Homework 10 (Optional) - MATH 141

Due Date: Wednesday 12/08/2021, 11:59 PM

Instructions:

- Please provide complete answers/solutions for each question/problem.
- **If it involves mathematical computations, please provide your reasoning and/or detailed solutions.**
- There are two ways you can write your answers, a: by handwriting (either physically or digitally), or b: by typing on a template document with file type options, Word or RMarkdown, which can be downloaded from the course website.
- If you had handwritten your answers/solutions on a physical paper, make sure to label it properly and please scan your document using a scanner app for convenience. Suggestions: (1) "Tiny Scanner" for Android or (2) "Scanner App" for iOS.
- If a problem asks you to show your R code, R outputs, or R plots, please provide them as additional pages into your current homework pdf while labeling them properly. This means that, **if you have handwritten your homework solutions and saved it as pdf, you would need to merge the separate pdf which contains your R code, R outputs, or R plots. Note that all of the problems that require R does not require you to show your R code - unless the problem specifically says so.**
- If you have questions or concerns, please feel free to ask the instructor.
- **Please save your work as one pdf file, don't put your name in any part of the document, and submit it to the Gradescope page for this course. Your document upload will correspond to your name automatically in Gradescope.**

I. Inference for Comparing Many Means

The exercise problems shown below was taken and slightly modified from your textbook OpenIntro: Introduction to Modern Statistics Section 22.5

1. Student Performance Across Discussion Sections.

Consider the following problem statement.

> A professor who teaches a large introductory statistics class (197 students) with eight discussion sections would like to test if student performance differs by discussion section, where each discussion section has a different teaching assistant. The summary table below shows the average final exam score for each discussion section as well as the standard deviation of scores and the number of students in each section.

	Sec 1	Sec 2	Sec 3	Sec 4	Sec 5	Sec 6	Sec 7	Sec 8
Mean	92.94	91.11	91.80	92.45	89.30	88.30	90.12	93.35
SD	4.21	5.58	3.43	5.92	9.32	7.27	6.93	4.57
n	33.00	19.00	10.00	29.00	33.00	10.00	32.00	31.00

The ANOVA output below can be used to test for differences between the average scores from the different discussion sections.

	df	sumsq	meansq	statistic	p.value
section	7	525.01	75.00	1.87	0.0767
Residuals	189	7,584.11	40.13	NA	NA
Total	196	8,109.12	NA	NA	NA

a. Conduct a hypothesis test to determine if these data provide convincing evidence that the average score varies across some (or all) groups.

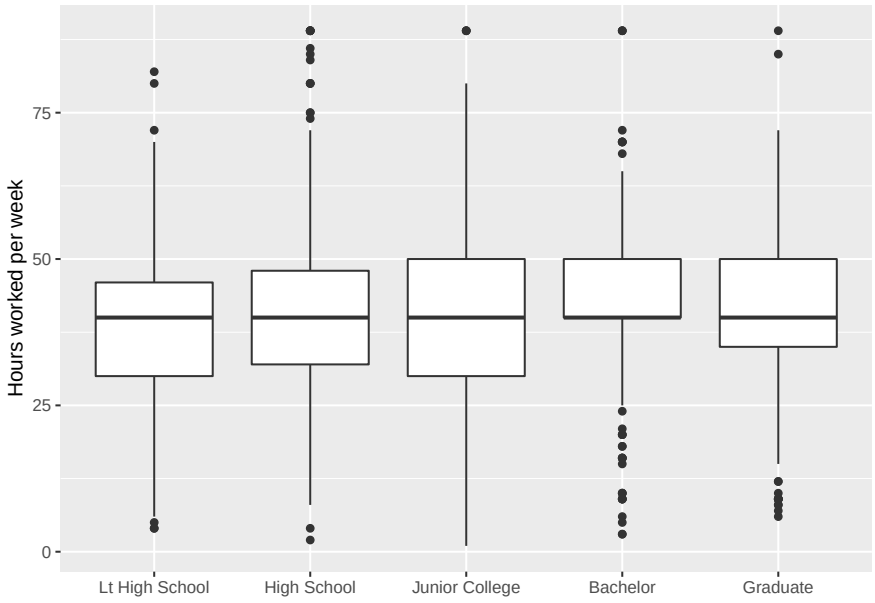
b. Check conditions and describe any assumptions you must make to proceed with the test.

2. Work Hours and Education.

Consider the following problem statement.

