Using Logistic Regression: A Case Study

Impact of Course Length and Use as a Predictor of Course Success

Presented by:
Keith Wurtz, *Dean, Institutional Effectiveness, Research & Planning*
Benjamin Gamboa, *Research Analyst*

Session Objectives

- Learn some of the advantages of using Logistic Regression
- Briefly learn how to conduct a Logistic Regression Analysis
- Learn some strategies for sharing the results with Faculty, Managers, and Staff

Advantages of Using Logistic Regression

- Logistic regression models are used to predict dichotomous outcomes (e.g.: success/non-success)
- Many of our dependent variables of interest are well suited for dichotomous analysis
- Logistic regression is standard in packages like SAS, STATA, R, and SPSS
- Allows for more holistic understanding of student behavior

Advantages of Using Logistic Regression

- The candidate predictor variables **do not** have to be…
  - Normally distributed
  - Linearly related
  - Have equal variances
- Candidate predictor variables can be…
  - Continuous
  - Dichotomous

Consider the Following when Setting-Up LR Analysis

- Setting up the Database
- Dummy coding
- Controlling for the number of predictor variables
- Multicollinearity
- Missing Cases (not discussed here, see Wurtz, 2008 and Harnell, 2001)

Setting-Up the Database

- Are summer terms included in the analysis?
- Which grade is most informative to the research question being examined?
  - first grade earned in the course,
  - the highest grade, or
  - the most recent grade
- How many years and/or what terms should be included in the analysis?
- Are status dates related to the analysis?

Dummy Coding

- In order to meet assumptions of LR independent variables need to be interval, ratio, or dichotomous
- Dummy coding allows the research to transfer a nominal (e.g.: ethnicity) or ordinal variable (e.g.: age categories) to a dichotomous variable

Dummy Coding Example

Ethnicity	Dummy Code	Value Labels
Asian	1	Asian Students
African American	0	All other Students
Hispanic	0	
Native American	0	
Caucasian	0	
Unknown	0	

Dummy Coding Example

Ethnicity	Dummy Code	Value Labels
Asian	0	All other Students
African American	1	African American Students
Hispanic	0	All other Students
Native American	0	
Caucasian	0	
Unknown	0	

Controlling for the Number of Predictor Variables

- The decision about whether there are too many predictor variables is related to the number of cases
- A high number of predictor variables can lead to a model that over fits the data
  - Over fits the data – model is too complex in relation to the number of candidate predictors and the number of cases
- One technique for controlling for the number of candidate predictors is to test for multicollinearity

Multicollinearity

- Multicolinearity – the Independent Variables are very highly correlated (r >= .80) with each other
- LR assumes that IVs are not correlated with each other

