StatMethodsHW2

Joshua Burkhart

January 17, 2016

BMI 651: HW2

Based on the readings and in-class discussion, conduct EDA and regression diagnostics to determine if there are any violations of assumptions.

Provide detailed script (10pts)

See the source .Rmd document.

Write-up with the key figures and tables as well as description of any issues and how you handled them in order to analyze the data (20 pts) Note: you should be able to describe any output or figures without jargon!!!

See below output.

Load Data

```
data <- read.csv("~/SoftwareProjects/StatisticalMethodsInCompBio/HW2/HIV.txt",sep="")
```

Missing Data

```
sapply(data, function(x) sum(is.na(x)))
```

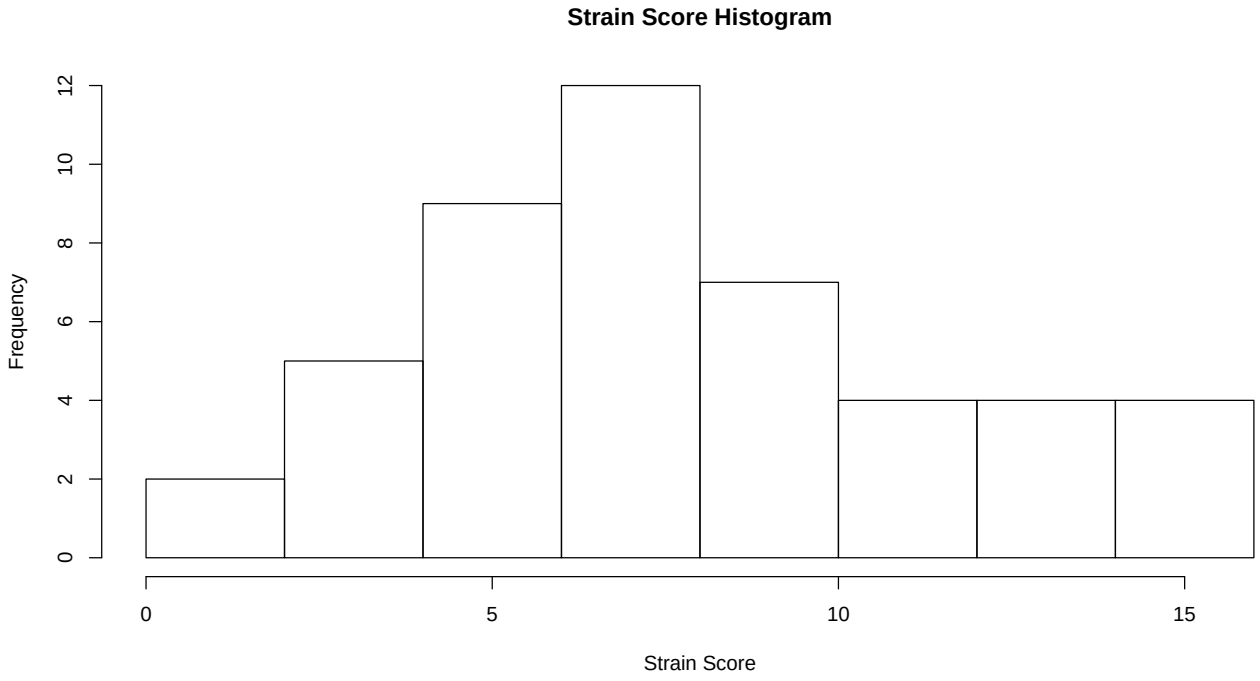
```
StrainScore   ViralLoad
          0           1
```

One missing ViralLoad is reported, violating an assumption of no missing or invalid data. We'll just remove it for now.

```
data_rm_na <- na.omit(data)
```

