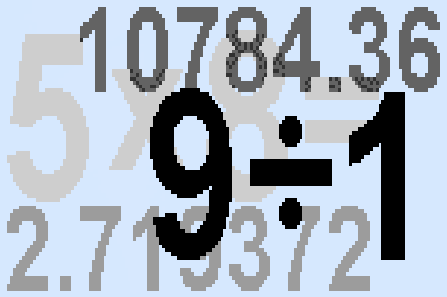
A Step-by-Step Introduction to Building a Student-at-Risk Prediction Model Using SPSS

******

http://www.unr.edu/ia/research
http://www.uhwo.hawaii.edu/academics/oie/research-and-presentations/

Serge Herzog, PhD
Director, Institutional Analysis
Consultant, CRDA StatLab
University of Nevada, Reno
Reno, NV, serge@unr.edu

John Stanley, MEd
Director, Institutional Research
University of Hawaii – West Oahu
Kapolei, HI, jstanley@hawaii.edu



AIR Forum 2017
Washington D.C May 29th – June 2nd

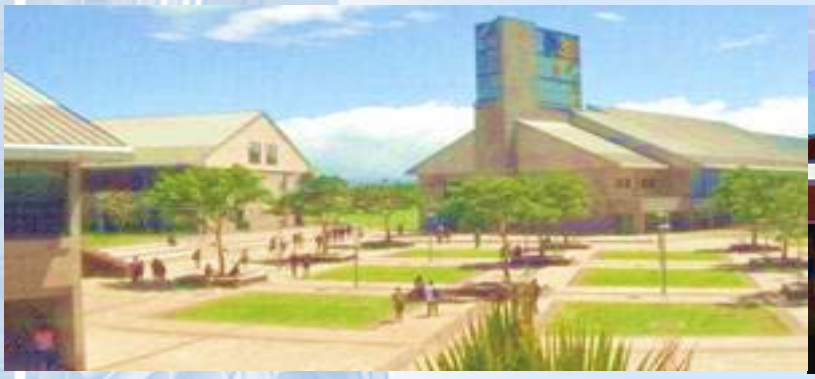
Workshop Objectives

1. Develop a conceptual understanding of how predictive models developed by an IR office can improve institutional effectiveness;
2. Learn how to set up a matriculation system (or census warehouse) data file in SPSS that can be used to develop a predictive statistical model to identify students at risk;
3. Learn how to use historical data to 'automatically' develop predictor coefficients to estimate (score) the dropout risk for students in future cohorts; and
4. Learn how to translate the student dropout risk into a relative percentile risk score to assist student support services with 'actionable' information.

Two Institutions, One Mission

University of Hawai'i West O'ahu

University of Nevada, Reno



PacAIR
Pacific Association for Institutional Research

AIR
Association for Institutional Research

RMAIR
ROCKY MOUNTAIN ASSOCIATION FOR INSTITUTIONAL RESEARCH

Challenges for Institutional Research

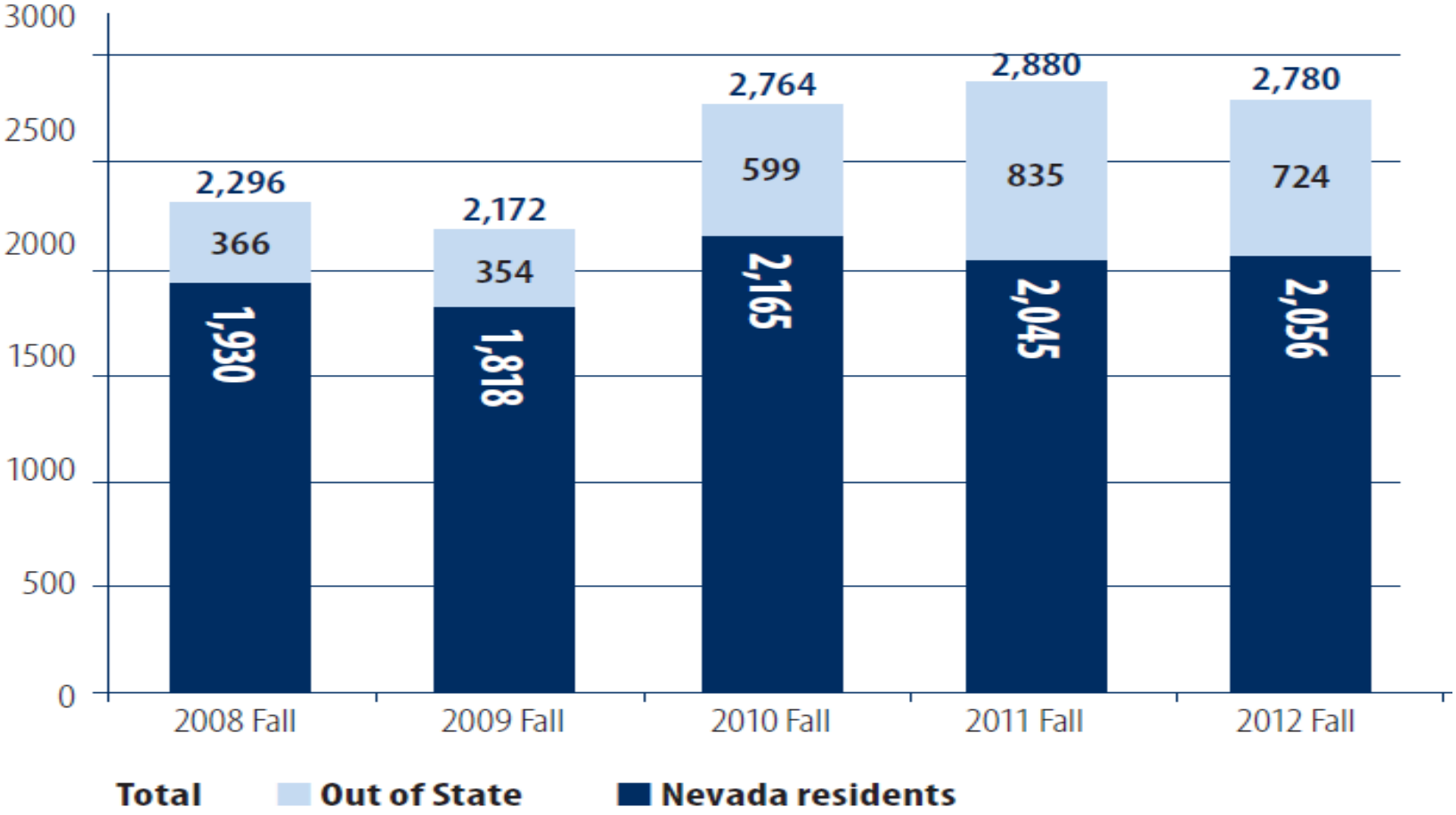
- Compliance vs. Self-Improvement
- Developing a culture of evidence
- From reporting to analysis
- Converting results into ‘actionable’ statements
- From ‘data silos’ to integrated warehouse
- Leverage technology, stay abreast of tech
- Follow highest standards, best practices
- Know your customers, mission
- Empower staff, continuous honing of skills

The Institutional Context

- Student success: a strategic imperative
- Performance-based state funding impending
- Dwindling state support for higher education
- Tuition-revenue maximization
- Reputation and marketing
- Effective senior-management support by IR
- K-16 Education Collaborative
  - High school transcript study
  - High school gateway curriculum
  - Reversing the tide of college remediation

The Institutional Context

New Freshmen Enrollment



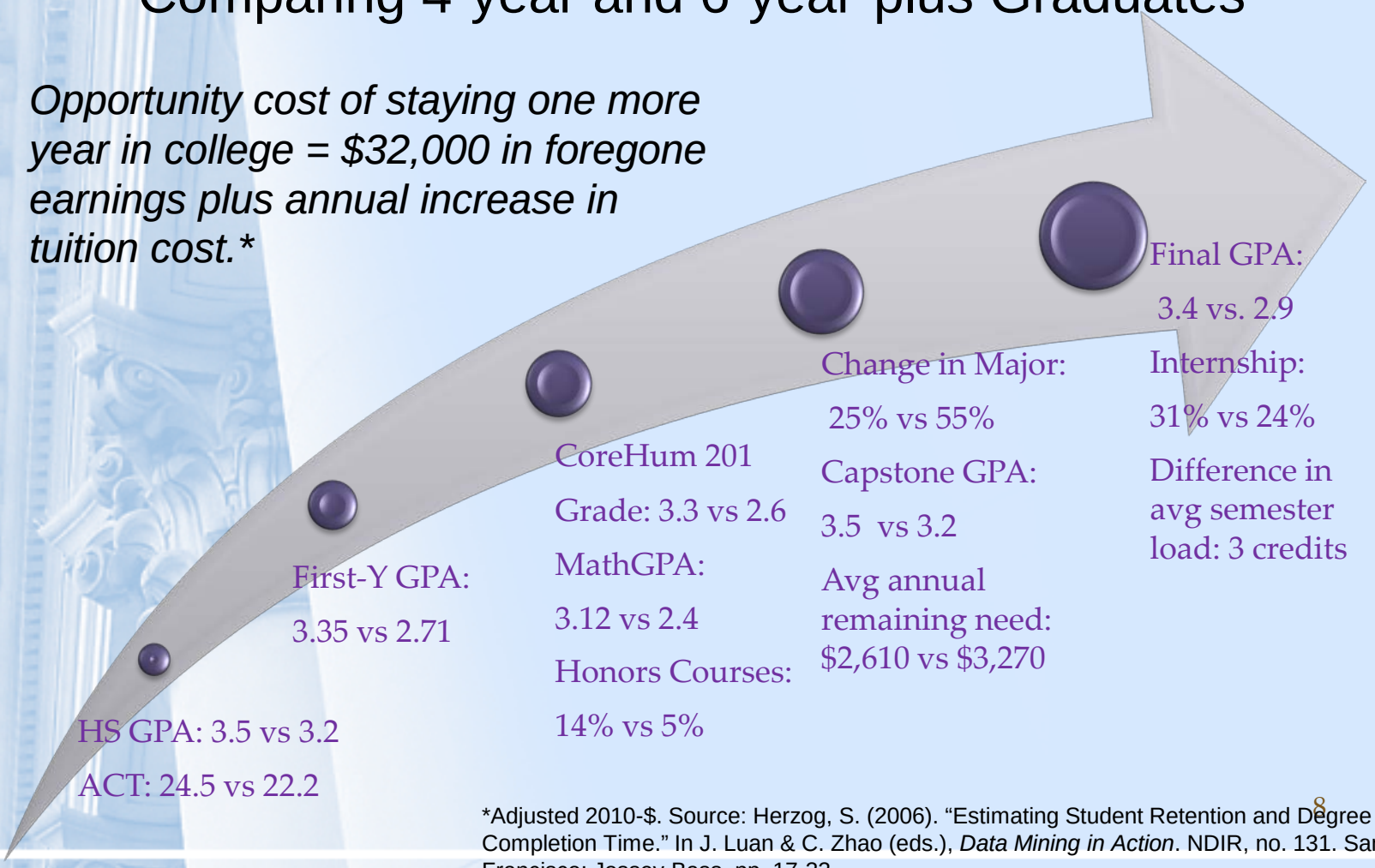
Examples of Actionable Findings

- Study abroad enhances academic performance
  - http://www.cis.unr.edu/IA_Web/research/USACConfOct2010.pdf
- Impact of classroom facilities/schedule on learning
  - Smaller rooms are preferable
  - After-2pm courses associated with lower performance
  - http://onlinelibrary.wiley.com/doi/10.1002/ir.224/abstract
- Student financial aid to maximize retention
  - Tuition discounts for middle-income students
  - More academic support for low-income students
  - http://www.uark.edu/ua/der/EWPA/Research/School_Finance/1802.html
- Effect of high school environment on freshmen success
  - http://www.uark.edu/ua/der/EWPA/Research/Achievement/1808.html

Raising Graduation Rates

Comparing 4-year and 6-year-plus Graduates

*Opportunity cost of staying one more year in college = $32,000 in foregone earnings plus annual increase in tuition cost.**



*Adjusted 2010-$. Source: Herzog, S. (2006). "Estimating Student Retention and Degree Completion Time." In J. Luan & C. Zhao (eds.), *Data Mining in Action*. NDIR, no. 131. San Francisco: Jossey-Bass, pp. 17-33.

