Linear Regression Diagnostic Plots

Validity Conditions for Linear Regression Models

Linearity: residuals vs. predicted graph does not show strong evidence of curvature or other patterns.

Independence: responses can be considered independent of each other.

Normality: the histogram of the residuals is approximately symmetric with no large outliers.

Equal Variance: residuals vs. predicted graph show a constant width.

```
library(tidyverse)
setwd("C:/Users/Bryan.Jonas/OneDrive - West Point/AY 20-2 MA256/Working Directory")
```

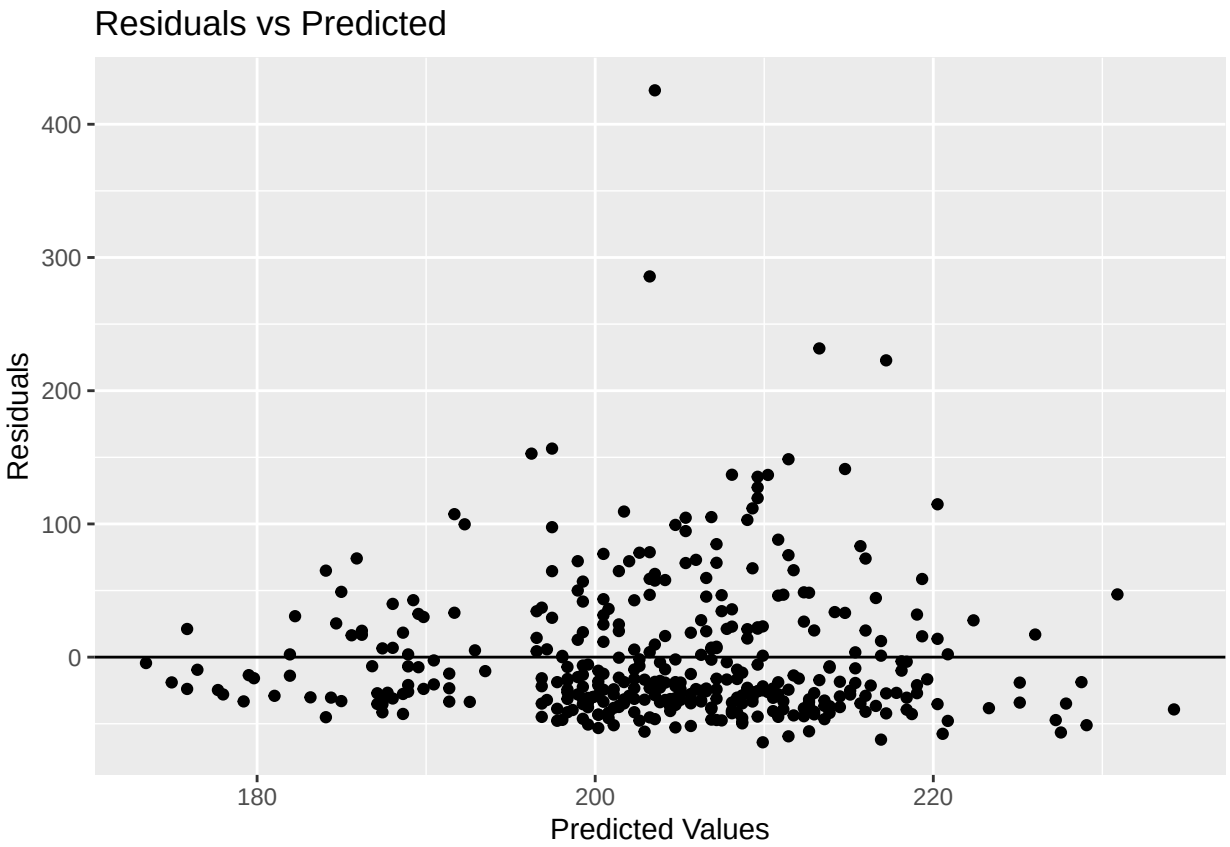
Linearity

Let's start by loading some data, creating our linear model, and creating the plot necessary to assess this validity condition.

```
IOCT <- read_csv('IOCT_tab_data.csv')

IOCTModel <- lm(IOCT_Time ~ APFT_Score, data = IOCT)

IOCT %>%
  ggplot(aes(x = IOCTModel$fitted.values, y = IOCTModel$residuals))+
  labs(x = "Predicted Values", y = "Residuals", title = "Residuals vs Predicted")+
  geom_point()+
  geom_hline(yintercept = 0)
```



From this plot we need to assess whether we see evidence of curvature or other patterns. It is immediately clear that we have a few predictions that have fairly significant residual values. What we want to assess, however, is if there is a patterns to the residuals that may demonstate an underlying relationship that our linear model is not taking into account. From this graph, I see no specific curve or underlying pattern that may demonstrate an non-linear relationship between IOCT Time and APFT Score.

