Stat 245 – Test 1

Trey Tipton

October 01, 2021

```
nerds <- readr::read_csv('https://sldr.netlify.app/data/nerds.csv')
```

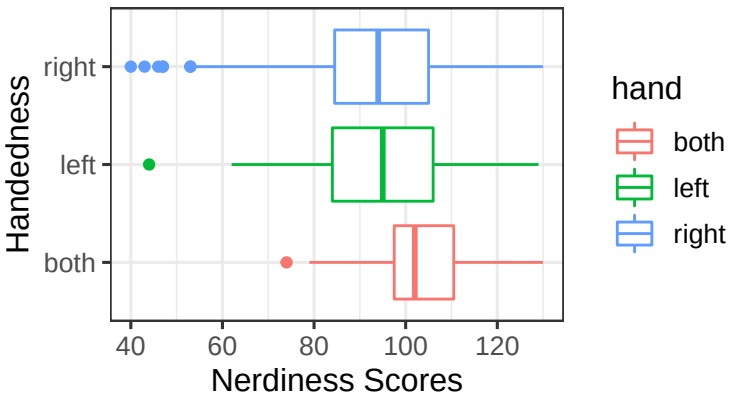
Choose Response and Predictor Variables

The question we are attempting to answer is if there is an association between nerdiness score and handedness. After looking at the "nerds" data set, it seems clear that the score is our response variable. After all, one's gender does not affect their age or handedness (theoretically) and vice versa. The nerdiness score is our response variable because it is what we are attempting to predict based on the other variables.

Our predictor variables are gender, age, and handedness. The ID number is simply a way to identify the test subjects and therefore should not be a predictor of score given that the sample was taken randomly and ID number is nominal, not ordinal. Gender is a predictor as someone may be more likely to be "nerdy" if they are male rather than female or vice versa, and the same goes for being younger or older and right handed or left handed.

Graph

```
gf_boxplot(hand ~ score, data = nerds, color = ~hand, xlab = "Nerdiness Scores", ylab = "Handedness")
```



From these boxplots we can compare the nerdiness score of people that are right, left, or both hand dominant: we learn that people that are right and left handed are similarly distributed in their Nerdiness Scores. The right hand dominant group has a slightly larger range and more outliers, however, this is expected because there are more right-handed people in this data set. Interestingly, the distribution for ambidextrous people (both right and left handed) is slightly moved to the right; the median and IQR are both higher than that of the other two, but there is a lot less data for both-handed people, so that could be why.

Fit the Model

```
nerds_lr <- lm(score ~ age + gender + hand,
data = nerds)
summary(nerds_lr)
```

```
##
## Call:
## lm(formula = score ~ age + gender + hand, data = nerds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.590  -8.872   0.095  10.938  36.065
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 101.72050    3.51626  28.929  < 2e-16 ***
## age          -0.04309    0.04989  -0.864  0.38814
## gendermale    2.64597    1.43918   1.839  0.06654 .
## genderother   8.60614    3.05610   2.816  0.00504 **
## handleft     -7.33667    3.92977  -1.867  0.06245 .
## handright    -8.18260    3.35186  -2.441  0.01496 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.62 on 536 degrees of freedom
## Multiple R-squared:  0.03308,    Adjusted R-squared:  0.02406
## F-statistic: 3.667 on 5 and 536 DF,  p-value: 0.002847
```

Model Equation:

$y_{score} = 101.72050 - .04309x_{age} + 2.64597I_{GenderMale} + 8.60614I_{GenderOther} - 7.33667I_{HandRight} - 8.18260I_{HandLeft} + \epsilon$

where y_{score} is the Nerdiness Score and x_{age} is the age and where:

$I_{GenderMale} = 1$ if the gender is "Male", 0 otherwise,

$I_{GenderOther} = 1$ if the gender is "Other", 0 otherwise,

$I_{HandRight} = 1$ if they are right handed, 0 otherwise,

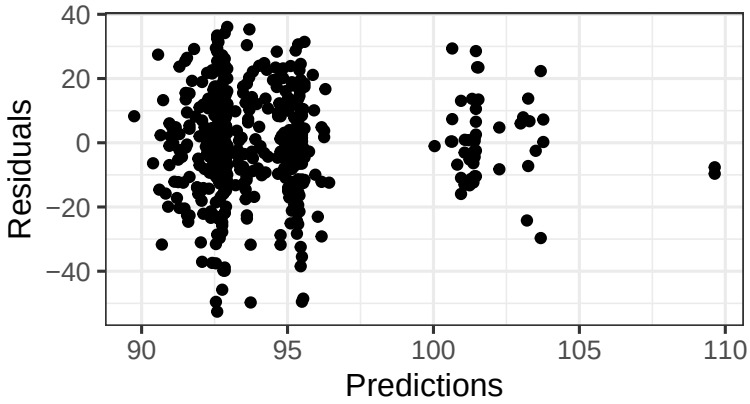
$I_{HandLeft} = 1$ if they are left handed, 0 otherwise.

