M349R (Unique 54230)

**Instructor:** Gustavo Cepparo

Project 1

Juan Acosta

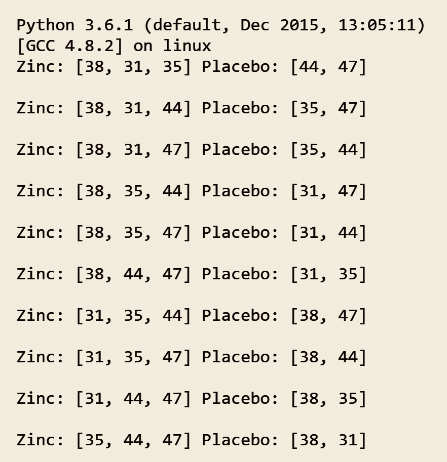
JA45384

Fall 2018

Problem 1

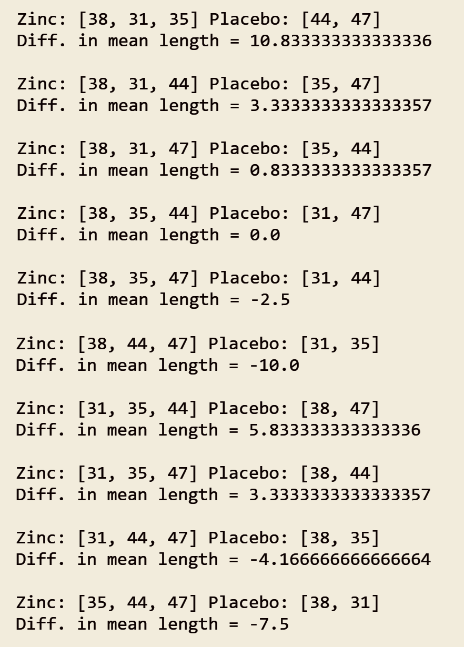
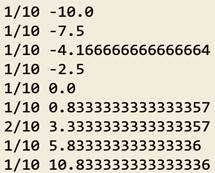
Does a zinc supplement taken at the onset of cold symptoms reduce the time one has a cold? Five volunteers (Dan, Hal, Joe, Kay, and Zoe) agree to take part in an experiment. Three are assigned completely at random to receive the zinc supplement and the other two receive the placebo. The experiment is double-blind. The results are (times are duration of cold in hours):

(31, 35, 38) - (47, 44)

(a)There are 10 possible ways the five subjects can be assigned to the two groups, with the zinc group having size 3 and the placebo group size 2. List these.

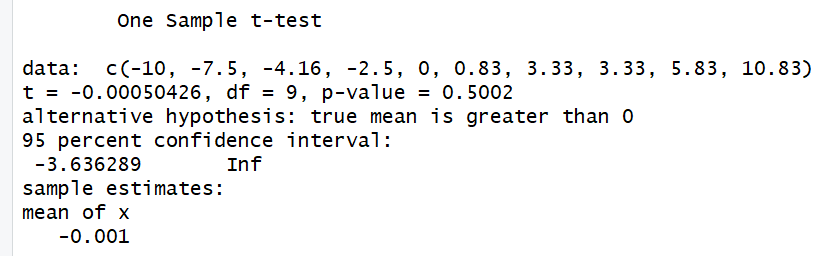
(b)For each, determine the difference in mean length of cold (mean for the placebo group minus mean for the zinc group). Combine any duplicates and make a table of the possible mean differences and the corresponding probability of each under the null hypothesis of no treatment effect. (Each of the 10 possible assignments of subjects to treatments has probability 1/10 under the null hypothesis.) This is the permutation distribution.

H0: Diff. in mean length = 0

H1: Diff in mean length > 0

P(Diff) | Diff. Means

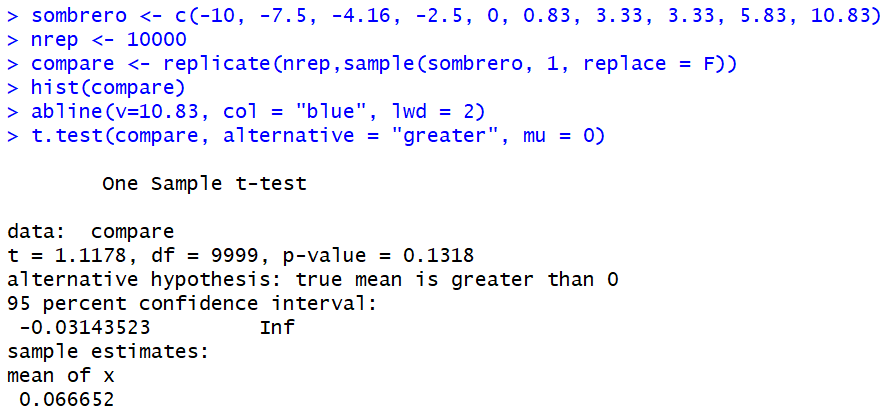
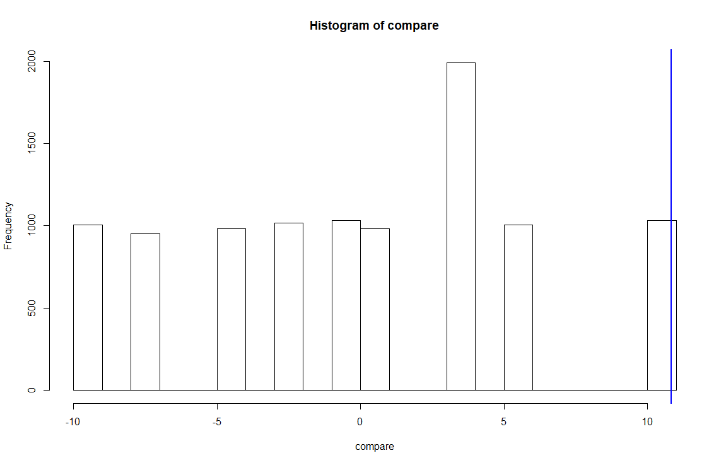
(c)Compute the P-value of the data. Assume the alternative hypothesis is that the mean duration of a cold is less for zinc.



