# Homework 3

**BUAN/MKT 6337 Section 001/002**

The dataset for this exercise is available in VideoGamesSales_Main.7bdat. 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 |
| **X360** | Microsoft XBOX 360 |
| **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

1. Develop a regression model (using **proc reg**) that links global sales to video game reviews. Explore ways in which the model fit could be improved through suitable changes to the model specification and variables.
   a. First, use proc freq to create a frequency table of 3 variables: platform, genre, and rating.
   b. In data step, create categorial variables for platform, genre, and rating using if/else statements. Also create a variable for the age of the game relative to year 2013.
   c. Run a regression with all relevant X variables. Report the adjusted R-squared.
   d. Now, generate natural log of the following variables: global sales, critic_score, critic_count, user_socre, user_count.
   e. Run a regression with the log of Y variable and report adjusted R-squared.

f. 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? Also interpret coefficients.

2. For the model you constructed in Q.1 part g), verify whether the various regression assumptions discussed in class are satisfied. If an assumption is violated, discuss how it can be handled, and implement the solution if possible. Discuss whether the solution had a practically significant impact on your model results. If no practical solution is possible, acknowledge it.

3. Replicate your final model in Q.1 part g) using **proc glm** and **class** commands. Note that **proc glm** allows you quickly check many basic results. However, it does not provide many diagnostic outputs that **proc reg** provides.