MA206, Lesson 20 - Multiple Linear Regression

Review: What is the form of our linear regression model with 1 explanatory variable?
$\hat{y} = \beta_0 + \beta_1 X_1$

If our model is valid, how do we expect our residuals to be distributed?
We expect our residuals to be normally distributed and centered at 0 with a constant variance.
$\varepsilon \approx \mathcal{N}(0, \sigma^2)$

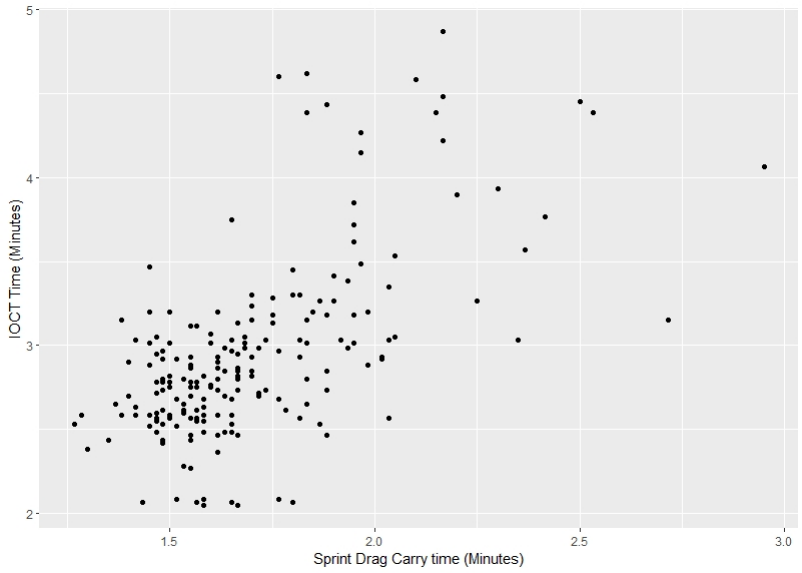
What are the validity conditions for our regression line?
Linearity - The residuals vs. predicted values graph does not show any strong evidence of patterns.
Independence - The responses can be considered independent of one another.
Normality - The histogram of the residuals is approximately symmetric with no large outliers.
Equal Variance - the residuals vs. predicted values graph shows a constant width.

How would we interpret the affects of variable X_1 on y, given the equation $\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$?
"For every unit increase in X_1, we expect y to change by β_1, on average, after adjusted for the variable X_2."

1) Use the **ACFT2.csv** dataset to examine more inferences we can make with regards to IOCT time.

a) We want to assess if the Sprint Drag Carry may have an association with IOCT time. Generate a scatterplot between SDC_Raw and $IOCT\_time$ and comment on the Form, Direction, Strength, and any unusual observations.

Form - It appears to be linear.
Direction - It has a positive direction.
Strength - It looks to be moderately strong with more variation towards later times.



b) Report the correlation between the Sprint Drag Carry and IOCT time. Does this coincide with your prevoius assessment?

cor(SDC, IOCT) = 0.6197. This coincides with our visual assessment above.

c) Generate a linear model using the Sprint Drag Carry as your explanatory variable and the IOCT time as the response. Write down the resulting equation. How do you interpret the slope?

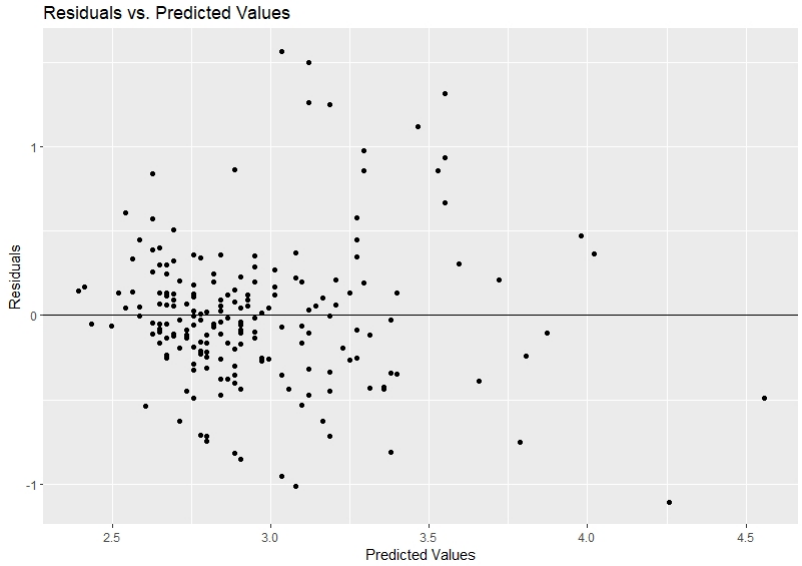
$\widehat{IOCT} = 0.7605 + 1.2873 \times SDC$