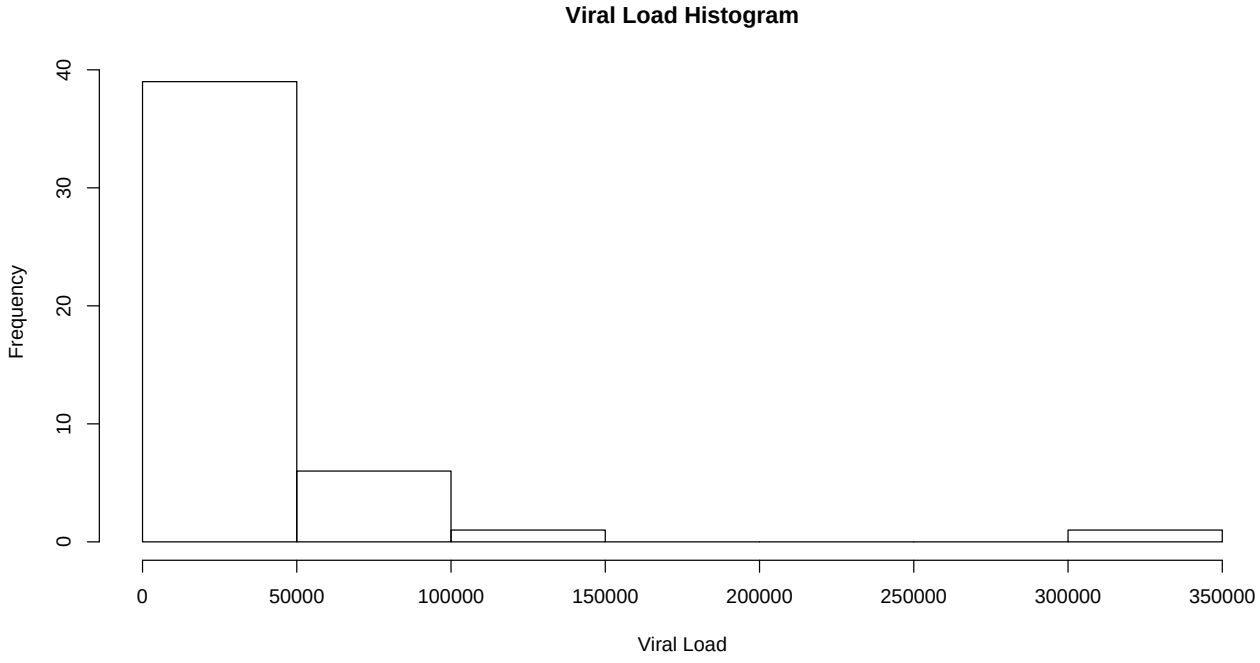
Distributions

```
hist(data_rm_na$StrainScore,
     xlab="Strain Score",
     main="Strain Score Histogram")
```



Unimodal wth a slight skew, looks close enough to a Gaussian…

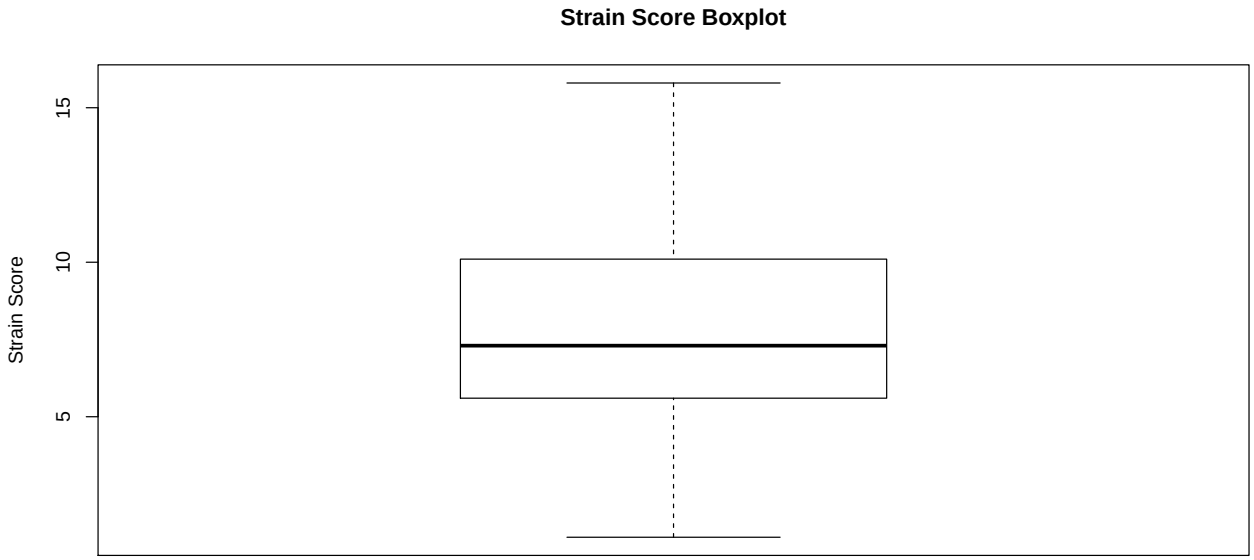
```
hist(data_rm_na$ViralLoad,
     xlab="Viral Load",
     main="Viral Load Histogram")
```



