Paper 10663-2016

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

Yan Zhang, Yoav Bergner, Educational Testing Service;

Ryan Baker, Columbia University

ABSTRACT

Massive Open Online Courses (MOOCs) have attracted increasing attention in educational-data-mining research. MOOC platforms provide free high 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. The SAS system offers multiple procedures to perform logistic regression, with each procedure having different features and functions. In this study, we applied two approaches - frequentist and Bayesian logistic regressions to MOOC data. Specifically, PROC SURVEYLOGISTIC was used for the frequentist approach and PROC GENMOD for the Bayesian analysis.

Our results suggest that MOOC students with higher course engagement and higher motivation are more likely to complete courses. Specifically, three predictors – the number of times of viewing discussion forum posts, viewing instructors' lectures, and watching lectures' videos are the most powerful predictors for course completion.

INTRODUCTION

Massive Open Online Courses (MOOCs) offer free online courses to students worldwide who lack access to elite universities for higher education. However, the overall completion rates for MOOCs courses are very low (2% ~ 10%). The goal of this study was to investigate which components (e.g., students' course engagement) can facilitate higher students' completion. In this study, we applied two different approaches, frequentist and Bayesian, to analyze MOOC data and compared the results.

Frequentist and Bayesian analyses are different approaches in statistics. The frequentist approach assumes data are random sample and parameters are fixed while the Bayesian approach assumes both data and parameters are random. The Bayesian analysis uses prior knowledge of the parameters and the conditional distribution of the data given the parameters to be employed to find the posterior distributions of the parameters given the observed data.

In this study, we want to introduce the general procedures of these two approaches to SAS users by using the MOOC data and then compared the results. PROC SURVEYLOGISTIC with Jackknife variance estimation method was used for the frequentist approach while PROC GENMOD with the BAYES statement for the Bayesian approach.

DATA SOURCES

The data used were collected from an 8-week Coursera course, "Big Data in Education" (BDE), offered by Ryan Baker's laboratory at Teachers College, Columbia University from October to December 2013. Three different types of data files were used: website clickstream logs, survey questionnaires, and homework and exam grades. The clickstream logs include time stamps, the destination URLs students visited, multiple unique IDs that identify web visitors, forums, posts, comments, and post threads. A pre-

course survey questionnaire was voluntarily filled out before students enrolled in this course. It contains 13 questions, including gender, age, educational level, motivation, and self-estimated probability of finishing this course. There were eight weekly announcements to remind students about new course materials, lecture videos, and quiz-due time.

Each event in the clickstream logs gave an indication of whether the students were involved in the course learning activities, watching lecture videos and viewing forum posts. Seven features (predictors) related to learning activities and forum participations were extracted from the aggregated clickstream logs. For example, the predictor *viewpost* meant the number of times students viewed discussion forum posts during the 8-week course periods.

In this study, the response variable *Grade* was derived from the final scores which were based on eight homework assignments and one final exam. If a student finished at least one quiz or final exam, he/she would have a final score. If a student had his/her final score higher than 70 out of 100, the response variable *Grade* was treated as 1, otherwise as 0. Only the students which had final scores and finished survey questionnaires were included for further analysis. Activities beyond the eight weeks were excluded from this study. All the statistical analyses were conducted with SAS 9.3.

DATA PREPROCESSING

This study included six continuous variables and seven categorical variables shown in Table 1. We followed the following three steps for data preprocessing. For categorical variables, the question items with multiple categories were re-coded to [0, 1] or [1, 2] because we wanted to ensure that the variables were measured at the interval level.

Variable	Range	Meaning
Q4	(0, 1)	Gender
Q5	(2, 8)	Age levels
Q7	(1, 16)	Current job sector
Q8	(1, 5)	The reasons for enrolling this course
Q10	(0, 10)	Self-estimated probability of finishing the course
Q11	(1, 2)	Native English speaker
viewpost	(0, 3347)	Number of times viewing forum posts
viewlecture	(0, 828)	Number of times viewing or downloading lecture handouts
watchvideo	(0, 11871)	Number of times watching instructors' videos
threads	(0, 1)	1 if students posted a thread, otherwise 0
thread_T	(1, 56)	First day a student posted a new thread
enter_day	(0, 56)	First day to log in BDE course
numthread	(1, 17)	Number of new threads a student posted in discussion forum
grade	(0, 1)	Response variable. 1: complete, 0: fail

