NASCAR consistently collects data on driver performance (<https://www.nascar.com/stats/>). Including in this data are the season statistics from 2007-2022 with variables such as wins, average start, average mid race, average finish, average position, pass difference, green flag passes, green flag passed, quality passes, percent of quality passes, number of fastest laps, laps in top 15, percent of laps in top 15, laps led, percent of laps led, total laps, driver rating, and points. The goal of this quiz is to use some of the available explanatory variables to predict the driver rating. See <https://www.nascar.com/stats/glossary> for more information about each of the variables.

* 1. (1 pt) Run the first two (provided) chunks of code provided in the file. Briefly describe what the second chunk does to the data.
  2. (1 pt) Using proper notation, write down the population model predicting driver rating using the percent of laps in top 15, percent laps led, average start, and average finish.
  3. (1 pt) Fit a linear regression model for predicting median sale pricefrom median household income. Report that fitted equation below using proper notation.
  4. (1 pt) Are there any issues with multicollinearity in this model? Support your answer with the appropriate metrics (and explanation).
  5. (2 pts) Fit a model to predict driver rating using Percent of Laps Led that produces parallel regression lines for the two levels of *WinIND*. Report the fitted equation below (using proper notation).
  6. (2 pts) Interpret the intercept for the line where *WinIND* takes the value of 0. Be specific about what this means for the context of the application. (e.g., don’t use the phrase *WinIND*…instead include its meaning in the interpretation)
  7. (2 pts) Is there a statistically significant difference (or shift) in the intercept between the two levels of *WinIND*? Provide the details for the appropriate hypothesis test. (Note, you do not need to manually recreate calculation for this question that R already provides.)

1. Consider again the NASCAR data from Quiz 5.

NASCAR consistently collects data on driver performance. Including in this data are the season statistics from 2007-2022 with variables such as wins, average start, average mid race, average finish, average position, pass difference, green flag passes, green flag passed, quality passes, percent of quality passes, number of fastest laps, laps in top 15, percent of laps in top 15, laps led, percent of laps led, total laps, driver rating, and points. The goal of this quiz is to use some of the available explanatory variables to predict the driver rating. See <https://www.nascar.com/stats/glossary> for more information about each of the variables.

We will now consider using a driver’s percent of laps in top 15 and their average finish to predict their rating.

* 1. (2 pts) Using proper notation, record the population form of the complete second order model using these variables of interest.
  2. (2 pts) Is there evidence that the interaction is needed in this model (given the other components of the model are already in it)? Include all the relevant pieces of the appropriate hypothesis test. (Note: You do not need to redo any calculations that R has already performed for you.)
  3. (3 pts) Is there evidence that the avgMidRace variable is needed in this analysis? Provide the details of the appropriate hypothesis test. While you may use R to calculate the results, provide the details of the test by showing which numbers would be plugged in where in the appropriate formula(s).

1. NASCAR consistently collects data on driver performance. Including in this data are the season statistics from 2007-2022 with variables such as wins, average start, average mid race, average finish, average position, pass difference, green flag passes, green flag passed, quality passes, percent of quality passes, number of fastest laps, laps in top 15, percent of laps in top 15, laps led, percent of laps led, total laps, driver rating, and points. See <https://www.nascar.com/stats/glossary> for more information about each of the variables. These data are available in the **nascar** object loaded in the Rmd file associated with the exam**.**
   1. (2 pts) Fit a quadratic regression model using average finish to predict driver’s rating. Using proper notation, report the fitted equation here.
   2. (2 pts) Fit a simple linear model using the natural logarithm of driver’s rating as the response (and keeping their average finish as the explanatory variable). Using proper notation, report the fitted equation here.
   3. (2 pts) Which model (a) or (b) is best? Explain.
   4. (2 pts) Do you have any additional concerns with your “best model”. Explain.
   5. (2 pts) Use the model from part a to construct a 95% confidence interval for the average driver’s rating for NASCAR racers that have an average finish of 10. If using R, provide the command(s) you used to get the interval. If using “by hand” show your work. (If doing a combination of both, show both.)
   6. (2 pts) Use the model from part b to construct a 95% prediction interval for the average driver’s rating for NASCAR racers that have an average finish of 10. If using R, provide the command(s) you used to get the interval. If using “by hand” show your work. (If doing a combination of both, show both.)
      * Tip: The “invariance property” that we discussed when working with confidence intervals in logistic regression applies to prediction intervals too.