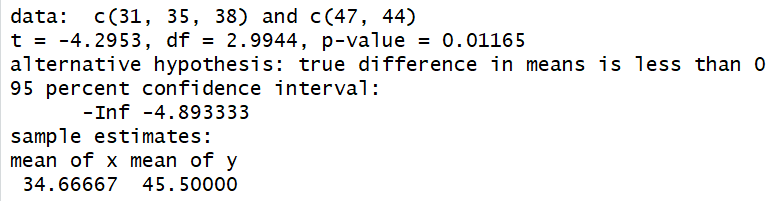


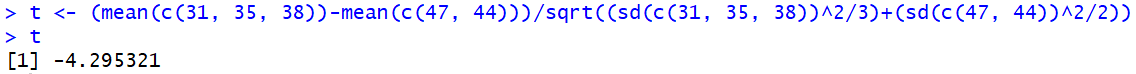
(d)In this example, is it possible to demonstrate significance at the 5% level using the permutation test? Explain.

Statistical significance, no. Because of the low number of data points (10) the permutation test will not yield a normal distribution, and it will be heavily defined by the probability distribution shown in part (B). Moreover, the value observed is the most extreme point already in the set of possible combinations, therefore it is the only value that will appear in the histogram as being equal or more extreme. The p-value would then be 1/10 which is not sufficient to prove statistical significance at the 5% level. Nonetheless, the data allows us to predict that there might be practical significance.

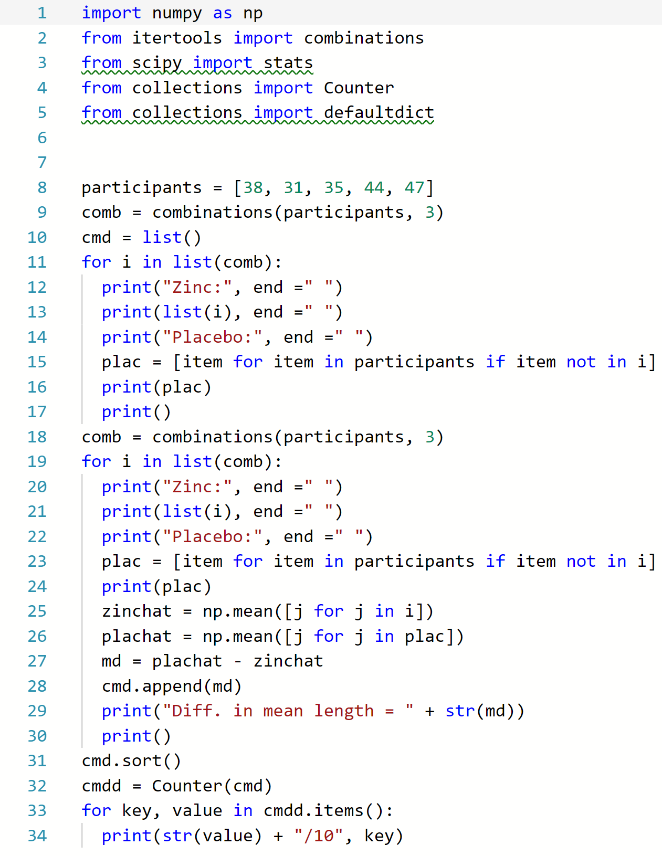
Permutation test, with t test and histogram:

(e)Assume that cold duration is Normally distributed for both zinc and the placebo. Use the two-sample tprocedure to test the hypotheses. (Use SAS and byhand for two sample t-test).

t.test(c(31, 35, 38), c(47, 44), alternative = "less", var.equal = FALSE)