Outcome distribution looks funny (bimodal with strong skew) and violates an assumption of normally distributed data.

Outliers

```
ssbp <- Boxplot(data_rm_na$StrainScore,
        ylab="Strain Score",
        main="Strain Score Boxplot")
```

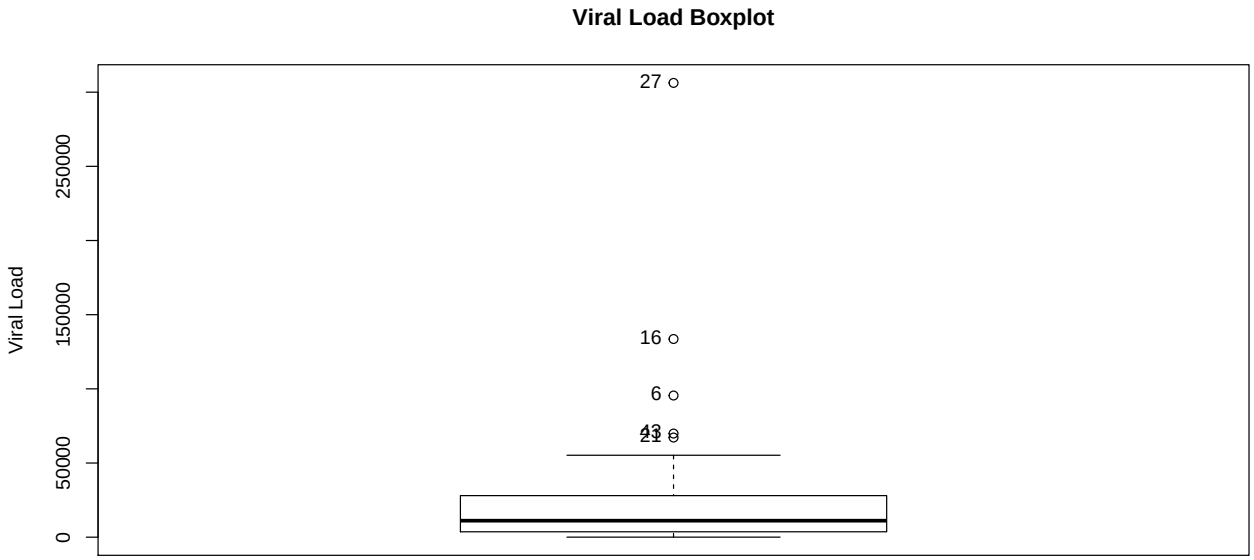



```
ssbp #outliers reported
```

```
NULL
```

No outliers reported for input.

```
vlbp <- Boxplot(data_rm_na$ViralLoad,
        ylab="Viral Load",
        main="Viral Load Boxplot")
```



```
vlbp #outliers reported
```

```
[1]  6 16 21 27 43
```

Five outliers reported in output. Row Numbers: 27, 16, 6, 21, and 43.