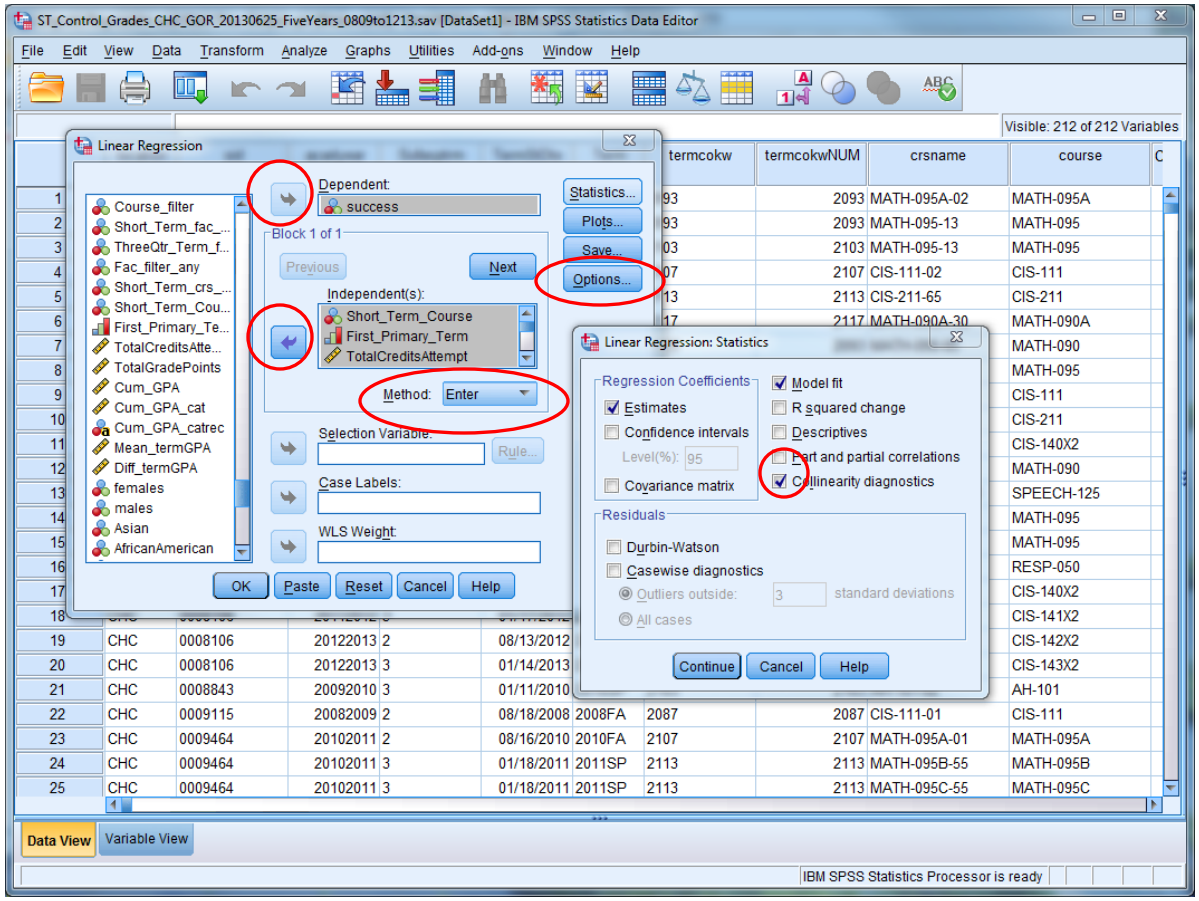
Setting Up Multicollinearity Test

- Independent variables:
  - Categorical (dummy-coded):
    - Ethnicity
    - Gender
    - Placement test results
    - First primary term of enrollment
    - Course length
  - Continuous:
    - Age at beginning of term
    - Normalized prior cumulative GPA
    - Prior credits attempted
    - Prior grade points earned

Setting Up Multicollinearity Test

- Multiple regression
  - In SPSS, select Analyze > Regression > Linear
  - Pull over dependent variable: course success (GOR of A, B, C or P/CR)
  - Pull over candidate predictor variables
  - Select "Enter" method
  - Open Statistics dialog box, check Collinearity diagnostics

Setting Up Multicollinearity Test



Multicollinearity Results

- Multicollinearity detected ($\beta \geq 1.0$ and/or tolerance ≤ 0.01) between:
  - Gender (both male and female)
  - Caucasian students
  - Certain student assessment placements in reading, math & English
- These variables had a moderate to high intercorrelation and were not be used as candidate predictors in further analysis.

Multicollinearity Results

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	.038	.112		.340	.734		
	Cumulative GPA of prior terms	.147	.004	.301	34.948	.000	.967	1.034
	Short term courses	.076	.009	.076	8.897	.000	.988	1.012
	Indicator of first term	.004	.023	.001	.169	.866	.995	1.005
	Asian	.131	.037	.063	3.549	.000	.226	4.429
	AfricanAmerican	.077	.036	.044	2.161	.031	.170	5.894
	Hispanic	.037	.033	.039	1.099	.272	.058	17.143
	NativeAmerican	.042	.043	.013	.977	.328	.426	2.348
	Age at start of the term	.002	.000	.033	3.802	.000	.962	1.040
	females	.138	.108	.151	1.282	.200	.005	194.519
	males	.128	.108	.140	1.187	.235	.005	194.466
	Caucasian	.072	.033	.079	2.167	.030	.054	18.608

a. Dependent Variable: Success Rate

Note: this is an abridged linear regression result from SPSS output window for illustration purposes only.