Improving the Bottom Line

- Rise in freshmen retention by 4 percentage points due to better at-risk forecasting
  - AY 2010-11 *additional net tuition revenues* = ***$215,119*** (for 94 NV,19 WUE, excl OS students) for one cohort in one year, without OS $!
  - Downstream *cumulative additional net tuition revenues* result in $ millions!
- Incentive for student to speed up graduation
  - Opportunity cost per year in foregone earnings = ***$32,000*** per year (published constant 2010-$)

Relevant Previous Research

- Allison, P. (2012). Logistic regression for rare events. *Statistical Horizons*, retrieved at http://www.statisticalhorizons.com/logistic-regression-for-rare-events
- Caison, A. L. (2006). Analysis of institutionally specific retention research: A comparison between survey and institutional database methods. *Research in Higher Education* 48(4): 435-451.
- DesJardins, S. T. (2002). An analytical strategy to assist institutional recruitment and marketing efforts. *Research in Higher Education* 43(5).
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters* 27: 861-874.
- Herzog, S. (2005). "Measuring determinants of student return vs. dropout/stopout vs. transfer: a first-to-second year analysis of new freshmen." *Research in Higher Education,* 46(8): 883-928.
- Herzog, S. (2006). "Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression." In J. Luan & C. Zhao (eds.), *Data Mining in Action: Case Studies of Enrollment Management. New Directions for Institutional Research,* no. 131. San Francisco, CA: Jossey-Bass.
- Hosmer, D. & Lemeshow, S. (2000). *Applied Logistic Regression (Second Edition).* New York: John Wiley & Sons, Inc.
- Pascarella, E. T., & Terenzini, P. T. (2005). *How College Affects Students: Volume 2, A Third Decade of Research.* San Francisco, CA: Jossey-Bass.

Impact of this At-Risk Forecasting Model

- *University Retention Rates Hold Steady As States Balance Access with Success*. Scripps Howard Foundation Wire, April 15, 2011.
- *Managing Talent: HCM and Higher Education*. Campus Technology Magazine, October 2010, Vol. 24 Number 2, pp. 36-42.
- *From Data to Information: Business Intelligence and Its Role in Higher Education Today*. University Business Magazine, January 2009, pp. 25-27.
- Consulting services to IR offices at institutions in Arizona, California, Hawaii, and Texas.

At-Risk Forecasting Model

- Identify at-risk freshmen students after initial matriculation for *early* intervention program
- Develop regression model to predict dropout risk of future cohort
  - Determine baseline retention to maximize correct classification
  - Identify statistical outliers to get trimmed dataset
  - Chose model with optimal balance in correct classification
- Dropout risk scoring for new freshmen
  - Transformation of the logit(*p*) into probability scores
  - Automated classification and probability score with SPSS
  - Decile grouping of scored students
- Reporting of dropout risk via secure online access

- Data sources
  - Matriculation system (Peoplesoft, data warehouse)
  - New student survey (in PS starting fall 2011)
- Student cohorts
  - New full-time first-year students (incl. advanced standing)
  - Historical cohorts: fall 2011-15 (training set, N = 4,446)
  - Predicted cohort: fall 2016 (holdout set, N = 986)
  - Excluding ~ 10% of students without entry survey data
- Data elements (predictors) at start of first semester
  - Student socio-demographics (personal, parent attributes)
  - Academic preparation (high school GPA, test scores)
  - Financial aid profile (unmet need, aid type received, income)
  - Student motivation (proxy variables)
  - Student social integration (on-campus experiences)
  - Student academic experience (credit load, math/English)

- Student socio-demographics (10 predictors)
  - Age19Plus, Male, Hisp, Blk, OS, OSDisc, Non-Local, MotherEd, FatherEd, Pell
- Academic preparation (2 predictors)
  - *HSPrep (HS Core GPA/Test Score Index)*, AdvStanding
- Financial aid profile (8 predictors)
  - Unmet, Loans, Merit, Inc38827 Inc77464 Inc125776 Inc125776up; FAComplete
- Student motivation (2 predictors)
  - EdGoal, FirstChoice
- Student social integration (5 predictors)
  - LLC, CampWork, OnCampus, PlanWorkNo, PlanWorkFT
- Student academic experience (6 or 7 predictors)
  - Crs13to15, Crs16up, NoEngl, NoMath, DistEd, Undeclared, MidtermGPA (if available)

Data Management Tasks

- Exploratory data analysis
  - Variable selection (bivariate regression on outcome variable)
  - Variable coding (*continuous* vs. dummy/binary)
  - Missing data imputation
  - Derived variable(s)
    - HSPrep = (HSGPA*12.5)+(ACTM*.69)+(ACTE*.69)
- Logistic regression model
  - Preliminary model fit (-2LL test/score, pseudo R^2, HL sig.)
  - Check for outliers with diagnostic tools (Cook's, Std Residuals)
  - Check correct classification rate (CCR) for enrollees vs. non-enrollees (i.e. model sensitivity vs. specificity) using baseline probability and Receiver Operating Characteristics (ROC) curve

Data Management Tasks

- Imputation example: HS Preparation index score for cases with missing core GPA or test score
  - Regress core GPA and test score on each other
  - Use regression coefficients to estimate GPA/test score, respectively
  - Run HSPrep index equation for new cases

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.167	.027		79.054	.000
	ACT_COMP	.060	.001	.419	51.618	.000

a. Dependent Variable: HS_CORE_GPA

Data Management Tasks

- Determine persistence rate of your historical cohorts (fall 2011 through fall 2015): (Set TrainingSpring, TrainingFall = 1)
  - Fall-to-Spring

SprRetention

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	975	21.9	21.9	21.9
	1	3471	78.1	78.1	100.0
	Total	4446	100.0	100.0	

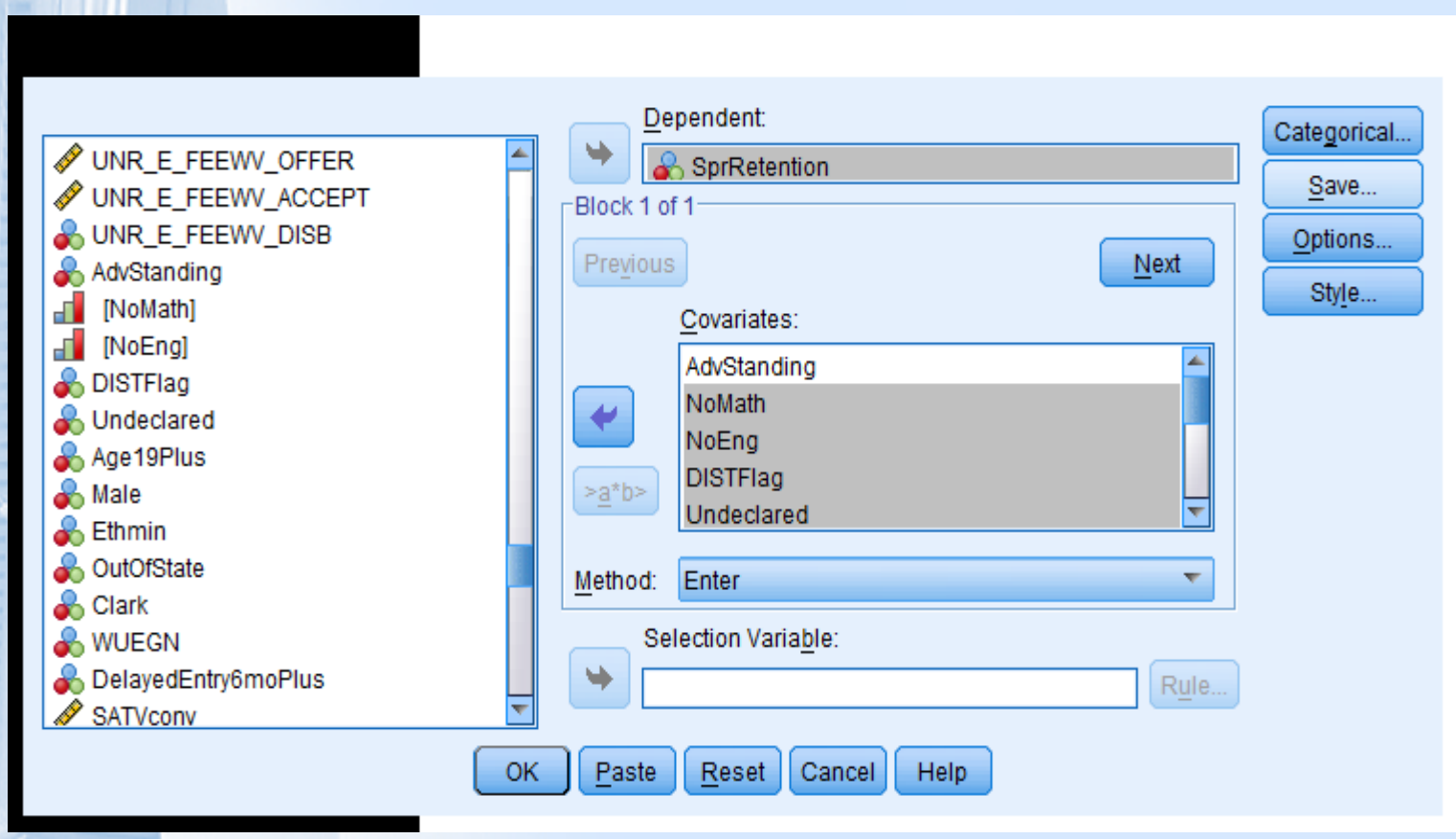
  - Fall-to-Fall

FallRetention

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	1294	29.1	29.1	29.1
	1	3152	70.9	70.9	100.0
	Total	4446	100.0	100.0	

