```
{r setup, include=FALSE} knitr::opts_chunk$set(echo = TRUE)
```

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

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

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)
```

```
{r, fig.height=3, fig.width=5} plot(x,y) plot(math_as_cr$residuals) hist(math_as_cr$residuals,
breaks = 100) plot(math_as_cr$residuals^2) qqnorm(y)
```

```
normies <- s_omitted %>%
  mutate(r = math_as_cr$residuals) %>%
  filter(r^2 < 2500)

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)
```

```
plot(normies$critical_reading, normies$math)
plot(normie.model$residuals)
plot(log(normie.model$residuals^2))
qqnorm(normie.model$residuals)
```

```
ggplot(normies, aes(x = critical_reading,
                    y = math,
                    color = normies$r^2)) +
  geom_point() +
  geom_smooth() +
  geom_abline(intercept = normie.model$coefficients[1],
              slope = normie.model$coefficients[2],
              color = "white") +
  theme_dark() +
  labs(title = "outliers removed")
```

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

Even with the outliers removed, it seems that there is a strong slope component correlating the data but there is a wide variation in intercept… There is a linear relationship, but it's not good enough to predict scores with certainty.