SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

Applying Frequentist and Bayesian Logistic Regressions to MOOCs Data in SAS: a Case Study

Yan Zhang, PhD, Educational Testing Service

#SASGF

Applying Frequentist and Bayesian Logistic Regressions to MOOCs Data in SAS: a Case Study

Yan Zhang[1], Yoav Bergner[1], Ryan Baker[2]

[1]Educational Testing Service; [2]Columbia University

INTRODUCTION

Massive Open Online Courses (MOOCs) platforms provide free higher education courses to Internet users worldwide. However, MOOCs have high enrollment but notoriously low completion rates.

The goal of this study is to use frequentist and Bayesian logistic regressions to investigate whether and how students' engagement, intentions, education levels, and other demographics are conducive to MOOC course completion.

The original data used in this study came from an online 8-week course titled "Big Data in Education" taught within the Coursera platform (MOOC) by Teachers College, Columbia University.

The datasets for analysis were created from three different sources – clickstream logs, a voluntary pre-course survey, and homework assignments.

METHODS

Data Processing

1) Data transformation

```
DATA one;  SET one;  watchvideo=log(watchvideo+0.01);   viewpost=log(viewpost+0.01); RUN;
```

2) Detect multicollinearity

```
PROC REG DATA= MOOC_data;
  MODEL grade = Q4_1 -- Q11_1 watchvideo viewlecture viewpost thread_T numthread /TOL VIF COLLIN;
RUN;
```

Frequentist approach

```
PROC SURVEYLOGISTIC data=trainset varmethod=jackknife (outjkcoefs=three) ;
  Model grade(ref='0')= watchvideo viewpost viewlecture;
   store logit_jackknife;          RUN;
```

Bayesian approach

```
PROC GENMOD DATA=trainset DESCENDING;
  MODEL grade=watchvideo viewpost viewlecture/link=logit dist=b;
   BAYES seed=15 nbi=1000 nmc=21000 outpost=post cprior=jeffreys diagnostics=all statistics=(summary
interval);
   STORE logit_bayes;
 RUN;
```

RESULTS

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-17.5684	4.0891	18.4590	<.0001
Q4	1	1	0.1686	0.2840	0.3524	0.5528
Q5	0	1	0.2498	0.2828	0.7800	0.3771
Q7	0	1	-0.0661	0.4714	0.0196	0.8885
Q8	0	1	-0.1011	0.3076	0.1080	0.7424
Q10	0	1	-0.4362	0.5122	0.7252	0.3944
Q11	1	1	-0.1728	0.2932	0.3473	0.5556
threads	0	1	0.6671	1.0249	0.4237	0.5151
enter_day		1	-0.0639	0.1634	0.1532	0.6955
numthread		1	0.9284	0.8818	1.1084	0.2924
thread_T		1	0.00797	0.0519	0.0236	0.8780
watchvideo		1	-0.5742	0.2108	7.4222	0.0064
viewpost		1	1.3919	0.3367	17.0919	<.0001
viewlecture		1	3.0101	0.9126	10.8798	0.0010

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	
Intercept	1	-15.6794	1.6187	-18.8520	-12.5068
watchvideo	1	-0.5439	0.1029	-0.7454	-0.3423
viewpost	1	1.3128	0.1982	0.9244	1.7012
viewlecture	1	2.8035	0.4029	2.0138	3.5932
Scale	0	1.0000	0.0000	1.0000	1.0000

Table 1. Maximum likelihood estimates of logistic regression by by frequentist (left) and Bayesian (right) logistic regressions.

Table of F_grade by I_grade

F_grade(From: grade)	I_grade(Into: grade) 0	1	Total
0	371 85.48 96.36 96.87	14 3.23 3.64 27.45	385 88.71
1	12 2.76 24.49 3.13	37 8.53 75.51 72.55	49 11.29
Total	383 88.25	51 11.75	434 100.00

Table of grade by grade1

grade	grade1 0	1	Total
0	372 85.71 96.62 96.88	13 3.00 3.38 26.00	385 88.71
1	12 2.76 24.49 3.13	37 8.53 75.51 74.00	49 11.29
Total	384 88.48	50 11.52	434 100.00

Table 2. Contingency tables between predicted and observed grade values for frequentist (left) and Bayesian (right) logistic regressions.

Applying Frequentist and Bayesian Logistic Regressions to MOOCs Data in SAS: a Case Study

Yan Zhang[1], Yoav Bergner[1], Ryan Baker[2]

[1]Educational Testing Service; [2]Columbia University

RESULTS CONTINUED

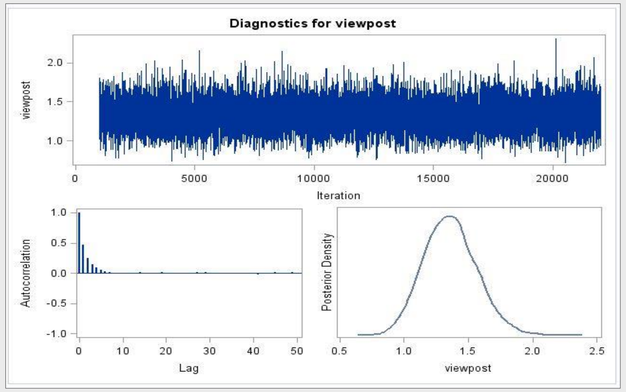

Figure 1. Diagnostic plots for the variable viewpost

CONCLUSIONS

The frequentist and Bayesian approaches gave the similar results on accuracy, recall and precision of our logistic models. One possible explanation is that the sample size of our data was not small enough to show the benefits of Bayesian logistic regression.

Our MOOC results suggest that the students with higher course engagement and forum activities will more likely complete MOOC courses.

REFERENCES

Zhu, M., Bergner, Y., Zhang, Y., Baker, R., Wang, Y., Paquette, L., & Barnes, T. (2016).Longitudinal Engagement, Performance, and Social Connectivity : a MOOC Case Study Using Exponential Random Graph Models. In *the Proceeding of the 6th International Learning Analytics and Knowledge Conference (LAK '16)*. Edinburgh, UK.

Efron,B. and Tibshirani,R.J.(1993) An Introduction to the Boot-strap. Chapman and Hall, New York.

Anscombe, F. J. (1949) *The transformations of Poisson, binomial, and negative binomial data*, Biometrika 35: 246—254.