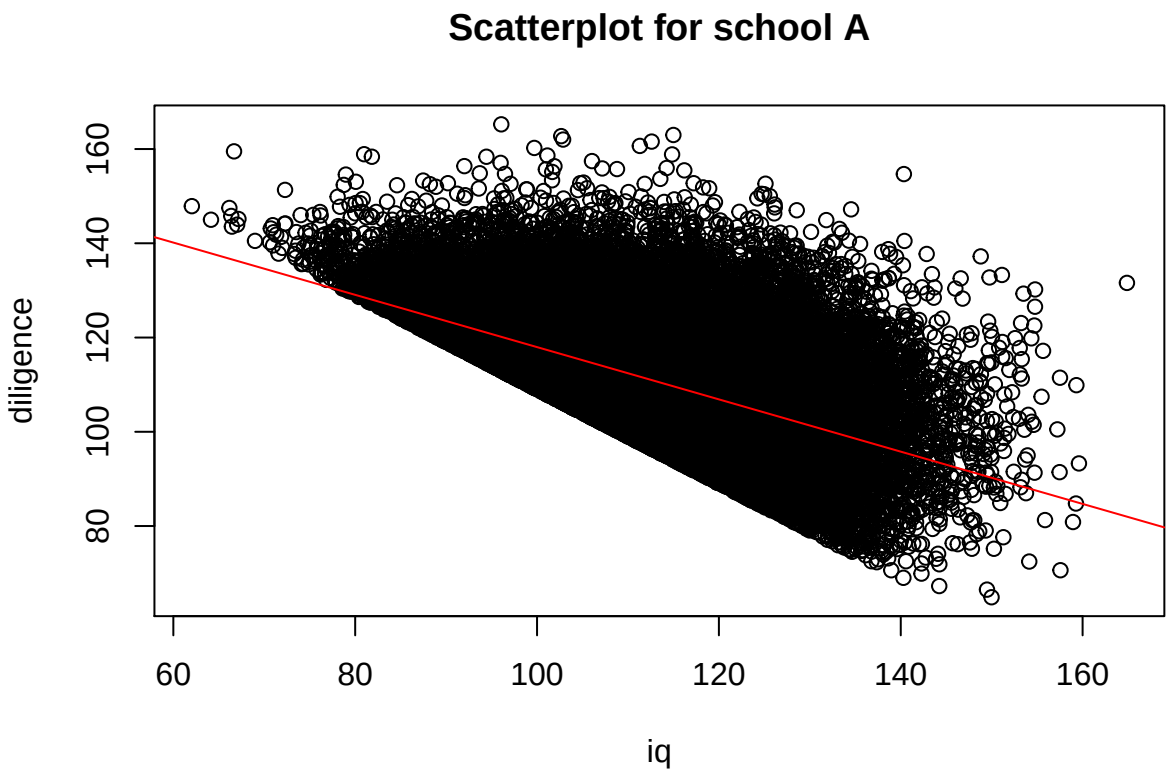

```
## [1] 115874   9138
```

The density plot for the residuals seems symmetric and bell-shaped.
The qqplot for the residuals fall on a straight line.
From these observations, the residuals appear Gaussian.