Independence

This validity condition requires no graph at all but rather an assessment of the dataset used. Can we reasonably assume the independence of our observation or, to be specific in this case, is the data recorded from one cadet independent from another. I would assume (hope) this to be true as I don't believe it's reasonable to think that the IOCT Time or APFT Score of one cadet has any impact on another.

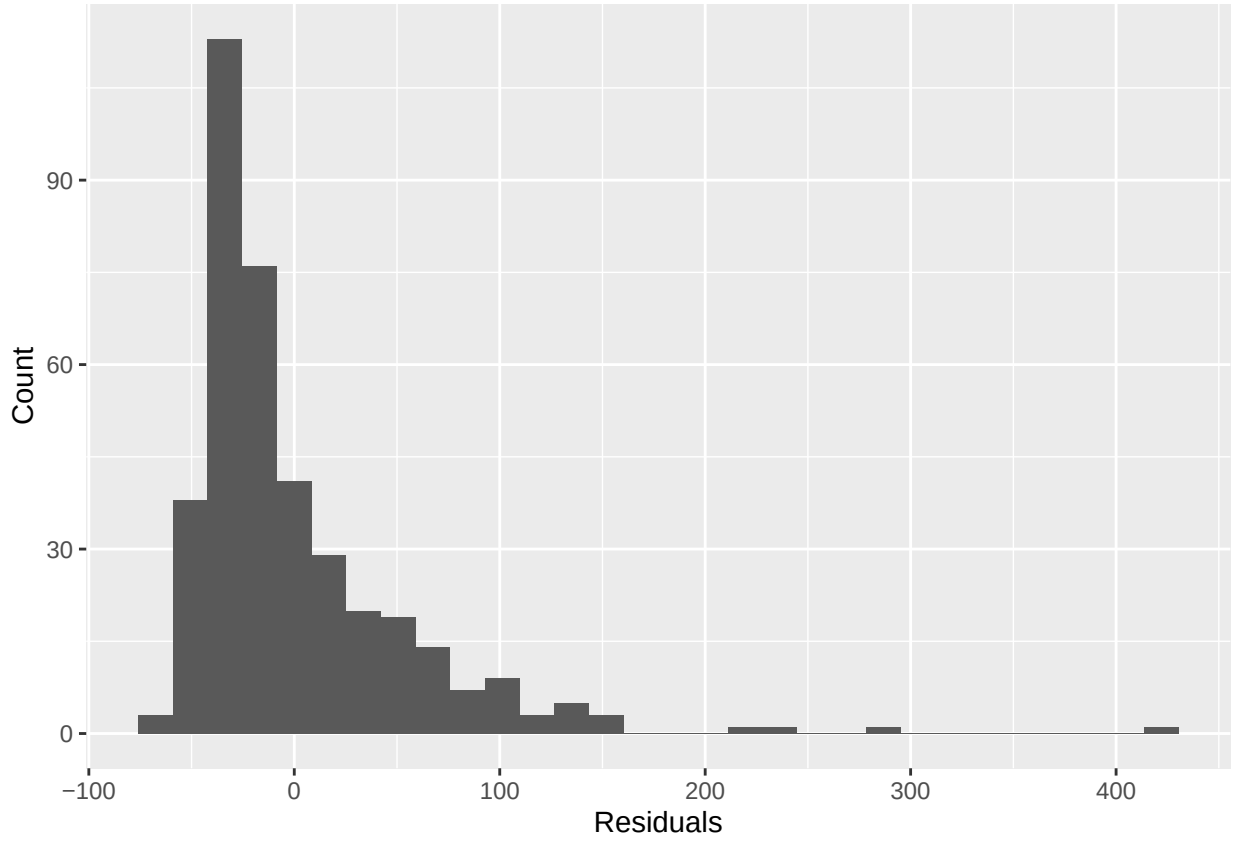
Normality of Residuals

```
IOCT %>%
  summarise(Mean = mean(IOCTModel$residuals),
            SD = sd(IOCTModel$residuals))
```

```
## # A tibble: 1 x 2
##       Mean    SD
##      <dbl> <dbl>
## 1 8.71e-16  53.8
```

```
IOCT %>%
  ggplot(aes(x = IOCTModel$residuals))+
  labs(x = "Residuals", y = "Count")+
```

```
  geom_histogram()
```



