Homework 3

February 22, 2023

Before getting into start working on this problem set, among other things, please get yourselves familiarized with how to use for loops in R.

1 (R) (Weak) Law of Large Numbers

- (a) Consider a continuous random variable $X_i \sim Uniform[0,2]$. What is $E[X_i]$ and $Var(X_i)$?
- (b) Consider the X_i defined above in (a) for $i = 1, 2, ..., n$, where each $X_i \perp\!\!\!\perp X_j$ whenever $i \neq j$. Consider the sample mean

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i.$$

  What is $E[\bar{X}_n]$ and $Var(\bar{X}_n)$? (The answer might be a function of n).

Now repeat the following (c)-(e) for $n = 1, 2, 3, 5, 10, 50, 100, 1000, 3000$. Use the for loops to execute (c)-(e).

- (c) (R) Generate a size n vector of independent $Uniform[0,2]$ random variables and calculate its sample mean $\bar{X}_n$.
- (d) (R) Take $|\bar{X}_n - E[X_i]|$ and report the value.
- (e) (R) Now consider a continuous transformation $f(x) = 2x^2 - 5x + 1 + \frac{1}{3x}$. Take

$$|f(\bar{X}_n) - f(E[X_i])|$$

  and report the value.

- (f) What happens to the reported value in (d) and (e) as n increases? Discuss.

2 (R) The Central Limit Theorem

- (a) Consider a continuous random variable $X_i \sim Uniform[0,2]$. What is $E[X_i]$ and $Var(X_i)$?
- (b) Consider the X_i defined above in (a) for $i = 1, 2, ..., n$, where each $X_i \perp\!\!\!\perp X_j$ whenever $i \neq j$. Consider the sample mean

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i.$$

  What is $E[\bar{X}_n]$ and $Var(\bar{X}_n)$? (The answer would be a function of n).

- (c) Consider the transformation

$$Y_n := \sqrt{n}\left(\bar{X}_n - E[X_i]\right).$$

  What is $E[Y_n]$ and $Var(Y_n)$?

- (d) Consider the transformation

$$Z_n := \sqrt{n}\frac{(\bar{X}_n - E[X_i])}{\sqrt{Var(X_i)}}.$$

  What is $E[Z_n]$ and $Var(Z_n)$?

Now repeat the following for $n = 1, 2, 3, 5, 10, 50, 100, 1000, 3000$. Use the for loops to execute (e)-(m).

- (e) (R) Generate $t = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, ..., 2500$ $n \times 1$ vectors of independent $Uniform[0,2]$ random variables and calculate its sample mean $\bar{X}_n^t$ respectively for each t. Denote this size 2500 vector as

$$\mathbf{v}_n^{2500} \equiv \left(\bar{X}_n^1, \bar{X}_n^2, ..., \bar{X}_n^{2500}\right)$$

  for now.

- (f) (R) Calculate the mean and variance of $\mathbf{v}_n^{2500}$ and report.
- (g) (R) Recall that each element of $\mathbf{v}_n^{2500}$ is composed of $\bar{X}_n^t$ for $t = 1, ..., 2500$. Now, for each $t = 1, 2, ..., 2500$, take the transformation

$$Y_n^t := \sqrt{n}\left(\bar{X}_n^t - E[X_i]\right)$$

and denote the transformed vector as

$$\mathbf{y}_n^{2500} \equiv \left(Y_n^1, Y_n^2, ..., Y_n^{2500} \right).$$

That is, subtract the $E[X_i]$ (that you calculated in (a)) from each element of $\mathbf{v}_n^{2500}$, and then multiply it by $\sqrt{n}$, and then denote it by $\mathbf{y}_n^{2500}$.

- (h) (R) Calculate the mean and variance of $\mathbf{y}_n^{2500}$ and report.
- (i) (R) Plot the histogram of $\mathbf{y}_n^{2500}$ and report.
- (j) (R) Recall that each element of $\mathbf{v}_n^{2500}$ is composed of $\bar{X}_n^t$ for $t = 1, ..., 2500$. Now, take the transformation

$$Z_n^t := \sqrt{n} \frac{\left(\bar{X}_n^t - E[X_i] \right)}{\sqrt{Var(X_i)}}$$

and denote the transformed vector as $\mathbf{z}_n^{2500}$. That is, subtract the $E[X_i]$ (that you calculated in (a)) from each element of $\mathbf{v}_n^{2500}$, then multiply it by $\sqrt{n}$, and divide it by $\sqrt{Var(X_i)}$ ($Var(X_i)$ you calculated in (a)), and then denote it by

$$\mathbf{z}_n^{2500} \equiv \left(Z_n^1, Z_n^2, ..., Z_n^{2500} \right).$$

- (k) (R) Calculate the mean and variance of $\mathbf{z}_n^{2500}$ and report.
- (l) (R) Plot the histogram of $\mathbf{z}_n^{2500}$ and report.
- (m) What happens to the reported values in (f), (h), (k) and histograms in (i) and (l) as n increases? Discuss.

3 (R) WLLN with Simple Regression

In this exercise, you will generate datasets for simple regression yourself, and then try to estimate the model parmaters to examine the properties of simple regression OLS estimators as the sample size n grows.

Repeat the following for $n = 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 30, 50, 75, 100, 250, 500, 1000, 2000, 3000$. Use the for loops as necessary.