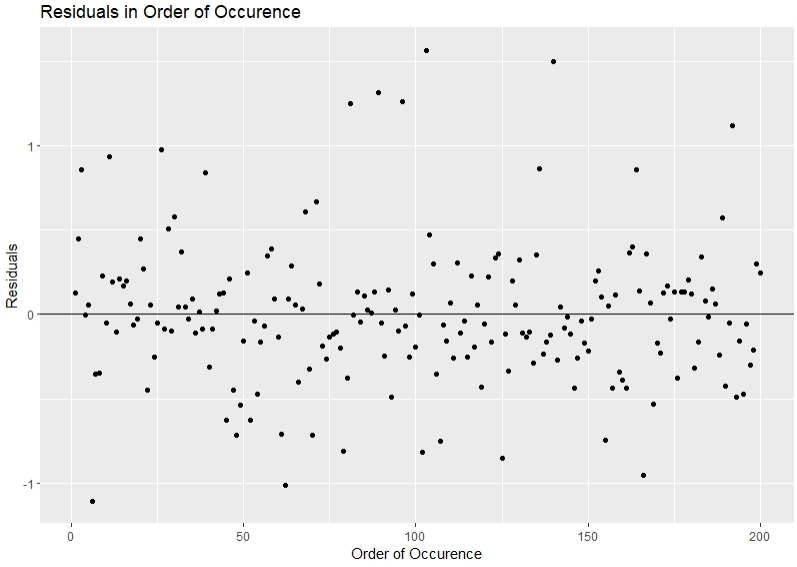
For every additional minute in the Sprint Drag Carry, the expected time to complete the IOCT will increase by 1.2873 minutes.

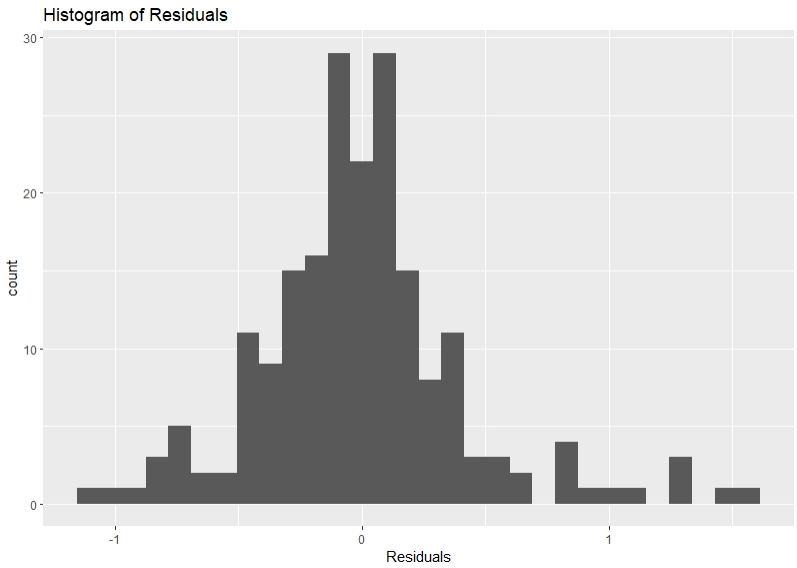
d) Report the R^2 and explain what it means, in words.

$R^2 = 0.384$. Roughly 38% of the variation in our data can be explained by our model (changes in Sprint Drag Carry time).

e) Validate the validity conditions and report on your findings.