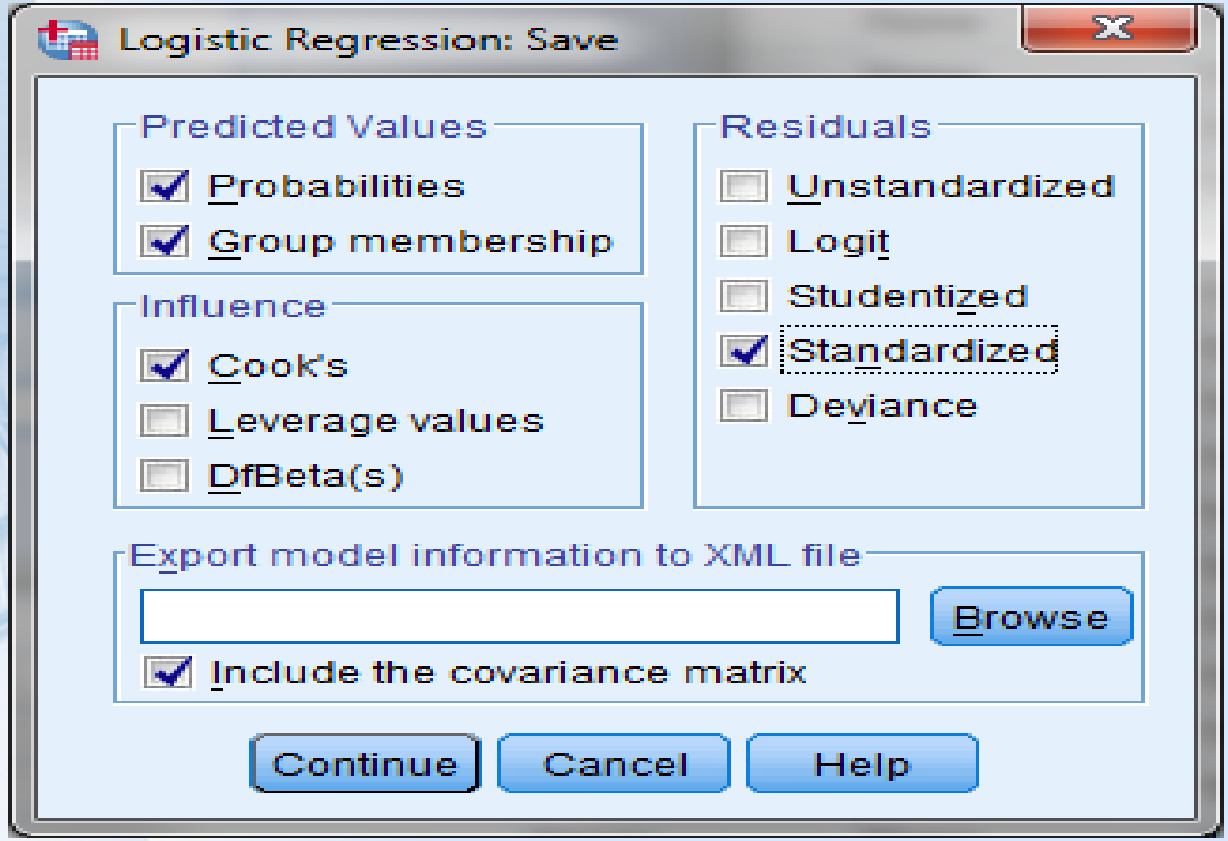
SPSS Menu Tasks

- Select *Analyze*, *Regression, Binary*



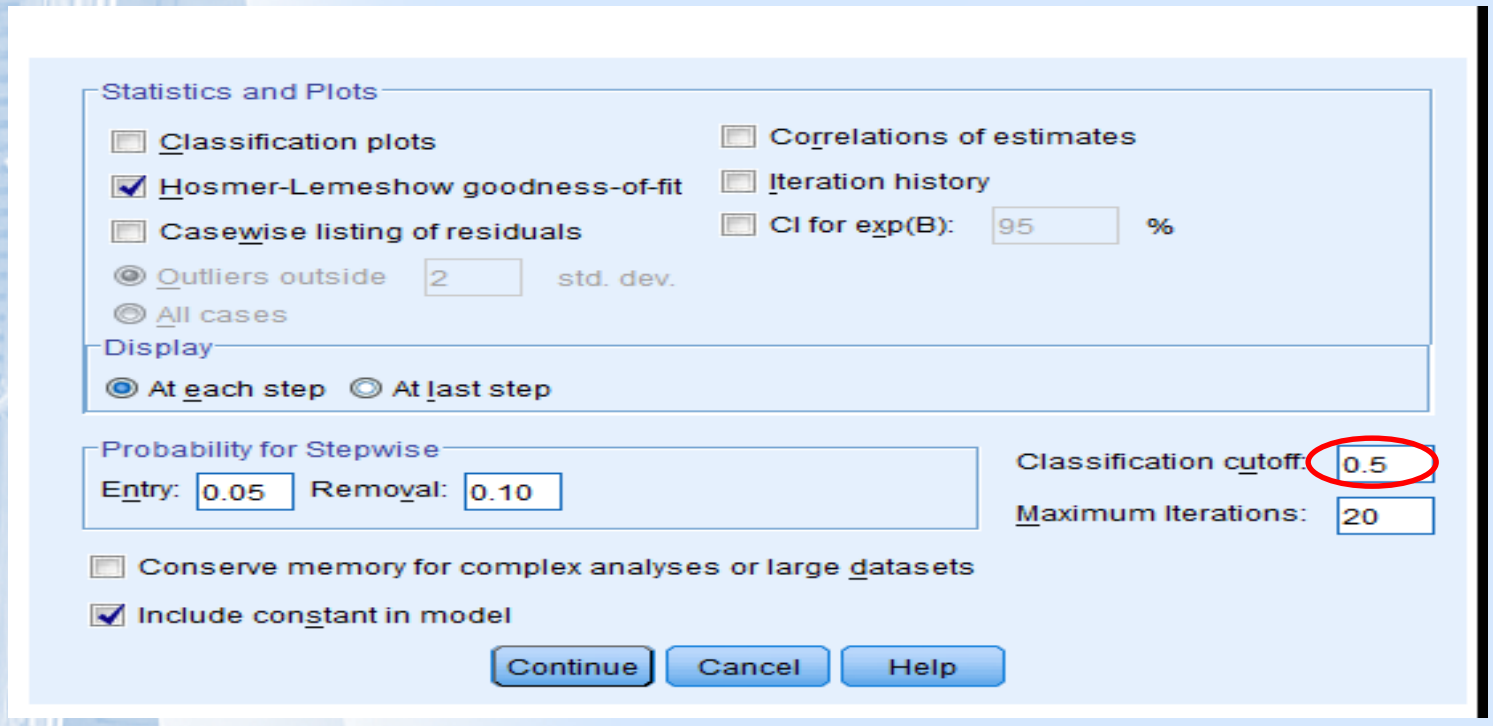
SPSS Menu Tasks

- Select *Analyze*, *Regression*, *Binary*, *Save*



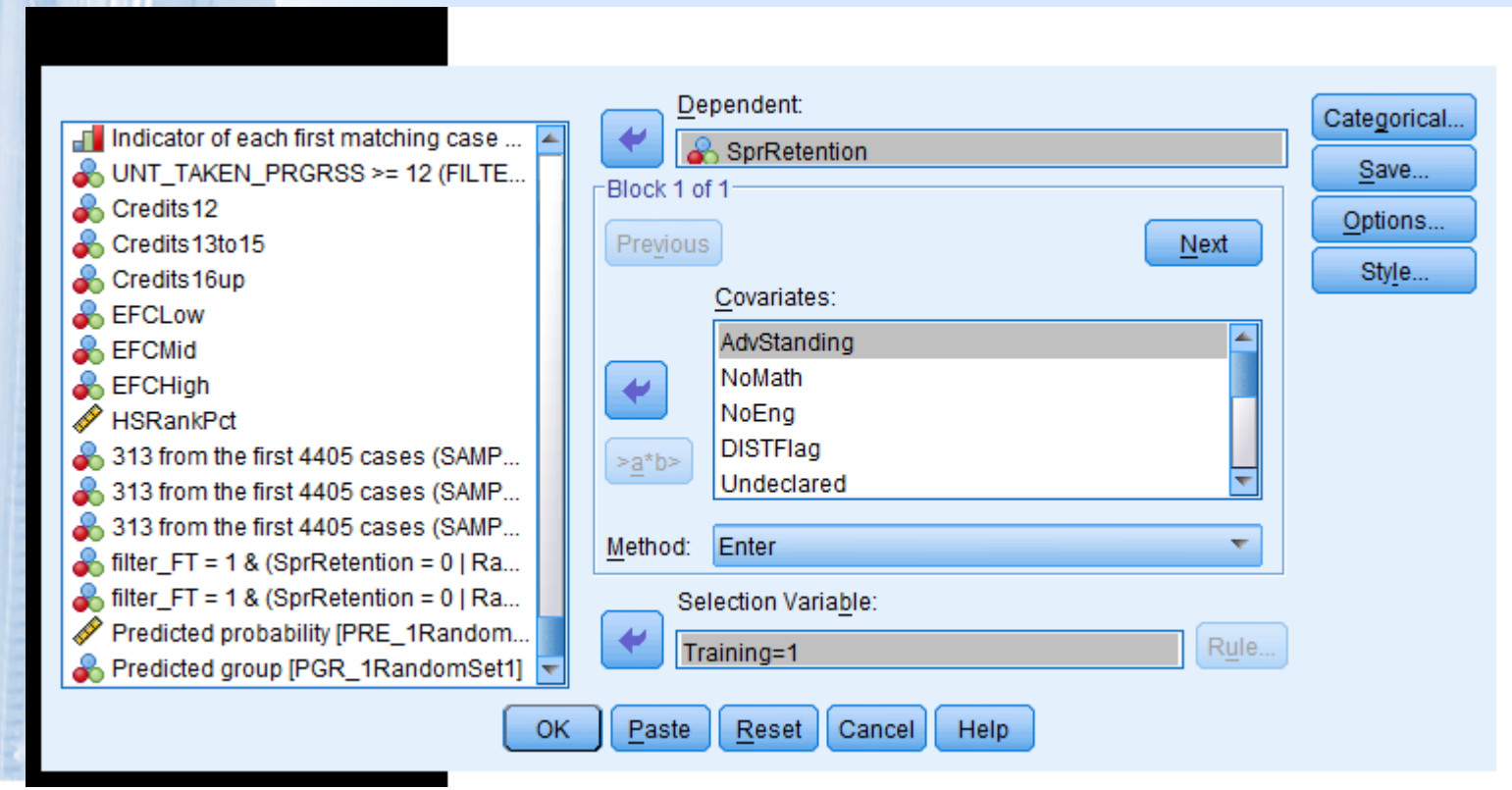
SPSS Menu Tasks

- Select *Analyze*, *Regression, Binary*
  - Under *Options,* select *HL goodness-of-fit*
  - Reset *classification cutoff* from 0.5 (default) to historical rate



SPSS Menu Tasks

- Select *Analyze*, *Regression, Binary*
  - Under Selection Variable, select *Training* variable, click Rule, insert *1*
  - Click Paste (inserts syntax in syntax window)



SPSS Menu Tasks

- Select *Analyze*, *Regression, Binary*
  - Click Paste (creates syntax in new window)
- Edit syntax as needed to re-specify parameters, re-estimate the dropout risk
- Or use syntax provided in SPSS file

```
DATASET ACTIVATE DataSet1.
LOGISTIC REGRESSION VARIABLES SprRetention
  /SELECT=TrainingSpring EQ 1
  /METHOD=ENTER AdvStanding NoMath NoEngl DistEd Undeclared Age19plus
Male Hisp Blk OS NonLocal WUE OnCampus CampWork Pell Unmet Loans Merit
FirstChoice EdGoalGrad MoEd4yrColl FathEd4yrColl PlanWorkFT PlanWorkNo LLC
Crs13to15 Crs16up HSPrep Inc38827 Inc77464 Inc125776 Inc125776up
FAComplete
  /SAVE=PRED PGROUP COOK ZRESID
  /PRINT=GOODFIT
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.781).
```

SPSS Output File

- Correct classification rate (CCR) for historical data is ~65%, for fall 2016 cohort it is ~66%.
- To improve CCR, check and exclude outlier cases

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			SprRetention		Percentage Correct	SprRetention		Percentage Correct
	Observed		0	1		0	1	
Step 1	SprRetention	0	637	338	65.3	102	97	51.3
		1	1235	2236	64.4	240	547	69.5
	Overall Percentage				64.6			65.8

a. The cut value is .781

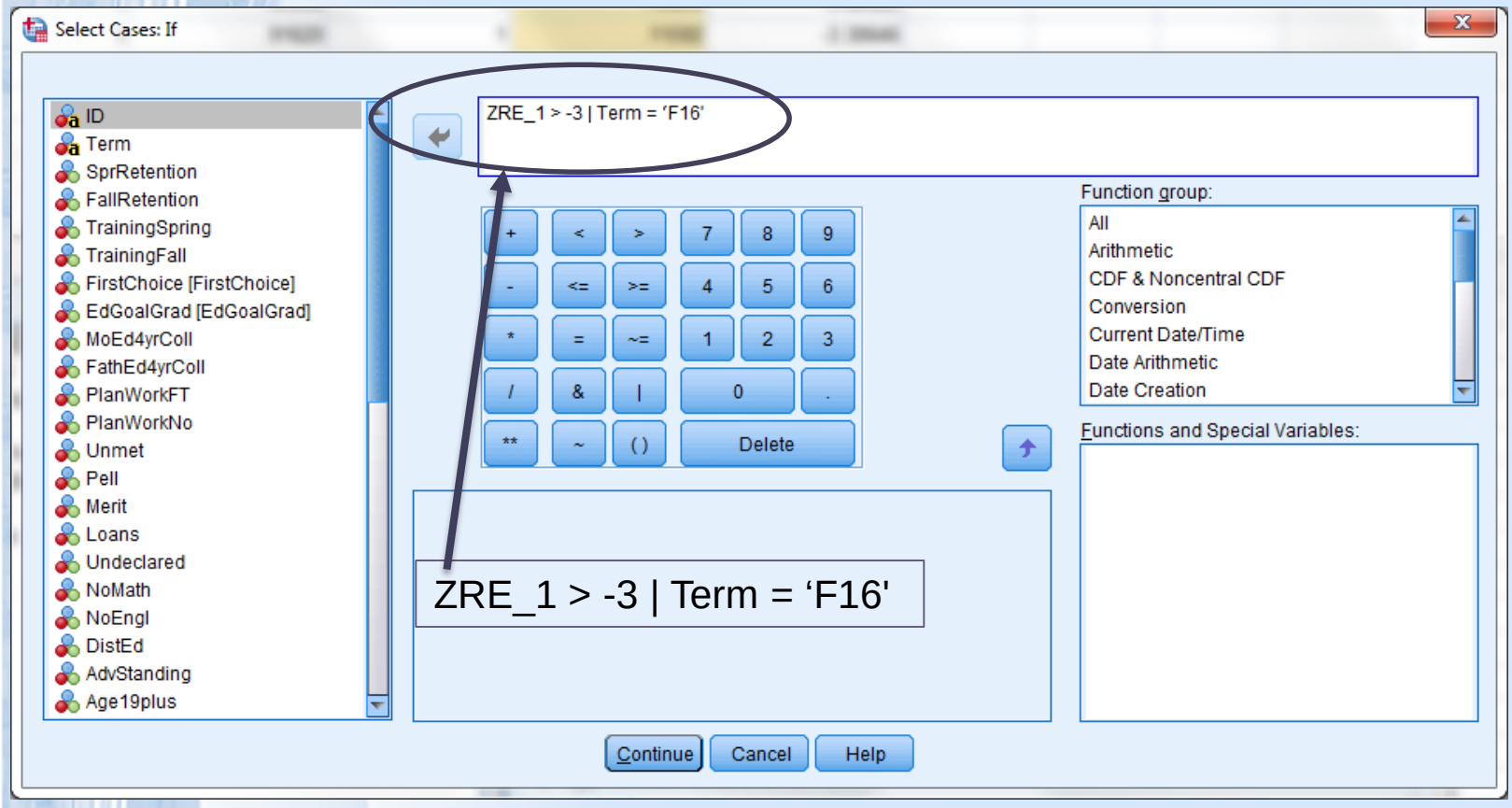
b. Selected cases TrainingSpring EQ 1

c. Unselected cases TrainingSpring NE 1

Identify and Exclude Outlier Cases

- Exclude Mahal(anobis Distance) [optional]
- Examine Cook's distance (COO_) and standardized residuals (ZRE_) for training data
- Exclude cases with
  - Cook's distance greater than 1, or visual separation
  - Standardized residuals greater |3|
- More stringent exclusion rules
  - Cook's distance greater than 4/n=number of cases
  - Standardized residuals greater |2|

Excluding Outlier Cases



Results from Trimmed Data

- Cut value adjusted to .792 to reflect trimmed training data
- Overall CCR at ~67% both historical and fall 2016 cohorts
- R-square = .21, but HL reached significance (<.05)
- Improve CCR by including Mid-Term Grades

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			SprRetention		Percentage Correct	SprRetention		Percentage Correct
	Observed		0	1		0	1	
Step 1	SprRetention	0	621	291	68.1	96	103	48.2
		1	1164	2307	66.5	224	563	71.5
	Overall Percentage				66.8			66.8

a. The cut value is .792

b. Selected cases TrainingSpring EQ 1

c. Unselected cases TrainingSpring NE 1

Results with Mid-Term Grades

- Include ‘mid term’ variable in syntax window
- Select all cases, no outlier exclusions: Cut value at 0.781
- Overall CCR at 82% for fall 2016 cohort
- R-square = .44, but HL reached significance (<.05)
- BUT, mediocre CCR for fall 2016 dropout students (58.6%)

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c,d]		
			SprRetention		Percentage Correct	SprRetention		Percentage Correct
	Observed		0	1		0	1	
Step 1	SprRetention	0	706	269	72.4	109	77	58.6
		1	618	2853	82.2	99	687	87.4
	Overall Percentage				80.0			81.9

a. The cut value is .781

b. Selected cases TrainingSpring EQ 1

c. Unselected cases TrainingSpring NE 1

d. Some of the unselected cases are not classified due to either missing values in the independent variables or categorical variables with values out of the range of the selected cases.