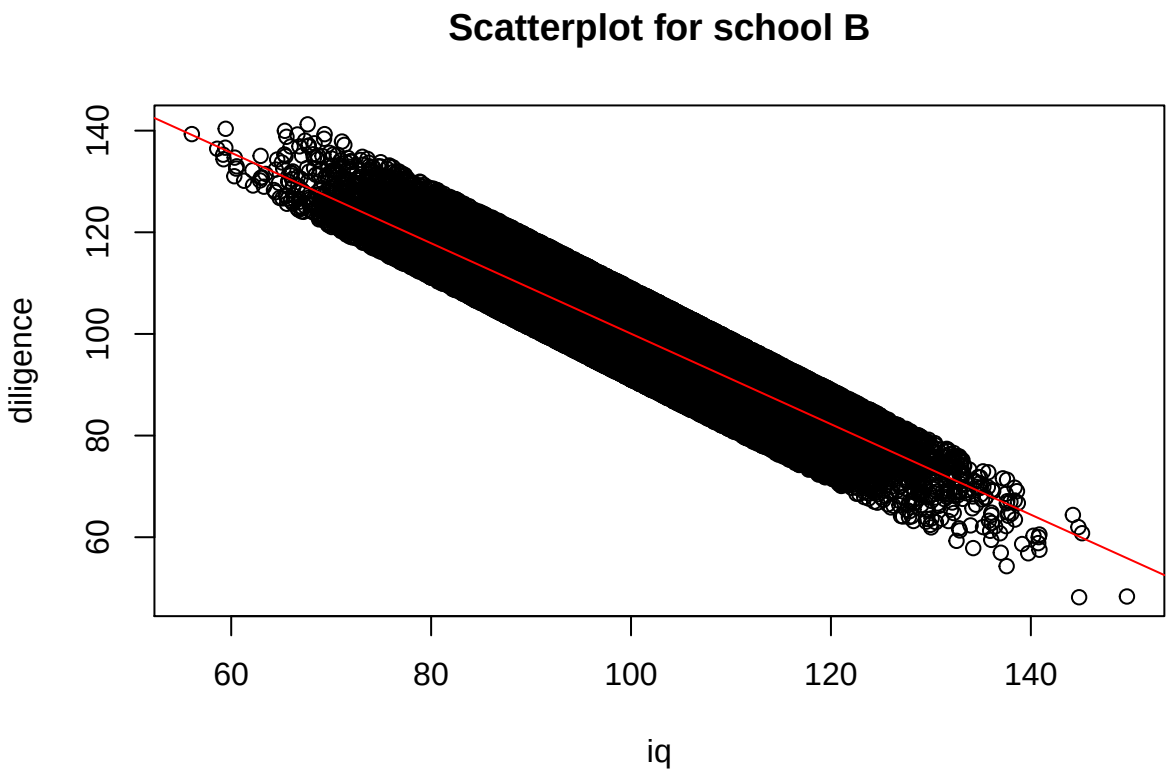
The results for Part 3 do not agree with those from Part1 and Part2

3)

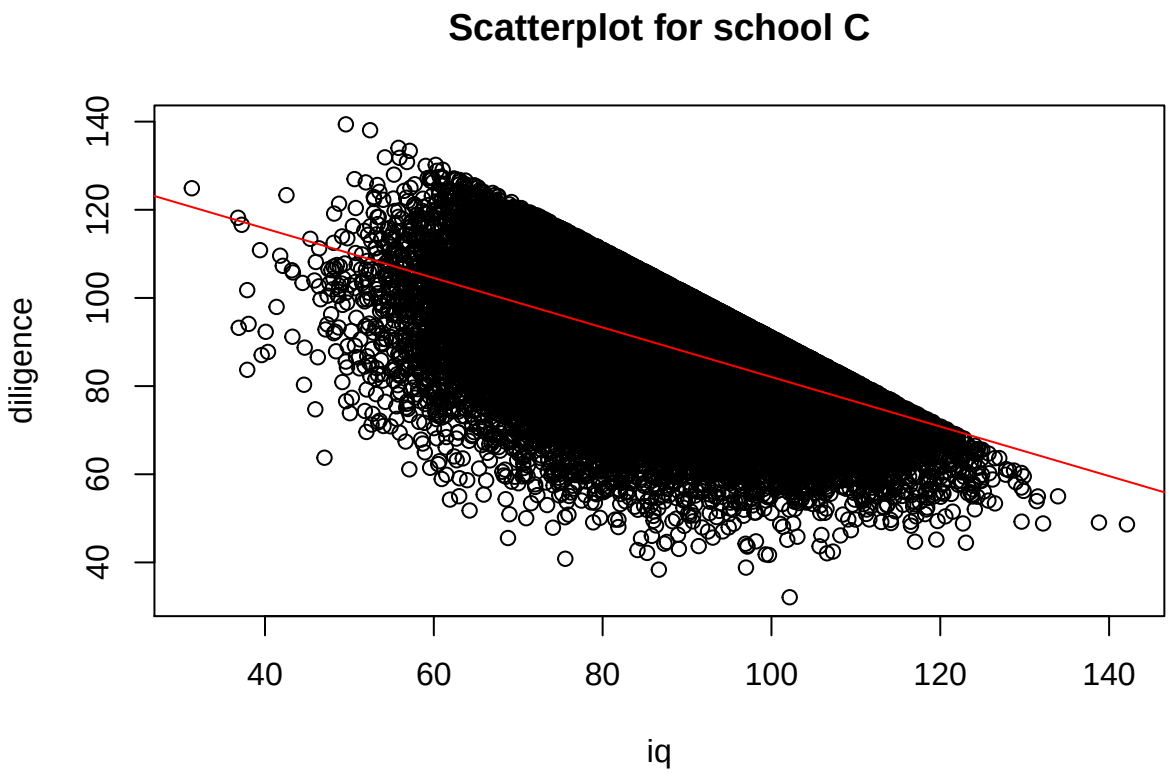
```
with(schoolA, plot(V1,V2, xlab = "iq" , ylab = "diligence", main = "Scatterplot for school A"))
abline(fit, col="red")
```