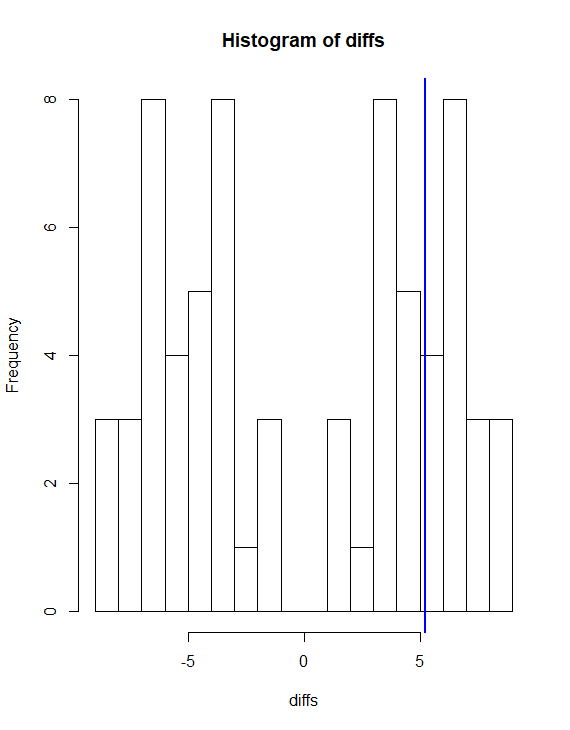
Confirming t value from t.test equals calculated by hand.

Code for Problem 1.

Problem 2

The Weeds among the Corn is completely randomized experiment. Two treatments—one weed per meter and nine weeds per meter—were assigned completely at random to eight plots of land. Each was assigned to four plots. Here are the yields of corn (bushels per acre):

(166.2, 157.3, 166.7, 161.1) – (162.8, 142.4, 162.8, 162.4)

Explain why you decided that two-sample t procedures probably may not be accurate. However, the permutation test can be used. The display below shows the distribution of all possible mean differences for the 70 different ways treatments could be assigned to plots. Use the display below to estimate the P-value of these data. (4 pts each)

8 data points is not sufficient to provide a meaningful two-sample t procedure, and provide statistical significance. Furthermore, it is hard to visualize 8 data points mentally to try and determine practical significance. On the other hand, because the allocation of values is not “maximized” for one of the two categories, then using a permutation test will probably yield some insight into the probability of this specific distribution being “extreme”.

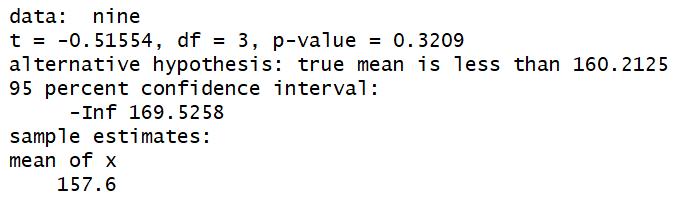
mean(166.2, 157.3, 166.7, 161.1) – mean(162.8, 142.4, 162.8, 162.4) = 5.224999999999966

p value estimate = 1/4

Use R and as a test statistic use the mean of Nine weeds/meter in order to test a one sided hypothesis. (16 points)

H0: Mean of nine weeds is equal to 160.2125

H1: Mean of nine weeds is less than 160.2125



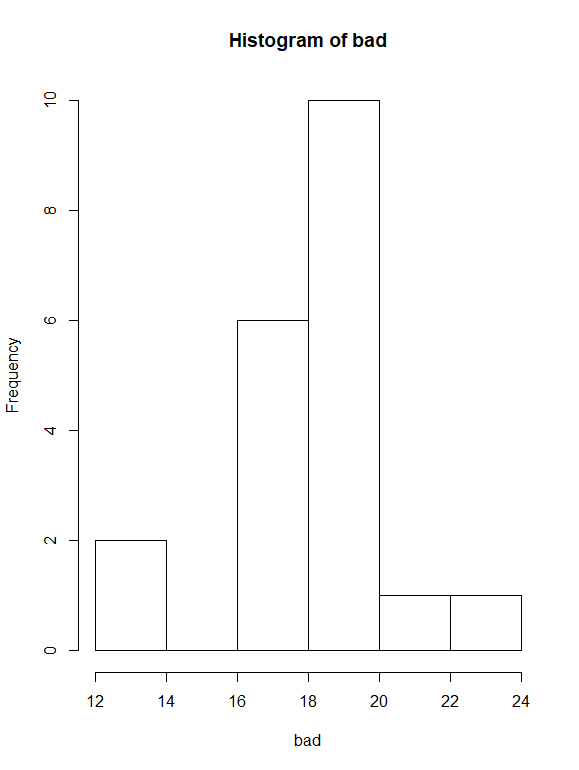
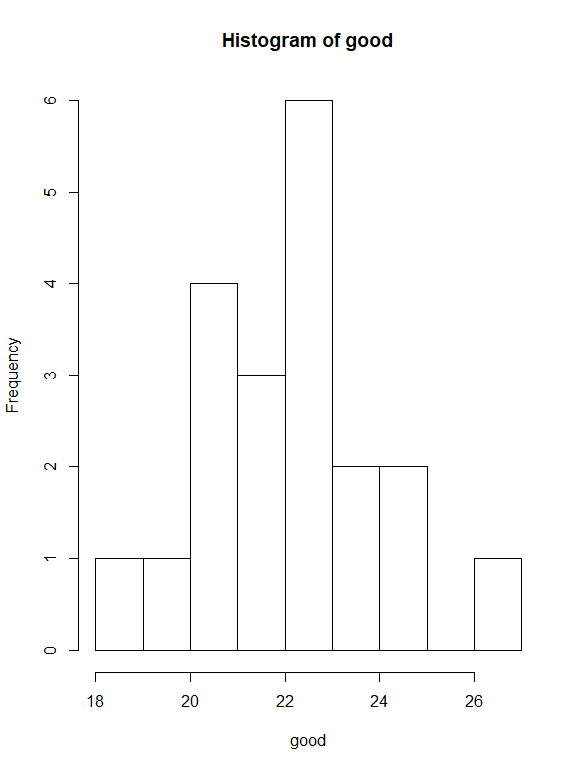
Problem 3