Residuals vs. Predicted Values
Residuals
1
0
-1
2.5
3.0
3.5
4.0
4.5
Predicted Values


Residuals in Order of Occurence
Residuals
1
0
-1
0
50
100
150
200
Order of Occurence

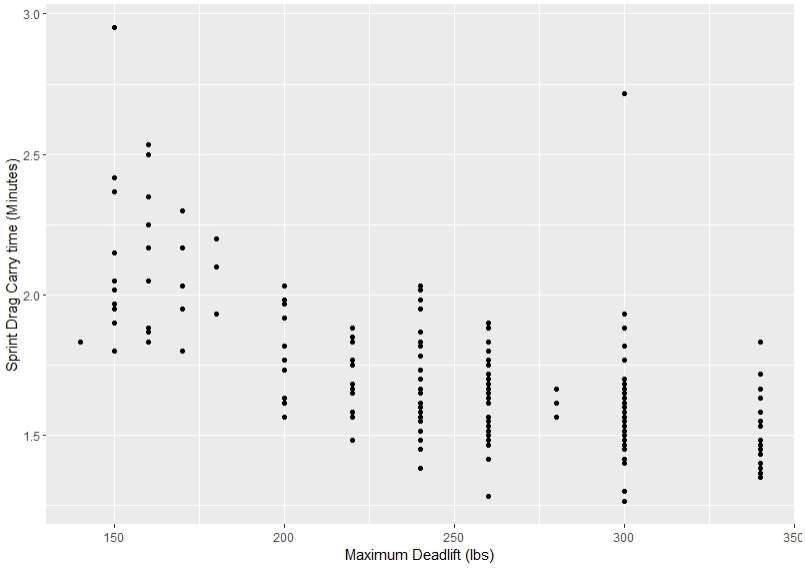


There are no obvious patterns indicating it is nonlinear (chart 1)
There is no obvious pattern in the order of occurrence, supporting independence (chart 2)
It appears to be roughly normal and centered on zero with some slight right skew (chart 3)
There does appear to be some above average outliers (5 points), but overall equal variance (chart 1)

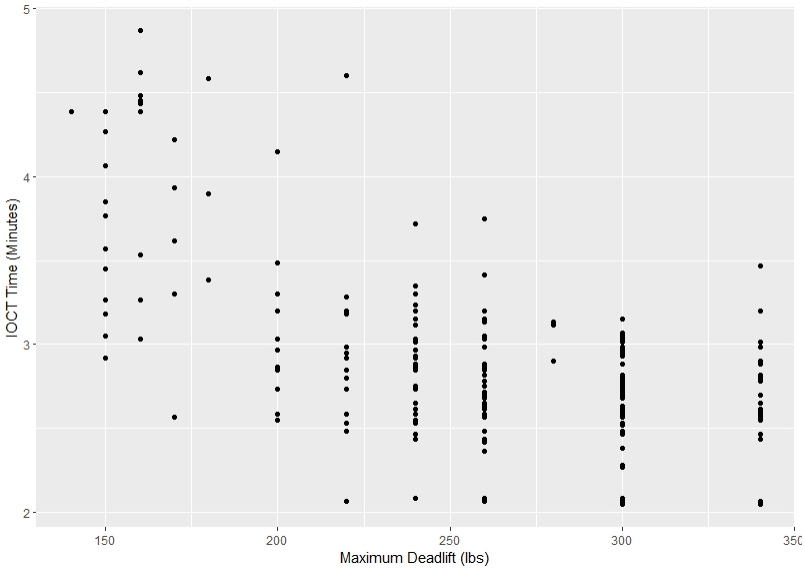
2) Now we want to assess if the Maximum Dead Lift is a potential confounding variable.

a) Generate two scatterplots comparing MDL_Raw with SDC_Raw as well as comparing MDL_Raw with $IOCT\_time$. Comment on the Form, Direction, and Strength for both plots.

Sprint Drag Carry versus Maximum Deadlift
Form - Linear
Direction- Negative
Strength - Moderate



Maximum Deadlift versus IOCT Time
Form - Linear
Direction- Negative
Strength - Moderate



b) Based on the previous graphs, might Maximum Deadlift be a confounder? Explain.

It is possible, as it seems to be similarly correlated with both our explanatory variable (SDC) and our response variable (IOCT Time).

c) Report the correlation between the three pairs of variables. Do these results coincide with your

previous assessments of Form, Direction, and Strength?

cor(SDC, MDL) = -0.6538
cor(SDC, IOCT) = 0.6197
cor(MDL, IOCT) = -0.5968
These coincide with our visual assessments above.

d) Generate a linear model using Sprint Drag Carry and Maximum Deadlift to explain IOCT Time. Write the resulting equation and interpret the slope for Sprint Drag Carry.

$\widehat{IOCT} = 2.3625 - 0.0032MDL + 0.8326SDC$
For every additional minute in the Sprint Drag Carry, the expected time to complete the IOCT will increase by 0.8326 minutes, on average, adjusting for Maximum Deadlift.

e) Did the coefficient for Sprint Drag Carry change between your models? Discuss what that means.

