# The University of Texas at Dallas

**School of Management**

**BUAN 6383/MIS 6386** **Modeling for**
**Syam Menon** **Business Analytics**

## Project I

# Objective

- Learn to build customized models

# Instructions

- **Due Date: See Syllabus**
  eLearning will stop accepting submissions after the due date, and late submissions will not be accepted
- **Submit one report per group via eLearning as a Microsoft Word document**
  - The report should be named **project-I-group*xx*.docx**
    (for example, group 05 should name the report project-I-group05.docx)
  - Clearly identify your group number and all group members on the cover page
  - A professional quality report is expected – messy or hard-to-read reports will be penalized
- **Submit all the code you have developed as a jupyter notebook**
  - The file should be named **project-I-group*xx*.ipynb**
  - (for example, group 05 should name the notebook project-I-group05.ipynb)
  - If you prefer, you can submit separate jupyter notebooks for each question If you choose to do so, the files should be named **project-I-group*xx*-p*yy*q*zz*.ipynb** (for example, group 05 should name the notebook
  - for Part 2, question 3 project-I-group05-p02q03.ipynb)
  - Clearly identify which question each part of the code is for, and what it is supposed to do
  - Clear, detailed comments are required; I should be able to run the codes you submit
- **This project counts for 120 points**

# Data Sets

- `billboard.csv`
- `khakichinos.csv`
- `books.csv`

# Part I: Replicating Models from Class

1. **The Poisson Model**
   Consider the example related to billboard exposures from class. The associated data is in the file `billboard.csv`. Write code to estimate the parameters of the **Poisson model** using maximum likelihood estimation (MLE). Report your code, the estimated parameters and the maximum value of the log-likelihood. Predict the number of people with 0, …, 23 exposures based on the Poisson model. Explain how the predicted values are obtained using the case of 2 exposures (show your calculations). Graph the original and predicted number of exposures (number of people on the y-axis and the numbers of exposures on the x-axis).
2. **The NBD Model**
   Next, write code (for the same dataset) to estimate the parameters of the **NBD model** using MLE. Report your code, the estimated parameters and the maximum value of the log-likelihood. Evaluate the NBD model vis-à-vis the

Poisson model; explain which is better and why. Predict the number of people with 0, …, 23 exposures based on the NBD model. Explain how the predicted values are obtained using the case of 2 exposures (show your calculations). Graph the original and predicted numbers of exposures.

3. **The Poisson Regression**
   Now consider the **khakichinos** example from class; The associated data is in the file `khakichinos.csv`. Estimate all relevant parameters for **Poisson regression** using MLE. Report your code, the estimated parameters and the maximum value of the log-likelihood. Predict the number of people with 0, …, 23 exposures based on the Poisson regression. Explain how the predicted values are obtained using the case of 2 exposures (show your calculations). Graph the original and predicted numbers of exposures.

4. **The NBD Regression**
   Consider the **khakichinos** example again. Estimate all relevant parameters for **NBD regression** using MLE. Report your code, the estimated parameters and the maximum value of the log-likelihood. Evaluate the NBD regression vis-à-vis the Poisson regression; explain which is better and why. Predict the number of people with 0, …, 23 exposures based on the NBD regression. Explain how the predicted values are obtained using the case of 2 exposures (show your calculations). Graph the original and predicted numbers of exposures.

5. For each of the models above, can you provide some managerial takeaways?

# Part II: Analysis of New Data

`books.csv` contains information on customer purchases from `amazon.com` and `barnesandnoble.com` in 2007 (see variable `domain`). Various variables on customer characteristics are also in the dataset. Information available on these variables is below. There are a few other variables in the dataset – the date of each purchase (`date`), the product purchased (`product`), the number of copies purchased (`qty`), and the price paid (`price`); we will not use `date`, `product`, and `price` for this project.

| Variables | Description |
|---|---|
| `education` | ordered categorical (range: 0 to 5; higher values → higher education level) |
| `age` | ordered categorical (range: 1 to 11; higher values → older) |
| `income` | ordered categorical (range: 1 to 7; higher values → higher income) |
| `region` | categorical (values: 1 to 4) |
| `race` | categorical (values: 1 to 5) |
| `country` | binary |
| `child` | binary (1 → children in the household; 0 → no children in the household) |
| `hhsz` | numeric (household size; range: 1 to 6) |

Suppose you are working for Barnes and Noble, and would like to understand the factors that affect customer purchasing behavior there. In particular, you are interested in the questions below. Your objective is to leverage the modeling skills you have learned so far from this class to answer these business questions.

