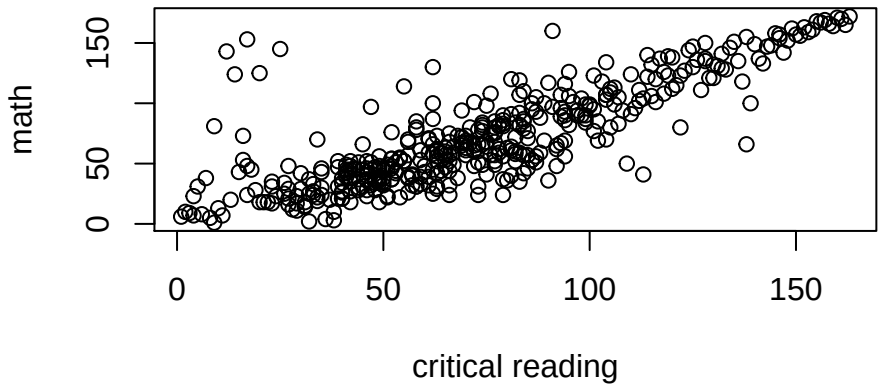
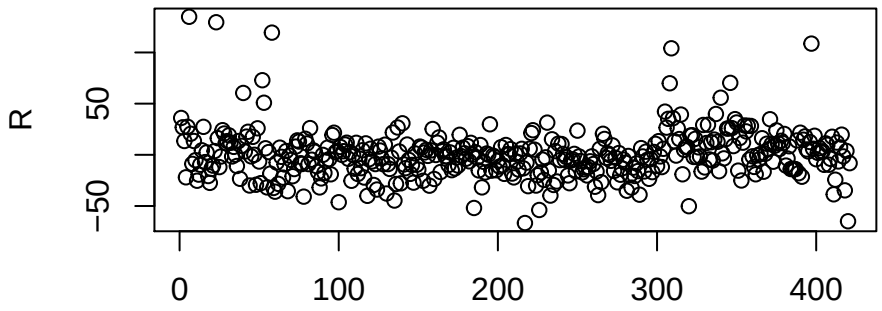
```r
library(RCurl)
library(dplyr)
library(ggplot2)
library(lmtest)

x <- getURL("https://raw.github.com/cmm6/data605/main/2012_SAT_Results.csv",
            .opts=curlOptions(followlocation = TRUE))

sat_scores <- read.csv(text = x, header=TRUE)
colnames(sat_scores) <- c('dbn', 'school_name',
                          'num_test_takers',
                          'critical_reading',
                          'math','writing')

sat_scores$num_test_takers <- as.numeric(sat_scores$num_test_takers)
