Lesson 19: ACFT Problem (solution)

Is there an association between the Standing Power Throw and Maximum Deadlift raw scores?

We are going to determine if an association exists between `spr\_raw` and `mdl\_raw` scores.

The relationship we are modeling looks like this:

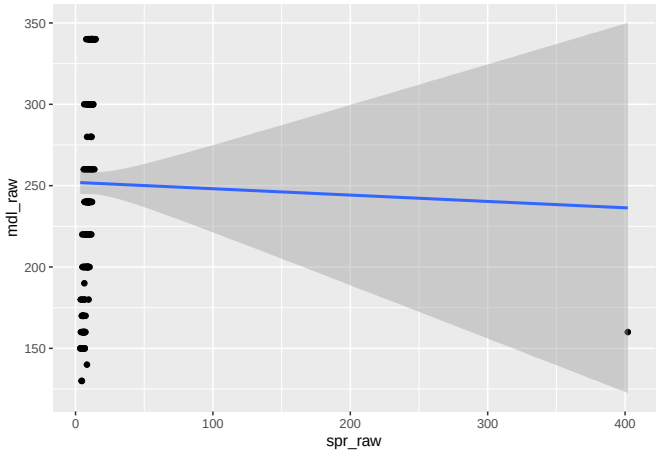
$$\widehat{mdl} = \beta_0 + \beta_1 \cdot spr$$

1. Provide the hypotheses in symbols and words:

$H_0 : \beta_1 = 0$. My null hypothesis is that there is no association between 'spr_raw' and 'mdl_raw'
$H_A : \beta_1 \neq 0$. My null hypothesis is that there is an association between 'spr_raw' and 'mdl_raw'.

2. Plot the association, including the regression line.

```
df %>%
  ggplot(aes(x = spr_raw, y = mdl_raw)) +
  geom_point() +
  geom_smooth(method = "lm")
```



3. Is there anything unusual about the data?

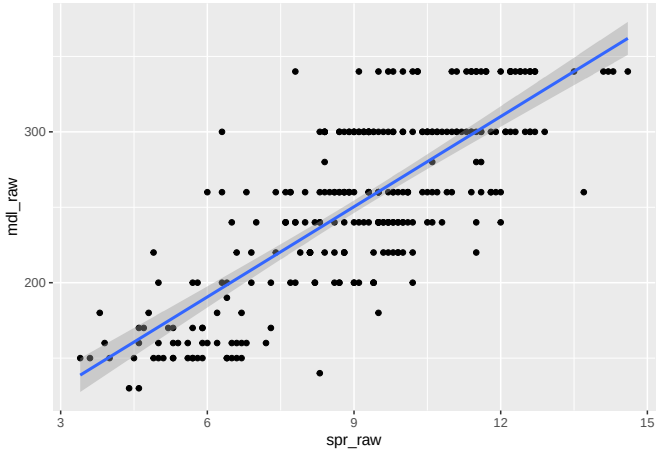
Yes. Outlier.

4. Remove the outlier using the `filter( )` command. How many observations were dropped?

```
df <- df %>% filter(spr_raw < 100)
```

5. Create a new scatter plot with updated dataframe.

```
df %>%
  ggplot(aes(x = spr_raw, y = mdl_raw)) +
  geom_point() +
  geom_smooth(method = "lm")
```



6. Create a linear model to test your hypotheses.

```
lrmodel <- df %>%
  lm(mdl_raw ~ spr_raw, data = .)
summary(lrmodel)
```

```
##
## Call:
## lm(formula = mdl_raw ~ spr_raw, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -96.394 -24.950  -2.174  23.672 113.572
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  70.9488     8.7017   8.153 1.08e-14 ***
## spr_raw      19.9332     0.9299  21.436  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.03 on 290 degrees of freedom
## Multiple R-squared:  0.6131, Adjusted R-squared:  0.6117
## F-statistic: 459.5 on 1 and 290 DF,  p-value: < 2.2e-16
```

7. Write out the linear model with the coefficients determined by the linear regression.

$\widehat{mdl} = 70.95 + 19.93 \cdot spr$

8. Interpret the intercept term.

With a 'spr_raw' of 0, we can expect an 'mdl_raw' of 70.95lbs. This does not make sense because we are extrapolating outside of our data.