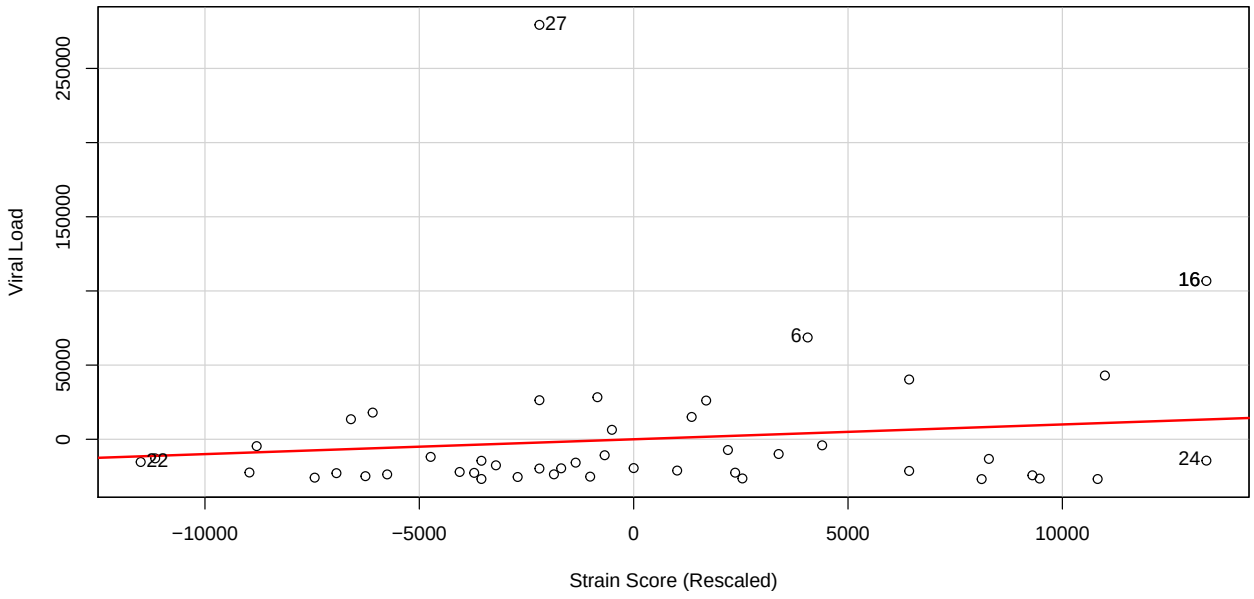
```
model <- lm(data_rm_na$ViralLoad ~ data_rm_na$StrainScore)
summary(model)
```

```
Call:
lm(formula = data_rm_na$ViralLoad ~ data_rm_na$StrainScore)

Residuals:
   Min     1Q Median     3Q    Max
-37612 -21935 -14397   5530 281570

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)               13520      17207   0.786    0.436
data_rm_na$StrainScore     1691       1972   0.858    0.396

Residual standard error: 50140 on 45 degrees of freedom
Multiple R-squared:  0.01609,   Adjusted R-squared:  -0.005778
F-statistic: 0.7357 on 1 and 45 DF,  p-value: 0.3956
```

In the context of a linear model, Strain Score does not appear to be a strong predictor of Viral Load ($p = 0.396$).

Leverage

```
leveragePlot(model,
             term.name="data_rm_na$StrainScore",
             id.n=3,
             xlab="Strain Score (Rescaled)",
             ylab="Viral Load")
```