Multicollinearity Results

- Multicollinearity was not detected in the remaining variables:
  - Ethnicity (other than Caucasian)
  - First primary term of enrollment
  - Age at the beginning of the respective term
  - Normalized prior cumulative GPA
  - Prior credits attempted
  - Prior grade points earned
  - Course length
- These variables were suitable independent variables for use in logistic regression analysis.

Formula to Control for Overfitting the Model

- $p < m/10$
- Where p is the number of candidate predictor variables and m is the number of cases in each group of the dependent variable
- If m was 981 then p equals 98

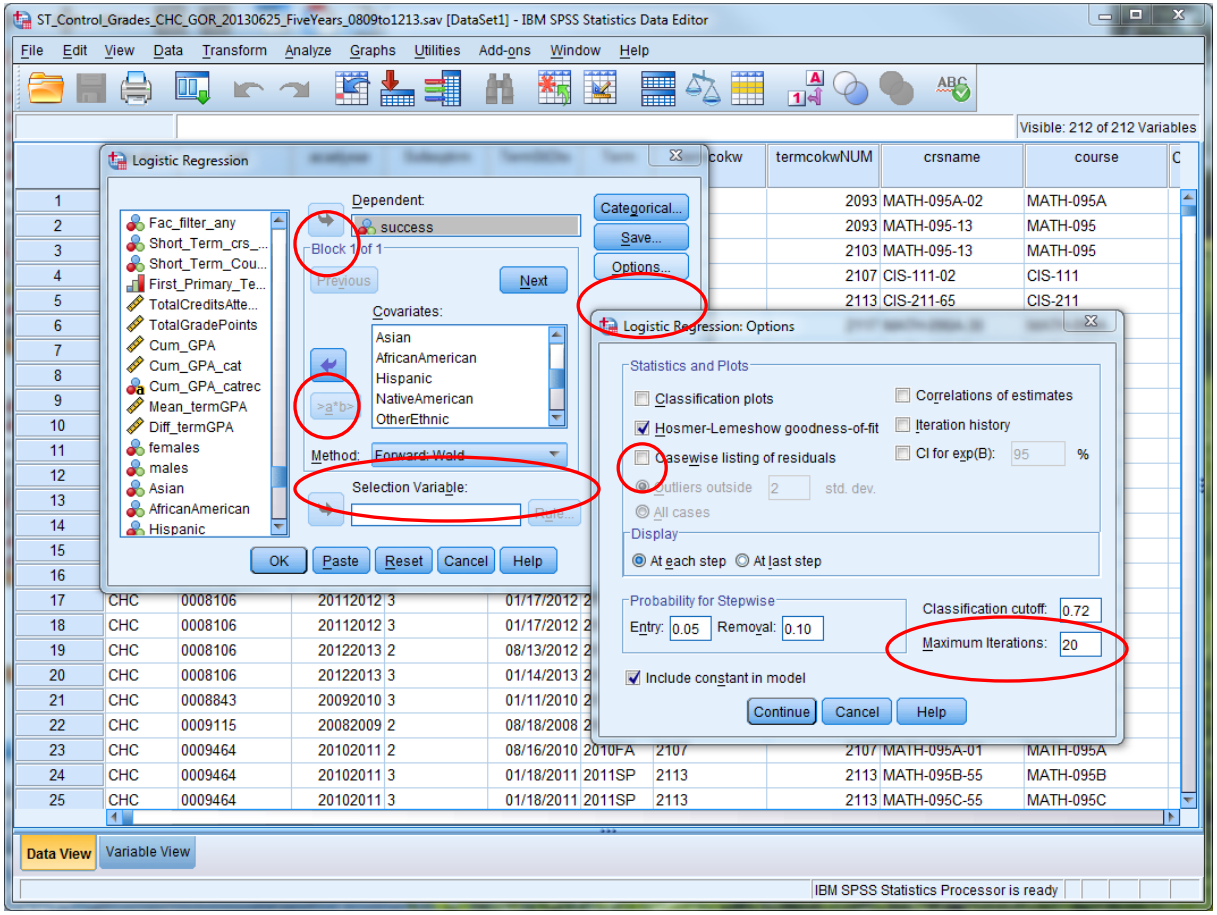
Need to Consider the Following when Conducting LR Analysis