s_omitted <- sat_scores[- grep("s", sat_scores$math),]

x <- as.numeric(s_omitted$critical_reading)
y <- as.numeric(s_omitted$math)

math_as_cr <- lm(y ~ x)
summary(math_as_cr)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.620 -13.769  -1.266  10.843 134.850
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.30624    2.57867   0.894    0.372
## x            0.93198    0.03159  29.506   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.49 on 419 degrees of freedom
## Multiple R-squared:  0.6751, Adjusted R-squared:  0.6743
## F-statistic: 870.6 on 1 and 419 DF,  p-value: < 2.2e-16
```

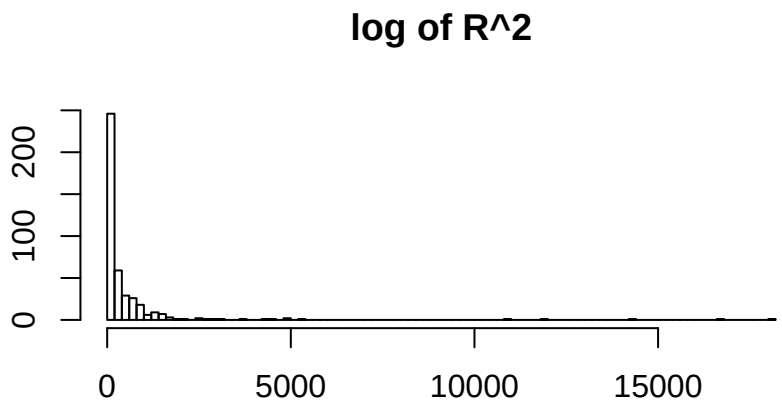
```r
plot(x, y, xlab="critical reading", ylab="math")
```



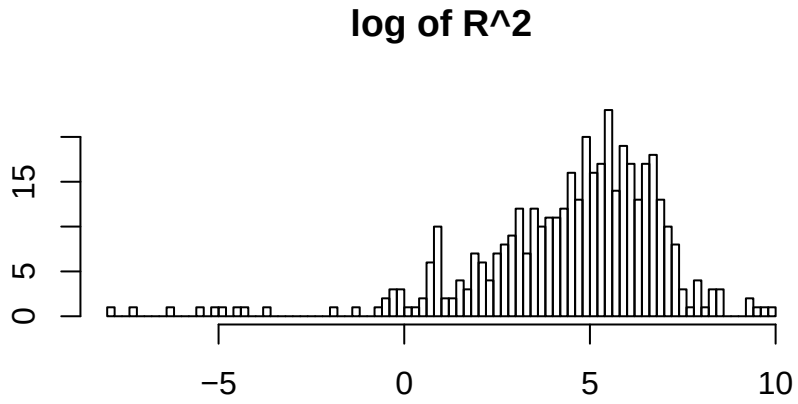

```r
plot(math_as_cr$residuals, xlab = "", ylab = "R")
```



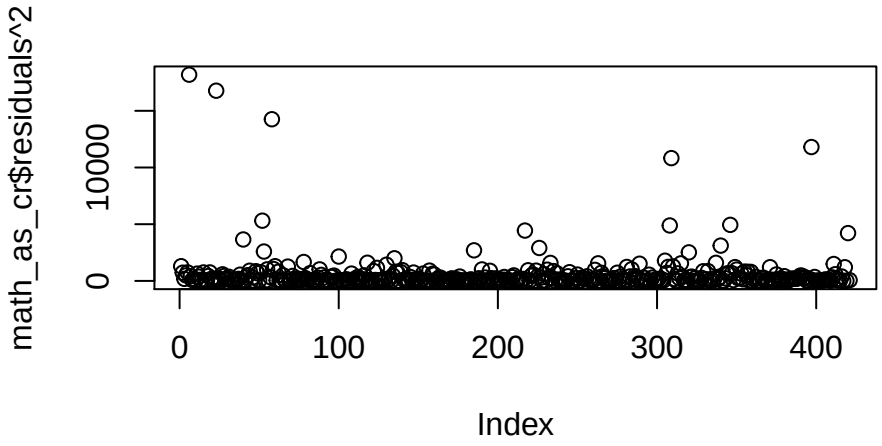

```r
hist(math_as_cr$residuals^2, breaks = 100, main = "log of R^2", xlab = "", ylab="")
```



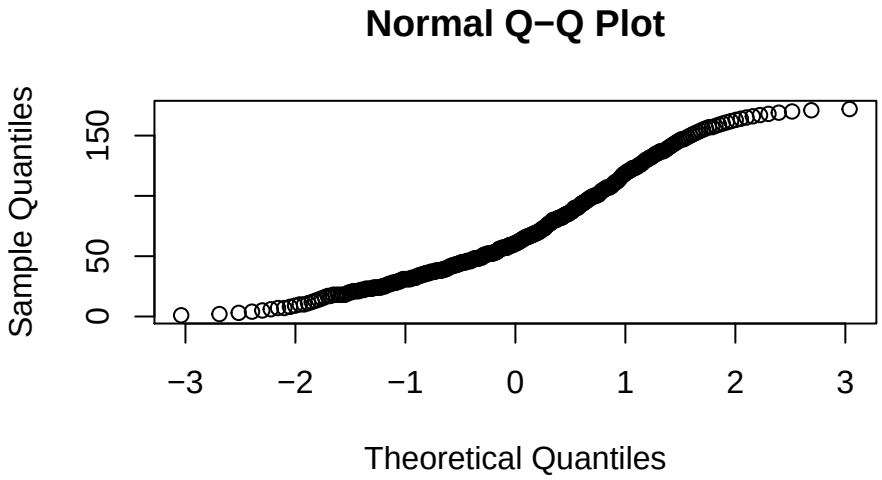

```
hist(log(math_as_cr$residuals^2), breaks = 100, main = "log of R^2", xlab = "", ylab="")
```




```
plot(math_as_cr$residuals^2)
```



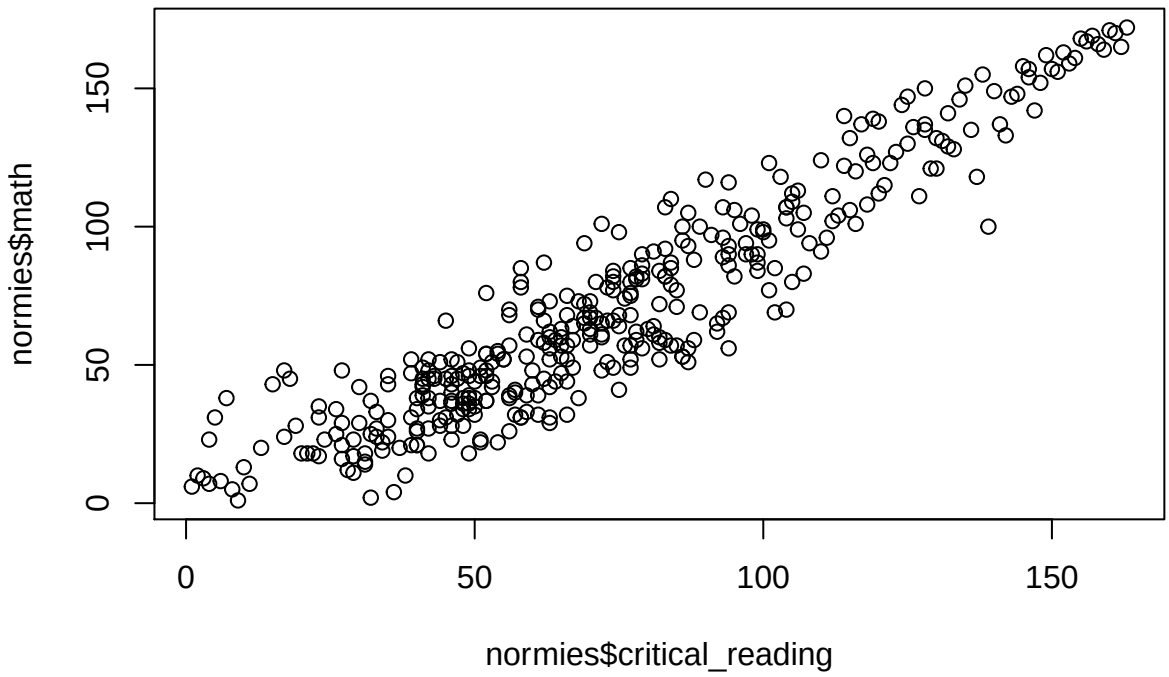

