The dataset **ironman1819.csv** contains data on female finishers of the Lake Placid Ironman Triathlon from 2002 to 2021. The motivation for this data analysis is to explore the relationship between bike times and run times (in minutes) in order to gain insights into the performance patterns of the athletes. For this activity, we will specifically focus on times from Canadian finishers in the years 2018 and 2019.

Model 1: Bike Times

1. Fit and report the least squares regression equation for predicting **Run Time** using **Bike Time** (in minutes).
2. Examine residual plots for this model. Do you have any concerns about the appropriateness of this linear model?

There is no curvature or other trends in residual vs. fits plot – Linearity  
There is no fanning or funneling in residuals vs. fits plot – Equal Variance  
The Normal Probability plot follows the expected linear trend, but there are some data that may be of concern

Overall, there are no concerns about the appropriateness of a linear model

1. Test (include all steps) if there is evidence that **Bike Time** (in minutes) is a useful predictor of **Run Time**?

Step 1:

Step 3: Test Statistic

t = 8.83  
df = 41  
p-value = 0.000

Step 2: Conditions

Already checked in Question 2

Step 4: Conclusion  
  
Reject the Null

There is very strong evidence that Bike Tome is a useful predictor of Run Time

Model 2: Swim Times

1. Fit and report the least squares regression equation for predicting **Run Time** using **Swim Time** (in minutes).
2. Examine residual plots for this model. Do you have any concerns about the appropriateness of this linear model?

There is no curvature or other trends in residual vs. fits plot – Linearity  
There is no fanning or funneling in residuals vs. fits plot – Equal Variance  
The Normal Probability plot follows the expected linear trend, but there are some data that may be of concern

Overall, there are no concerns about the appropriateness of a linear model

1. Construct and interpret a 95% confidence interval for the population slope relating **Run Time** and **Swim Time**.

(2.5104 , 4.6476)

Conditions: Already checked in Question 5

n = 43  
df = 41  
t\* = 2.02

With 95% confidence, when the Swim Time increases by 1 minute, the predicted Run Time will increase between 2.5105 and 4.6476 minutes

1. Based on your confidence interval, is there evidence that **Swim Time** is a useful predictor of **Run Time**? Explain briefly.

Yes. Because 0 (meaning not useful) is not contained in the confidence interval

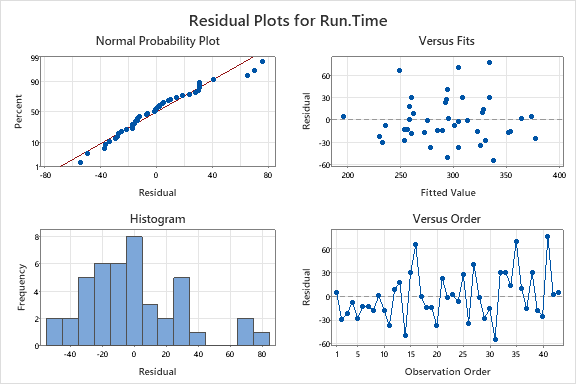
Model 3: Both

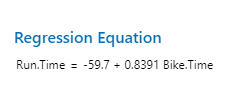
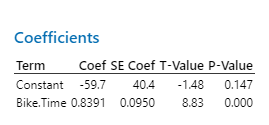
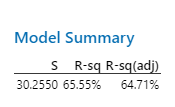
1. Now put both **Bike Time** and **Swim Time.** (in minutes) in the model as predictors of **Run Time**. Report the resulting equation below. This is a *multiple linear regression model*.
2. Predict the **Run Time** of a triathlete with a Bike Time of 385 minutes and a Swim time of 71 minutes
3. Contrast the output from this multiple linear regression model with the output from Models 1 and 2. What differences do you notice? Why might this be?

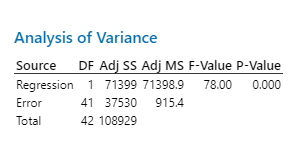
Answers may vary  
  
Sample Response: The biggest difference from the new output is that Swim Time is no longer significant, but when it was in a model by itself, it was a useful predictor. Possible reasons for this is that Swimming and Biking are also correlated with one another, so you only need one of them to predict Run Times, and Bike Times just happens to do a better job.

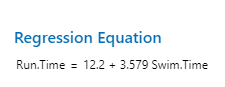
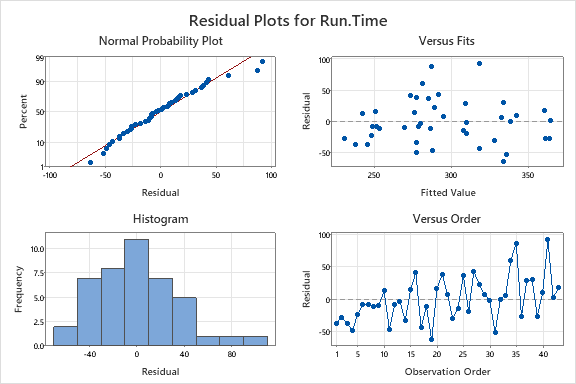
1. Which model would you recommend using and why?

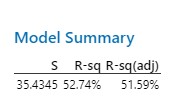
Answers will likely involve comparing R-Squared and/or significance values.

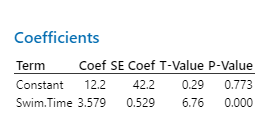
  
Model 1 Output

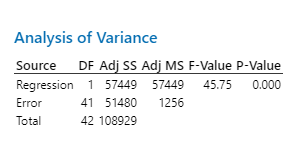


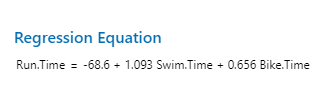


Model 2 Output







Model 3 Output