Results with Mid-Term Grades

- Select all cases, no outlier exclusions: Cut value at 0.781
- **Change classification cutoff value to 0.87**
- Overall CCR down (72.4%), but more balanced CCR
- Nearly 70% CCR for dropout cases in predicted (fall '16) cohort)

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c,d]		
			SprRetention		Percentage Correct	SprRetention		Percentage Correct
	Observed		0	1		0	1	
Step 1	SprRetention	0	795	180	81.5	128	58	68.8
		1	1116	2355	67.8	210	576	73.3
	Overall Percentage				70.9			72.4

a. The cut value is .870

b. Selected cases TrainingSpring EQ 1

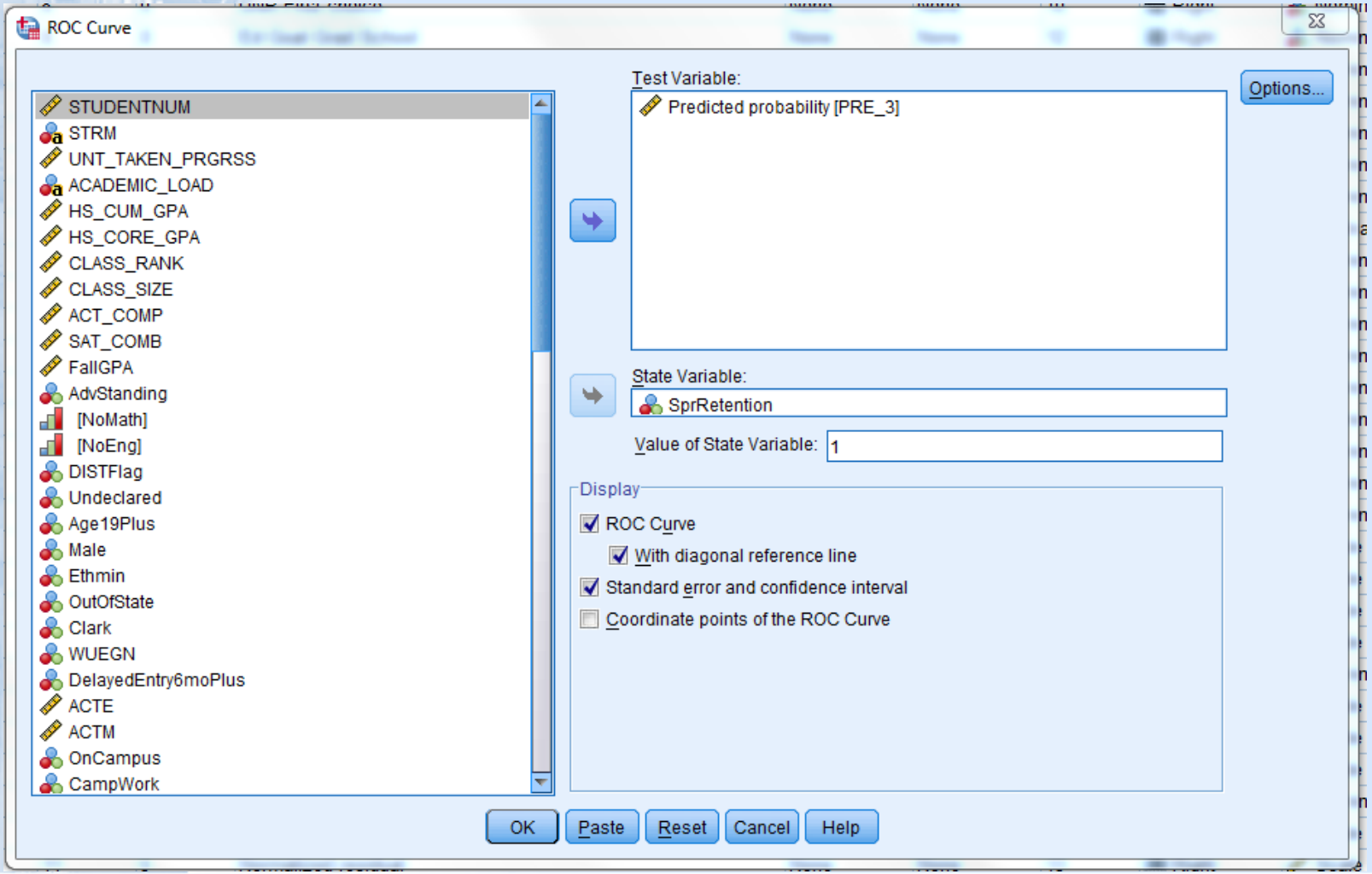
c. Unselected cases TrainingSpring NE 1

d. Some of the unselected cases are not classified due to either missing values in the independent variables or categorical variables with values out of the range of the selected cases.

Pondering Results

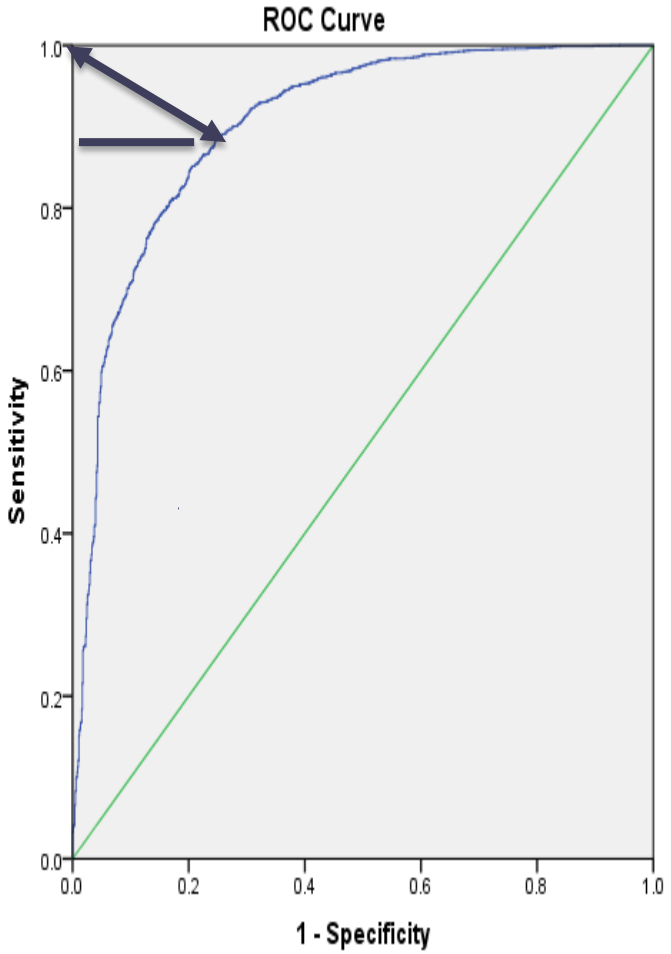
- Outlier removal improves prediction accuracy, but exclusion of too many cases may bias results
- Midterm prediction, including midterm grades, yields higher prediction accuracy without exclusion of outlier cases
- Thus, prediction accuracy is a balancing act between waiting for more pertinent data (e.g. midterm grades) and excluding outlier cases for better model fit but possibility of biasing results
- When excluding outlier cases, examine how many are removed (keep number of excluded outliers below 5% of total cases; check coefficient of determination R-square, Hosmer-Lemeshow alpha level preferably > 0.05)

Determine Balanced CCR: ROC Charts



Determine Balanced CCR: ROC Charts

- Simultaneous measure of sensitivity (true positive) and specificity (true negative) for all possible cutoff values
- Calculate area under the ROC curve (exercise)
- Area under the ROC: .901 (all case data)
- Suggested cutoff point to maximize overall CCR is around 0.901. (associated CCR for dropout = 73.1%)



Assess Prediction Accuracy

- Compare results from full-data model with results from trimmed-data model
- Determine the best cut value (classification) based on re-adjusted baseline probability versus ROC-curve derived probability level
- Evaluate relative cost of (in-)accurate prediction of retained students (sensitivity) versus dropout students (specificity)
- Usually, err in favor of accurate identification of students at risk of dropping out, without sacrificing too much accuracy for retained

Translate Dropout Risk

- Copy retention probability for fall 2016 cohort to new file (to eliminate all other cases)
- Group into deciles using binning function:
  - Transform, Visual Binning, Make 9 cutpoints, Label 'Deciles', check 'reverse scale'
- Note bottom high-risk deciles with far lower retention probability (run decile avage)
- Identify cusp of probability border between predicted dropouts and persisters and corresponding decile groups
- Identify priority decile groups near the cusp for student assistance

Sample Data for Advisors

R Number	Last Name	First Name	Email Addr	Age	College	Dept	Major	Dropout Risk Decile (10=highest; 1=lowest)	Relative Spring Retention %tile
				18	LBA	ART	BA-AHI	9	14.92
				18	LBA	ANTH	BA-AN	8	28.52
				18	LBA	ANTH	BA-AN	7	36.80
				18	LBA	ANTH	BA-AN	7	39.18
				18	LBA	ANTH	BA-AN	6	46.87
				18	LBA	ANTH	BA-AN	4	66.48
				19	LBA	ANTH	BA-AN	1	92.42
				18	LBA	ANTH	BA-AN	1	95.57

Sample Data for Advisors

Gender	Ethnicity	Credits	Resident	State/Cnty	HS GPA	ACTE	ACTM	Has Pell$ (1=yes)	Has Loan$ (1=yes)	Clark Cnty Resi (1=yes)
F	AS	12	NV	NWA	3.10	24	18	1	0	0
F	WH	15	NV	NCL	3.23	21	18	0	1	1
M	WH	16	WU	CA	3.19	23	20	0	0	0
M	WH	17	WU	OR	3.23	24	17	0	0	0
F	WH	16	NV	NWA	3.18	17	17	1	0	0
F	WH	15	NV	NDO	3.47	30	21	0	0	0
M	WH	15	NV	NWA	3.65	26	25	1	0	0
F	AS	16	NV	NCL	3.90	30	28	0	0	1

Unbalanced Data

- Proportion of dropouts is usually much smaller than proportion of retained students
- Number of cases in rare event (dropout) should be sufficient to yield *minimum* 10:1 ratio with number of predictors (preferably 30:1 ratio)
- Check standard errors in coefficient results table ("Variables in the Equation) for inflated values
- Check variance inflation factor (VIF) in collinearity diagnostics (must run linear regression) to determine which predictor(s) to remove if ratio well below 10:1 or run *Exact Logistic Regression* (see example at http://www.ats.ucla.edu/stat/stata/dae/exlogit.htm)
- Suggested VIF threshold: 2.5 (R-sq = .60) (see Paul Allison, *Statistical Horizons*, Sept. 10, 2012)

Exercise

- Estimate fall-to-fall dropout risk for 2016 cohort, using 2011 through 2015 cohorts

Case Summaries

Term		FallRetention	TrainingFall
F11	N	734	734
	Mean	.66	1.0000
F12	N	749	749
	Mean	.72	1.0000
F13	N	834	834
	Mean	.74	1.0000
F14	N	1021	1021
	Mean	.70	1.0000
F15	N	1108	1108
	Mean	.71	1.0000
F16	N		986
	Mean		.0000
Total	N	4446	5432
	Mean	.71	.8185

Exercise

- Estimate fall-to-fall dropout risk for 2016 cohort using 2011 through 2015 cohorts
- Set cutoff value = 0.709. All cases included.
- Check/exclude outliers, re-run model

Classification Table[a]

			Predicted[c]					
			Selected Cases[b]			Unselected Cases[d,e]		
			FallRetention		Percentage Correct	FallRetention		Percentage Correct
	Observed		0	1		0	1	
Step 1	FallRetention	0	882	412	68.2	0	0	.
		1	643	2509	79.6	0	0	.
	Overall Percentage				76.3			.

a. The cut value is .709

b. Selected cases TrainingFall EQ 1

c. There are no unselected cases. Therefore, no unselected cases are classified.

d. Unselected cases TrainingFall NE 1

e. Some of the unselected cases are not classified due to either missing values in the independent variables or categorical variables with values out of the range of the selected cases.