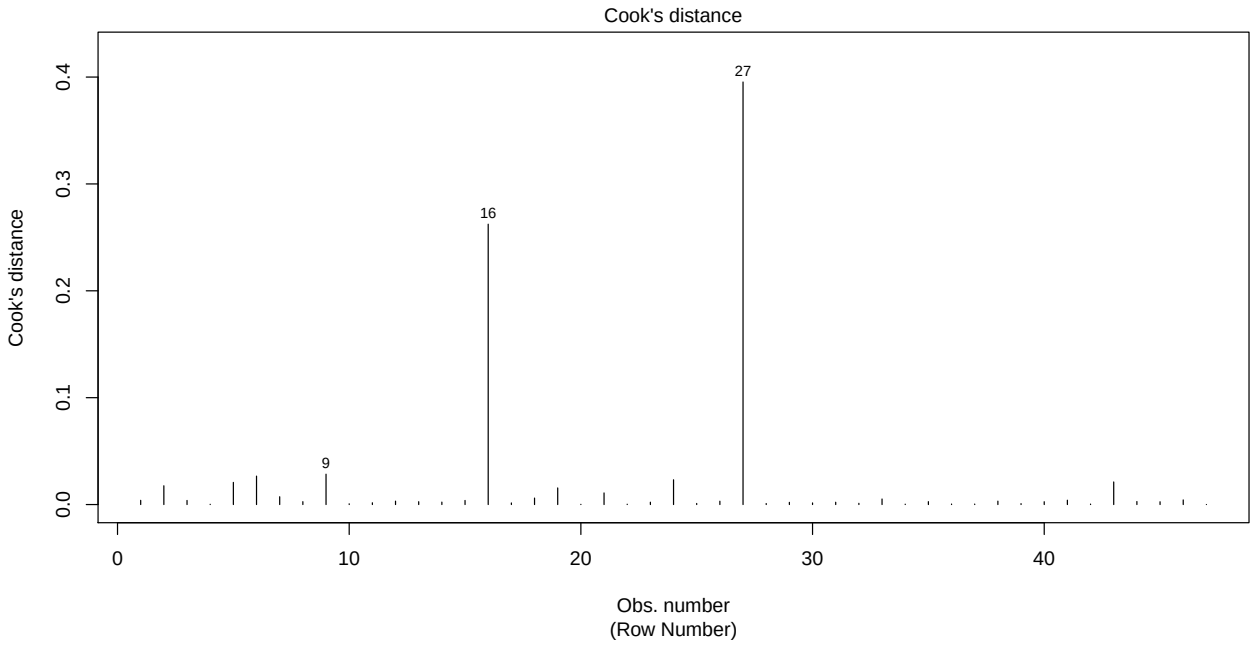



```
27 16  6 16 24 22
27 16  6 16 24 22
```

The datapoints with highest leverage are from row numbers 27, 16, 24, 6, and 22.

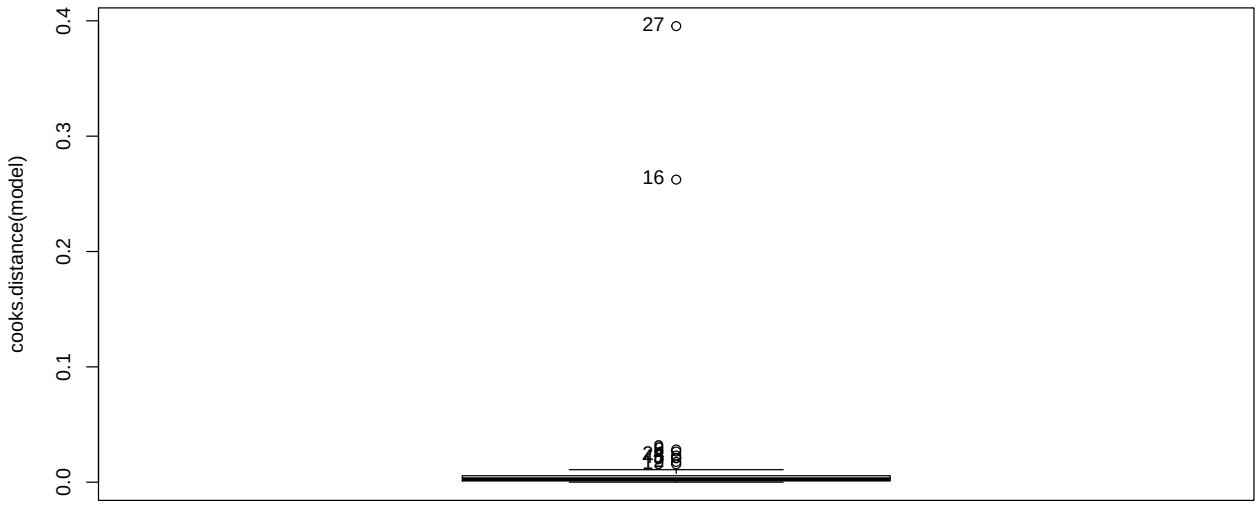
Influence

```
plot(model,
     which=4,
     sub.caption="(Row Number)") #plot.lm 4 is Cook's Distance
```