- Setting Up Logistic Regression
- Setting the Cutoff Value
- Selecting the best model
- Interpreting the individual predictors
- Interpreting the odds ratios when they are negative

Setting Up Logistic Regression

- Logistic Regression
  - In SPSS, select Analyze > Regression > Binary Logistic
  - Pull over dependent variable: course success (GOR of A, B, C or P/CR)
  - Pull over candidate predictor variables
  - Select "Forward: Wald" method
  - Open Options dialog box,
    - Check Hosmer-Lemeshow goodness-of-fit test
    - Set Classification cutoff value to current average

Setting Up Logistic Regression



Setting the Cutoff Value

- The cutoff value is the probability of obtaining a 1 (e.g.: course success)
- The cutoff value directly impacts the results generated for the classification tables
- The default is set at .50
- Set the cutoff value to match the current probability of success
- Example: If trying to increase success in an English course and the success rate is 61%, set the cutoff value as .61

Selecting the Best Model

- An acceptable model would have an overall percentage correct greater than cutoff value
- None of the models reached this threshold
- Model 2 had the highest overall percentage correct of 66.2%
- Provided the best predictors of student success, because it had the highest overall percentage correct of all eleven models

Interpreting Logistic Regression Re

Classification Table[a]

Observed			Predicted		
			Success Rate		
			Not Successful	Successful	Percentage Correct
Step 1	Success Rate	Not Successful	2261	1366	62.3
		Successful	2925	5938	67.0
	Overall Percentage				65.6
Step 2	Success Rate	Not Successful	2324	1303	64.1
		Successful	2923	5940	67.0
	Overall Percentage				66.2
Step 3	Success Rate	Not Successful	2330	1297	64.2
		Successful	2936	5927	66.9
	Overall Percentage				66.1
Step 4	Success Rate	Not Successful	2408	1219	66.4
		Successful	3133	5730	64.7
	Overall Percentage				65.2
Step 5	Success Rate	Not Successful	2402	1225	66.2
		Successful	3126	5737	64.7
	Overall Percentage				65.2

Note: this is an abridged logistic regression classification table result from SPSS output window for illustration purposes only.

Interpreting Logistic Regression Results

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	90.336	8	.000
2	69.443	8	.000
3	75.928	8	.000
4	23.662	8	.003
5	19.670	8	.012
6	22.099	8	.005
7	19.768	8	.011
8	21.032	8	.007
9	17.570	8	.025
10	24.190	8	.002
11	33.170	8	.000

- Hosmer & Lemeshow goodness-of-fit statistic is statistically significant suggesting that the model is not reliable.

Interpreting the Best Predictors

- Although the model is not determined to be reliable, examining the predictor variables included—and the predictor variables not included—in Model 2 provides insight into what possible relationships may or may not exist.
- Two best predictors in this model are cumulative prior GPA and course length.
- Student is two times more likely to succeed for every 1 point increase in that student's prior cumulative GPA.
- Student enrolled in a compressed course is one and a half times more likely to succeed than a student enrolled in a traditional-length course.