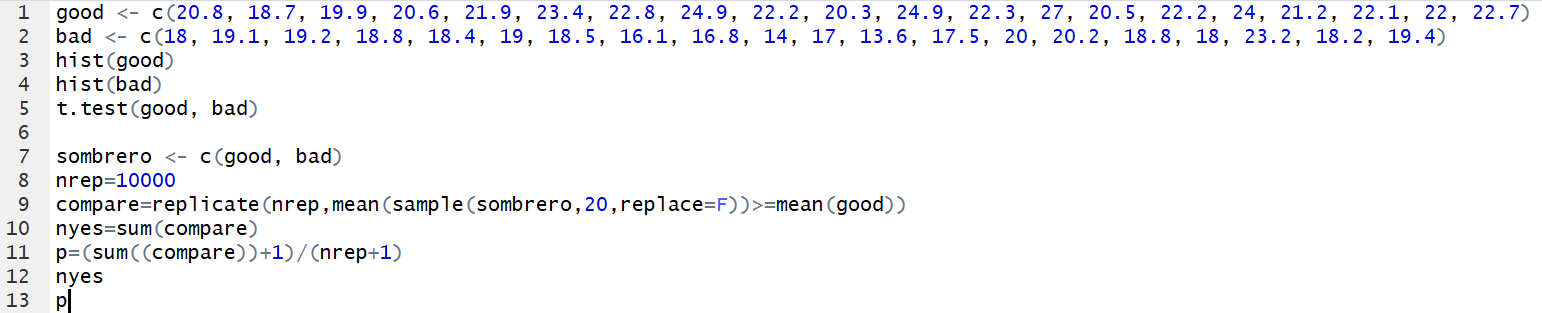
Researchers gave 40 index cards to a waitress at an Italian restaurant in New Jersey. Before delivering the bill to each customer, the waitress randomly selected a card and wrote on the bill the same message that was printed on the index card. Twenty of the cards had the message, “The weather is supposed to be really good tomorrow. I hope you enjoy the day!” Another 20 cards contained the message, “The weather is supposed to be not so good tomorrow. I hope you enjoy the day anyway!” After the customers left, the waitress recorded the amount of the tip (percent of bill) before taxes. Here are the tips for those receiving the good-weather message:

good <- c(20.8, 18.7, 19.9, 20.6, 21.9, 23.4, 22.8, 24.9, 22.2, 20.3, 24.9, 22.3, 27, 20.5, 22.2, 24, 21.2, 22.1, 22, 22.7)

The tips for the 20 customers who received the bad weather message are:

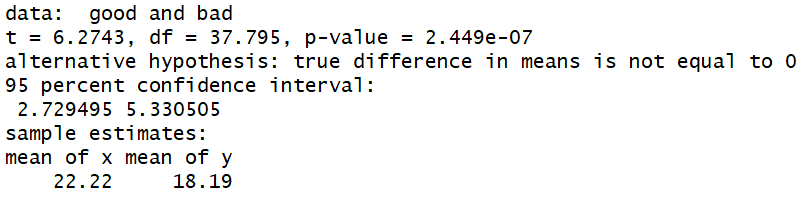
bad <- c(18, 19.1, 19.2, 18.8, 18.4, 19, 18.5, 16.1, 16.8, 14, 17, 13.6, 17.5, 20, 20.2, 18.8, 18, 23.2, 18.2, 19.4)

(a)Make stemplots or histograms of both sets of data. Because the distributions are reasonably symmetric with no extreme outliers, the t procedures will work well. (Use R to graph) (4 pts)

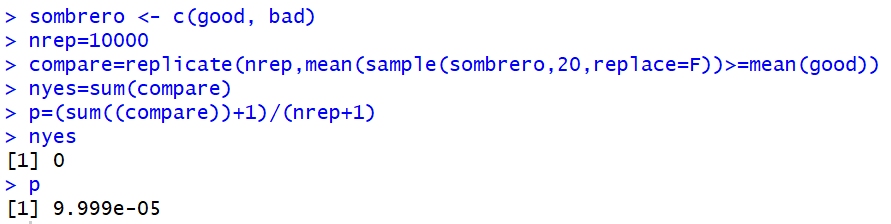
(b)Is there good evidence that the two different messages produce different percent tips? State hypotheses, carry out a two-sample t test, and report your conclusions. (Use SAS for a two sample t-test) (14 pts)

H0: diff in means = 0

H1: diff in means ≠ 0



(c) Use R in order to run a randomization test (one sided). (12 pts)