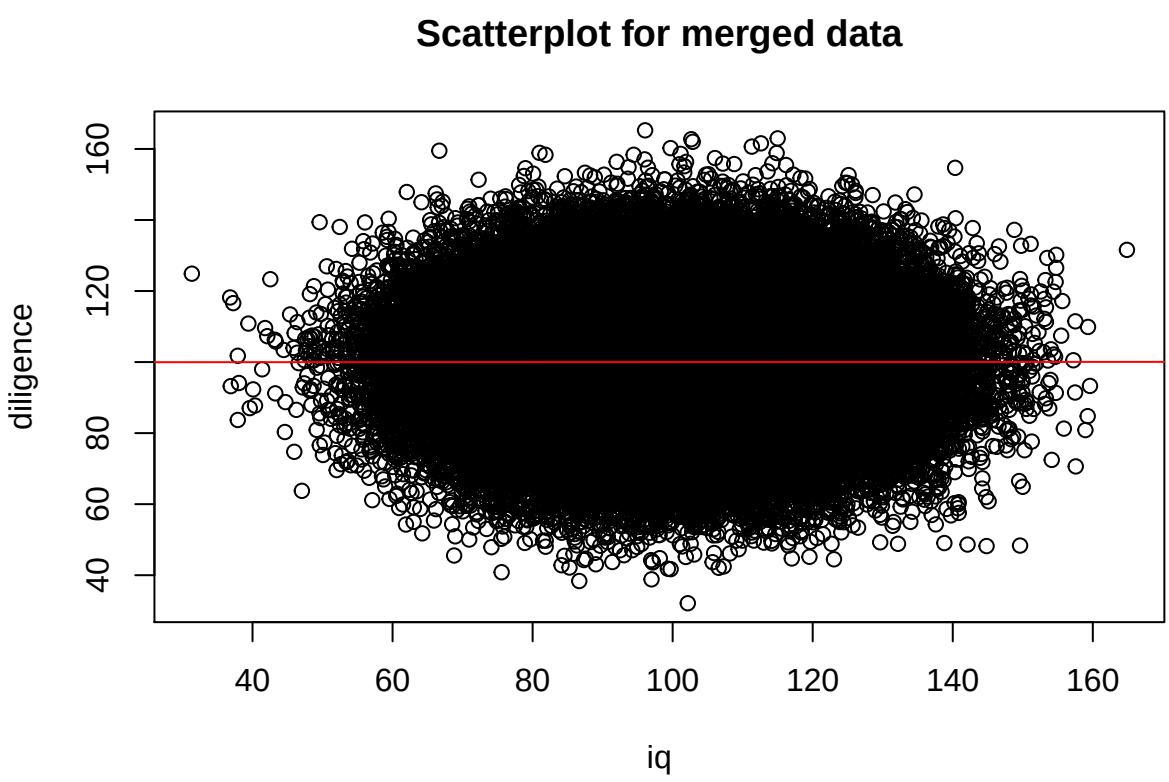

```
with(schoolB, plot(V1,V2, xlab = "iq" , ylab = "diligence", main = "Scatterplot for school B"))
abline(fit2, col = "red")
```




```
with(schoolC, plot(V1,V2, xlab = "iq" , ylab = "diligence", main = "Scatterplot for school C"))
abline(fit3, col = "red")
```




```
with(merged_data, plot(V1,V2, xlab = "iq" , ylab = "diligence", main = "Scatterplot for merged data"))
abline(fit4, col="red")
```



From the scatterplots, we can see that for school A, school B and school C, best fit lines with negative slopes are obtained, suggesting that as iq increases, diligence decreases.(explaining why we got a p-value smaller than 0.05 for Part 1 and Part 2)

On the scatterplot of the merged dataset, we get a best fit line with a slope of 0. This means that iq has no effect on diligence, explaining why we got a p-value greater than 0.05 for Part 3.

This is most likely due to the fact that the datasets for A, B and C are not independent. In fact, from the scatterplots, it appears that the datasets for A,B and C are distinct and non-overlapping.