- (a) (R) Generate a $n \times 1$ column vector of $Uniform[0, 12]$ random variable and denote it as x.

- (b) (R) Generate a $n \times 1$ column vector of $Uniform\,[-4, 4]$ random variable and denote it as u.
- (c) (R) Generate the y vector using the following formula:

$$y_i = 3 + 2x_i + u_i$$

for each $i = 1, 2, 3, ..., n$. That is, i'th row of x and u corresponds to i'th observation.

- (d) (R) Now you have a Monte-Carlo dataset of size n. Estimate the β in the following model

$$y_i = \alpha + \beta x_i + u_i$$

using OLS. (Recall the formula $\frac{\widehat{Cov}(x_i, y_i)}{\widehat{Var}(x_i)}$.) What is the calculated value of $\hat{\beta}_{OLS,n}$? Report.

- (e) What happens to $\left|\hat{\beta}_{OLS,n} - 2\right|$ as n increases? Discuss.

4 (R) CLT with Simple Regression

In this exercise, you will generate datasets for simple regression yourself, and then try to estimate the model parmaters to examine the properties of simple regression OLS estimators as the sample size n grows.

Repeat the following for $n = 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 30, 50, 75, 100, 250, 500, 1000, 2000, 3000$. Use the for loops as necessary.

- (a) (R) Generate a $n \times 1$ column vector of $Uniform\,[0, 12]$ random variable and denote it as x.
- (b) (R) Generate a $n \times 1$ column vector of $Uniform\,[-4, 4]$ random variable and denote it as u.
- (c) (R) Generate the y vector using the following formula:

$$y_i = 3 + 2x_i + u_i$$

for each $i = 1, 2, 3, ..., n$. That is, i'th row of x and u corresponds to i'th observation.

- (d) (R) Now you have a Monte-Carlo dataset of size n. Estimate the β in the following model

$$y_i = \alpha + \beta x_i + u_i$$

using OLS. (Recall the formula $\frac{\widehat{Cov}(x_i, y_i)}{\widehat{Var}(x_i)}$.) Save it in the memory.

- (e) (R) Repeat (a)-(d) for 2,500 times. You must have 2,500 $\hat{\beta}_{OLS,n}$ estimates in the memory at the end of this sub-question. Denote this size 2,500 vector by $\mathbf{b} = \left(\hat{\beta}^1_{OLS,n}, \hat{\beta}^2_{OLS,n}, ..., \hat{\beta}^{2500}_{OLS,n}\right)'$.
- (f) (R) Calculate the variance of $\mathbf{b}$ and report.
- (g) (R) Subtract 2 from $\mathbf{b}$ and multiply $\sqrt{n}$ on each element of $\mathbf{b}$ and denote this as $\mathbf{c} = (c_1, c_2, ..., c_{2500})$, i.e.,

  $$c_i = \sqrt{n}\left(\hat{\beta}^i_{OLS,n} - 2\right).$$

  Draw the histogram of $\mathbf{c}$ and report the histogram.
- (h) What happens to the reported values in (f) and the histogram in (g) as n grows large? Discuss.

5 (R) Video Game Sales Regression

The dataset for this exercise is available in VideoGamesSales_Main.csv. This dataset contains information on the global sales and critic and user review ratings for videogames launched between 2001 and 2012 (from www.vgchartz.com). The variables are:

Name of the game

Videogame platform on which it was released.

	Platform
DS	Nintendo DS
GBA	Nintendo Game Boy Advance
GC	Nintendo Game Cube
PC	Personal Computer
PS2	Sony PlayStation 2
PS3	Sony PlayStation 2
PSP	Sony PlayStation Portable
Wii	Nintendo Wii
XB	Microsoft XBOX

Videogame Genre (e.g., Action, Sports, Shooter etc.)

Publisher

Developer

Rating: E = Everyone, E10+ = Everyone 10+, T = Teen, M = Mature

Global Sales (Millions of units)

Year of release

Critic Score (0 – 100): Average critic rating

Critic Count : Number of critic ratings
User Score (0 – 10): Average user rating
User Count: Number of user ratings

- Your task is to develop a regression model (using lm()) that links global sales to video game reviews, and explore ways in which the model fit could be improved through suitable changes to the model specification and variables.

  (a) (R) First, create a frequency table of 3 variables: platform, genre, and rating.

  (b) (R) Create categorial variables for platform, genre, and rating using data.table. Also create a variable for the age of the game relative to year 2013 (Note that these games were released before 2013).

  (c) (R) Run a regression with all relevant X variables. Report the adjusted R-squared.

  (d) (R) Now, generate natural log of the following variables: global sales, critic_score, critic_count, user_socre, user_count as ln_[original_variable_name].

  (e) (R) Run a regression with the log of Y variable and report adjusted R-squared.

  (f) (R) Run a regression with the log of Y variable as well as log of X variables generated in part d). Report adjusted R-squared.

  (g) Which model (out of part c, e, and f) offers the highest adjusted R-squared? What would be the economic reasoning on why that particular model provides the best fit?

  (h) Interpret the parameter estimates for each of 'genre' in plain English.

  (i) Interpret the parameter estimate for 'rating' in plain English.

  (j) Interpret the parameter estimate for 'ln_user_count' in plain English.