ST 352

**Lab Assignment 3**

***31 points***

***Due 11:59 PM on Friday, October 25***

## Problem 1: (24 points)

During Week 4, you participated in an activity to test your reaction time. For this problem, we’ll investigate the relationship between age and reaction time. *In particular, we’ll investigate if there is evidence to indicate a person’s age helps to predict their reaction time.*

The data from the activity are stored in the **reaction** data set on Canvas. There are two variables in the data set:

* age (in years)
* time (average of three reaction times in milliseconds)

Use the data in the data set to answer the following questions on a separate document.

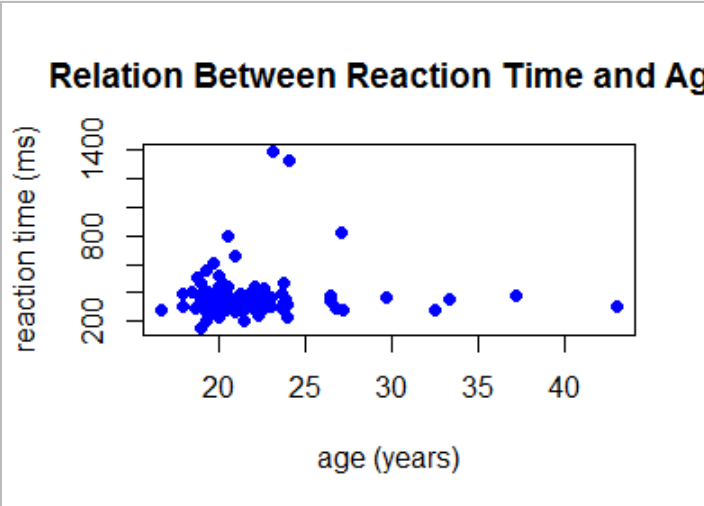
# Assessing conditions

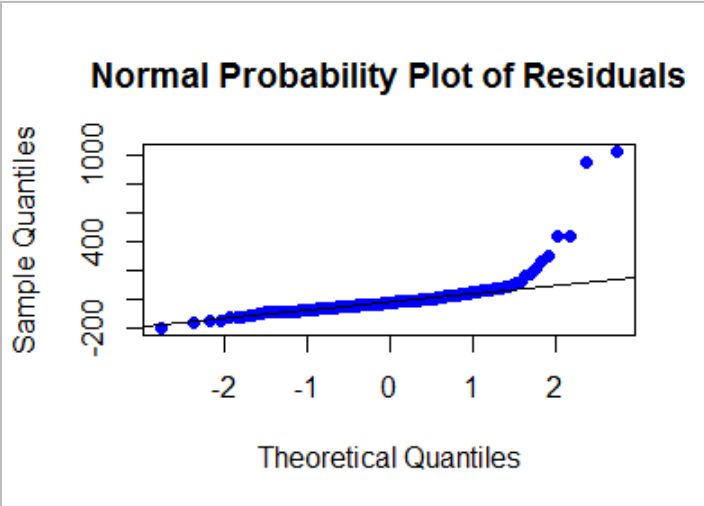
Questions are only about the conditions that can be assessed graphically. Obtain each of the following graphs in R. (Use the R code from the Lab 3A Notes to generate these plots.)

* Scatterplot
* Residual plot
* Normal probability plot of the residuals

1. ***(2 points)*** Copy and paste all three properly labeled plots here.

**(2 points for correct graphs with labels. If missing 1 graph, - 1 point. If missing 2 graphs, -1.5 points)**



1. ***(2 points)*** Are there any outliers or other deviations from the pattern? Which plot(s) are you using to assess the outlier condition?

There seem to be a couple of outliers. There are a couple of students who had unusually high reaction times. An argument can also be made there are some outliers with higher ages – some a bit older than a traditional student. All of this is a bit subjective, so an answer of no outliers can be accepted as well.

(1 pt for stating whether or not there are outliers and 1 pt for reference to scatterplot or residual plot)

1. ***(2 points)*** Is the linearity condition satisfied? Explain. In your explanation, reference the plot(s) you are using to assess the linearity condition.

Using the scatterplot (or residual plot), there is no strong curvature, although the outlier and clusters may make it hard to assess this (and other) conditions. We’ll accept an answer of “linear” or “non-linear” as long as there is correct support.

