Data Analytics in Business

Zhaohu (Jonathan) Fan

Information Technology Management

Scheller College of Business

Georgia Institute of Technology

August 22, 2024

About Me

- Ph.D. in business analytics.
- Applied statistician, business practitioner.
- Foodie and more...

Previous experiences with R/RStudio

- I have never heard of R/have heard of R before, but have never used it.
- I have used R before in a class or at work.

If you need a quick refresher, please take a look at the HTML file called `Introduction to R and RStudio`.

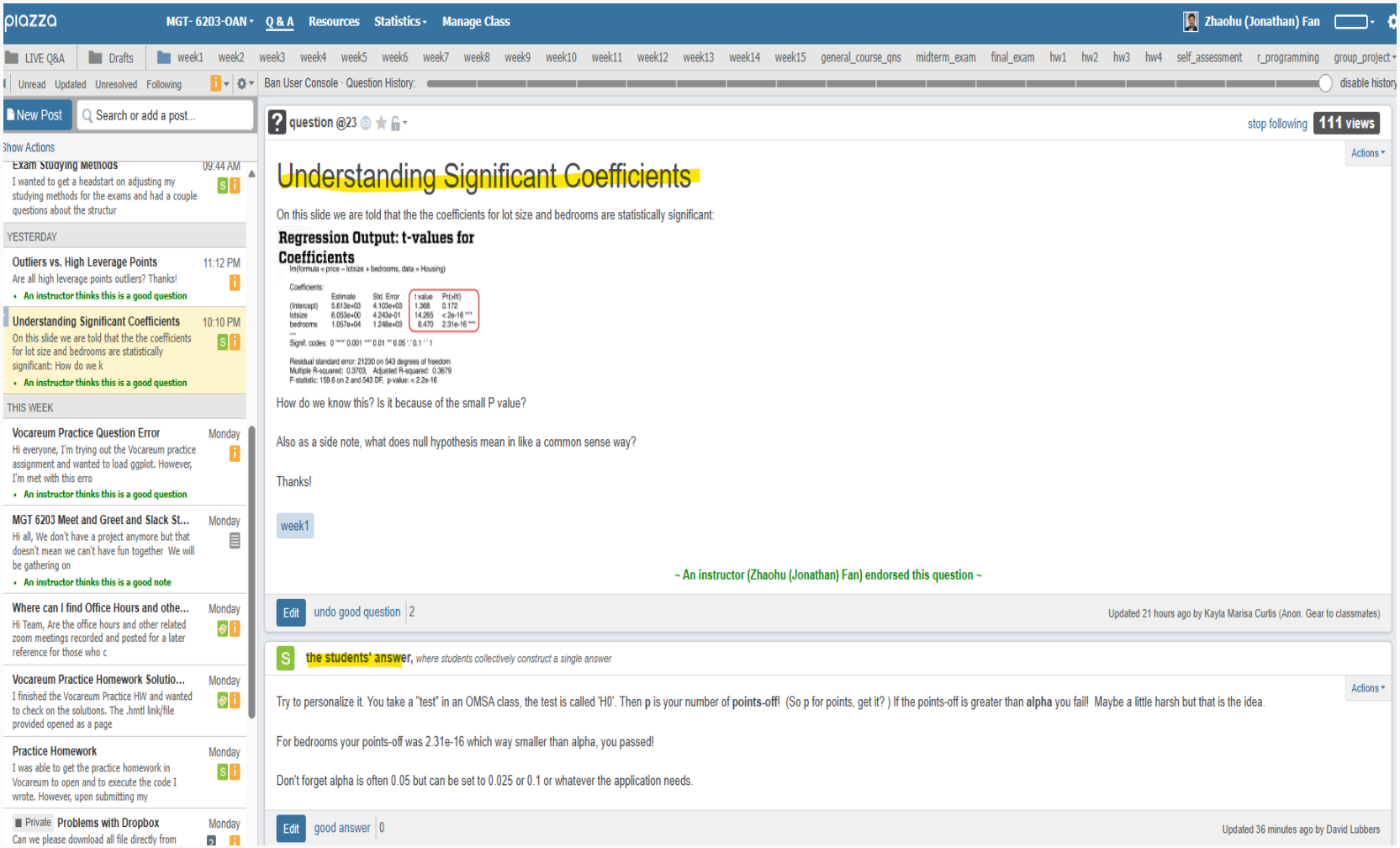
Meet our Teaching Assistants (TAs) Team

- **Ronak Patel** (Head TA)
  - **Evan Jones** (Lead TA for OMSA students in Canvas and Course Projects)
  - **Maria Fernanda Romero Creel** (Lead TA for MicroMaster students in EdX)
  - **James Brad Ashworth** (Lead TA for Course Projects)
  - **Xinyue Zhao** (Lead TA for Homework)
  - **Eli Colasante** (Lead TA for Homework)
  - **Plus 16 other talented TAs, each eager to assist and inspire!**
  - You'll have the opportunity to meet and interact with the teaching team during regular office hours(i.e., instructor's office hours and TA's office hours) and on our course's **Piazza Forum.**

Piazza

- **Primary Interaction Platform:** Piazza is our main platform for class interactions. Our teaching team is active here and looks forward to engaging with you.
    - **Active Participation Encouraged:** Since this is an online course, we highly encourage you to participate actively in discussions on Piazza.
    - **Search for Teammates! Check out post #5 titled 'Search for Teammates!' on our course's Piazza discussion board**
    - **Introduce Yourself:** Please make your first post on Piazza by introducing yourself to the class. In your post, include:
        - Your location
        - Your occupation
        - Your educational background
        - A list of your hobbies or special interests
        - What makes you unique