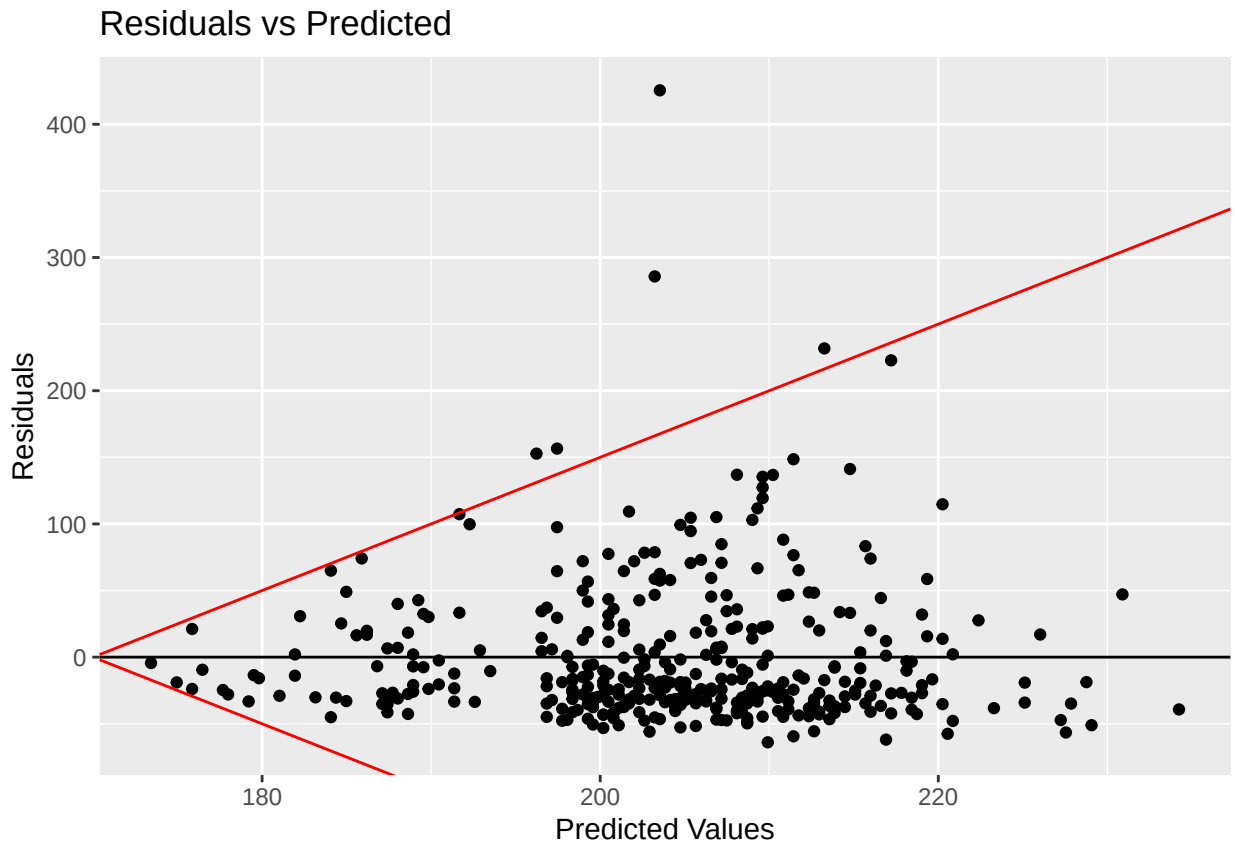

```
shapiro.test(IOCTModel$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  IOCTModel$residuals
## W = 0.76298, p-value < 2.2e-16
```

Here I have created a histogram of our residuals as well as conducted a Shapiro-Wilk test of a normality. From the plot, I have concerns about some significant outliers to the right of the main concentration of the residuals. Some of these outliers are over four standard deviations from the mean and one is over eight standard deviations from the mean. Combine this with the p-value of the Shapiro-Wilk test stating that we have significant evidence (at any reasonable significance level) that the distribution of our residuals is not the normal distirbution, I would have concerns about the validity of this condition.

Equal Variance of Residuals

```
IOCT %>%
  ggplot(aes(x = IOCTModel$fitted.values, y = IOCTModel$residuals))+
  labs(x = "Predicted Values", y = "Residuals", title = "Residuals vs Predicted")+
  geom_point()+
  geom_hline(yintercept = 0)+
  geom_abline(slope = 5, intercept = -850, col = 'red')+
  geom_abline(slope = -5, intercept = 850, col = 'red')
```



Taking another look at this graph, we can see that we do not have an equal variance of residuals and there appears to be some "fanning out" of the residuals as our predicted values increase. Note: There is no magic formula to these red lines on this plot. They are only there to demonstrate what I am talking about when I discuss "fanning out."