(1.5 pts for explanation, 1/2 pt for reference to scatterplot or residual plot.

* Note: if student supports their answer by saying the “spread” is similar (or different) in the residual plot, 1.5 points should be deducted as the spread of the residuals is not used to assess the linearity condition.
* If a student uses the normal probability plot of the residuals to assess this condition, deduct 2 points.

1. ***(2 points)*** Is the constant variation condition satisfied? Explain. In your explanation, reference the plot(s) you are using to assess the constant variation condition.

This may be less subjective than the linearity condition. Using the residual plot, the cluster of points with lower ages has more spread in their reaction times than students older than 30. Even though this seems like the most reasonable answer, look for proper support.

(1.5 pts for explanation, 1/2 pt for using residual plot)

* If student tries to use normal probability plot to assess this condition, deduct 2 points.

1. ***(2 points)*** Is the normality condition satisfied? Explain. In your explanation, reference the plot(s) you are using to assess the normality condition.

Using the normal probability plot, most of the residuals are on or close to the reference line, supporting that the normality condition is satisfied. There may be some argument that the points at the right side of the plot “tail off” of the line and get too far away from the line to be comfortable saying the condition is satisfied. Again, we’ll look for support – give full credit for reasonable support.

(1.5 points for reasonable explanation, ½ point for referencing normal probability plot)

* **Deduct two points if student uses any other plot to assess this condition**

1. ***(2 points)*** Is a transformation necessary in this example? Explain.

Answers should depend on how they answered questions 3 through 5 above. If a student thought that at least one of the linearity, constant variation, and/or normality conditions was violated, they should answer “yes” to this question with the support that at least one condition was violated. If a student thought all three conditions were satisfied, they should answer “no” to this question with the support that all conditions were satisfied.

(2 points for reasonable explanation)

* Note: a student should NOT say a transformation is necessary because of the outlier. If they do, 1 point should be deducted. They could mention that conditions may not be satisfied because of the outlier, but a transformation is not performed just because there is an outlier.

# Inference in Regression

**Regardless of your answer to #6, answer the remaining questions of this problem assuming no transformation is necessary.**

Obtain output from the linear regression model. (Reminder, you generated a linear model in order to create the residual plot.) Refer to the Lab 3A Notes for the R code to obtain output from the linear regression model. *Do not include the output in your assignment!*

1. ***(2 points)*** Write the least-squares regression equation. Define the terms in the equation.

where *x* is the age of a person (years) and is the predicted reaction time (milliseconds)

(1 point for correct equation, 1 point for defining terms.)

* **If students had y-intercept and slope reversed, deduct ½ point.**
* **If student does not include units in their definition of the variables, deduct 1/4 point**
* **If a student does not include “predicted” when defining**  , deduct 1/4 point.

1. Is there evidence that a person’s age can be used to predict their reaction time?
   1. ***(2 points)*** State the null and alternative hypotheses in notation and words.

H­0: β1 = 0, which implies age does NOT help to predict (or “explain”) reaction time

HA: β1 ≠ 0, which implies age does help to predict (or “explain”) reaction time

(1 point for notation, 1 point for explanation in words)

* 1. ***(1 point)*** *From the regression output,* report the appropriate test-statistic with degrees of freedom.

t-stat = 1.019

df = 167

(1/2 point each)

Note: F-statistic (1.038 with 1 and 167 df) can be used here instead of t-statistic.

* 1. ***(2 points)*** State a conclusion in the context of the problem. Support your conclusion with a p-value.

**There is not enough evidence to indicate that age helps to predict reaction time (p-value = 0.3099).**

**(1 point for correct adjective to describe the evidence to say alternative hypothesis is true, ½ point for conclusion in context of alternative hypothesis, and ½ point for using correct p-value.)**

# Prediction, Prediction intervals, and Confidence Intervals for the mean of the response variable

To predict the response variable for a certain value of the explanatory variable, use the **predict()** command in R. (See the Lab 3A Note for more details.)

9) ***(1 point)*** Use R to predict the reaction time of a 22 year old. Report your answer here.

**3633.619 milliseconds**

The same **predict()** command can also be used to obtain either a prediction interval or a confidence interval for the mean of the response variable. See the Lab 3B Notes for more details.