9. Interpret the slope term.

On average, every one unit (meter) increase in Standing Power Throw is associated with an increase of Maximum Deadlift by 19.93lbs, holding all else equal.

10. Conclusion of your hypothesis test at a 5% significance level.

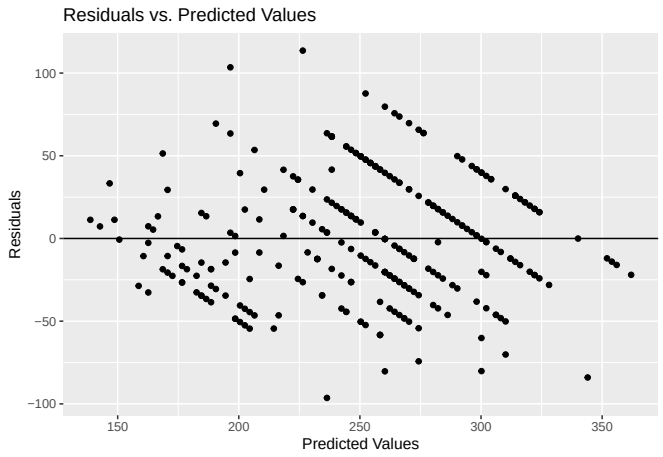
I reject my null at the 5% significance level that there is no association between ‘spr_raw‘ and ‘mdl_raw‘.

11. Interpret the R-squared value.

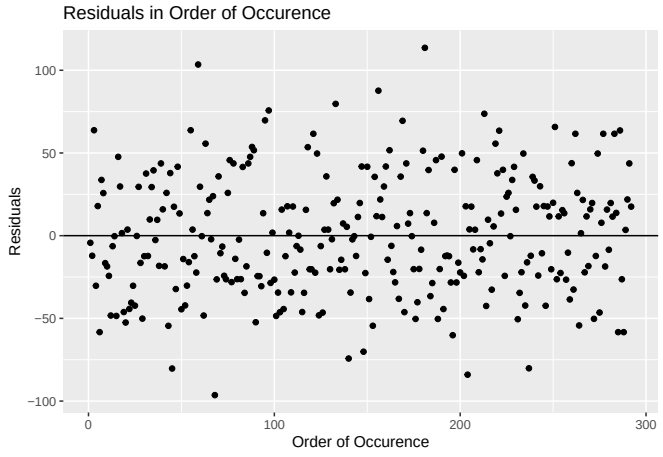
About 61% of the variation in ‘mdl_raw‘ is explained by ‘spr_raw‘.

12. Check the 4 Validity Conditions (L.I.N.E.). Are any not met?

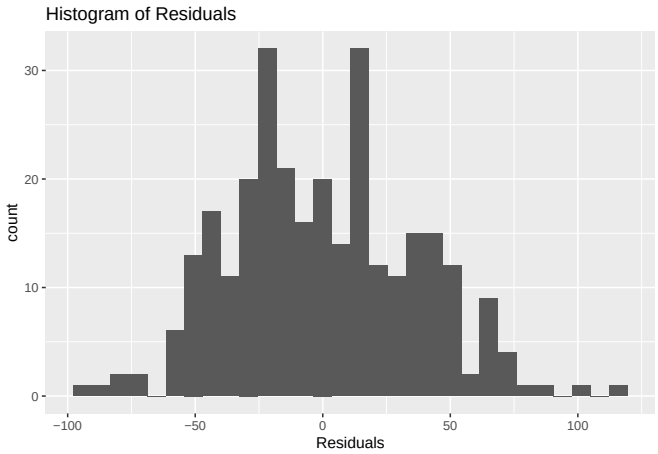
```
lrmodel %>%
  fortify(lrmodel$model) %>%
  ggplot(aes(x = .fitted,
             y = .resid)) +
  geom_point()+
  geom_hline(yintercept = 0) +
  labs( x = "Predicted Values",
        y = "Residuals",
        title = "Residuals vs. Predicted Values")
```




```
lrmodel %>%
  fortify(lrmodel$model) %>%
  mutate(row = row_number()) %>%
  ggplot(aes(x = row,
             y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  labs(x = "Order of Occurence",
       y = "Residuals",
       title = "Residuals in Order of Occurence")
```



```
lrmodel %>%
  fortify(lrmodel$model) %>%
  ggplot(aes(x = .resid)) +
  geom_histogram() +
  labs(x = "Residuals",
       title = "Histogram of Residuals")
```



While there are "bands" in this data, that is due to the limited number of 'mdl_raw' values. Overall, the plot appears to be somewhat linear but variance is not constant! There does not appear to be independence issues and the histogram of the residuals seems fairly normal/bell-shaped.