Yes, our slope decreased from 1.29 to 0.83. This indicates that controlling for Maximum Dead Lift changes our estimated impact from Sprint Drag Carry, or SDC is *counfounded* by MDL.

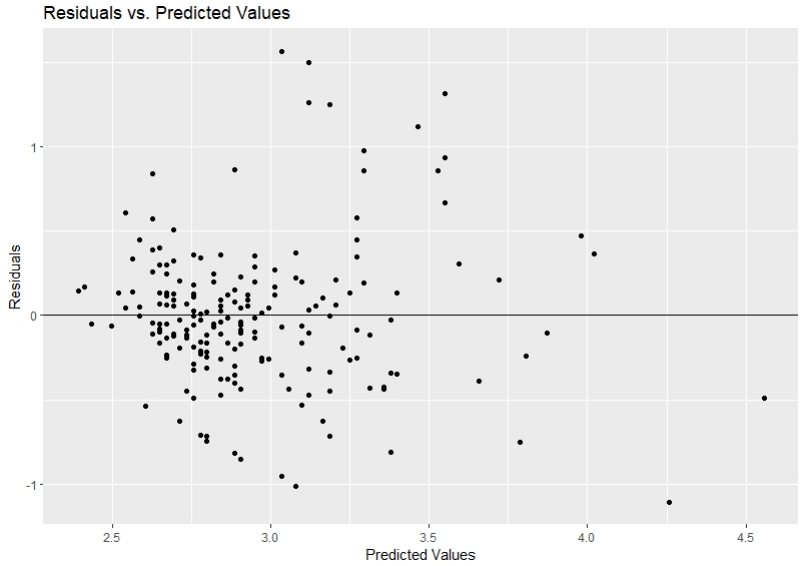
f) Report the R^2 and explain what it means.

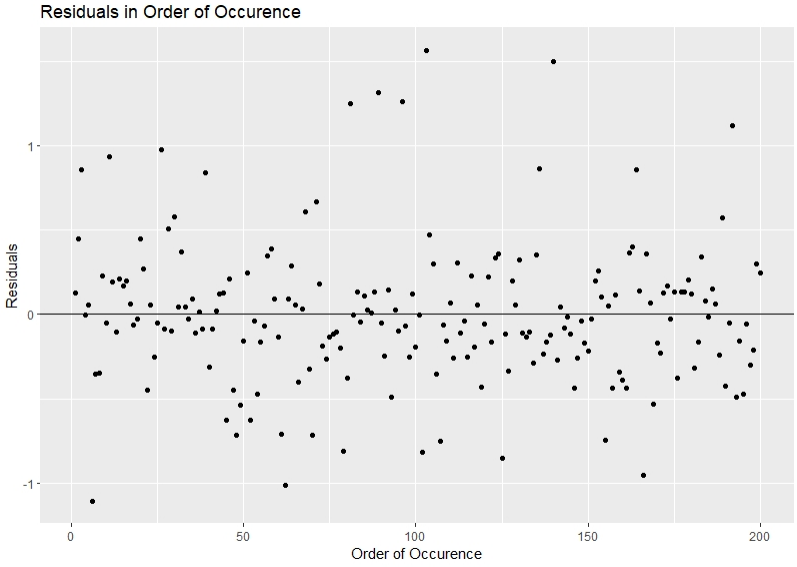
$R^2 = 0.4482$. Roughly 45% of the variation in our IOCT times can be explained by our model (changes in Sprint Drag Carry and Maximum Deadlift).

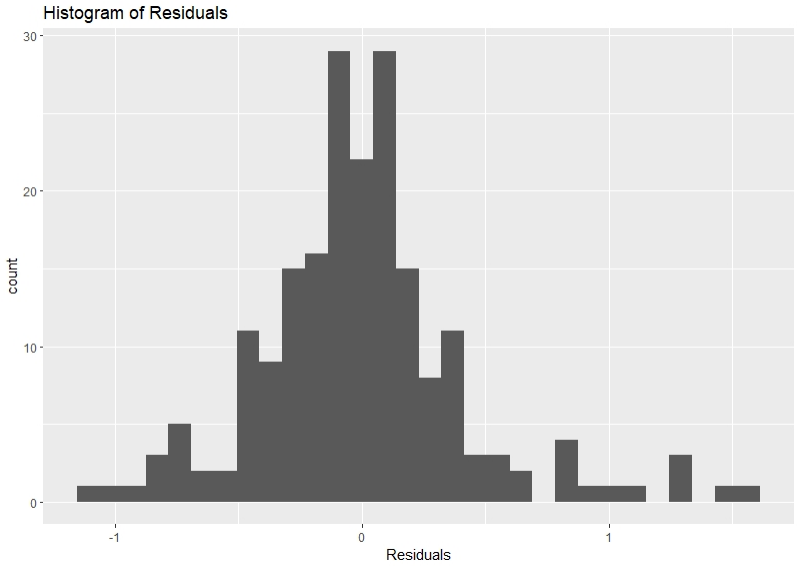
g) What are the p-values associated with each coefficient (β_1 and β_2) and what do they mean?