We can see that data from rows 16 and 27 appear highly influential.

```
cdbp <- Boxplot(cooks.distance(model))
```



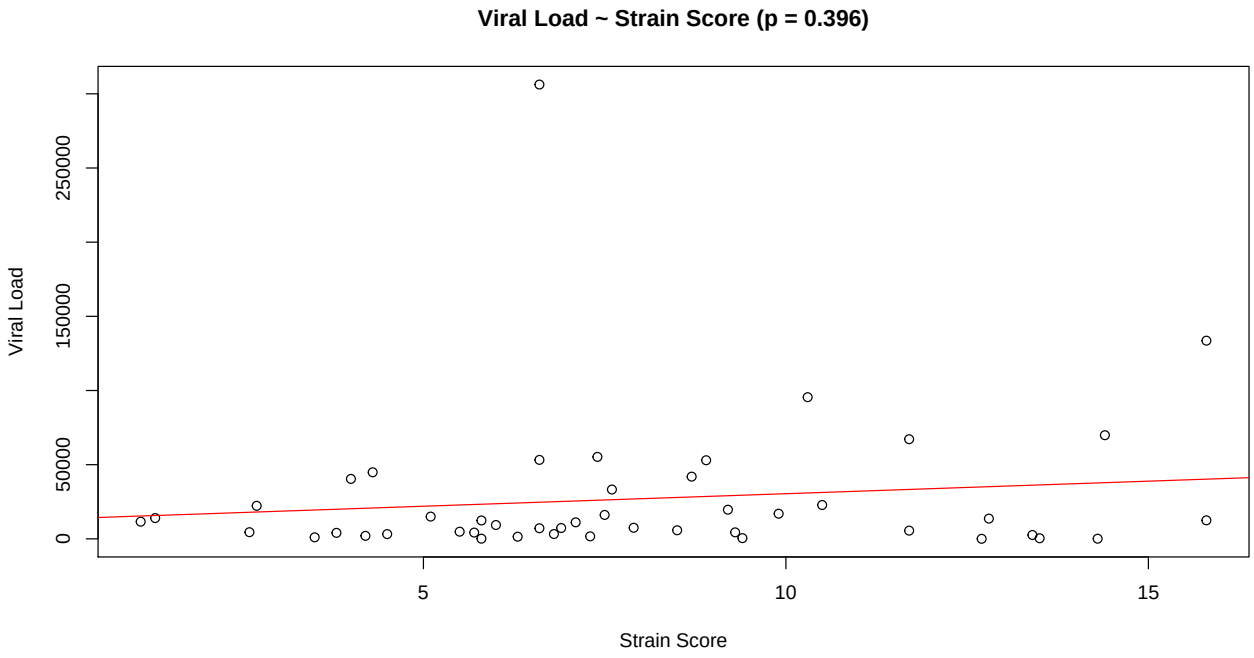

```
cdbp
```

```
[1]  2  5  6  9 16 19 24 27 43
```

This boxplot presents the distribution of influential points and labels data from rows 2, 5, 6, 9, 16, 19, 24, 27, 43 as outliers.

Model

```
plot(data_rm_na$StrainScore,data_rm_na$ViralLoad,
     xlab="Strain Score",
     ylab="Viral Load",
     main="Viral Load ~ Strain Score (p = 0.396)")
abline(model, col="red")
```



The model does appear to fit the data but we must remember that it does not do so significantly.

Rescue?

```
data_rm_na[27,]
```

```
   StrainScore ViralLoad
28         6.6    306251
```

```
range(data_rm_na[-27,][2])
```

```
[1]     40 133599
```

Data Row 27 reports a Viral Load almost triple the range for all other points. If we assume this is a mistake, we can remove it and rerun our model.

