ILLINOIS

STAT107 Data Science Discovery

Lab: Regression

Man Fung (Heman) Leung

Fall, 2021

University of Illinois at Urbana-Champaign

Collaboration

- Please work in a group of 2–4 students
  - collaboration is important in data science!
  - meet new friends and discuss :)
  - let us know if you have any questions

Comment: CLT

- Common/potential mistakes
    - 1.3: I let you go if you did not print mean/std. The solution does not do so as well and the rubric does not mention this issue
    - 2.2: some of you did not simulate $n = 1000$ per replication, or $r = 100, 10000$ replications (2 points deducted). Some of you did not write any for loop (4 points deducted). This puzzle worth 10 points
    - List Comprehension: one of you used this in 2.2. Great job! I did not know this originally
    - 3.3: some of your answers are not accurate enough (but they are not wrong). For example, we do not use CLT to calculate average. Instead, we use CLT to understand the variability of our average estimate

Comment: CLT

- Common/potential mistakes
    - Poisson: one of you mentioned that this was fish in French in your reflection. Yes :P I learn this from a friend in high school
    - Game: one of you mentioned Genshin Impact's wishing as an application of CLT. Indeed, I tried using CLT to assess if the "SS" probability of some card game is accurate as well
    - $n = 30$: it can be mathematically justified but at the end it is still a rule of thumb
    - Machine learning: some of you wondered if CLT can be used in any machine learning problem. The answer is YES! This is also related to one of my research areas
- Running the test cases successfully do not imply full score

Final remark

- Data science in real world
  - data collection
  - data cleansing
  - modelling/prediction
  - assumption/interpretation
- Resume tips
  - table with transparent border in Word (or use LaTeX)
  - one page only (reduce font size/margin properly if you need)
  - three points (max) per experience

Today's lab: Regression

- Main page
- Retrieve the lab using git
- Complete the notebook
  - 2.1: type `model.` and check the box in IDE to see how to access variables like intercept $\hat{\beta}_0$ and coefficient(s) $\hat{\beta}_1$
  - 2.2: the *p*-value is `2*scipy.stats.t.sf(abs(TEST_STAT), df=DEG_FREE)`
  - 3.1: use `MLB[["ERA"] + ["WAR"]]` (+ instead of ,) for multiple regression
  - 3.2: get predicted win by `model.predict(MLB[["ERA"] + ["WAR"]])`. The actual win is `MLB["W"]`
- Submit your work. Feel free to:
  - ask us questions
  - leave whenever you finish the lab

Checking completion

Default total number of cells: 59

- 0.0 in cell 4
- 1.0 in cell 6
- 1.1 in cell 9
- 1.2a in cell 12, 14, 16
- 1.2b in cell 18, 20
- 1.2c in cell 22, 24, 26 (reflection)
- 1.2d in cell 28
- 2.1 in cell 31, 33, 34, 36
- 2.2 in cell 38, 40
- 2.3 in cell 43 (reflection)
- 2.4 in cell 45, 46
- 3.1 in cell 48, 50
- 3.2 in cell 52
- 3.3 in cell 55
- 3.4 in cell 58 (reflection)