STAT107 Data Science Discovery

Lab: Regression

Man Fung (Heman) Leung

Spring, 2022

University of Illinois at Urbana-Champaign

Collaboration

ILLINOIS

- 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
- Attendance form
  - you can come up if you do not want to use this form
  - submit before you leave the lab

Announcement

ILLINOIS

- This is our last lab. No lab next week
- The last office hour will be held next Wed
- If you miss any lab notebook, you can (late) submit on or before 8 May (Sun) 23:59
  - To be fair to other students, a 20% penalty will be imposed in general
  - You can only submit if you completely miss the corresponding notebook
  - This only applies to my sections
  - Please email me stating which lab(s) you submit
- You can get 20 extra credit by completing the physical notes
  - the instructors said they would announce the detail after staff meeting yesterday
- Please fill in the ICES evaluation. I value your feedback

Comment: lab_clt

- Check email for score decomposition
- **LLN vs CLT**: they are both related to the behavior of sample average when sample size is large. However, CLT additionally tells how variable the sample average is at a particular (large) sample size
- **CLT does not guarantee the sample is normal**: it only states that the mean (or other summary statistics) of a large sample is normal
- 2.2: -0.5 if you use the wrong range. Note that `range(1,100)` is $1, 2, \ldots, 99$ in Python
- 2.3: -0.5 if your answer does not accurately describe the shape of the histograms

Comment: lab_clt

- 3.2: -0.5 if you plot columns other than `claims`. No points deducted if you do not separate the histograms. However, you should do that in practice since combining them is hard to read here
- 3.3: -0.5 if your answer does not demonstrate why CLT is necessary or contains misconception (see my elaborations in bold in previous page)

Extra credit: lab_clt

- 1 point for creating Cauchy average function
- 2 points for the three simulations
  - 1 point off if results are not reproducible (e.g., no `seed()`)
- 2 points max for the reflection question
  - All the other parts of the extra credit must be done to receive points for this
  - 1 point for each of the following observations/explanations
    - Dispersion: the range of simulated values is changing
    - Central tendency: the histograms do not look normal
    - Explanation: the moments of Cauchy random variable is undefined, so CLT does not apply
    - Other reasonable unique answers are accepted

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

ILLINOIS

- Main page
- Hints
  - 2.1: fit `model` from `LinearRegression()` first. 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)` or `2*scipy.stats.t.cdf(-abs(TEST_STAT), df=DEG_FREE)`
  - 3.1: use `MLB[["ERA", "WAR"]]` for predictors in 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