Exercise

- Excluding Z-residuals (> +/-3), 95 cases (2.1%)
- CCR improved to 78.6% from 76.3%
- R-square 0.47 (cut value adjusted to .723)

Classification Table[a]

Observed			Predicted[c]					
			Selected Cases[b]			Unselected Cases[d,e]		
			FallRetention		Percentage Correct	FallRetention		Percentage Correct
			0	1		0	1	
Step 1	FallRetention	0	876	330	72.6	0	0	.
		1	599	2546	81.0	0	0	.
	Overall Percentage				78.6			.

a. The cut value is .723

b. Selected cases TrainingFall EQ 1

c. There are no unselected cases. Therefore, no unselected cases are classified.

d. Unselected cases TrainingFall NE 1

e. Some of the unselected cases are not classified due to either missing values in the independent variables or categorical variables with values out of the range of the selected cases.

Translate Dropout Risk

- Copy retention probability for fall 2016 cohort to new file (to eliminate all other cases)
- Group into deciles using binning function:
  - Transform, Visual Binning, Make 9 cutpoints, Label ‘Deciles’, check ‘reverse scale’
- Note bottom high-risk deciles with far lower retention probability (run decile avage)
- Identify cusp of probability border between predicted dropouts and persisters and corresponding decile groups
- Identify priority decile groups near the cusp for student assistance
- Send student record file with predicted probability, predicted outcome, decile group to student assistance/advising personnel

Community College Data Set Details

Mimic* dataset based on data from:

- ~ 4,300 student enrollment
- Open access
- Large % of under-represented, low income, and first generation students
- 60% male
- Average age is 26 years old
- 66% part-time enrollment
- Over half of academic programs are vocational/ career technical
- 18% grad rate (150%)
- 72% fall-to-spring retention first-time freshmen; 50% fall-to-fall retention

*The CC Dataset used in this class has been de-identified, randomized, and altered for instructional and sharing purposes. These "mimic" data do not match actual institutional data, trends, or outcomes.

Community College Data Set Details

- Data Sources
  - Matriculation system (Banner, data warehouse)
- Student cohorts
  - New first-year students (part-time and full-time)
  - Historical cohorts: fall 2013-15 (training set, N=2,243)
  - Predicted cohort: fall 2016 (holdout set, N=626)
  - Newest cohort: fall 2017 (holdout set #2, N=702)
- Data elements (predictors) at start of first semester
  - Student socio-geo-demographics (age, gender, ethnicity, geographic proximity to campus, residency, military)
  - Academic preparation (Compass test scores, high school attended, remediation/ developmental courses needed)
  - Financial aid profile (unmet need, pell)
  - Student motivation proxies (degree audit logins, educational goals survey responses)
  - Student academic experience (credit load, math/English enrollment, major type)

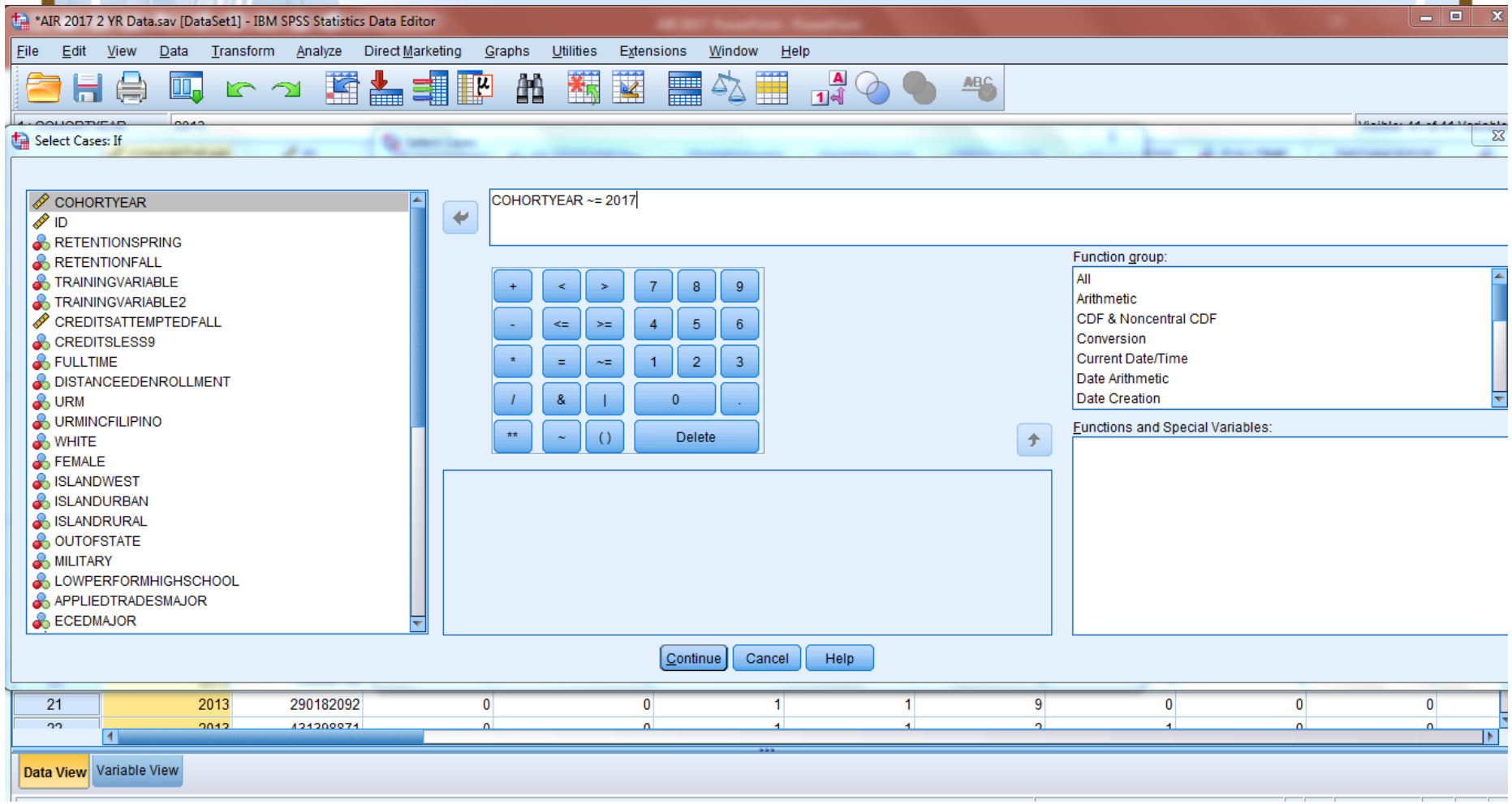
35 predictor variables in the data set

- Student socio-demographics (12 predictors)
  - *AGE, AGE19PLUS, FEMALE, URM, URMINCFILIPINO, WHITE, ISLANDWEST, ISLANDURBAN, ISLANDRURAL, OUTOFSTATE,MILITARY, LOWPERFORMHIGHSCHOOL*
- Academic preparation (9 predictors)
  - *COMPASS READING, COMPASS WRITING, COMPASSANYMATHHIGHEST, REMEDIAL/ DEVELOPMENTAL/ COLLEGELEVEL (Math/English) FLAGS,*
- Financial aid profile (2 predictors)
  - PERCENTUNMETNEED, PELL
- Student motivation (4 predictors)
  - EDGOAL1, EDGOAL2, STARUSAGE, STARUSAGEAVERAGEFLAG,
- Student academic experience (8 predictors)
  - CREDITSATTEMPTED, CREDITSLESS9, FULLTIME, DISTANCEEDENROLL, ECED MAJOR, APPLIEDTRADESMAJOR, *ANYMATHENROLL, ANYENGLISHENROLL*

Step 1: Filter out the 2015 data

Select *Data*, Select Cases, *If condition…*

COHORTYEAR ~= 2017



CC Data: SPSS Menu Tasks

- Select *Analyze*, *Regression, Binary*
  - *Use same menu options learned in the UNR example.*
  - Click Paste (creates syntax in new window).
- From here on, we will edit syntax as needed to re-specify parameters, re-estimate the dropout risk

```
DATASET ACTIVATE DataSet1.
LOGISTIC REGRESSION VARIABLES RETENTIONSPRING
  /SELECT=TRAININGVARIABLE EQ 1
  /METHOD=ENTER CREDITSATTEMPTEDFALL DISTANCEEDENROLLMENT URM FEMALE
ISLANDRURAL OUTOFSTATE
    LOWPERFORMHIGHSCHOOL ECEDMAJOR AGE19PLUS EDGOAL1 PELL
PERCENTUNMETNEED STARUSAGE COMPASSREADING
    COMPASSANYMATHHIGHEST REMEDIALMATH REMEDIALENG
  /SAVE=PRED PGROUP COOK ZRESID
  /PRINT=GOODFIT
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

SPSS Output File

- R-square = .255 ; HL test sig. = .103
- Null model correct classification rate (CCR) for spring dropout is nil in both training and holdout data (0.0%)

Classification Table[a,b]

			Predicted					
			Selected Cases[c]			Unselected Cases[d]		
			RETENTIONSPRING		Percentage Correct	RETENTIONSPRING		Percentage Correct
	Observed		0	1		0	1	
Step 0	RETENTIONSPRING	0	0	626	.0	0	177	.0
		1	0	1617	100.0	0	449	100.0
	Overall Percentage				72.1			71.7

a. Constant is included in the model.

b. The cut value is .500

c. Selected cases TRAININGVARIABLE EQ 1

d. Unselected cases TRAININGVARIABLE NE 1

Here, we calculated the baseline fall-to-spring retention rate

SPSS Menu Tasks

- Select *Analyze*, *Regression, Binary*
  - Click Paste (creates syntax in new window)
- Edit cut value in syntax to reflect baseline probability of spring retention (i.e. 72.1%)