Table 1. A list of variables used in this study

First, for continuous variables, outlier were removed based on diagnostic criteria (e.g. Cook's D) provided by PROC LOGISTIC. The RESCHI= option names Pearson Chi-square residuals and the H= options specifies the diagonal element of the Hat matrix. The Cook's D can be calculated by cookd= ((chires**2)*hatdiag)/(5*(1-hatdiag)**2).

```
PROC LOGISTIC DATA=MOOC_data;
  CLASS threads Q4 - Q11;
  MODEL grade (ref='0') = Q4-Q11 threads-viewlecture/INFLUENCE IPLOTS;
  OUTPUT OUT=LOGOUT p=yhat reschi=chires resdev=devres
difchisq=difchisq h=hatdiag;
RUN;
```

Second, in order to stabilize the variance of the continuous variables with high maximum values, a simple log transformation was applied to three variables – *watchvideo*, *viewpost*, and *viewlecture*. To handle zero values for the log transform, a small positive constant 0.01 was added to the three variables.

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

Third, the REG procedure was used to diagnose multicollinearity among all the predictors since to our knowledge it is the only SAS procedure that can calculate VIF and TOL. All other results from this procedure should be ignored. The VIF and TOL options in MODEL statement were used to calculate the variance inflation factor and tolerance. The predictors with high VIF values (VIF>5) were removed from the model. In PROC REG, dummy variables were created for categorical variables (e.g. *Q4_1*, *Q4_2*). All the variables were included in further analysis since all of them had VIF< 4. After removing duplicates, missing data and outliers, around 1200 students were included in further analysis.

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

LOGISTIC REGRESSION ANALYSIS

The data was split into two data sets - train set (70%) and test set (30%) by the SURVEYSELECT procedure for cross validation. The METHOD=SRS option with REP=1 means simple random sampling without replacement.

```
PROC SURVEYSELECT DATA = one METHOD = SRS OUTALL REP = 1 SAMPRATE = 0.7
SEED = 12345 OUT = OUTALL;
 ID _all_;
RUN;
```

FREQUENTIST APPROACH

Since the number of the independent variables in this study was small, we didn't apply any automated feature selection methods (e.g., stepwise). Instead, we ran the logistic regression with all the variables included and selected variables based on their standard errors and Wald Chi-square statistics. The SURVEYLOGISTIC procedure was applied to the data. To avoid and correct bias on regression coefficients, jackknife variance estimation was applied. The VARMETHOD = jackknife option invokes the delete-1 jackknife variance estimation method. The reference levels for categorical variables were set to zero. All the significance levels were set at 0.05.

```
PROC SURVEYLOGISTIC data=trainset varmethod=jackknife;
 class threads Q4 Q5 Q7 Q8 Q10 Q11;
 Model grade(ref='0')= Q4 Q5 Q7 Q8 Q10 Q11 threads enter_day numthread
thread_T watchvideo viewpost viewlecture;
RUN;
```

Results of frequentist logistic regression

The regression coefficient estimates, standard errors and Chi-square statistics were shown in Table 2.

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

Table 2, Maximum likelihood estimates of frequentist logistic regression.

Three variable *viewpost, watchvideo,* and *viewlecture* had significant regression coefficients, p<0.005. Thus, these three variables were selected to fit the logistic model. We reran logistic regressions with only these three variables. The STORE statement saves the statistical parameters and the results obtained from PROC SURVEYLOGISTIC.

The SCORE statement in PROC PLM uses stored estimates to make predictions for a new dataset - *testset*. The ILINK option uses the inverse of the link function to obtain the probability of the predicted grade variable. The results showed the three variables had significant coefficient estimates, p<0.0001.

```
PROC SURVEYLOGISTIC data=trainset varmethod=trainset;
  Model grade(ref='0')= watchvideo viewpost viewlecture;
  store logit_jackknife;
RUN;
PROC PLM source=logit_jackknife;
  score data=testset out=logit_jackknife_test predicted=pred
lclm=lower uclm=upper/ilink;
RUN;
data logit_jackknife_test;
      set logit_jackknife_test;
      if pred>=0.5 then grade1=1;
      else if pred<0.5 then grade1=0;
RUN;
```

The mean squared error (MSE) for the fitted logistic model was 0.044, and was 0.100 for the intercept-only model. PROC FREQ was run to calculate Goodman and Kruskal's lambda for the association between predicted and observed values of *grade*. The value of lambda was equal to 0.55.