Logistic Regression Results

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Cum_GPA	.723	.023	1015.505	1	.000	2.061	1.971	2.155
	Constant	-1.061	.063	279.234	1	.000	.346		
Step 2[b]	Short_Term_Course	.440	.048	84.594	1	.000	1.553	1.414	1.706
	Cum_GPA	.734	.023	1030.557	1	.000	2.083	1.991	2.178
	Constant	-1.208	.066	335.723	1	.000	.299		
Step 3[c]	Short_Term_Course	.433	.048	81.251	1	.000	1.541	1.403	1.693
	TotalGradePoints	.002	.000	36.679	1	.000	1.002	1.001	1.003
	Cum_GPA	.683	.024	808.590	1	.000	1.981	1.889	2.076
	Constant	-1.224	.066	344.462	1	.000	.294		
Step 4[d]	Short_Term_Course	.461	.048	90.973	1	.000	1.585	1.442	1.743
	TotalCreditsAttempt	-.059	.004	176.937	1	.000	.943	.935	.951
	TotalGradePoints	.022	.002	197.331	1	.000	1.023	1.019	1.026
	Cum_GPA	.432	.029	215.278	1	.000	1.540	1.454	1.632
	Constant	-.472	.083	32.149	1	.000	.624		
Step 5[e]	Short_Term_Course	.460	.048	90.462	1	.000	1.584	1.440	1.741

Note: this is an abridged variable coefficients result from SPSS output window for illustration purposes only.

Interpreting Odds Ratios when they are Negative

- The "TotalCreditsAttempt" variable had a negative regression coefficient (i.e. β), -.059
- Students who attempted a lower number of units were more likely to successfully complete their courses
- Using the inverse Odds-Ratio (i.e. 1/logg odds) allows researcher to calculate the impact of the predictor variable on the outcome
- The inverse odds-ratio was 1 / .943 which equals 1.06
- Students were only slightly more likely to successfully complete their courses if they were enrolled in less units

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 4[a]	Short_Term_Course	.461	.048	90.973	1	.000	1.585	1.442	1.743
	TotalCreditsAttempt	-.059	.004	176.937	1	.000	.943	.935	.951
	TotalGradePoints	.022	.002	197.331	1	.000	1.023	1.019	1.026

Strategies for Sharing the Information with Faculty, Managers, and Staff

- Know your audience
- Always start with the limitations
- Avoid using the following words to describe any part of the research: logistic regression, multicollinearity, dichotomous, etc.
- If appropriate, point out the difference in language when the results are described: relationship versus causation
- Mention that the full report that describes all of the methodology and limitations is available, but share the results in summary/visual form
- Discuss possible implications of the results and remind the audience that the results need to inform the discussion, not make the decision

Questions/Discussion

References

DesJardins, S.L. (2001). A comment on interpreting odds-ratios when logistic regression coefficients are negative. *The Association for Institutional Research*, *81*, 1-10. Retrieved October, 15, 2006 from http://airweb3.org/airpubs/81.pdf

George, D., & Mallery, P. (2006). *SPSS for windows step by step: A simple guide and reference (6th ed.)*. Boston: Allyn and Bacon.

Harrell, F.E. (2001). *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer Science+Business Media, Inc.

RP Group (2013). *Suggestions for California Community College Institutional Researchs Conducting Prerequisite Research*. Retrieved January 29, 2014 from http://www.rpgroup.org/sites/default/files/RPGroupPreqreqGuidelinesFNL.pdf

Tabachnick, B.G. & Fidell, L.S. (2007). *Using Multivariate Statistics (5th ed.)*. Boston: Pearson Education.

Wetstein, M. (2009, April). *Multivariate Models of Success*. PowerPoint presentation at the RP/CISOA Conference, Tahoe City, CA. Retrieved January 28, 2014 from http://www.rpgroup.org/sites/default/files/Multivariate%20Models%20of%20Success.pdf

Wurtz, K. A. (2008). A methodology for generating placement rules that utilizes logistic regression. *Journal of Applied Research in the Community College*, *16*, 52-58.