> The General Social Survey collects data on demographics, education, and work, among many other characteristics of US residents. NORC 2010 Using ANOVA (ANalysis Of VAriance), we can consider educational attainment levels for all 1,172 respondents at once. Below are the distributions of hours worked by educational attainment and relevant summary statistics that will be helpful in carrying out this analysis.

Educational attainment	Mean	SD	n
Lt High School	38.66942	15.81423	121
High School	39.59707	14.97125	546
Junior College	41.39175	18.10361	97
Bachelor	42.54941	13.61731	253
Graduate	40.84516	15.50540	155



a. Write hypotheses for evaluating whether the average number of hours worked varies across the five groups.

b. Check conditions and describe any assumptions you must make to proceed with the test.

c. Below is the output associated with this test. What is the conclusion of the test?

term	df	sumsq	meansq	statistic	p.value
`degree`	4	2,006.16	501.54	2.19	0.07
`Residuals`	1,167	267,382.16	229.12	NA	NA

II. Inference for Multiple Linear Regression

The exercise problems shown below was taken and slightly modified from your textbook OpenIntro: Introduction to Modern Statistics Section 25.5.

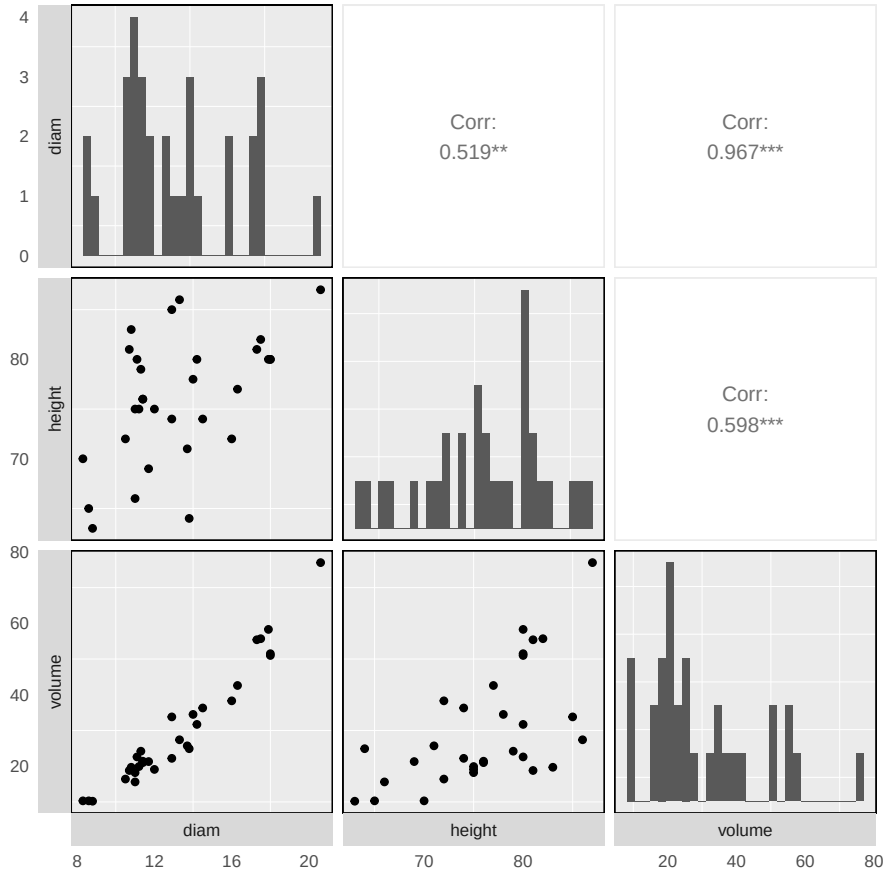
1. Cherry Trees.

Consider the following problem statement.

> Timber yield is approximately equal to the volume of a tree, however, this value is difficult to measure without first cutting the tree down. Instead, other variables, such as height and diameter, may be used to predict a tree's volume and yield. Researchers wanting to understand the relationship between these variables for black cherry trees collected data from 31 such trees in the Allegheny National Forest, Pennsylvania. Height is measured in feet, diameter in inches (at 54 inches above ground), and volume in cubic feet. Hand 1994

The `cherry` data used in this exercise can be found in the **openintro** R package.

The plots below show the distribution of each of these variables (on the diagonal) as well as provide information on the pairwise correlations between them.



