The University of Texas at Dallas

School of Management

BUAN 6383/MIS 6386
Syam Menon

Modeling for
Business Analytics

Homework 01

Objective

- Gain an understanding of Principal Component Analysis

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 **hw01-group*xx*.docx**
    (for example, group 05 should name the report hw01-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 **hw01-group*xx*.ipynb**
  - (for example, group 05 should name the notebook hw01-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 **hw01-group*xx*-p*yy*q*zz*.ipynb** (for example, group 05 should name the notebook
  - for Part 2, question 3 hw01-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 homework counts for 70 points**

Data Sets

- `boxOffice.csv`

Internet Buzz and the Movie Box Office

In his undergraduate honors thesis, Versaci (2009) investigated what he termed “internet buzz variables” to see whether they provide any additional predictive information towards a movie's box office revenues (beyond movie characteristics like genre, actors, budget, etc.). `boxOffice.csv` contains this data involving 62 movies (all wide-released movies between November 7, 2008 and April 3, 2009); the variables available (along with their descriptions) are in the table below. We will conduct the analysis ignoring the “buzz” variables (`addict`, `cmngsoon`, `fandango`, and `cntwait3`) first.

Variables	Description
box	domestic opening weekend box office revenues ($)
G PG PG13	binary variables indicating MPAA rating code (if all 3 are 0, the movie is rated R)
budget	production budget (in millions of $)
starpowr	star power rating based on the Forbes 2009 Star Currency list (range: 0 to 10)
sequel	binary (1 → sequel; 0 → not a sequel)
action comedy animated horror	binary variables indicating movie genre (if all 4 are 0, the movie genre was drama)
addict	number of trailer views at traileraddict.com
cmngsoon	number of message board comments at comingsoon.net
fandango	sum of “can't wait” and “don't care” votes at fandango.com
cntwait3	percent of “can't wait” votes at fandango.com

1. Plot histograms of the continuous variables (box, budget, starpwr) to see if any transformations are needed. Are any of them skewed? Apply a log-transformation to all the skewed variables.
2. Run a linear regression of box office revenues on the “traditional” variables (i.e., using all the independent variables (except the “buzz” variables). If any variables were transformed, be sure to use the transformed versions of those variables. What are the R^2 and adjusted- R^2 values? Which variables (if any) are significant at the 0.10 level, based on the t-statistics and associated probabilities (p > |t|)?
3. Run another linear regression using only the variables that were significant (again, ignoring the “buzz” variables). What are the R^2 and adjusted- R^2 values? Are all the variables still significant at the 0.10 level?
4. Plot histograms of the four “buzz” variables. Are any of them skewed? Apply a log-transformation to all the skewed variables.
5. Run a linear regression of box office revenues on *all* the independent variables, including the “buzz” variables (transformed as needed). What are the R^2 and adjusted- R^2 values? Which variables (if any) are significant at the 0.10 level, based on the t-statistics and associated probabilities (p > |t|)?
6. Run another linear regression using only the variables that were significant. What are the R^2 and adjusted- R^2 values? Are all the variables still significant at the 0.10 level?
7. Compare the models developed so far – which of these would you choose, and why?
8. Apply Principal Component Analysis to just the 4 “buzz” variables. If you transformed any of them, make sure you use the transformed versions. Also make sure that you standardize the variables first. What are the eigen values associated with each component? How many principal components are selected using (i) Kaiser's Rule, and using “explained variance" thresholds of (ii) 60%, (iii) 70%, (iv) 80% and (v) 90%?
9. Run a linear regression using all the “traditional” independent variables (if transformed, use the transformed versions) *and* all 4 principal components (the only variables you should not use here are the four “buzz” variables). What are the R^2 and adjusted- R^2 values? Which variables (if any) are significant at the 0.10 level? In particular, are any of the principal components significant? What can you say about this model vis-à-vis the other models built so far?
10. Now run regressions using the number of principal components based on (i) Kaiser's Rule and “explained variance” thresholds of (ii) 60%, (iii) 70%, (iv) 80% and (v) 90% (if any of the models are identical, point this out and run it only once). Compare all the regression models involving the principal components (including the one involving all four components). Which of these would you recommend, and why?

11. Now apply Principal Component Analysis to the 4 “buzz” variables *and* the other continuous variables (`budget` and `starpowr`). Again, use transformed versions of the variables if any were transformed and standardize the variables first. What are the eigen values associated with each component? How many principal components are selected using (i) Kaiser's Rule, and using “explained variance" thresholds of (ii) 60%, (iii) 70%, (iv) 80% and (v) 90%?

12. Next, run regressions using the number of principal components based on (i) Kaiser's Rule and “explained variance” thresholds of (ii) 60%, (iii) 70%, (iv) 80% and (v) 90%. Compare these regression models and explain which one you would recommend, and why?

13. Are the “buzz” variables helping build a better model? How about PCA?

14. Did you learn anything surprising while doing these analyses? Can you provide some managerial takeaways?