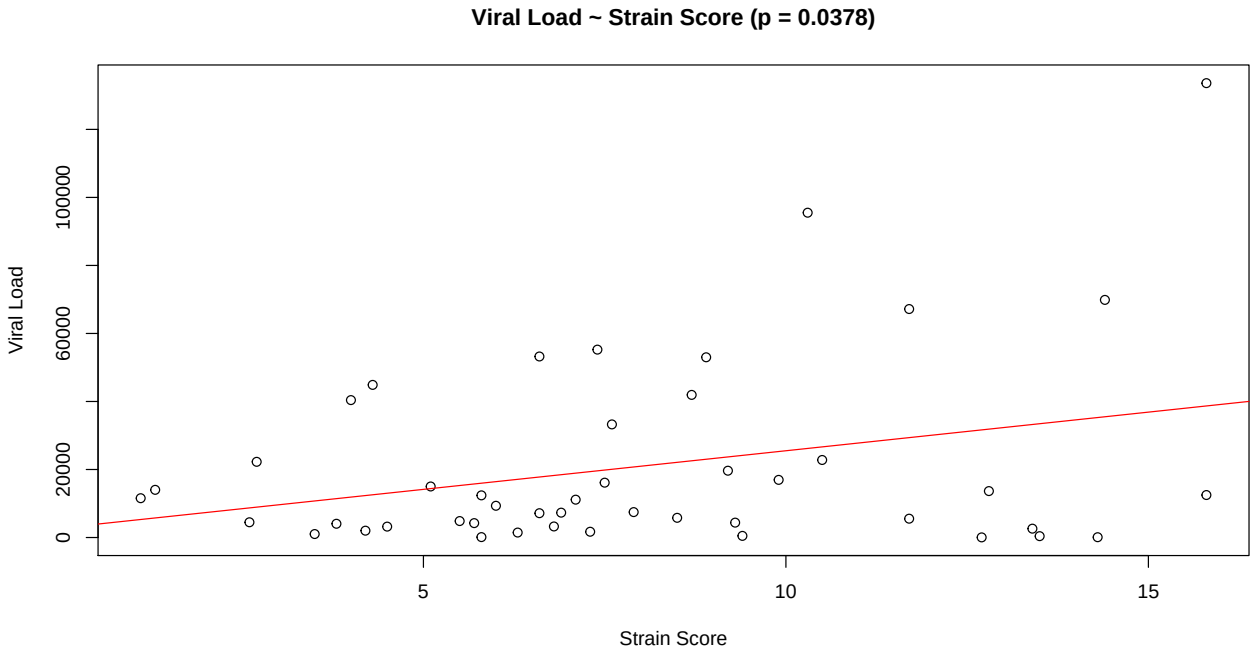
```
data_rm_na_no27 <- data_rm_na[-27,]
model_no27 <- lm(data_rm_na_no27$ViralLoad ~ data_rm_na_no27$StrainScore)
summary(model_no27)
```

```
Call:
lm(formula = data_rm_na_no27$ViralLoad ~ data_rm_na_no27$StrainScore)

Residuals:
   Min     1Q Median     3Q    Max
-35186 -15803  -8064  11966  94917
```

```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                    2802       9296   0.301   0.7646
data_rm_na_no27$StrainScore    2271       1060   2.142   0.0378 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26930 on 44 degrees of freedom
Multiple R-squared:  0.09443,   Adjusted R-squared:  0.07384
F-statistic: 4.588 on 1 and 44 DF,  p-value: 0.03777
```

Strain Score becomes a significant predictor of Viral Load after removing row 27.

```
plot(data_rm_na_no27$StrainScore,data_rm_na_no27$ViralLoad,
     xlab="Strain Score",
     ylab="Viral Load",
     main="Viral Load ~ Strain Score (p = 0.0378)")
abline(model_no27, col="red")
```



We must be cautious in accepting this result as row 27 may not have been a mistake after all. Viral Loads over 1 million have been reported[1].

References

[1] http://www.catie.ca/en/fact-sheets/transmission/hiv-viral-load-hiv-treatment-and-sexual-hiv-transmission