The AUC (also called c-statistic) of the fitted model was 0.973, significantly different from AUC=0.5 (p<0.0001). In general, the high AUC suggests that the fitted logistic model has high predictive power. However, since we were interested in the *grade=1* events (~10%), besides the AUC criteria, it was necessary to investigate the accuracy, sensitivity and positive predicted values of the fitted models.

We applied the fitted model to *testset* by the SCORE statement in PROC PLM. The contingency table between predicted and observed grade values was shown in Table 3. The *I_grade* variable was for predicted values and *F_grade* for observed values. The accuracy of this logistic regression model was 94%, sensitivity (recall) 73%, specificity 97%, positive predicted value (precision) 76%.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-15.6794	1.8515	71.7121	<.0001
watchvideo	1	-0.5439	0.1172	21.5187	<.0001
viewpost	1	1.3128	0.2039	41.4692	<.0001
viewlecture	1	2.8035	0.4874	33.0876	<.0001

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 3. Maximum likelihood estimates (left panel) and the contingency table between predicted and observed *grade* values (right panel) for the frequentist logistic regressions.

BAYESIAN APPROACH

We used the three selected variables - *viewpost*, *viewlecture*, and *watchvideo* to build the Bayesian logistic model. Compared with the LOGISTIC procedure, the GENMOD procedure offers a convenient way to run Bayesian logistic analysis by adding the BAYES statement. The prior information for all three variables used Jeffreys' prior. A sample code was provided below:

```
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 of Bayesian logistic regression

The trace plot, autocorrelation, and kernel density plots of all six variables suggested that Markov chains successfully converged and the mean of the Markov chain had stabilized. One example for the variable *viewpost* was shown in Figure 1.

The Bayesian estimates for three variables were shown in Table 4. The MSE of the fitted Bayesian model was 0.044, and was 0.100 for the intercept-only Bayesian model. The Goodman and Kruskal's lambda was equal to 0.56.

The Bayesian estimates and the standard errors were the same as those from the frequentist approach. The contingency table between predicted (*grade1*) and observed (*grade*) values was almost the same as that from frequentist approach shown in Table 4. The Bayesian approach showed a very small increase (<1%) in accuracy and sensitivity.

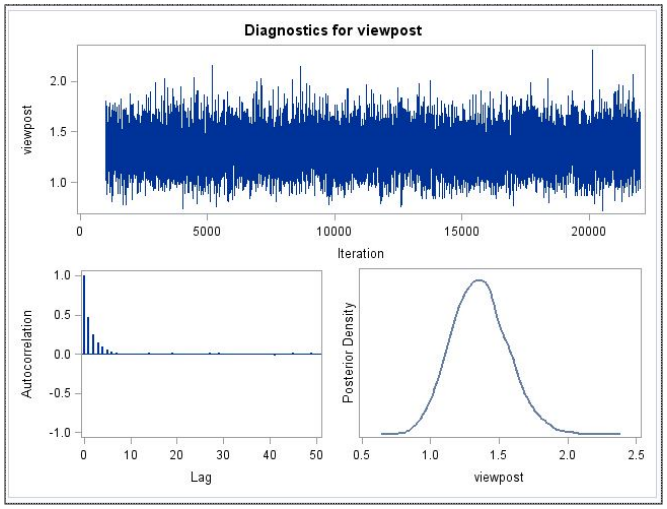

Figure 1. Diagnostic plots for the variable viewpost

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 of grade by grade1			
	grade1		
grade	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 4. Maximum likelihood estimates (left panel) and the contingency table between predicted and observed *grade* values (right panel) for the Bayesian logistic regression.

CONCLUSION

In this study, we applied the frequentist and Bayesian approaches to build logistic regressions from MOOC data. The two approaches gave the same results for the logistic models. One possible explanation is that the sample size of our data was not small enough to show the benefits of the Bayesian logistic regression. Unlike Bayesian analysis, the frequentist approach doesn't require prior information and is computationally inexpensive. The frequentist approach works well for a large sample size.

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

ACKNOWLEDGEMENTS

The work has supported by NSF (DRL-1418219). I would like to thank Shelby Haberman, Steven Holtzman and Chunyi Ruan for their valuable editing advice.

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.

Contact information

Your comments and questions are valued and encouraged. Contact the author:

Yan Zhang, PhD
Data Analyst and Research Technologies
Educational Testing Service
660 Rosedale Rd, Princeton, NJ 08540
Phone: (609) 734-5889
yzhang002@ets.org

Trademarks