The adjusted R-Squared value is 0.02406. This is an abnormally low R Squared value, indicating that our data may not have much of an association between the variables or that some condition may not be met. Let's check our conditions.

Conditions

Lack of Non-Linearity

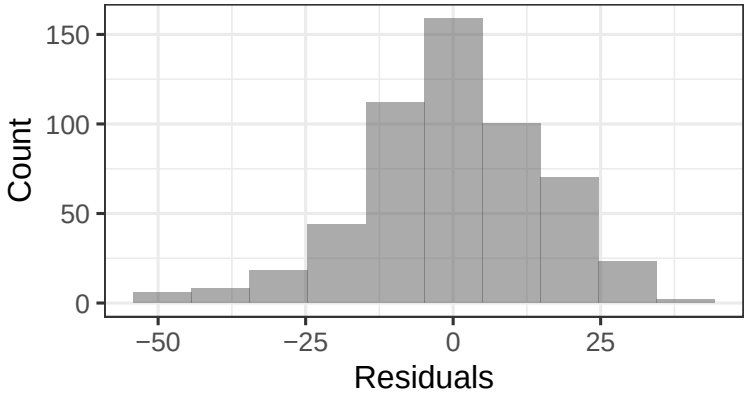
```
nerds2 <- nerds %>%
  mutate(preds = predict(nerds_lr), resids = resid(nerds_lr))

gf_point(resids~ preds, data = nerds2) %>%
  gf_labs( x = "Predictions", y = "Residuals")
```



This residuals vs. fitted plot checks the lack of non-linearity condition. This condition does not seem to be met as there is a trend in the graph. There is clear vertical lines and the data does not seem to be scattered randomly.

Normality

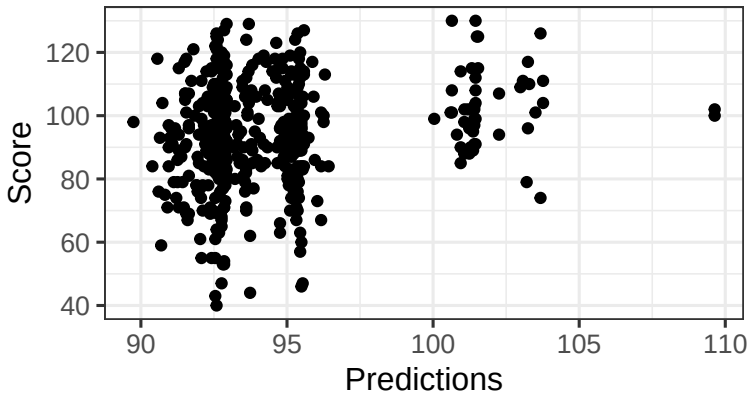
```
gf_histogram(~resids, data = nerds2, bins = 10) %>%
  gf_labs(x = "Residuals", y = "Count")
```



The condition we are checking here is normality of the residuals. The residuals are unimodal and normally distributed. This condition is met.

Conclusion

```
gf_point(score~ preds, data = nerds2) %>%
  gf_labs( x = "Predictions", y = "Score")
```



Since one of our LINE conditions are not met (lack of non-linearity), it is impossible to determine whether there is an association between Nerdiness Score and Handedness. However, given our new prediction plot and that the LINE conditions are met, we do not see a linear trend in the scatter plot. This means that the predictors have little effect on the response variable. Our very low R squared value shows us something similar: age, gender, and handedness have little effect on a person's nerdiness score. Our boxplots showed us that the scores had similar distributions in regards to handedness. Our lack of conditions being met could be a result of there not being enough observations at each handedness since there are not as many left-handed or ambidextrous people as right-handed people (both in general and in the data set). However, given that our conditions are met and based on all of our graphs, I think it is fair to say that there is little to no association between someone's nerdiness score and their handedness.