        - **Engage with Classmates:** Have fun with the introduction and feel free to respond to at least one or two of your classmates' posts.
- **Support from Teaching Assistants:** Designated TAs are available daily to address any course-related inquiries. Don't hesitate to post any questions – we're all here to help!

Piazza (cont'd)



Piazza (cont'd)

"Give a man a fish, and you feed him for a day. Teach a man to fish, and you feed him for a lifetime."

Image courtesy of AI Image Generator, generated on August 16, 2024.

Piazza (cont'd)

- **Steps for effective learning:**
  - **Identify your challenge:** Pinpoint where you're stuck.
  - **Try to solve it:** Work on a solution yourself first.
  - **Seek feedback:** If unsure, discuss your approach with us.
- **Why it helps:**
  - Builds critical thinking skills.
  - Equips you to handle new challenges at work.

Four modules

Thanks to our wonderful professors from our Scheller School of Business who made the videos for our four modules:

- Professor Sridhar Narasimhan for **Analytics**
    - Professor Jonathan Clarke for **Finance**
    - Professor Frederic Bien for **Marketing**
    - Professor Bob Myers for **Operations**

Course expectations

- Watch the course videos for each module.
  - Complete the homework assignments.
  - Take the self-assessment quizzes.
  - **Although optional, attending office hours is strongly recommended for clarifying homework questions and reviewing exams. These sessions will be recorded and available on Canvas.**

TA and Instructor's Weekly office hours

- **Instructor Session:**
  - Start Date: Thursday, August 22nd
  - Time: 8:30 PM Eastern Time
  - Frequency: Every Thursday from 8:30 to 9:30 PM Eastern Time
- **TA Session:**
  - Start Date: Monday, August 19th
  - Time: 8:30 PM Eastern Time
  - Frequency: Every Monday from 8:30 to 9:30 PM Eastern Time
- **Joining Instructions:**
  - Platform: Zoom
  - Access: Click on "Zoom" in the left panel on the Canvas course page.
    - Recordings of office hours can be accessed through the "Office Hours Recordings" module on Canvas.

