Homework 1, STAT 252, AY 22-23, 2nd Sem
Course instructor: Michael Daniel Lucagbo
Due date: Match 31, 2023 (Friday), before midnight

Homework instructions

- Your homework must be submitted with a careful and concise write-up of the algorithms and the results. Email me your answers as portable document format (PDF) files. Any necessary codes should also be included in the file/s that you submit. However, a solution to a problem that consists only of software code and output will receive no credit.
- For the algorithms you write, use clear notations. Define the notations that you use.
- You may work individually or in groups of at most three members. Discussions with your classmates is encouraged. However, copying someone else's work or some other group's work is NOT allowed. If this happens, both parties will get a 0 on the assignment, and further proper disciplinary action will be taken.
- You are encouraged (but not required) to use R Markdown. For an introduction to R Markdown, you may watch this video: https://www.youtube.com/watch?v=DNS7i2m4sB0

---

1. Consider the data from Table 3.1 of Efron and Tibshirani (1993) shown below.

Table 3.1. *The law school data. A random sample of size $n = 15$ was taken from the collection of $N = 82$ American law schools participating in a large study of admission practices. Two measurements were made on the entering classes of each school in 1973: LSAT, the average score for the class on a national law test, and GPA, the average undergraduate grade-point average for the class.*

School	LSAT	GPA	School	LSAT	GPA
1	576	3.39	9	651	3.36
2	635	3.30	10	605	3.13
3	558	2.81	11	653	3.12
4	578	3.03	12	575	2.74
5	666	3.44	13	545	2.76
6	580	3.07	14	572	2.88
7	555	3.00	15	594	2.96
8	661	3.43			

   (a) Suppose that we *cannot* assume that the joint distribution of $(\text{LSAT}, \text{GPA})'$ is bivariate normal. Write down an algorithm that uses the bootstrap to estimate the standard error of the Pearson coefficient of correlation

$$r = \frac{\frac{1}{n-1}\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{S_x S_y},$$

   where S_x and S_y are the sample standard deviations of the X- and Y-variable, respectively.

   (b) Using $B = 2000$ bootstrap samples, implement your procedure in (a) and provide the following: (i) bootstrap estimate of the standard error of r (denoted by $\widehat{\text{se}}(r)$), (ii) 95% percentile confidence confidence interval for ρ (the true population correlation), (iii) a histogram showing the bootstrap distribution of the correlation.

   (c) Change the maximum value among the r_b^*s in part (b) to 100 times its value (note that in real life, this could be an encoding error). By how much does $\widehat{\text{se}}(r)$ change?

2. Use the same data set in #1. A biased but more robust estimate of the bootstrap standard error is

$$\widetilde{\text{se}}_\alpha(r) = \frac{r^*_\alpha - r^*_{1-\alpha}}{2z_\alpha},$$

where r^*_α, $\alpha \in (0,1)$, is the α-quantile of $r^*_1, \ldots, r^*_B$, and B = number of bootstrap samples, and z_α is the α-quantile of the standard normal distribution (for example, $z_{0.95} = 1.645$).

(a) Use $B = 2000$ bootstrap samples (you may use the same ones as in #1) and fill in the table below:

α:	.05	.10	.20	.50	.70	.90	.95
r^*_α:							

(b) Using the information in part (a), compute $\widetilde{\text{se}}_\alpha(r)$ for $\alpha = 0.95, 0.90, 0.85$.

(c) Change the maximum value among the r^*_b to 100 times its value, then re-compute $\widetilde{\text{se}}_\alpha(r)$ for $\alpha = 0.95, 0.90, 0.85$. Compare the old and new values of $\widetilde{\text{se}}_\alpha(r)$.

3. A researcher in a scientific foundation wished to evaluate the relation between intermediate and senior level annual salaries of research mathematicians (Y, in USD) and an index of publication quality (X_1), number of years of experience (X_2), and an index of success in obtaining grant support (X_3). The data for a sample of 24 intermediate and senior level research mathematicians follow (Neter et al., 1989).

i:	1	2	3	4	5	6	7	8	9	10	11	12
X_{i1}:	3.5	5.3	5.1	5.8	4.2	6.0	6.8	5.5	3.1	7.2	4.5	4.9
X_{i2}:	9	20	18	33	31	13	25	30	5	47	25	11
X_{i3}:	6.1	6.4	7.4	6.7	7.5	5.9	6.0	4.0	5.8	8.3	5.0	6.4
Y_i:	33.2	40.3	38.7	46.8	41.4	37.5	39.0	40.7	30.1	52.9	38.2	31.8

i:	13	14	15	16	17	18	19	20	21	22	23	24
X_{i1}:	8.0	6.5	6.6	3.7	6.2	7.0	4.0	4.5	5.9	5.6	4.8	3.9
X_{i2}:	23	35	39	21	7	40	35	23	33	27	34	15
X_{i3}:	7.6	7.0	5.0	4.4	5.5	7.0	6.0	3.5	4.9	4.3	8.0	5.0
Y_i:	43.3	44.1	42.8	33.6	34.2	48.0	38.0	35.9	40.4	36.8	45.2	35.1

Use the following multiple linear regression model is used to describe the relation:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i,\ i = 1, \ldots, 24$$

where the Xs are fixed and $\epsilon_i \sim \text{iid}\, N\left(0, \sigma^2\right)$, $i = 1, \ldots, 24$. Let $\hat{\beta}_j$ be the OLS estimator of β_j, $j = 0, 1, 2, 3$.

(a) Write down an algorithm that uses the bootstrap to estimate the standard error of $\hat{\beta}_2$.

(b) Using $B = 2000$ bootstrap samples, implement your procedure in (a) and compare your bootstrap estimate with $\sqrt{\text{MSE}\left((X'X)^{-1}\right)_{(22)}}$, which is the usual estimate for s.e. $\left(\hat{\beta}_2\right)$.

(c) The quantity β_1/β_3 is the ratio of the effects of the index of publication quality and index of success in obtaining granting support. If the ratio is equal to 1, it means the effects are equal. Suppose we estimate β_1/β_3 using $\hat{\beta}_1/\hat{\beta}_3$. Write down an algorithm that uses the bootstrap to compute a 95% percentile confidence interval for β_1/β_3.

(d) Using $B = 2000$ bootstrap samples (you may use the same ones as in part (b)), implement your procedure in (c) and provide the confidence interval. Is 1 inside the confidence interval? What does this mean?