10) For each of the following, use R to obtain the appropriate 95% interval. Report *and interpret* the appropriate interval.

***a) (2 points)*** A student wants to estimate the average reaction time for all 22-year old people.

95% Confidence Interval for the mean of the response (342.1541 ms , 385.084 ms)

We are 95% confident that the *average* reaction time for all 20-year-olds, is between 342.1541 and 385.084 milliseconds.

(1 point for correct interval, 1 point correct interpretation, which should include “average” reaction time for “all” individuals)

***b) (2 points)*** A student wants to wants to estimate the reaction time for a person who is 22 years old.

95% Prediction Interval (91.72 ms , 635.52 ms)

We are 95% confident that the reaction time for an individual who is 20 years old is between 91.72 and 635.52 milliseconds.

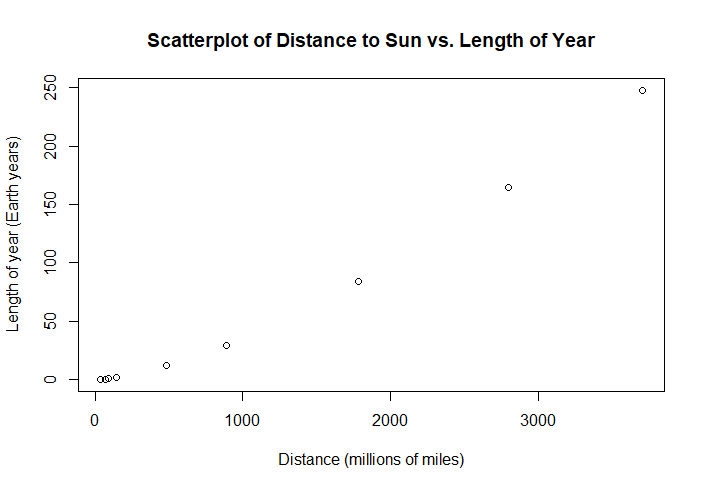
(1 point for correct interval, 1 point for “an individual” reaction time in interpretation)

## Problem 2: (7 points)

At a meeting of the International Astronomical Union (IAE) in Prague in 2006, Pluto was determined not to be a planet, but rather the largest member of the Kuiper belt of icy objects. Let's examine the relationship between the distance from the sun for the 9 sun-orbiting objects (including Pluto) and their length of years (earth years) for one complete orbit around the sun. The data are in the **planets** data set on Canvas. (***Distance***is the distance from the sun is in millions of miles and ***length*** is the length of a year in earth years.) Use these data to answer the following questions:

Use R to construct a scatterplot of the *length of year* (the response variable) against the *distance from sun*.

11) ***(2 points)*** Include the properly labeled scatterplot here and describe the relationship between these two variables.



**1 point for graph. 1 point for describing the relationship (positive, non-linear relationship)**

Use the proper code from the Lab 3B Notes to take the natural log of both variables to straighten the relationship. Obtain the output from the model where both variables are transformed and use that output

to answer the questions that follow. (Note: you do not have to assess the other conditions for this problem.)

12) ***(2 points)*** Write the least-squares regression equation. Define the terms in the equation.

ln() = -- 6.7971 + 1.4992\*ln(x)  
  
where is the predicted years in Earth years and x is the distance from the sun in millions of miles.  
(1.5 point for equation, ½ point for defining terms)

13) ***(1 point)*** Report the value of R-square. What does this value tell you about the relationship between ln(length of year) and ln(distance from the sun)?

**Multiple R-squared: 1**

**Since 100% of the variation in ln(year length) is explained by ln(distance to the sun), there is a perfect linear relationship between ln(year length) and ln(distance to sun)**

**(1/2 point for R2, 1/2 point for answer to second question)**

14) ***(2 points)*** It has been suggested that the asteroid belt between Mars and Jupiter may be the remnants of a failed planet. Suppose that failed planet was 285 million miles from the sun. Predict its length of year (in earth years) using the best-fitting model. You may do this by hand (showing work) or using R.

**ln() = -6.7971 + 1.4992\*ln(285)**

**ln() = 1.67711177**

**= e^1.67711177 = 5.3501 earth years**

**(Note: if student used R, answer is 5.350672)**

**(2 points for correct answer, partial credit can be given if arithmetic error or not back transforming)**