Grading breakdown

- Self-Assessment Quizzes: 10 %
- Homework Assignments: 30 % (3 assignments, each worth 10 %)
- Group Project 15 %
- Midterm Exam: 20 %
- Final Exam: 25 %

Here’s the breakdown of the attempt allowances for our coursework

- **Weekly Self-Assessment (SA):** You have 2 attempts.
- **HW Part 1 (Theoretical part):** You have 1 attempt.
- **HW Part 2 (Computation part):** You also have 1 attempt.
- As a side note: You may work on (computation and theoretical) part of the homework for as long as you like within the given window. **As long as you do not click "submit," you can enter and exit the assignment as many times as necessary during the time period that it is available.** Again, please note, you should only click "submit" when you are completely finished with the assignment and ready to submit it for grading.

(Sick) Extensions Policy

- **Documentation Required:** All requests must be accompanied by verifiable documentation of the illness or leave event. This must be an official or signed notice.
- **Ineligible Documentation:** Screenshots or pictures of at-home COVID tests do not qualify, as we cannot verify their authenticity.
- **Advance Notice:** Please advise us ahead of a deadline. Extensions cannot be granted after the fact.
- **Approval Criteria:** Requests that do not meet these criteria will not be approved.

Open for discussion

Module 1 (weeks 1-5): Analytics

Basics of Statistics and Regression - Covers statistical concepts and regression techniques

Business context (Background & Purpose)

- **Regression models** are invaluable in real-world applications, such as in manufacturing or logistics.
- **Practical implications:**
  - In fields like manufacturing or logistics, **fractional predictions (e.g., from `lm(Qty~Price, data=price)`) are not practical**.
  - It's common to round up to ensure sufficient quantities for production and shipping.
    - Rounding up to avoid shortages and meet customer expectations.

R demo

- **Practical implications:**
  - In fields like manufacturing or logistics, **fractional predictions (e.g., from `lm(Qty~Price, data=price)`) are not practical**.
  - **Round up to the next whole number means moving to the nearest larger whole number**
  - Example: 2541.298 rounds up to 2542.
    - Use the `ceiling()` function in R to round up numbers.

```
# Given value
value ← 2541.298
# Round up to the next whole number
rounded_value ← ceiling(value)
# Print the rounded value
print(rounded_value)
2542
```

Q: How many different rounding methods are outlined in our course?

A: Let's use the examples provided as a reference.

Rounding methods

- `XX.xxx` means round to three decimal places, like `0.534`.
- `XX.xxx%` refers to rounding percentages to three decimal places, like `53.432%`.
- **Round Up To Next Whole Number:**
  - 101.22 would be 102
  - 101.001 would be 102
- **Round Down To Next Whole Number:**
  - 101.22 would be 101
  - 100.9999 would be 100
- If the instructions specify to (simply) round to a whole number, then **traditional rounding rules** should be followed.
  - 101.22 would be 101
  - 101.51 would be 102

Steps in a regression analysis

- Step 1. State the problem
- Step 2. Data collection (more details!)
- Step 3. Model fitting & estimation
    - Model specification (linear? logistic?)
    - Model fitting (least squares)
    - Select potentially relevant variables
    - Model validation and criticism
    - Back to 3.1? Back to 2?

