CS 7200 - Fall 2023 Olga Vitek

Homework 1

Please follow the homework submission instructions provided on Piazza.

Due on Canvas before midnight on Tuesday, September 26 2023.
Each part of the problems 5 points
[Please note that interpretation is the major part of answer to each question (except for the questions that only ask for an implementation). Substantial points will be reduced for answers that provide plots or other outputs, but no interpretation]

1. *[Implementation question. The use of quantile-quantile plot libraries is not allowed. You can use a generator of Normal or Cauchy random variables]*
   (a) Overlay on a same graph the probability density functions of $\mathcal{N}(0,1)$, $\mathcal{N}(3,5)$ and $Cauchy(-2,2)$, and interpret the differences between the distributions. Please limit x axis as appropriate.
   (b) Implement a function, which takes as input a vector Y, and plots a Normal quantile-quantile plot.
   (c) Illustrate the implementation on a random sample of 100 $\mathcal{N}(0,1)$ random variables and interpret the use of the plot.
   (d) Illustrate the implementation on a random sample of 100 $\mathcal{N}(3,5)$ random variables and interpret the use of the plot.
   (e) Illustrate the implementation on a random sample of 100 $Cauchy(-2,2)$, random variables and interpret the use of the plot.
2. *[Conceptual question illustrating the concept of degrees of freedom].* Assume you have 10 observations, $x_1, x_2, \ldots, x_{10}$.
   (a) Suppose that you know the mean of all the 10 values, i.e. $\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i$, as well as 9 values $x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_9 - \bar{x}$. Show that you can use this information to obtain $x_{10} - \bar{x}$. Please show your work conceptually (i.e., using variables instead of numbers).
   (b) Write the expression of sample variance s^2 for $x_1, x_2, \ldots, x_{10}$. What are the degrees of freedom of s^2? Why?
3. *[Implementation question. We will use this function throughout the homework. The use of linear regression libraries is not allowed.]* Implement linear regression as a function, which takes as input two vectors X and Y, and returns the least squares estimates of the intercept and the slope of linear regression, the mean squared error of the residuals s^2, the degrees of freedom associated with s^2, and the t statistic and the p-value of testing H_0 that the slope is zero.

4. *[Implementation question that sets up simulations. We will use these simulations throughout the homework.]*

   (a) Simulate two datasets from a population that follows a relationship $Y = 5 + 3 \cdot X + \varepsilon,\ \varepsilon \overset{iid}{\sim} \mathcal{N}(0, 1)$ and make a scatterplot of each dataset. Ideally, set the seed of any random generators (in R, use `set.seed(seed)` before random generators); otherwise please hard code your randomly generated datasets so that we are both viewing the same data. In what aspects do the datasets look similar or different?

   (b) Using the code from the previous question, analyze these two datasets. Report the values of slopes in each dataset. Comment on why the outputs are different.

5. *[Simulation illustrating the Central Limit theorem. The use of linear regression libraries or quantile-quantile plot libraries is not allowed for this question. Use your implementations from the previous questions.]*

   (a) Read the article by Krzywinski & Altman "Importance of being uncertain" (posted on the course website). Interpret Figure 3 in the article.

   (b) We will reproduce this figure for the slope of linear regression. Consider linear regression $Y = 5 + 3 \cdot X + \varepsilon,\ \varepsilon \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 = 1)$. Let X only take two possible values, 0 and 2. Plot the sampling distribution of b_1 for $n = 3$ and $n = 100$ (consider 1,000 instances of random sampling). Overlay the true β_1 on each plot. *[Note: In other words, generate 1000 instances each of n=3 and n=100 samples, where X can only be 0 or 2. Parameter estimation will fail when all X values are the same. Skip all such samples.]*

   (c) Repeat (b) with $\sigma^2 = 100$.

   (d) Repeat (b) with $\varepsilon \overset{iid}{\sim} \chi^2_{10}$. [Note: the χ^2_{df} distribution has a single parameter, df which is the degrees of freedom - in this case 10.]

   (e) Summarize the results in (b)-(d). Comment on which distributions are wider or narrower, and why? What are the implications for inference of β_1?

6. *[Simulation illustrating interval estimation. The use of linear regression libraries is not allowed for this question. Use your implementation from the previous questions.]*

   (a) Read the article by Krzywinski & Altman "Error bars" (posted on the course website). Interpret Figure 2a in the article.

   (b) We will reproduce this figure for the slope of linear regression. Consider the same linear regression $Y = 5+3\cdot X+\varepsilon,\ \varepsilon \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 = 1)$ and X taking only two values, 0 and 2, as above. Simulate 100 datasets with n=3, and generate 100 95% confidence intervals for the slope. Overlay them next to the true sampling distribution of the slope as in the figure. Interpret the result: do the confidence intervals always contain the true slope? Why or why not?

   (c) Repeat (b), but for the 99% confidence intervals. Are the new intervals wider or narrower? Why?

(d) Make the scatterplot of one simulation in (b). Overlay on the scatterplot the 95% confidence interval for the mean $E\{Y_h|X_h = 1\}$, and the 95% prediction interval for the new observation $Y_{h(new)}|X_{h(new)} = 1$. State all the formulas used for calculation. Which of the intervals is wider? Why?

(e) Repeat (d), but for $X_{h(new)} = 5$. Are the intervals in this question wider or narrower than in (d)? Why?

7. *[Simulation illustrating properties of p-values. The use of linear regression libraries is not allowed for this question. Use your implementation from the previous questions.]*

(a) Consider a linear regression $Y = 5 + 0 \cdot X + \varepsilon$, $\varepsilon \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 = 1)$ (i.e., X has no association with Y), and X taking only two values, 0 and 2. Simulate 1,000 datasets from this process with $n = 3$, test $H_0 : \beta_1 = 0$, and plot the histogram of the resulting p-values. Interpret the probability distribution of the p-values when H_0 is true.

(b) Read the article "The fickle P value generates irreproducible results" by Halsey *et al* (posted on the course website). Interpret Figure 4.

(c) We will reproduce this figure for the slope of linear regression, in a situation when the slope is small relative to σ^2. Consider a linear regression $Y = 5+0.1\cdot X+\varepsilon$, $\varepsilon \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 = 1)$, and X taking only two values, 0 and 2. Simulate 1,000 datasets with $n = 3$, test $H_0 : \beta_1 = 0$, and plot the histogram of the resulting p-values. (Consider plotting the histogram on the $\log_{10}$ scale as in the article.) Repeat for the sample size of $n = 30$, 64 and 100. What do the plots say about our ability to reproducibly discover the association for each n?

(d) Repeat (c) for the sample sizes of $n = 10$, 30, 64 and 100, and for each sample size record the estimate of the slope for each "significant" hypothesis test among the 1,000 repetitions. For each sample size, plot the histogram of the estimates of the slopes among the significant tests, and overlay the true value. Based on these results, comment whether it is reasonable to only report the estimates of slope when we reject H_0.