13. Can we determine causation? Why or why not.

No, because this was an observational study and there was no random assignment.

Does adding the variable `sex` change the association between the Standing Power Throw (`spr\_raw`) and Leg Tuck (`ltk\_raw`) raw scores?

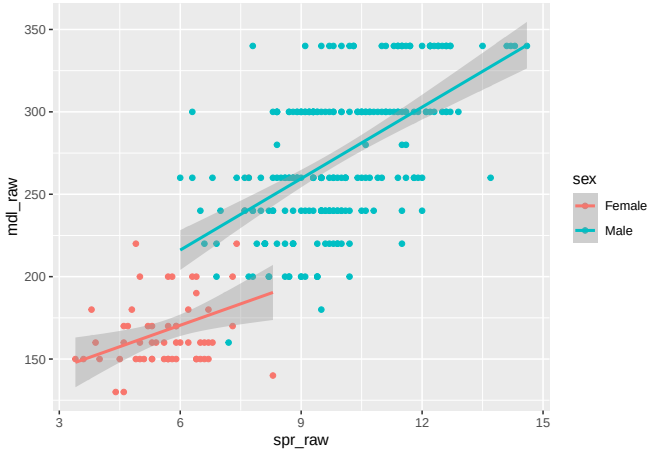
Now, we add `sex` as another variable to the model but also include it as an interaction with `spr\_raw`.

`Our updated model looks like this:`

$$\widehat{mdl} = \beta_0 + \beta_1 \cdot spr + \beta_2 \cdot sexmale + \beta_3(spr \cdot sexmale)$$

14. Plot the data **(including the interaction term)**, including the regression line.

```
df %>%
  ggplot(aes(x = spr_raw, y = mdl_raw, color = sex)) +
  geom_point() +
  geom_smooth(method = "lm")
```



15. Does there appear to be a different relationship between `mdl\_raw` and `spr\_raw` based on `sex`?

Maybe. The lines don't have a very different slope.

16. Conduct a linear regression with the updated model.

```
lrmodel <- df %>%
  lm(mdl_raw ~ spr_raw * sex, data = .)
summary(lrmodel)
```

```
##
## Call:
## lm(formula = mdl_raw ~ spr_raw * sex, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -86.729 -24.267  -1.634  27.365  97.850
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      118.529     24.401   4.858 1.95e-06 ***
## spr_raw            8.662      4.207   2.059   0.0404 *
## sexMale           10.844     28.003   0.387   0.6989
## spr_raw:sexMale    5.797      4.423   1.310   0.1911
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.84 on 288 degrees of freedom
## Multiple R-squared:  0.6611, Adjusted R-squared:  0.6576
## F-statistic: 187.3 on 3 and 288 DF,  p-value: < 2.2e-16
```

17. Interpret the coefficient on the interaction term.

The coefficient indicates that, for Males, a one-unit increase in 'spr_raw' results in an 8.66+5.797 increase in 'mdl_raw' , on average, holding all else constant.

18. Is this significant at the 5% level?

No, because our p-value (.1911) exceeds our significance level.

19. Interpret the intercept term.

With a zero ‘spr_raw‘ , Females will exhibit a ‘mdl_raw‘ of 119lbs.

20. Interpret the base slope term for `spr\_raw`.

For Females, a one-unit increase in ‘spr_raw‘ results in an 8.67 increase in ‘mdl_raw‘, on average, holding all else equal.