β_1 (MDL): p-value = $3.34e^{-6}$. The probability of observing a coefficient estimate at least as extreme as -0.0032 if the null hypothesis is true ($\beta_1 = 0$, no linear association) is 0.00000334.
β_2 (SDC): p-value = $3.71e^{-8}$. The probability of observing a coefficient estimate at least as extreme as 0.8326 if the null hypothesis is true ($\beta_2 = 0$, no linear association) is 0.0000000371.

h) Validate the validity conditions and report on your findings.


Residuals vs. Predicted Values
Residuals
1
0
-1
2.5
3.0
3.5
4.0
4.5
Predicted Values


Residuals in Order of Occurence
Residuals
1
0
-1
0
50
100
150
200
Order of Occurence



There are no obvious patterns indicating it is nonlinear (chart 1)
There is no obvious pattern in the order of occurrence, supporting independence (chart 2)
It appears to be roughly normal and centered on zero with some slight right skew (chart 3)
There does appear to be some above average outliers (2 points), but overall equal variance (chart 1)

3) We suspect the categorical variable *Sex* is confounding, so we want to include that in our analysis as well.

a) Generate a linear model using Sprint Drag Carry, Maximum Deadlift, and Sex to explain IOCT Time. Write the resulting equation and interpret the slope for Sprint Drag Carry.

$\widehat{IOCT} = 2.9299459 - 0.0011126(MDL) + 0.4855389(SDC) - 0.6348586(Sex = MALE)$
For every additional minute in the Sprint Drag Carry, the expected time to complete the IOCT will increase by 0.4855 minutes, on average, adjusting for Maximum Deadlift and Sex.

b) List the p-values for each of our three variable coefficients $(\beta_1, \beta_2, \beta_3)$ and discuss the implications.

β_1 **(MDL)**: When fixing for SDC and Sex, the probability of observing a coefficient at least as extreme as -0.00111 is 0.127. We would say this is weak evidence against the null hypothesis that there is no association.
β_2 **(SDC)**: When fixing for MDL and Sex, the probability of observing a coefficient at least as extreme as 0.4855 is 0.00113. We would say this is very strong evidence against the null hypothesis that there is no association.
β_3 **(Sex)**: When fixing for MDL and SDC, the probability of observing a coefficient at least as extreme as -0.6349 is $2.01e^{-8}$. We would say this is very strong evidence against the null hypothesis that there is no association.

Of note, when fixing for SDC and Sex, we find that MDL no longer gives statistically significant inference

about IOCT time.

4) The dataset found below lists different sets from Lego along with the price, number of reviews, number of pieces, and if it is licensed from a Disney, Marvel, DC, or Star Wars franchise. We want to determine if the number of pieces can accurately predict the price of Lego sets.

```
Legos <- read_csv("https://raw.githubusercontent.com/rslasater82/MA206Datasets/main/legos.csv")
```

a) What is the correlation between pairs for *pieces*, *reviews*, and *price*?

cor(pieces, reviews) = 0.4957703
cor(pieces, price) = 0.9381424
cor(reviews, price) = 0.4228281

b) Create a linear regression model using pieces and reviews to calculate price. Report your equation.

$\widehat{price} = 7.222043 + 0.087361(pieces) - 0.07298(reviews)$

c) Create a new model including Licence as a variable and report your equation.

$\widehat{price} - 11.508201 + 0.087231(pieces) - 0.075685(reviews) - 6.0967(Non - Licensed)$.

d) Based on the above equations, would you consider *Licensed* a confounding variable? Explain why.

No. The coefficients for pieces and reviews did not change when adding or removing *Licensed*, indicating that they are not directly related.