```
qqnorm(y)
```



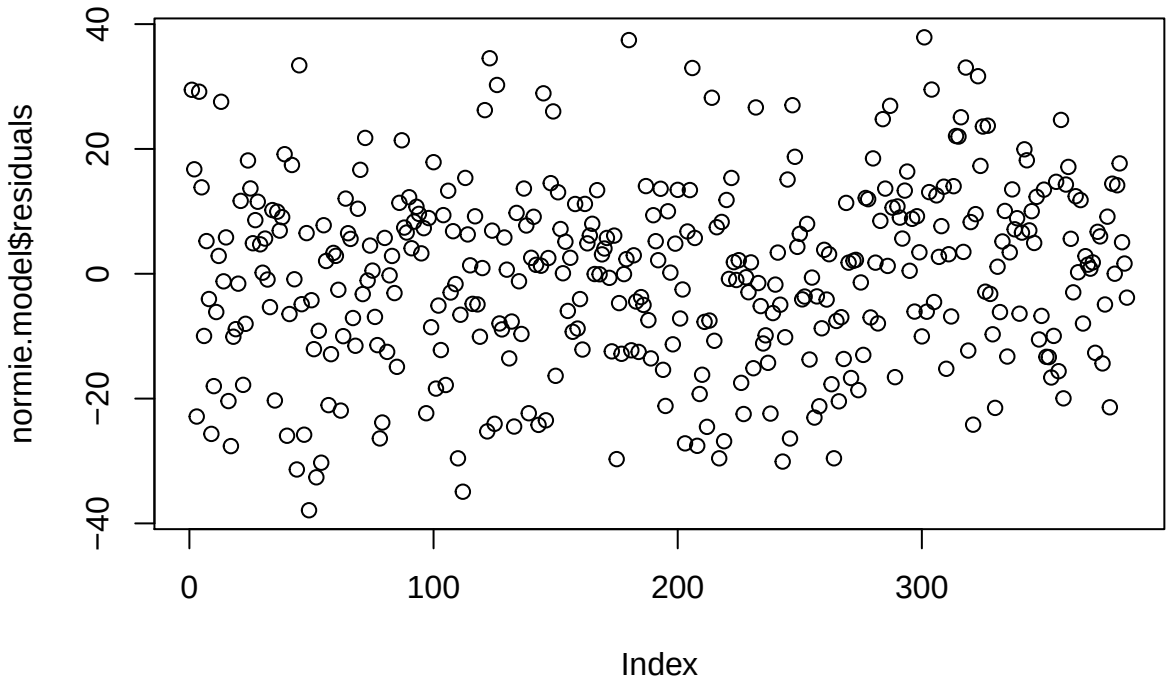

```
normies <- s_omitted %>%
  mutate(r = math_as_cr$residuals) %>%
  filter(r^2 < 1200)

normies$math <- as.numeric(normies$math)
normies$critical_reading <- as.numeric(normies$critical_reading)
normies$writing <- as.numeric(normies$writing)

y <- normies$math
x <- normies$critical_reading

normie.model <- lm(y ~ x)
summary(normie.model)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.890  -9.741   1.320   9.256  37.880
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.18652    1.66723   -4.31 2.08e-05 ***