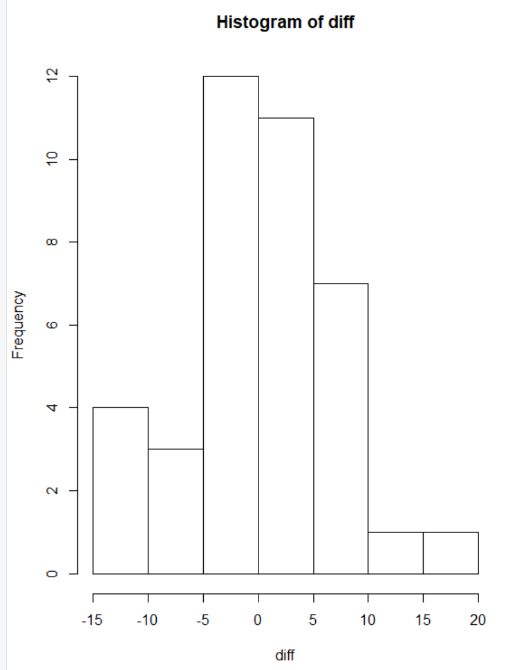


Fail to reject that the true mean diff is 0

Problem 4

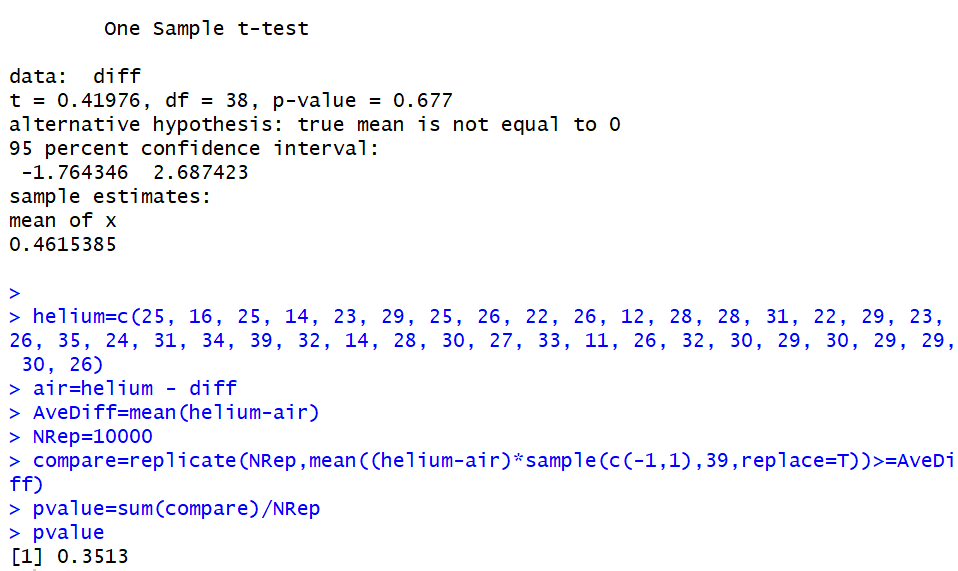
Does a football filled with helium travel farther than one filled with ordinary air? To test this, the Columbus Dispatch conducted a study. Two identical footballs, one filled with helium and one filled with ordinary air, were used. A casual observer was unable to detect a difference in the two footballs. A novice kicker was used to punt the footballs. A trial consisted of kicking both footballs in a random order. The kicker did not know which football (the helium-filled or the air-filled football) he was kicking. The distance of each punt was recorded. Then another trial was conducted. A total of 39 trials were run. Here are the data for the 39 trials, in yards that the footballs traveled. The difference (helium minus air) is the response variable.

(a)Examine the data. Is it reasonable to use the t procedures? (10 pts)

There are 39 data points which by our “rule of thumb” would mean that our data shouldn’t be “too” skewed, and it’s a little. Therefore, it would be reasonable to assume it would provide some useful insight into the behavior of the data.

(b)If your conclusion in part (a) is “yes,” do the data give convincing evidence that the helium-filled football travels farther than the air-filled football? (Use SAS to Test, 16 pts)