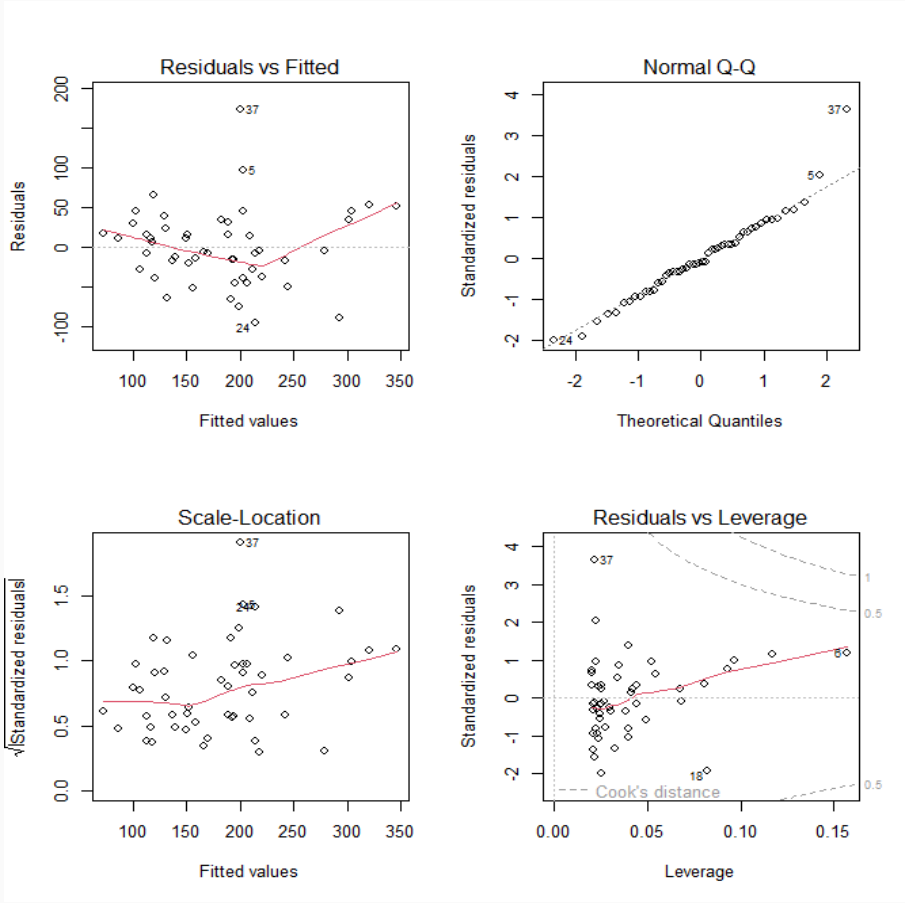
Assumptions of a regression (L.I.N.E)

- **L**inearity
  - The relationship between X and Y is linear.
- **I**ndependence of Errors
  - Error values are statistically independent.
  - Particularly important when data are collected over a period of time.
- **N**ormality of Error
  - Error values are normally distributed for any given value of X.
- **E**qual Variance (also called homoscedasticity)
  - The probability distribution of the errors has constant variance.

Assumptions of regression (L.I.N.E)

Code

```
plot(model1)
```



More examples of statistical relationships

- Simple linear regression: $Y = \beta_0 + \beta_1 X + \epsilon$
- Multiple linear regression: $Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \epsilon$
- Polynomial regression: $Y = \beta_0 + \sum_{i=1}^{p} \beta_i X^i + \epsilon$
- Logistic regression: $\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$
- Nonlinear regression: $Y = \frac{\beta_1 X}{(\beta_2 + X)} + \epsilon$
- and more...

Multiple linear regression

- Multiple linear regression: $Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon$
  - Is the linear regression as a whole effective and significant? **This leads to the test for significance of regression (or F-test).**
    - The F-test in multiple linear regression is used to assess whether the overall regression model is statistically significant.
  - Which specific regressors/independents seem important? **This leads to t-test.**
    - The t-test for each coefficient tests the null hypothesis that the coefficient is equal to zero (meaning the predictor does not have a statistically significant relationship with the response variable) against the alternative that it is not zero.

Test for significance of regression

- The test for significance of regression is also called F-Test or ANOVA test (analysis of variance test).
- It a test to determine if there is a linear relationship between the response and any of the regressor variables.
- The hypotheses are
  - $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$
  - $H1 : \beta_j \neq 0$ for at least one j

Interpreation of F Test

- **Reject null hypothesis** means the linear regression as a whole is significant in explaining the variation in response variable. At least one regressor is significant with nonzero slope.
- **Fail to reject null hypothesis** means the linear regression is not significant and all the slopes of regressors are zero.
- Confidence level of 95 $\%$ (significance level α =5 $\%$).

Test for significance of regression

- To decide whether to reject or not reject the null hypothesis, we look at the p-value of the F-test.
- To perform the F-test, we use `summary()`.

Determining the Significance of Individual Regressors

- To assess the significance of individual regressors, a t-test is performed for each coefficient.
- The null hypothesis for each t-test is $H_0 : \beta_j = 0$, where j is the index of the regressor.
- A low p-value for a t-test indicates that the corresponding regressor is significant.

What is the topic for next week?