## x            1.04372    0.02026   51.52  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.47 on 382 degrees of freedom
## Multiple R-squared:  0.8742, Adjusted R-squared:  0.8738
## F-statistic:  2654 on 1 and 382 DF,  p-value: < 2.2e-16
```

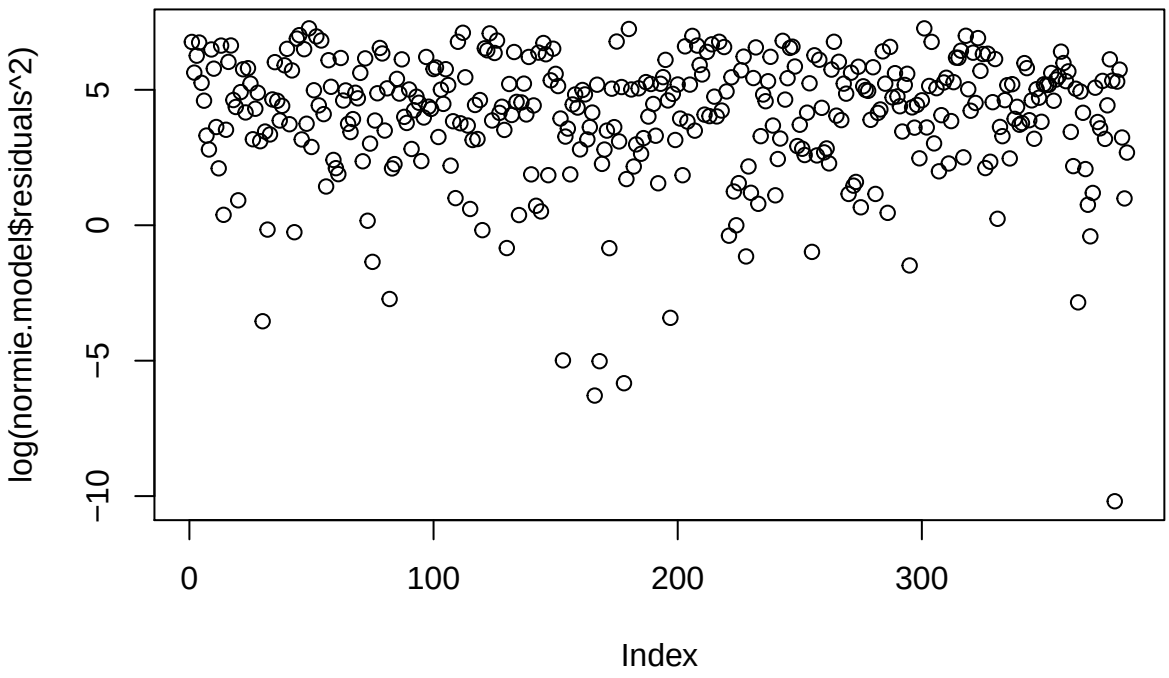
```
plot(normies$critical_reading, normies$math)
```



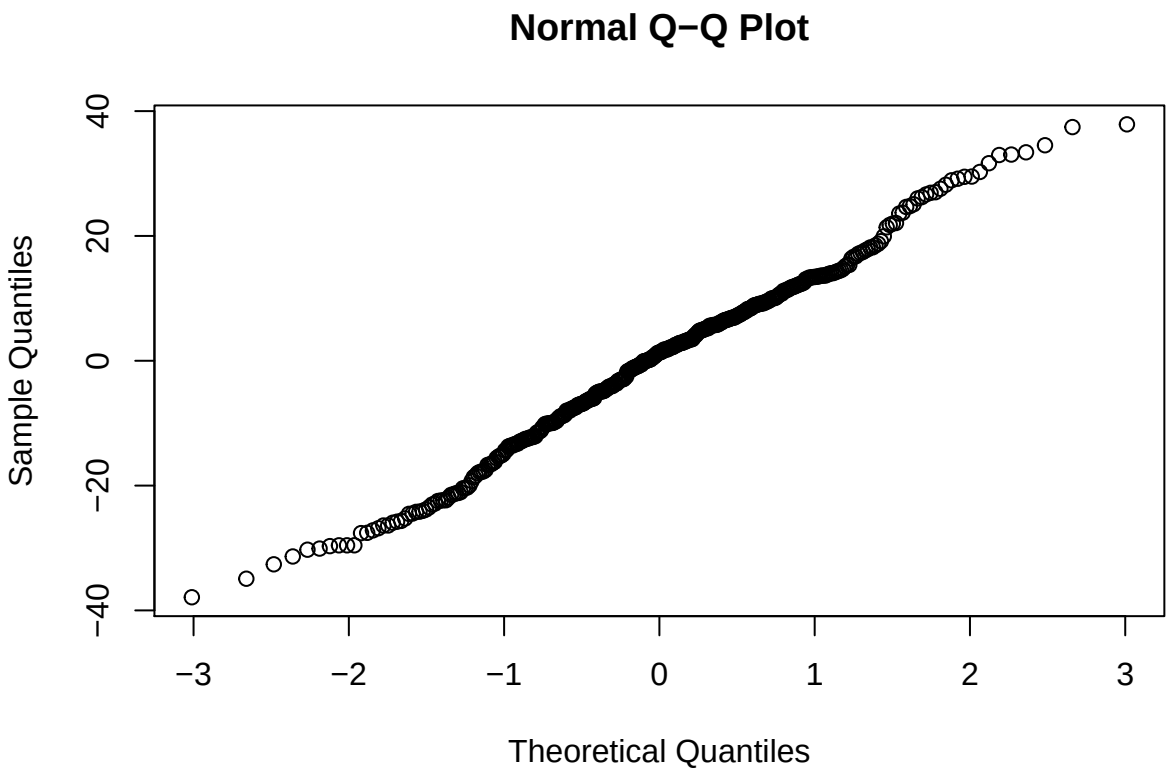

```
plot(normie.model$residuals)
```



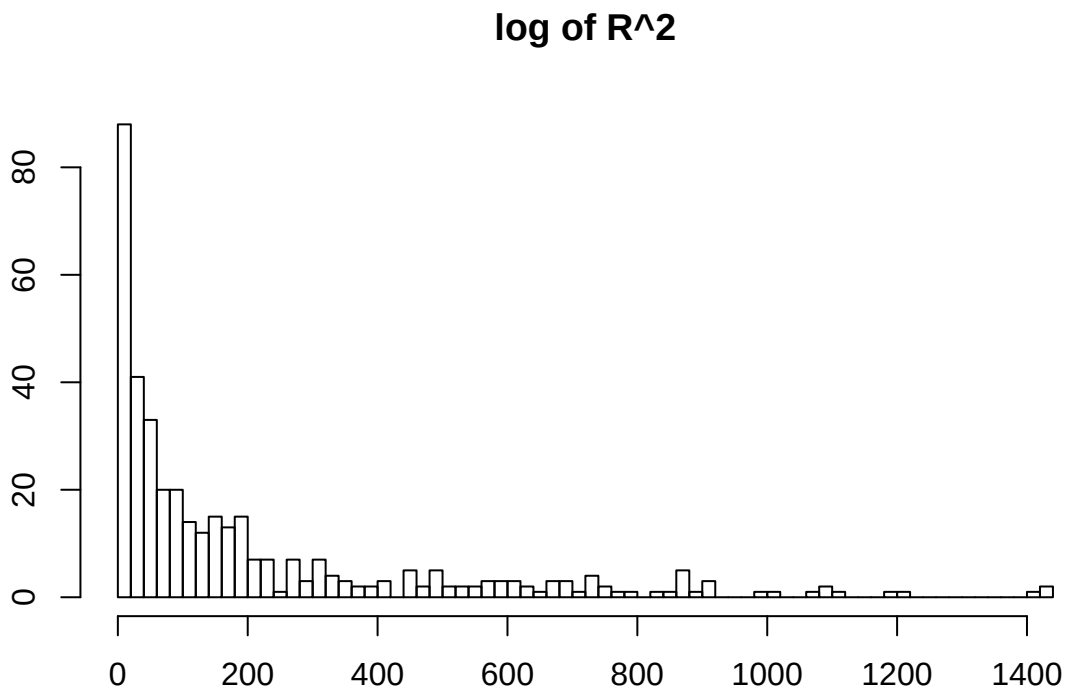

```
plot(log(normie.model$residuals^2))
```



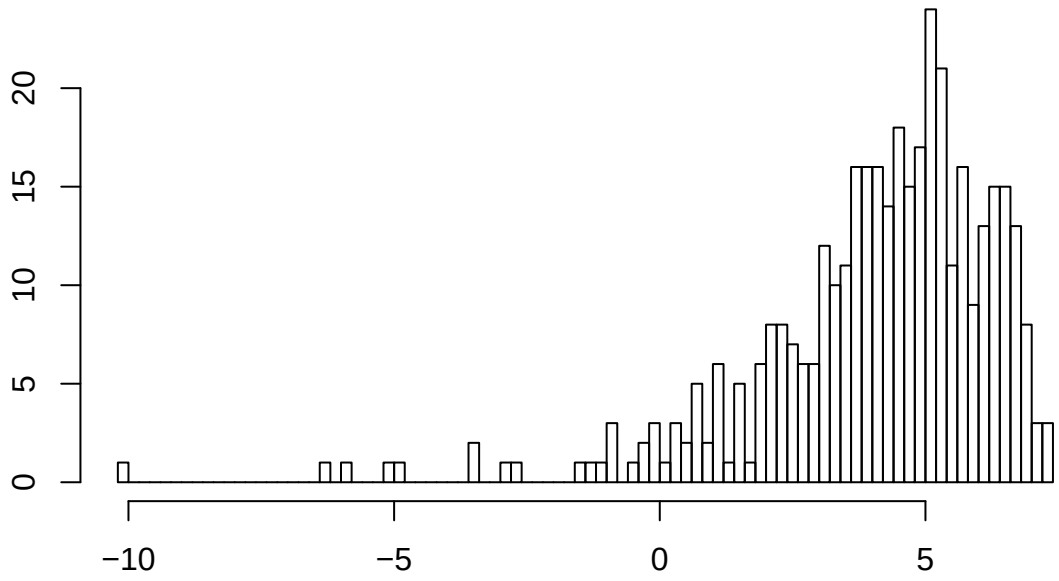