```
DATASET ACTIVATE DataSet2.
LOGISTIC REGRESSION VARIABLES RETENTIONSPRING
  /SELECT=TRAININGVARIABLE EQ 1
  /METHOD=ENTER CREDITSATTEMPTED DISTANCEEDENROLLMENT URM FEMALE
ISLANDRURAL OUTOFSTATE
    LOWPERFORMHIGHSCHOOL ECEDMAJOR AGE19PLUS EDGOAL1 PELL
PERCENTUNMETNEED STARUSAGE COMPASSREADING
    COMPASSANYMATHHIGHEST REMEDIALMATH REMEDIALENG
  /SAVE=PRED PGROUP COOK  ZRESID
  /PRINT=GOODFIT
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.721).
```

SPSS Output File

- R-square = .255 ; HL test sig. = .103
- CCR for spring dropout at 70% for training and 80% for holdout cohorts
- Good correct classification rate of dropout students
  - Check for outliers to seek further improvement

Classification Table[a]

Observed			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
			RETENTIONSPRING		Percentage Correct	RETENTIONSPRING		Percentage Correct
			0	1		0	1	
Step 1	RETENTIONSPRING	0	440	186	70.3	142	35	80.2
		1	490	1127	69.7	164	285	63.5
	Overall Percentage				69.9			68.2

a. The cut value is .721

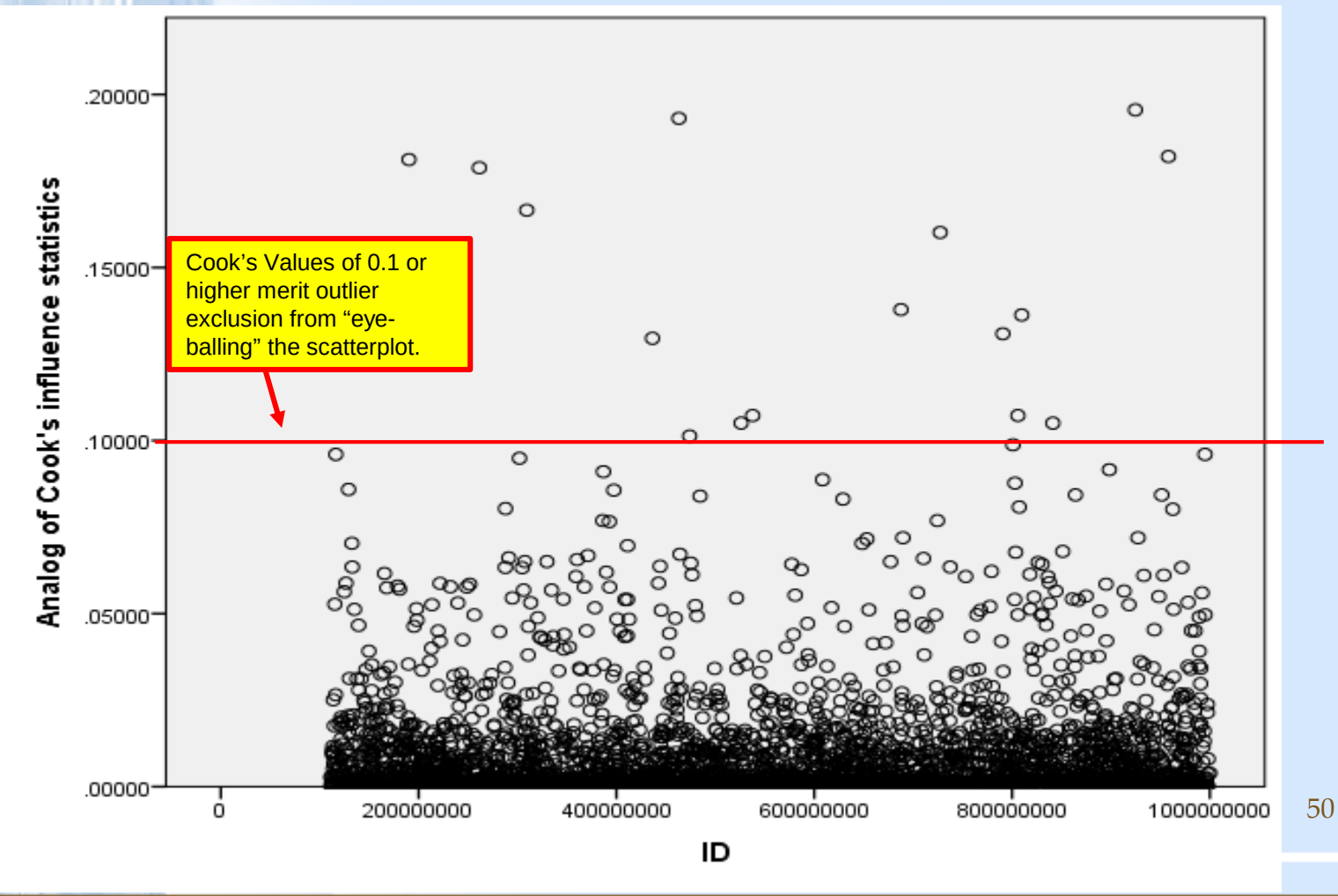
b. Selected cases TRAININGVARIABLE EQ 1

c. Unselected cases TRAININGVARIABLE NE 1

Identify Outlier Cases

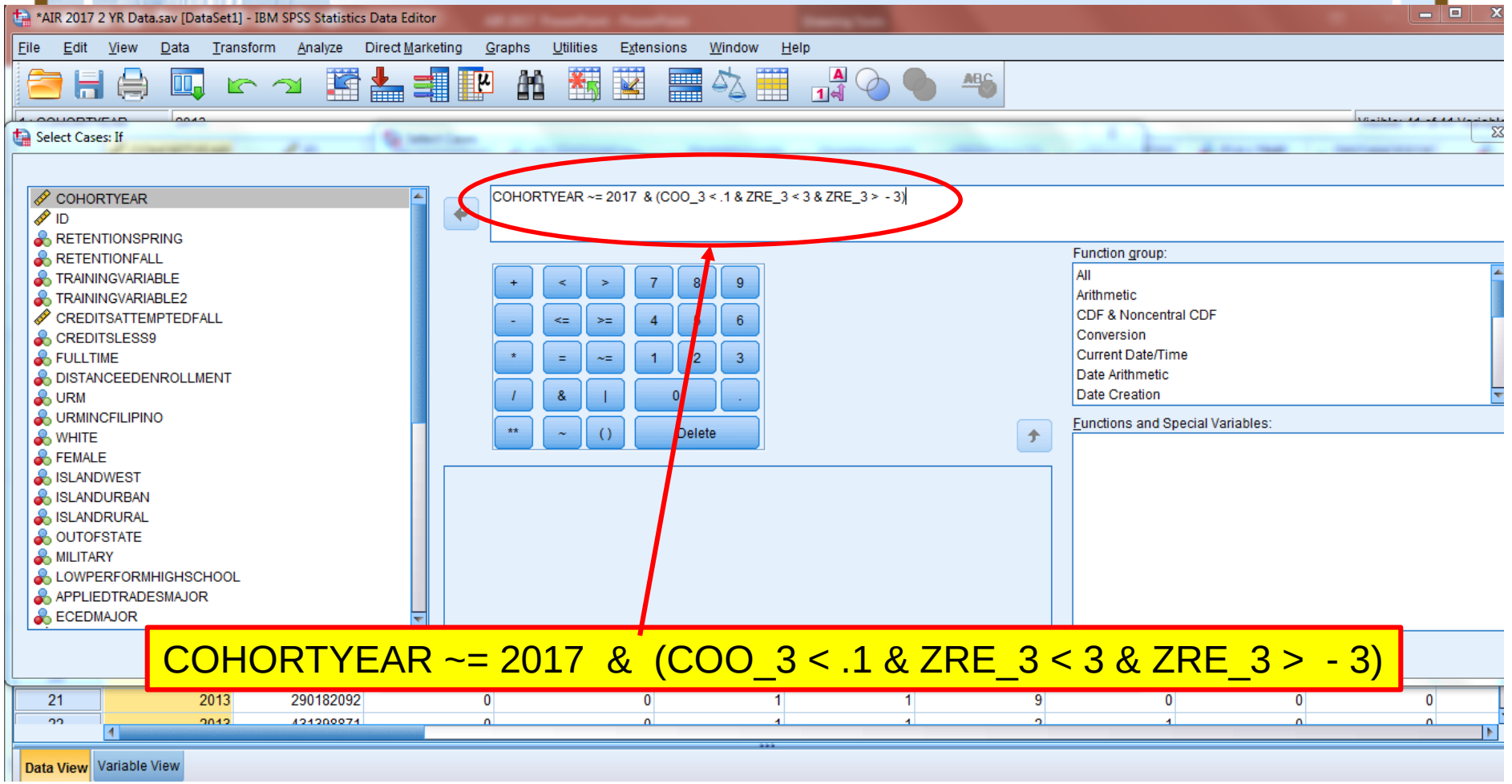
- Examine Cook's distance (COO_) and standardized residuals (ZRE_)
- Exclude cases with
  - Cook's distance greater than 1, or visual separation
  - Standardized residuals greater |3|
- More stringent exclusion rules
  - Cook's distance greater than 4/n=number of cases
  - Standardized residuals greater |2|

Identify Outlier Cases



SPSS Menu Tasks

- Exclude outliers via ‘select cases if’ function
- Use ‘filter_Trim (already included)



COHORTYEAR ~= 2017 & (COO_3 < .1 & ZRE_3 < 3 & ZRE_3 > - 3)

SPSS Syntax Version of Filter Tasks (fyi)

```
DATASET ACTIVATE DataSet1.
USE ALL.
COMPUTE filter_$=(COHORTYEAR ~= 2017  &  COO_3 < .1 & ZRE_3 < 3 &
ZRE_3 >  - 3).
VARIABLE LABELS filter_$ 'COHORTYEAR ~= 2017  &  (COO_3 < .1 & ZRE_3 <
3 & ZRE_3 >  - 3) (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
```

CC Data: SPSS Menu Tasks

- Run regression syntax again with the 0.721 baseline retention rate

```
DATASET ACTIVATE DataSet1.
LOGISTIC REGRESSION VARIABLES RETENTIONSPRING
  /SELECT=TRAININGVARIABLE EQ 1
  /METHOD=ENTER CREDITSATTEMPTEDFALL DISTANCEEDENROLLMENT URM FEMALE
ISLANDRURAL OUTOFSTATE
    LOWPERFORMHIGHSCHOOL ECEDMAJOR AGE19PLUS EDGOAL1 PELL
PERCENTUNMETNEED STARUSAGE COMPASSREADING
    COMPASSANYMATHHIGHEST REMEDIALMATH REMEDIALENG
  /SAVE=PRED PGROUP COOK ZRESID
  /PRINT=GOODFIT
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.721).
```

Calculate new baseline from trimmed data

- New baseline retention rate is .724 based on trimmed training data

Classification Table[a,b]

			Predicted					
			Selected Cases[c]			Unselected Cases[d]		
	Observed		RETENTIONSPRING		Percentage Correct	RETENTIONSPRING		Percentage Correct
			0	1		0	1	
Step 0	RETENTIONSPRING	0	0	614	.0	0	173	.0
		1	0	1611	100.0	0	441	100.0
	Overall Percentage				72.4			71.8

a. Constant is included in the model.

b. The cut value is .721

c. Selected cases TRAININGVARIABLE EQ 1

d. Unselected cases TRAININGVARIABLE NE 1

CC Data: SPSS Menu Tasks

- Re-run regression syntax AGAIN with the new baseline retention rate = 0.724