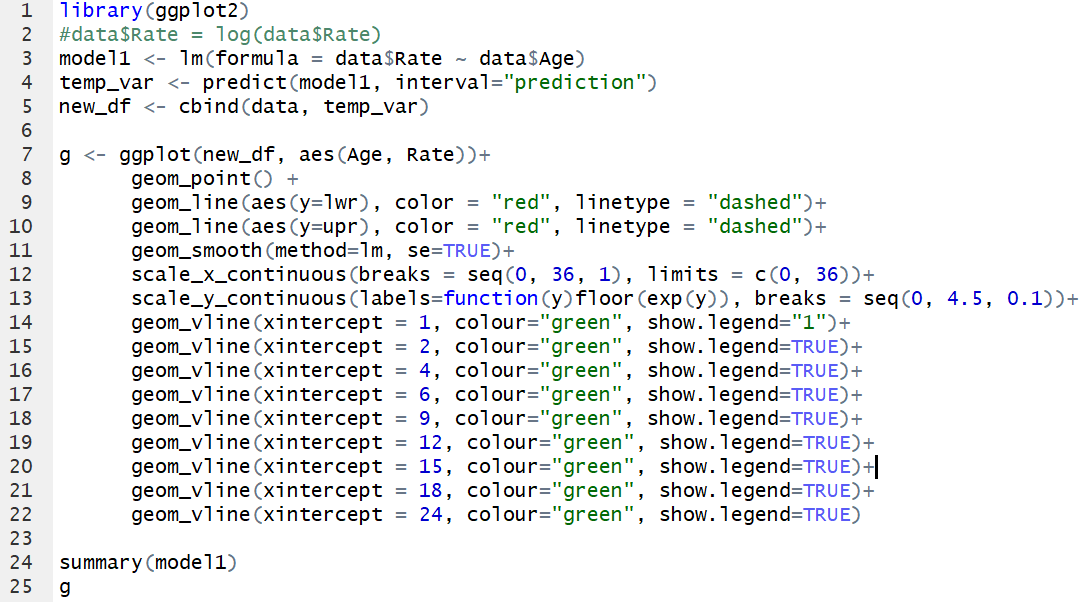
No, looking at the histogram and the t-test (below) the data appears to be pseudo normal, with a positive outlier (17). Therefore, looking at the mean (0.4) we can safely assume that without the outlier it would most likely be even closer to 0. Conclusively, the helium has a negligible effect on the trajectory of the ball.

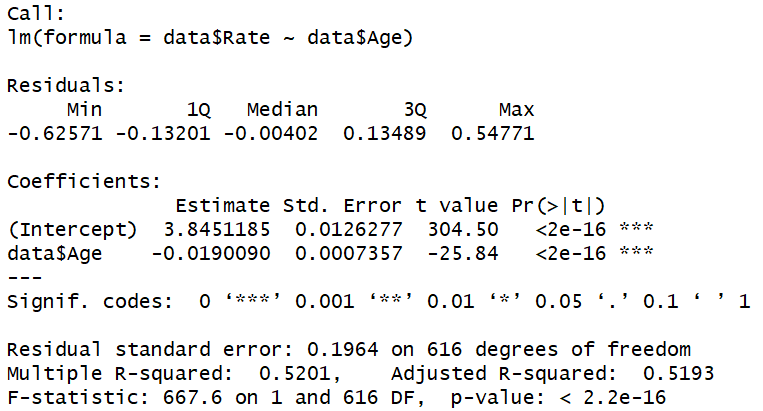
(c) Use run a randomization (Matched Pairs Test in R). (16 pts)

The p value of 0.35 confirms our previous results that we should fail to reject the null that the true mean is equal to 0.

Problem 5

A high respiratory rate is a potential diagnostic indicator of respiratory infection in children. To judge whether a respiratory rate is truly “high”, however, a physician must have a clear picture of the distribution of normal respiratory rates. To this end, Italian researchers measured the respiratory rates of 618 children between the ages of 15 days and 3 years. Analyze that data and provide a statistical summary. Include a useful plot or chart that a physician could use to assess a normal range of respiratory rate for children of any age between 0 and 3.

(units of age in months and respiratory rate in number of breaths per minute)