```
qqnorm(normie.model$residuals)
```



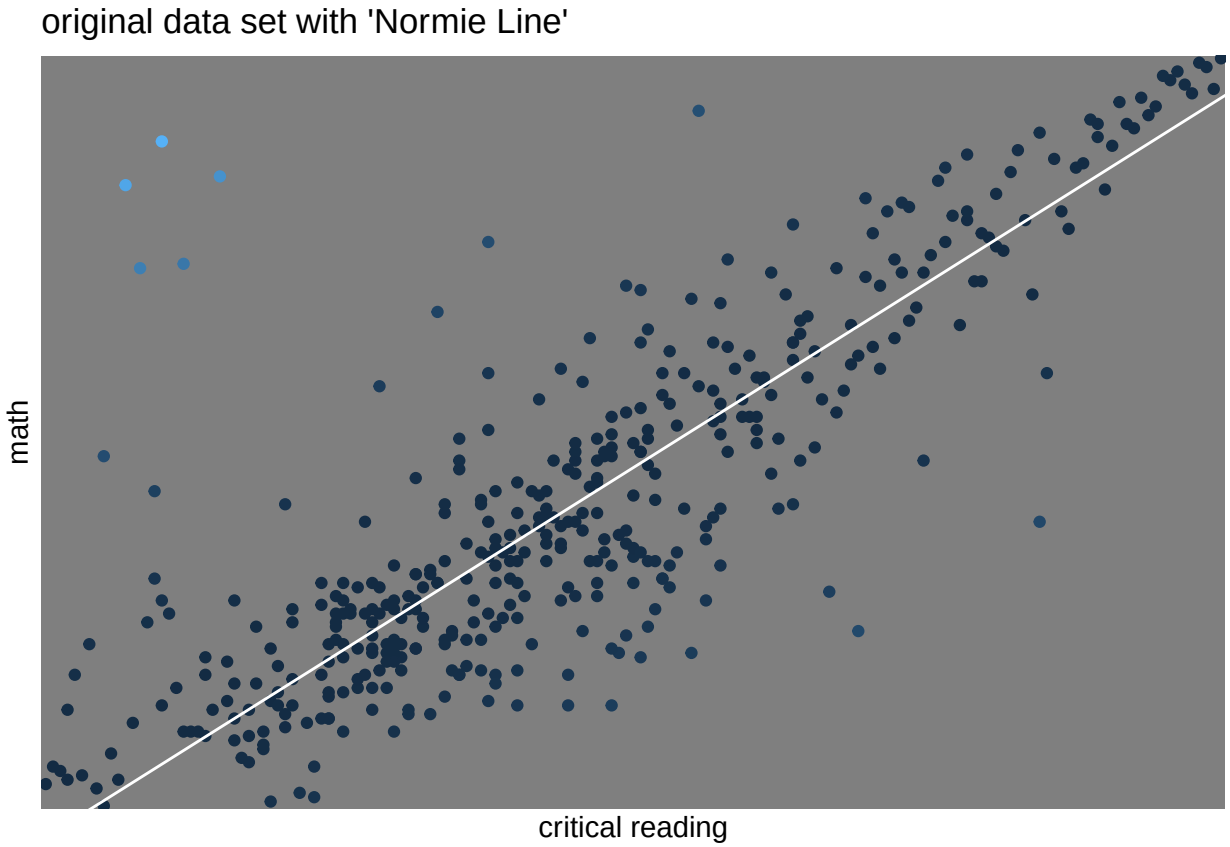

```
hist(normie.model$residuals^2, breaks = 100, main = "log of R^2", xlab = "", ylab="")
```




```
hist(log(normie.model$residuals^2), breaks = 100, main = "log of R^2", xlab = "", ylab="")
```




```
math_as_cr$math <- as.numeric(math_as_cr$math)
math_as_cr$critical_reading <- as.numeric(math_as_cr$critical_reading)
math_as_cr$writing <- as.numeric(math_as_cr$writing)

ggplot(s_omitted, aes(x = critical_reading,
                      y = math,
                      color = math_as_cr$residuals^2)) +
  geom_point() +
  geom_abline(intercept = normie.model$coefficients[1],
              slope = normie.model$coefficients[2],
              color = "white") +
  theme_dark() +
  labs(title = "original data set with 'Normie Line'",
       x = "critical reading",
       y = "math") +
  theme(legend.position = "none") +
  scale_x_discrete(breaks=NULL) +
  scale_y_discrete(breaks=NULL)
```




```
bptest(normie.model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  normie.model
## BP = 0.68361, df = 1, p-value = 0.4083
```

It's still not very good!