```
DATASET ACTIVATE DataSet1.
LOGISTIC REGRESSION VARIABLES RETENTIONSPRING
  /SELECT=TRAININGVARIABLE EQ 1
  /METHOD=ENTER CREDITSATTEMPTEDFALL DISTANCEEDENROLLMENT URM FEMALE
ISLANDRURAL OUTOFSTATE
    LOWPERFORMHIGHSCHOOL ECEDMAJOR AGE19PLUS EDGOAL1 PELL
PERCENTUNMETNEED STARUSAGE COMPASSREADING
    COMPASSANYMATHHIGHEST REMEDIALMATH REMEDIALENG
  /SAVE=PRED PGROUP COOK ZRESID
  /PRINT=GOODFIT
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.724).
```

Updated Results from Trimmed Data

- Cut value adjusted to .724 to reflect trimmed training data
- Dropout CCR at 72% for training, 82% for holdout data
- Overall CCR at ~70% for both training and holdout data
- R-square = .295, but HL reached significance (<.05)

Classification Table[a]

			Predicted					
			Selected Cases[b]			Unselected Cases[c]		
	Observed		RETENTIONSPRING		Percentage Correct	RETENTIONSPRING		Percentage Correct
			0	1		0	1	
Step 1	RETENTIONSPRING	0	437	177	71.2	141	32	81.5
		1	470	1141	70.8	154	287	65.1
	Overall Percentage				70.9			69.7

a. The cut value is .724

b. Selected cases TRAININGVARIABLE EQ 1

c. Unselected cases TRAININGVARIABLE NE 1

81.5% accuracy in identifying dropped students

Results from Trimmed Data

- Some false positives in Decile 1 for predicting retainers, but overall results suggest stability.

Contingency Table for Hosmer and Lemeshow Test

		RETENTIONSPRING = 0		RETENTIONSPRING = 1		
		Observed	Expected	Observed	Expected	Total
Step 1	1	144	159.891	79	63.109	223
	2	123	112.615	100	110.385	223
	3	90	87.850	133	135.150	223
	4	77	70.315	146	152.685	223
	5	53	56.206	170	166.794	223
	6	52	45.928	171	177.072	223
	7	38	36.500	185	186.500	223
	8	34	27.224	189	195.776	223
	9	3	13.452	220	209.548	223
	10	0	4.021	218	213.979	218

Results from Trimmed Data

- Parameter estimates results. 9 variables significant at .05 level.

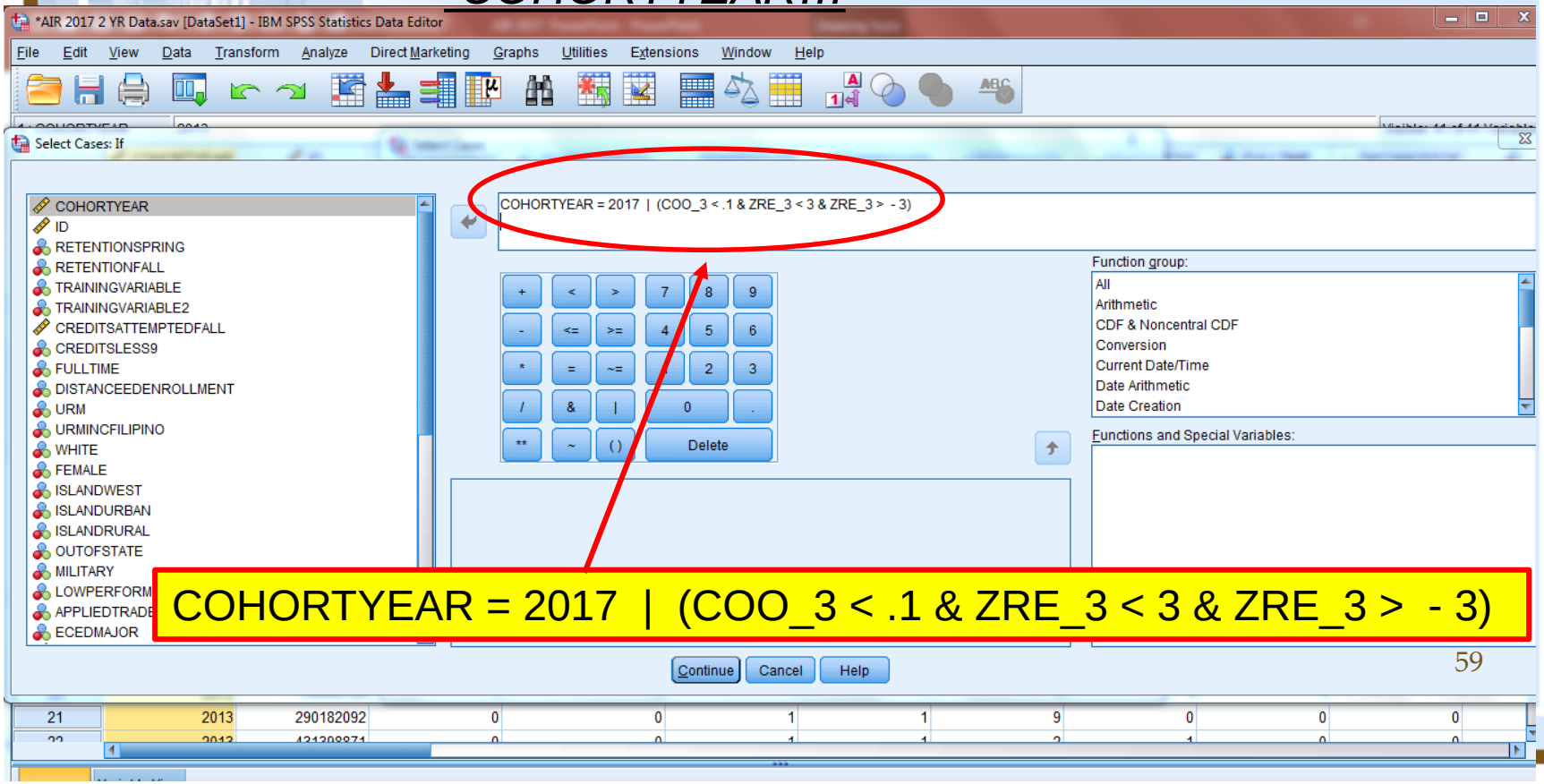
Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	CREDITSATTEMPTEDFALL	.152	.016	87.617	1	.000	1.164
	DISTANCEEDENROLLMENT	-.349	.189	3.402	1	.065	.706
	URM	-.643	.113	32.390	1	.000	.526
	FEMALE	-.062	.119	.269	1	.604	.940
	ISLANDRURAL	-.590	.201	8.629	1	.003	.554
	OUTOFSTATE	-.598	.268	4.998	1	.025	.550
	LOWPERFORMHIGHSCHOOL	-.472	.143	10.879	1	.001	.624
	ECEDMAJOR	.022	.272	.007	1	.935	1.023
	AGE19PLUS	-.287	.118	5.912	1	.015	.751
	EDGOAL1	1.655	.265	39.002	1	.000	5.235
	PELL	2.978	.271	120.323	1	.000	19.639
	PERCENTUNMETNEED	-4.428	.428	106.951	1	.000	.012
	STARUSAGE	.018	.022	.702	1	.402	1.019
	COMPASSREADING	-.002	.002	.420	1	.517	.998
	COMPASSANYMATHHIGHEST	.005	.003	1.818	1	.178	1.005
	REMEDIALMATH	-.164	.128	1.656	1	.198	.848
	REMEDIALENG	-.216	.136	2.507	1	.113	.806
	Constant	-.276	.284	.945	1	.331	.759

a. Variable(s) entered on step 1: CREDITSATTEMPTEDFALL, DISTANCEEDENROLLMENT, URM, FEMALE, ISLANDRURAL, OUTOFSTATE, LOWPERFORMHIGHSCHOOL, ECEDMAJOR, AGE19PLUS, EDGOAL1, PELL, PERCENTUNMETNEED, STARUSAGE, COMPASSREADING, COMPASSANYMATHHIGHEST, REMEDIALMATH, REMEDIALENG.

Final Step: "Go Live" and score the incoming cohort

- Update filter in menu: Select *Data*, Select Cases, *If condition…In the syntax, change "~=" to "=" for "COHORTYEAR…"*

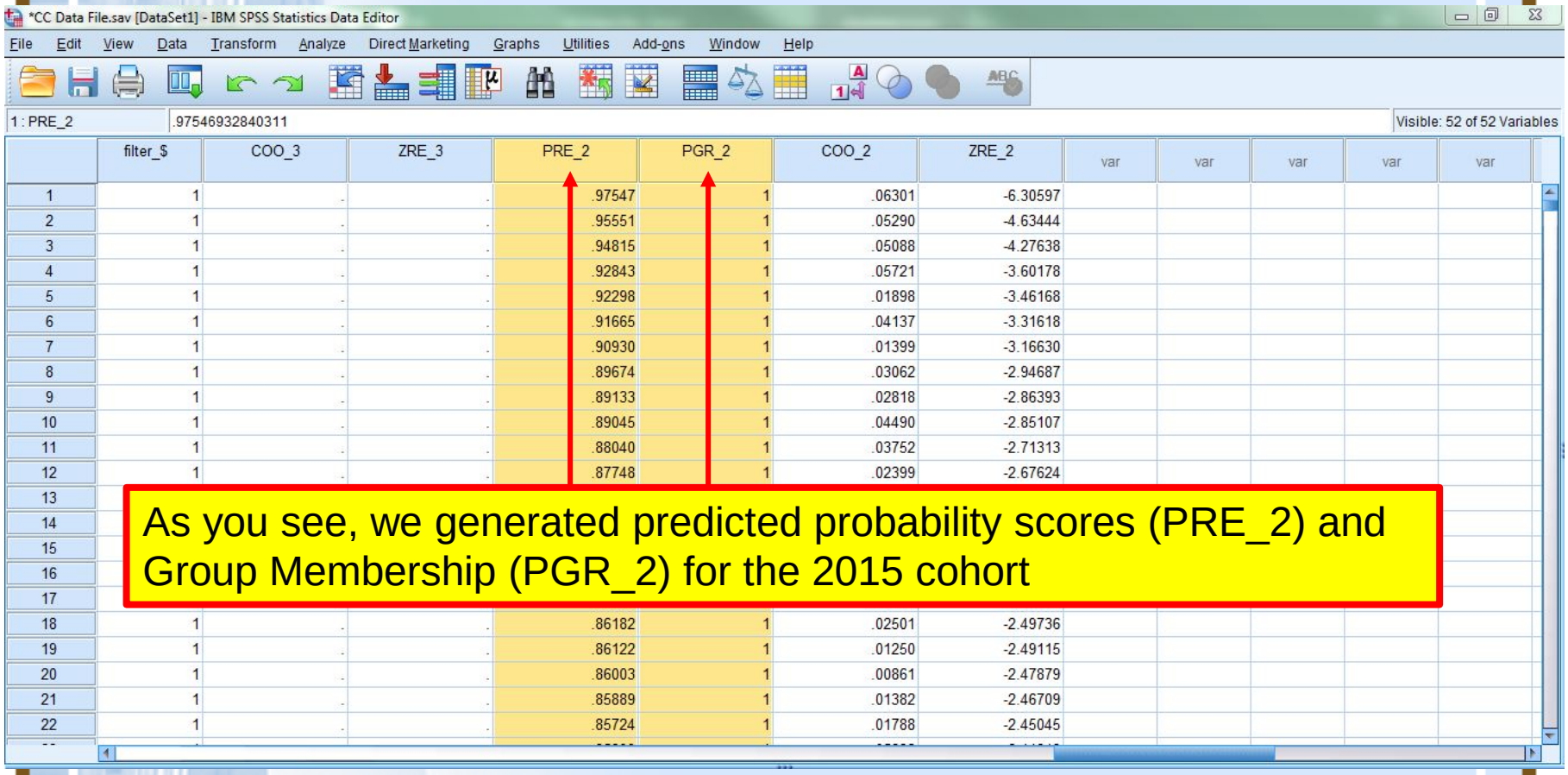


CC Data: SPSS Menu Tasks

- Re-run regression syntax AGAIN with the last baseline retention rate = 0.724
- Change "TRAININGVARIABLE2 EQ 1" to score the 2017 cohort.