1. Read `books.csv` and generate two new datasets –

   a. `books01.csv`, with the structure of the dataset used in the **billboard exposures** example (i.e., with only two columns – (i) the number purchases, and (ii) the number of people making the corresponding number of purchases), and

   b. `books02.csv`, with the structure of the dataset used in the **khakichinos** example, with a new column containing a count of the number of books purchased from `barnesandnoble.com` by each customer, while keeping the demographic variables (remember to drop `date`, `product`, and `price`).

   Print the first and last 10 records of both new datasets.

2. Develop a **Poisson model** using `books01.csv`. Report your code, the estimated parameters and the maximum value of the log-likelihood (and any other information you believe is relevant).
3. Develop a **Poisson model** using `books02.csv`, i.e., *by ignoring the independent variables available*. Report your code and confirm that the estimated parameters and the maximum value of the log-likelihood are identical to those obtained with the Poisson model developed using `books01.csv`.

   Predict the number of people with 0, …, 20, 20+ visits based on the Poisson model. Explain how the predicted values are obtained using the case of 2 exposures (show your calculations). Graph the original and predicted number of visits.

4. Develop an **NBD model** using `books01.csv`. Report your code, the estimated parameters and the maximum value of the log-likelihood (and any other information you believe is relevant).
5. Develop an **NBD model** using `books02.csv` (again, ignoring the variables available). Report your code, and confirm that the estimated parameters and the maximum value of the log-likelihood are identical to those obtained with the NBD model developed using `books01.csv`.

   Predict the number of people with 0, …, 20, 20+ visits based on the NBD model. Explain how the predicted values are obtained using the case of 2 exposures (show your calculations). Graph the original and predicted number of visits.

6. Calculate the values of (i) reach, (ii) average frequency, and (iii) gross ratings points (GRPs) based on the NBD model. Show your work.
7. Identify all independent variables with missing values. How many values are missing in each? Drop any variable with many missing values (specify how you are defining "many"). If the number of missing values are very few (again, specify how you are defining "few"), delete the rows involved. For the remaining variables (if any), replace the missing values with the means of the corresponding variables. Explain the steps taken; report your code.
8. Incorporate all the available customer characteristics and estimate all relevant parameters for **Poisson regression** using MLE. Report your code, the estimated parameters and the maximum value of the log-likelihood (and any other information you believe is relevant). What are the managerial takeaways – which customer characteristics seem to be important?

   Predict the number of people with 0, …, 20, 20+ visits based on the Poisson regression. Explain how the predicted values are obtained using the case of 2 exposures (show your calculations). Graph the original and predicted number of visits.

9. Estimate all relevant parameters for **NBD regression** using MLE. Report your code, the estimated parameters and the maximum value of the log-likelihood (and any other information you believe is relevant). What are the managerial takeaways – which customer characteristics seem to be important?

   Predict the number of people with 0, …, 20, 20+ visits based on the NBD regression. Explain how the predicted values are obtained using the case of 2 exposures (show your calculations). Graph the original and predicted number of visits.

10. Evaluate all the models developed using the log-likelihood ratio, AIC, and BIC. What are your recommendations on which model to use based on each of these criteria? Are the recommendations consistent? Explain why you are recommending the model you have selected. Are there any significant differences among the results from the models? If so, what exactly are these differences? Discuss what you believe could be causing the differences.

Briefly summarize what you learned from this project. This is an open-ended question, so please include anything you found worthwhile – relating to the modeling process, insights from the process and models, any managerial takeaways that were insightful to you, and so on.