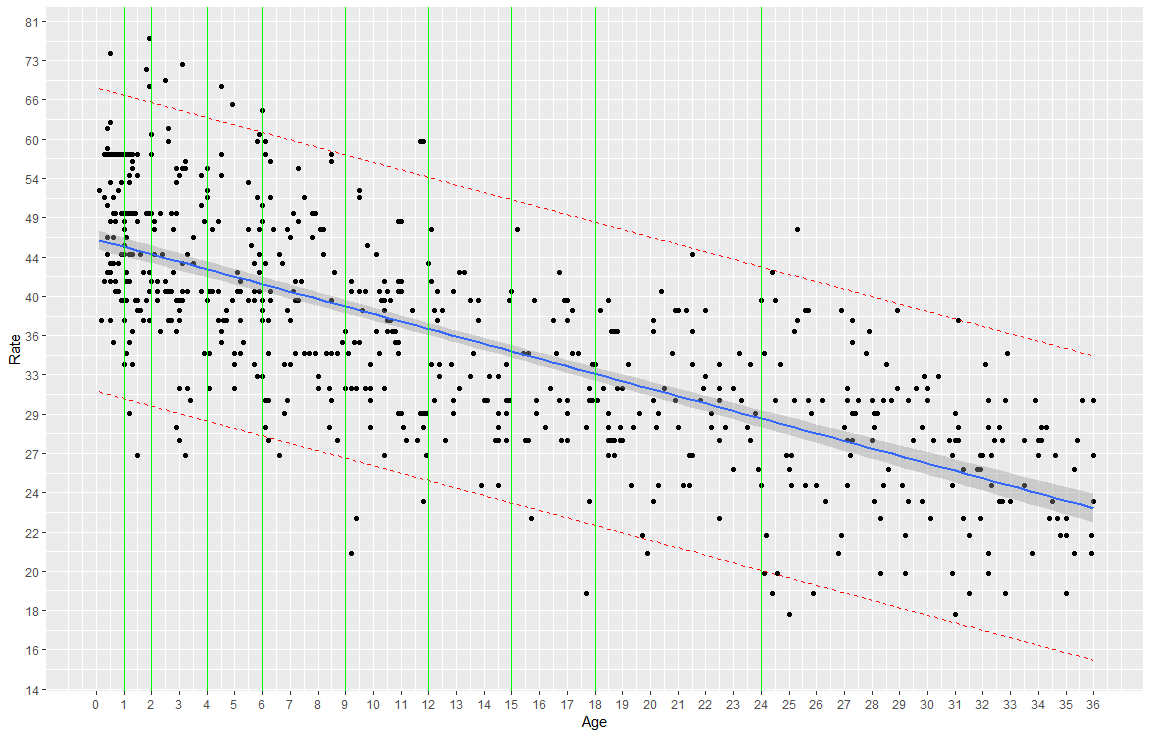
Log(Rate) = 3.845 – 0.019 \* Age

(0.013) (0.0007)

R2=0.52

N = 618

RMSE = 0.1964

Note: data is plotted log(Rate) vs Age (months). Y-Axis ticks are transformed with floor(exp(y)) therefore all Y values represent the true rate yet are slightly less than the true rate. To compensate for this, the prediction intervals are conservative at 95%. Green markers indicate recommended age for pediatrician visits for easy location and cross reference with rate.

Problem 6

The data for problem 6 are the average wine consumption rates (in liters per person) and the number of ischemic heart disease deaths (per 1000 men aged 55 to 64 years of age) for 18 industrialized countries. Do these data suggest that heart disease death rate is associated with average wine consumption? If so, how can that relationship be described? Do any countries have substantially higher or lower death rates with similar wine consumption rates? Analyze the data and write a brief report that includes a summary of statistical findings, a graphical display, and a section detailing the methods to answer the question of interest.

Pre-processing the data to achieve pseudo-linearity we define the model:

Log(Mortality) = 2.55 – 0.3556 log(Wine)

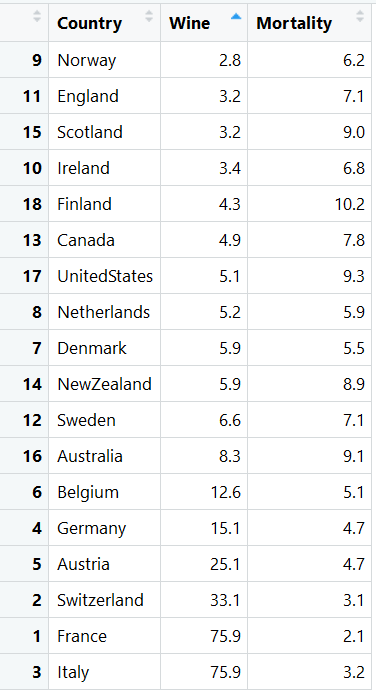
(0.1269)(0.05291)

R2 = 0.74

N = 18

RMSE = 0.2285

Yes, the data appears to be correlated. This correlation can be described by the R2 = 0.74



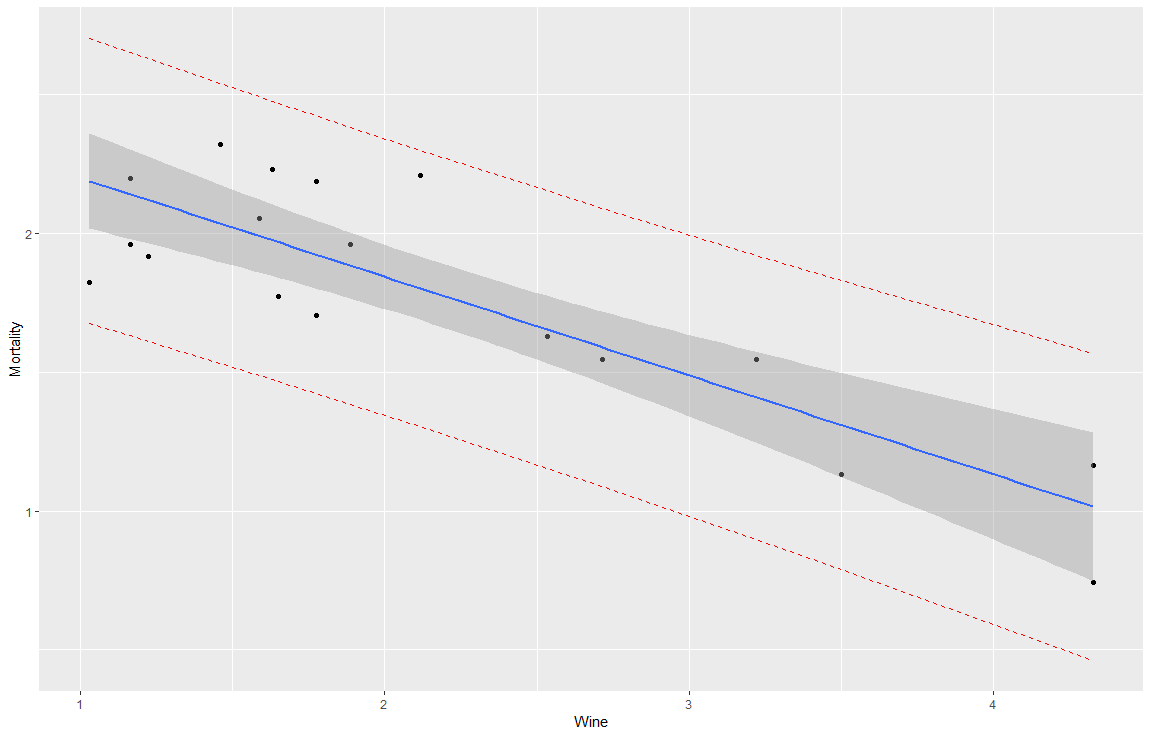
There are definitely some countries with similar wine consumption rates yet significantly different mortality rates. Such as:

New Zealand and Denmark

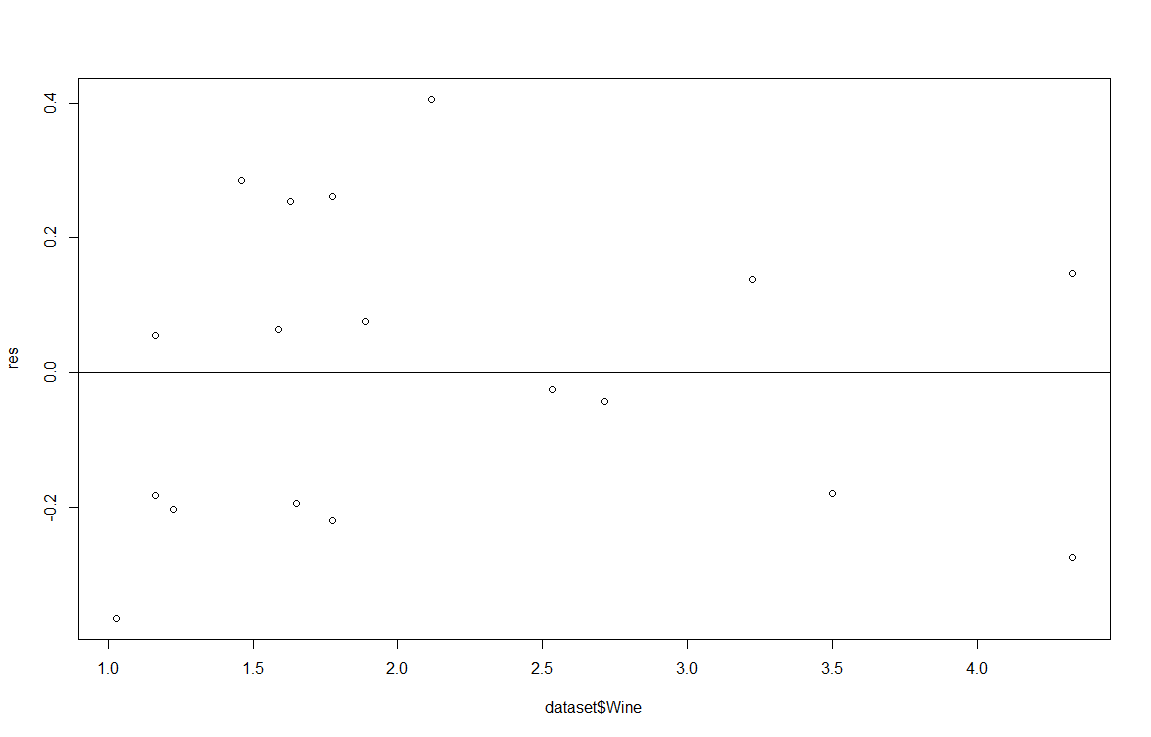
United States and Netherlands

England and Scotland

Analyze the data and write a brief report that includes a summary of statistical findings, a graphical display, and a section detailing the methods to answer the question of interest.

The dataset shows a correlation between the way wine consumption affects mortality. More specifically, the model created predicts that the countries with higher wine consumption have lower mortality rates for ischemic heart disease.

Nonetheless, there are some unaccounted differences across some countries with similar wine consumption which present higher or lower mortality rates, therefore implying that there are other factors that drive mortality rates for ischemic heart disease across countries. This can be more clearly seen by observing the cluster to the left side of the residuals plot.



This provides some insight to the fact that there might be some omitted variables, such as: cultural tradition, eating habits, genetic predisposition, etc. Which might also be driving the mortality rates for IHD. Or, there is no underlying correlation between the two.

Finally, observing the confidence interval of the linear model plotted and the spread of the data points, some interesting facts are useful to point out:

Only half of the values fall within the confidence interval lines

Most of the line is strongly driven by the 6 values beyond the 2.5 mark in the x axis, which only accounts for 1/3 of the population sampled

The first 2/3rds of the data does not appear to have any correlation whatsoever. And its residuals plot is clearly clustered.

Conclusively, this dataset seems to be a great example of correlation does not imply causation. As wine consumption does not appear to have a causal effect on IHD mortality rates.