```
DATASET ACTIVATE DataSet1.
LOGISTIC REGRESSION VARIABLES RETENTIONSPRING
  /SELECT=TRAININGVARIABLE2 EQ 1
  /METHOD=ENTER CREDITSATTEMPTEDFALL DISTANCEEDENROLLMENT URM FEMALE
ISLANDRURAL OUTOFSTATE
    LOWPERFORMHIGHSCHOOL ECEDMAJOR AGE19PLUS EDGOAL1 PELL
PERCENTUNMETNEED STARUSAGE COMPASSREADING
    COMPASSANYMATHHIGHEST REMEDIALMATH REMEDIALENG
  /SAVE=PRED PGROUP COOK ZRESID
  /PRINT=GOODFIT
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.724).
```

CC Data: SPSS Ouput



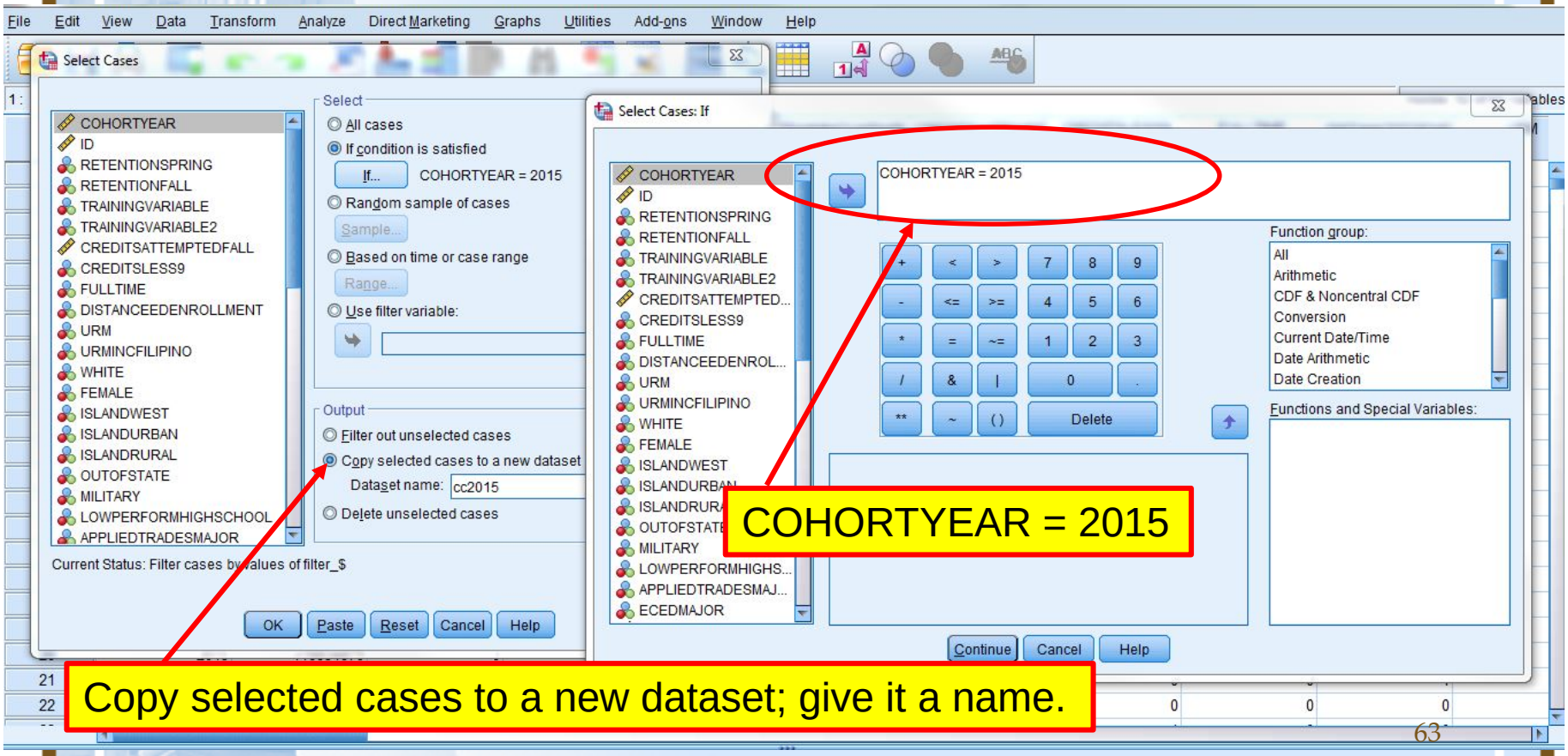
	filter_$	COO_3	ZRE_3	PRE_2	PGR_2	COO_2	ZRE_2	var	var	var	var	var
1	1	.	.	.97547	1	.06301	-6.30597					
2	1	.	.	.95551	1	.05290	-4.63444					
3	1	.	.	.94815	1	.05088	-4.27638					
4	1	.	.	.92843	1	.05721	-3.60178					
5	1	.	.	.92298	1	.01898	-3.46168					
6	1	.	.	.91665	1	.04137	-3.31618					
7	1	.	.	.90930	1	.01399	-3.16630					
8	1	.	.	.89674	1	.03062	-2.94687					
9	1	.	.	.89133	1	.02818	-2.86393					
10	1	.	.	.89045	1	.04490	-2.85107					
11	1	.	.	.88040	1	.03752	-2.71313					
12	1	.	.	.87748	1	.02399	-2.67624					
13												
14												
15												
16												
17												
18	1	.	.	.86182	1	.02501	-2.49736					
19	1	.	.	.86122	1	.01250	-2.49115					
20	1	.	.	.86003	1	.00861	-2.47879					
21	1	.	.	.85889	1	.01382	-2.46709					
22	1	.	.	.85724	1	.01788	-2.45045					

As you see, we generated predicted probability scores (PRE_2) and Group Membership (PGR_2) for the 2015 cohort

Last Step: Translate Dropout Risk in to deciles for easier interpretation by academic support office

- Convert retention probability to dropout risk deciles (1 = highest, 10 = lowest)
- Filter Data for “2017” cohort only.
- Group into deciles using binning function:
  - Transform, Visual Binning, Make 9 cutpoints, Label ‘Deciles’, check ‘reverse scale’
- Note bottom high-risk deciles with far lower retention probability (run decile avage)

Last Step: Filter 2017 cohort and create new dataset



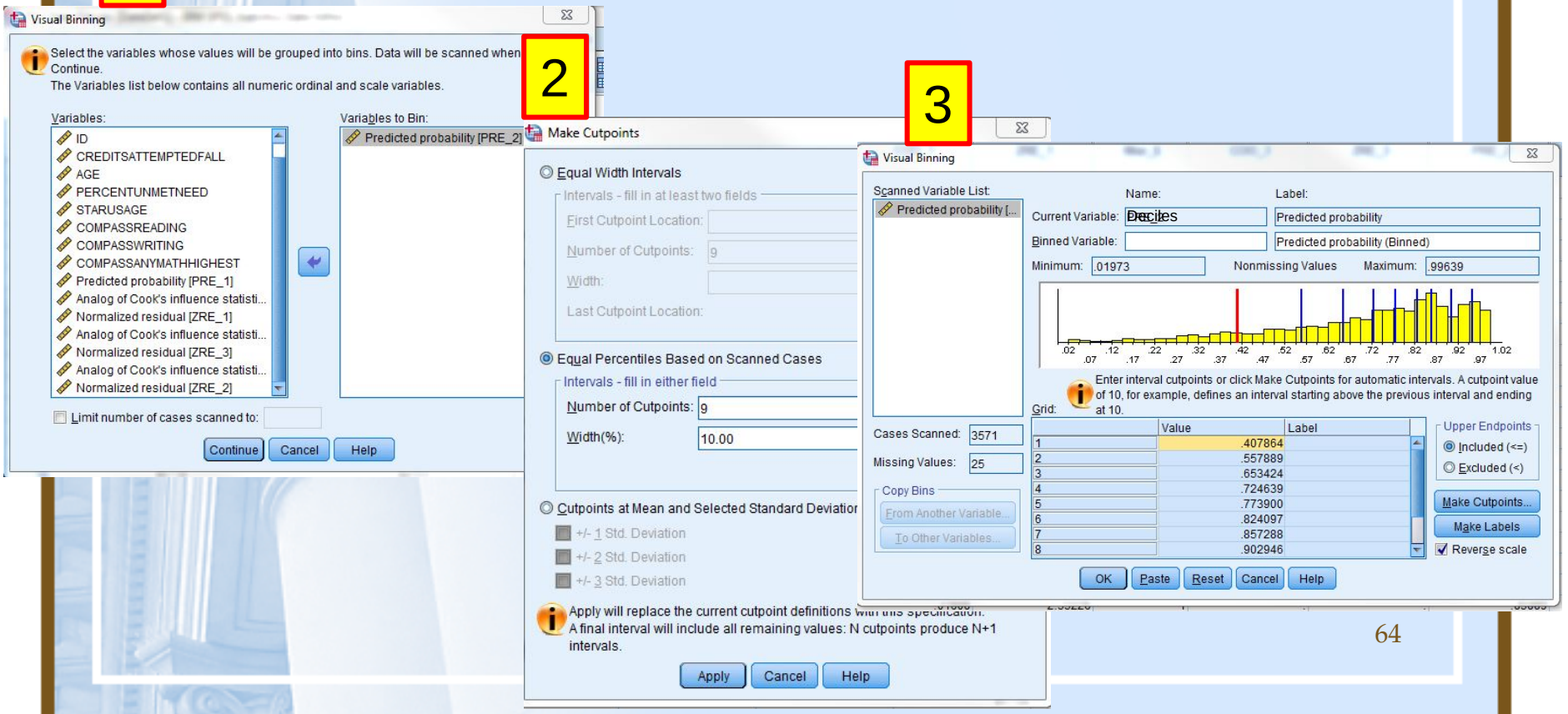
Last Step: Group into deciles using binning function:

-Transform, Visual Binning, Make 9 cutpoints on "PRE_2", Label 'Deciles', check 'reverse scale'

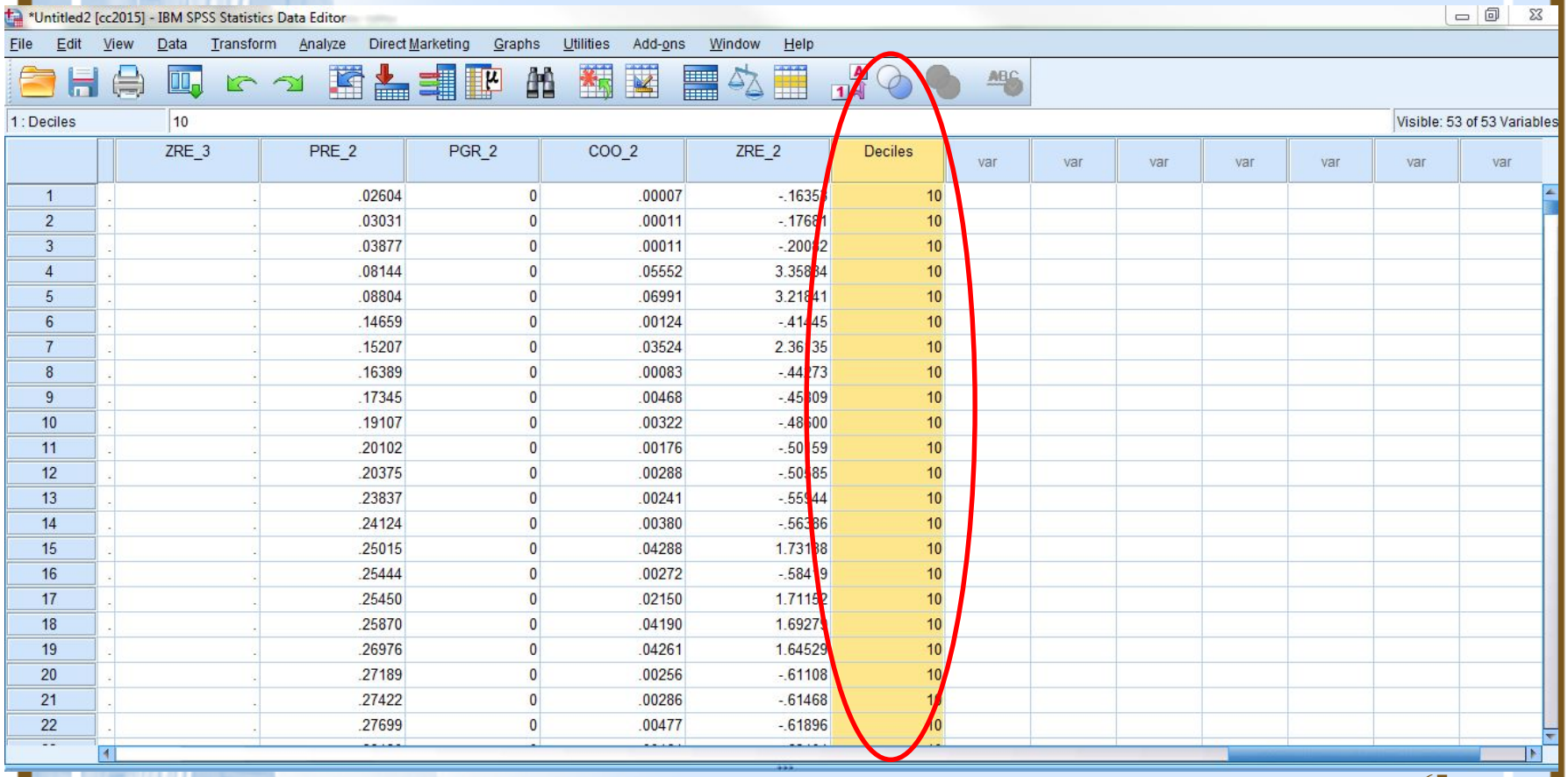
1

2

3



Now your new 2017 dataset has 10 deciles with an even distribution of low-to-high risk scores. Decile 10 is the highest risk



	ZRE_3	PRE_2	PGR_2	COO_2	ZRE_2	Deciles	var	var	var	var	var	var	var
1	.	.	.02604	0	.00007	-.16358	10						
2	.	.	.03031	0	.00011	-.17681	10						
3	.	.	.03877	0	.00011	-.20092	10						
4	.	.	.08144	0	.05552	3.35884	10						
5	.	.	.08804	0	.06991	3.21841	10						
6	.	.	.14659	0	.00124	-.41445	10						
7	.	.	.15207	0	.03524	2.36735	10						
8	.	.	.16389	0	.00083	-.44273	10						
9	.	.	.17345	0	.00468	-.45809	10						
10	.	.	.19107	0	.00322	-.48600	10						
11	.	.	.20102	0	.00176	-.50159	10						
12	.	.	.20375	0	.00288	-.50985	10						
13	.	.	.23837	0	.00241	-.55944	10						
14	.	.	.24124	0	.00380	-.56386	10						
15	.	.	.25015	0	.04288	1.73188	10						
16	.	.	.25444	0	.00272	-.58419	10						
17	.	.	.25450	0	.02150	1.71152	10						
18	.	.	.25870	0	.04190	1.69275	10						
19	.	.	.26976	0	.04261	1.64529	10						
20	.	.	.27189	0	.00256	-.61108	10						
21	.	.	.27422	0	.00286	-.61468	10						
22	.	.	.27699	0	.00477	-.61896	10						

Now that your data is ready, create a spreadsheet for delivery to your advisors/success coaches. Here is an example:

ID	LAST NAME	FIRST NAME	EMAIL	CURRENT CREDITS	RESIDENT	AP/ CLEP	HS GP A	WORK ON CAMP	1st YR EXP CLASS	% FIN NEED MET	STAR LOGINS	ADVISOR PREVIOUS CONTACT
001				15	HI	6	3.80	Y	Y	77%	5	Y
002				14	HI	0	3.33	N	Y	63%	3	N
003				12	CA	6	3.00	N	N	45%	0	N

ID	AGE	GENDER	ETHNICITY	COLLEGE	MAJOR	DEGREE	Ed Goal Specified	Relative Risk Value	Risk Level
001	18	F	CH	CA&H	ART	BA	Yes	14.92	LOW
002	18	F	HW	CSS	SOC	BA	Yes	36.88	MEDIUM
003	18	M	UNDEC	UNDEC	UNDEC	UNDEC	No	89.18	HIGH

Progress on Implementation at Honolulu Community College (2014)

- Delivering student dropout risk scores to HCC's Academic Success Center (via an Excel file).
- Training staff members on using the data.
- Academic advisors moving towards a proactive, targeted approach.

Summary

- Predicting students at-risk
  - Keep prediction model parsimonious
  - Keep prediction data for student advising intuitive and simple (actionable)
  - Triangulate prediction data with multiple sources of information
  - Use prediction data as component part of student dropout-risk assessment
  - Follow ‘best practices’ in IR and keep abreast of changes in analytical and data reporting tools
- Using prediction data for student advising
  - Embrace the use of available data
  - Ensure users conceptually understand what’s behind the data
  - Use data as a complementary piece of information when advising students
  - Timing can be critical in terms of student intervention as well as maximizing advising resources
- Stay abreast of new research on predictive analytics:
  - E.g. “Analytics in Higher Education” by J. Bichsel, Educause, 2012

****