21. Using this model (even with no significance), calculate the following:

- The `mdl\_raw` for a male cadet with an `spr\_raw` of 6m:

```
(118.529 + 10.84) + 6*(8.662 + 5.797)
```

```
## [1] 216.123
```

- The `mdl\_raw` for a female cadet with an `spr\_raw` of 6m:

```
(118.529) + 6*(8.662)
```

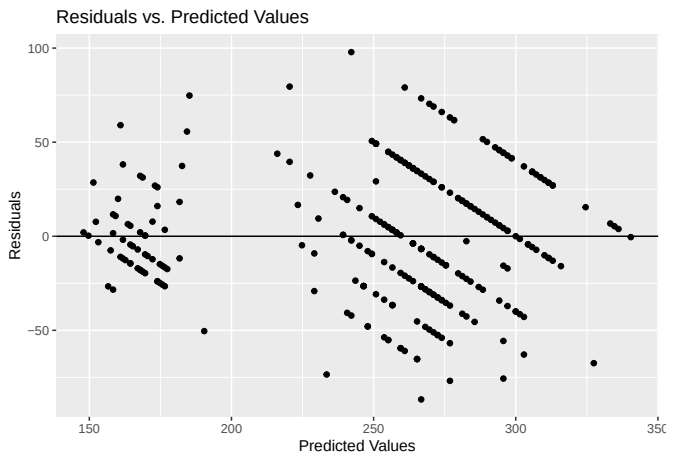
```
## [1] 170.501
```

22. Interpret the R-squared value.

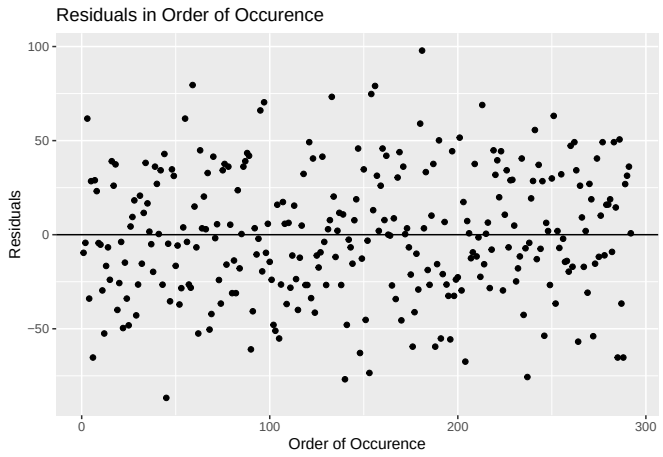
About 66% of the variation in ‘mdl_raw‘ is explained by ‘spr_raw‘ and sex and the interaction between the two.

23. Check the 4 Validity Conditions (L.I.N.E.). Are any not met?

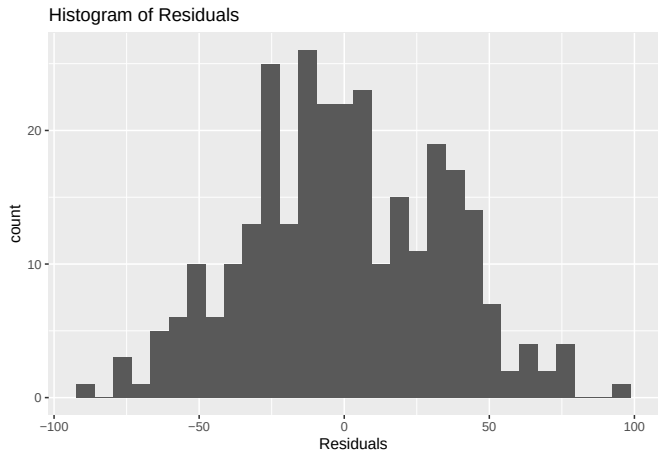
```
lrmodel %>%
  fortify(lrmodel$model) %>%
  ggplot(aes(x = .fitted,
             y = .resid)) +
  geom_point()+
  geom_hline(yintercept = 0) +
  labs( x = "Predicted Values",
        y = "Residuals",
        title = "Residuals vs. Predicted Values")
```



```
lrmodel %>%
  fortify(lrmodel$model) %>%
  mutate(row = row_number()) %>%
  ggplot(aes(x = row,
             y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  labs(x = "Order of Occurence",
       y = "Residuals",
       title = "Residuals in Order of Occurence")
```




```
lrmodel %>%
  fortify(lrmodel$model) %>%
  ggplot(aes(x = .resid)) +
  geom_histogram() +
  labs(x = "Residuals",
       title = "Histogram of Residuals")
```



The plots look very similar to before. However, there are two-groups of residuals now...