Provided below are three regression model outputs: `volume` vs. `diam`, `volume` vs. `height`, and `volume` vs. `height + diam`.

term	estimate	std.error	statistic	p.value
`(Intercept)`	-36.943	3.365	-10.978	<0.0001
`diam`	5.066	0.247	20.478	<0.0001

term	estimate	std.error	statistic	p.value
`(Intercept)`	-87.124	29.273	-2.976	0.006
`height`	1.543	0.384	4.021	0.000

term	estimate	std.error	statistic	p.value
`(Intercept)`	-57.988	8.638	-6.713	<0.0001
`height`	0.339	0.130	2.607	0.0145
`diam`	4.708	0.264	17.816	<0.0001

a. There are three variables described in the figure, and each is paired with each other to create three different scatterplots. Rate the pairwise relationships from most correlated to least correlated.

b. When using only one variable to model a tree's `volume`, is `diam`eter a significant predictor variable? Is `height` a significant predictor variable? Explain.

c. When using both `diam`eter and `height` to predict a tree's `volume`, are both predictor variables still significant? Explain.

2. Baby's Weight.

Consider the following problem statement.

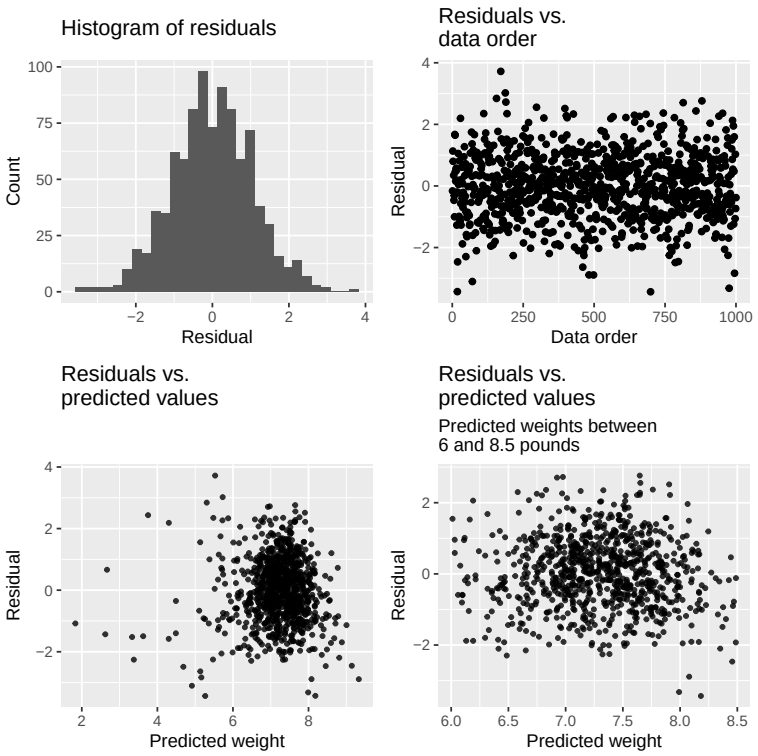
> US Department of Health and Human Services, Centers for Disease Control and Prevention collect information on births recorded in the country. The data used here are a random sample of 1,000 births from 2014. Here, we study the relationship between smoking and weight of the baby. The variable `smoke` is coded 1 if the mother is a smoker, and 0 if not. ICPSR 2014

The summary table below shows the results of a linear regression model for predicting the average birth weight of babies, measured in pounds, based on the smoking status of the mother.

The `births14` data used in this exercise can be found in the **openintro** R package.

term	estimate	std.error	statistic	p.value
`(Intercept)`	-3.82	0.57	-6.73	<0.0001
`weeks`	0.26	0.01	18.93	<0.0001
`mage`	0.02	0.01	2.53	0.0115
`sexmale`	0.37	0.07	5.30	<0.0001
`visits`	0.02	0.01	2.09	0.0373
`habitsmoker`	-0.43	0.13	-3.41	7e-04

Also shown below are a series of diagnostics plots.



a. Determine if the conditions for doing inference based on mathematical models with these data are met using the diagnostic plots above. If not, describe how to proceed with the analysis.

b. Using the regression output, evaluate whether the true slope of `habit` (i.e. whether the mother is a smoker) is different than 0, given the other variables in the model. State the null and alternative hypotheses, report the p-value (using a mathematical model), and state your conclusion.