An Overview of Logistic Regression

Christoph Maier
Coordinator of the Applied Research Lab

Stats For Lunch

December 8, 2009

Outline

References

SPSS Survival Manual, 3rd edition by Julie Pallant, McGraw Hill, 2007.
ISBN-13 978-033522366-4.

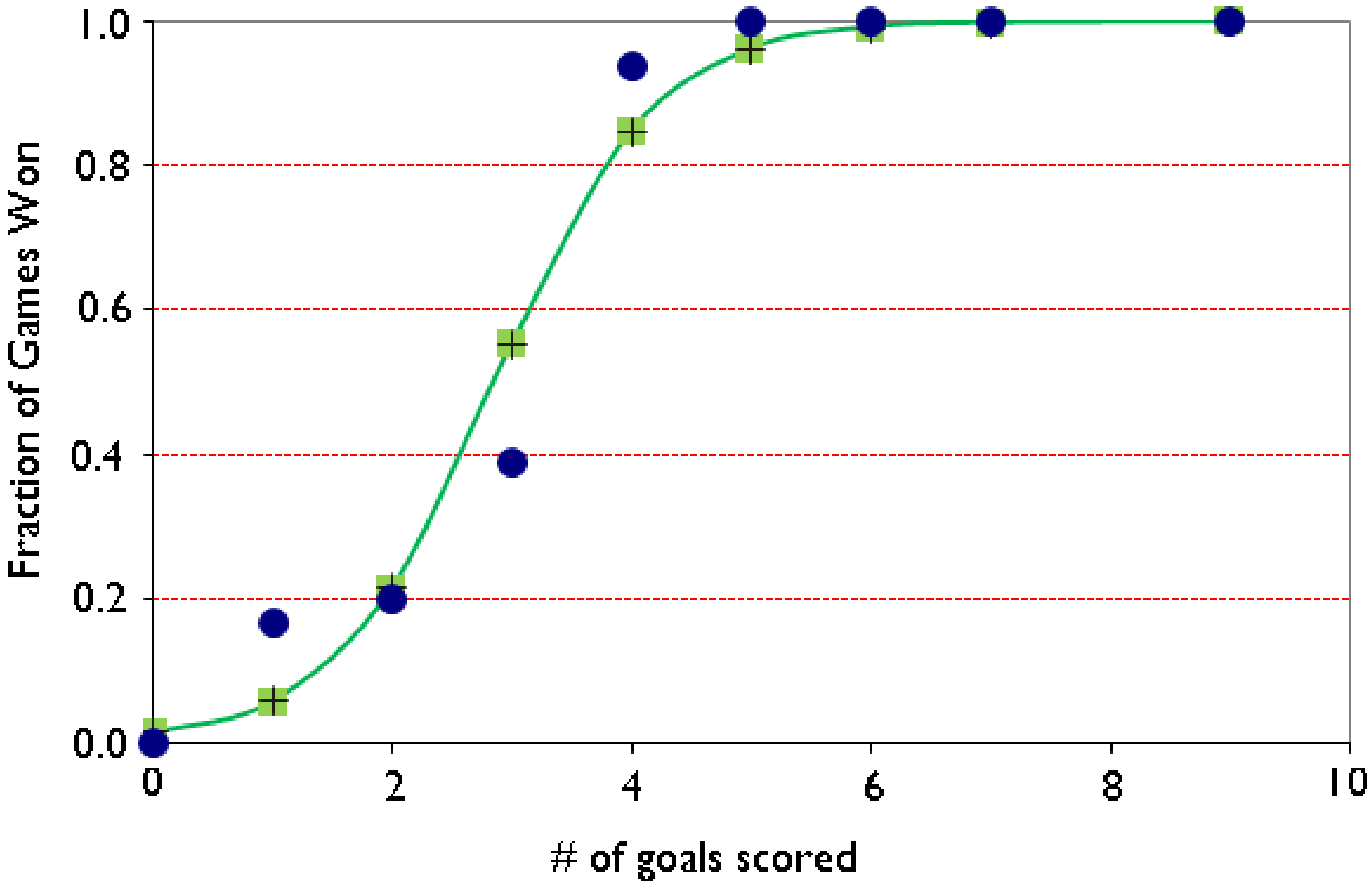
Discovering Statistics Using SPSS by Andy Field, Sage Publications, 2005.
ISBN 0-7619-4452-4.

http://faculty.chass.ncsu.edu/garson/PA765/logistic.htm

Pittsburgh Penguins Hockey Team 2008-2009
Predicting the Likelihood of Winning a Game

# of goals scored by Pittsburgh	# of games won	# of games played	Percentage of games won
0	0	3	0%
1	2	12	17%
2	3	15	20%
3	7	18	39%
4	15	16	94%
5, 6, 7,9	18	18	100%

Observed Likelihood and the Predicted Likelihood of Winning



Use SPSS to Estimate the Likelihood (Probability) of Winning

Important Fields in the Variable View Tab:

	Name	Type	Decimals	Label	Values	Measure
1	ID	Numeric	0	Game ID	None	Nominal
2	GoalsScored	Numeric	0	Goals Scored	None	Scale
3	Won	Numeric	0	Won game	{0, no}...	Nominal
4	HomeGame	Numeric	0	Home game	{0, no}...	Nominal

0 = No

1 = Yes

SPSS Data View Tab

Data View Tab:

	ID	GoalsScored	Won	HomeGame
1	1	4	1	0
2	2	1	0	1
3	3	1	0	1
4	4	3	1	1
5	5	3	0	1

Scored 3 goals

Won=0
so they lost the game

HomeGame=1 so it was a home game

From the SPSS Output

Variables in the Equation

		B	S.E.
Step 1[a]	GoalsScored	1.504	.328
	Constant	-4.308	1.001

a. Variable(s) entered on step 1: GoalsScored.

$$P(winning) = \frac{1}{1+e^{-(b_0 + b_1 \text{ NumGoals})}} = \frac{1}{1+e^{-(-4.308 + 1.504 \text{ NumGoals})}}$$

So when they score 3 goals the likelihood of their winiing the game

$$\frac{1}{1+e^{-(-4.308 + 1.504 \times 3)}} = .551$$

Multiple Regression vs Logistic Regression

Multiple Regression	**Logistic Regression**
Predicted values like the DV	DV=binary (yes/no) but your predict probability=likelihood [0,1]
Estimation by OLS=Ordinary Least Squares	by MLE=Maximum Likelihood Estimation (involves iterating)

Dummy or Indicator Variables

In multiple and logistic regression, you can not use nominal variables like scale variables.

Must create dummy variables to use in place of the nominal variable:

First Decide which level is the reference category

Then create dummy variables for all other levels

Each dummy variable is coded 0 = no and 1=yes

Example: Variable=Race

Race: Nominal variable with 4 levels

1=Caucasian	2=African American	3=Asian	4=Other
Reference Category	First Dummy Variable	Second Dummy	Third Dummy
	AfricanAm	Asian	OtherRace
	0=No 1=Yes	0=No 1=Yes	0=No 1=Yes

In SPSS

Race	AfricanAm	Asian	OtherRace
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1

How does the reference category work?
Race=1

AfricanAm=0 (no), Asian=0 (no) Otherrace=0 (no)

Caucasian=Not African American, not Asian, not other

Odds of an event occurring

$$\text{odds} = \frac{\text{probability of the event occurring}}{\text{probability of the event not occurring}}$$

Probability (likelihood) of contracting a certain disease by race

race	Caucasian (reference category)	African American	Other
Probability	.23	.17	.75
Odds	.23/.77=.3	.17/.83=.2	.75/.25=3

Odds Ratio

$$\text{odds ratio} = \frac{\text{odds of the target category}}{\text{odds of the reference category}}$$

race	**Caucasian (reference category)**	**African American**	**Other**
Probability	.23	.17	.75
Odds	.23/.77=.3	.17/.83=.2	.75/.25=3
Odds Ratio	**Reference**	**.2/.3 = .67**	**3/.3 = 10**

Interpretation

race	Caucasian (reference category)	African American	Other
Probability	.23	.17	.75
Odds	0.3	0.2	3
Odds Ratio	Reference	0.67	10

An individual from an other race is 3-times more likely to contract the disease than not to contract the disease

The odds of an African-American individual contracting this disease is 67% of the <u>odds of a Caucasian contracting the disease.</u>

The odds of an individual from a race other than Caucian or African American contracting the disease is 10 times <u>that of a Caucasian</u>

Odds Ratios for Continuous Variables

Suppose Odds ratio = 1.1 where

- Reference category= any year
- Target category= the next year

The odds of contracting the disease increases by a multiplicative factor of 1.1 every year.

- The target and the reference category can be reversed. Target category is the year before the reference category. Then the odds ratio = 1/1.1 = .909 . Recommended when odds ratio < 1.

Odds Ratios for Continuous Variables

For odds ratio of 1.1 per year

- If the odds is 0.8 for a 50 year old, then the odds for a 51 year old is $0.8*1.1 = 0.88$
- And the odds of a 52 year old is $0.88*1.1=0.8*(1.1)^2 = 0.968$
- … and the odds for a 60 year old is $.8*(1.1)^{10} = 2.07$

Interpretation of Odds Ratios for Continuous Variables

Odds ratio = 1.1 for age (in years)

The odds of contracting the disease increases by a factor of 1.1 per year

The odds of contracting the disease increases 10% per year.
(not by 10 percentage points!)

The odds of contracting the disease doubles every 7.3 years.

$$\frac{\ln(2)}{\ln(\text{odds ratio})} = \frac{\ln(2)}{\ln(1.1)}$$

Odds ratio = .4 for income (in thousands of $)

The odds of contracting the disease changes by a factor of .4 for every additional $1000 increase in salary

The odds of contracting the disease increases by a factor of 2.5 for every $1000 <u>drop</u> in income.

The odds of contracting the disease more than doubles for every $1000 <u>drop</u> in income.

Second Example

Predict the likelihood of Pittsburgh winning a game based on two predictors:

The number of goals they score in the game.
GoalsScored = scale variable

Whether the game is a home game.
Home = Nominal variable
where 0= no, not a not home (away game)
1=yes, a home game

Home is a nominal Variable

But it only has two levels so once you choose the reference category, there is only one level that must be converted to a dummy variable.

Reference category: 0= Away game

Dummy variable : Home 0=away 1=home

☺The original variable is the dummy variable.

Dummy variables coded 0 and 1, not 1 and 2.

Question # 1
Does at least one of these predictors significantly predict the likelihood of winning?

Omnibus Tests of Model Coefficients

Chi-square	df	Sig.
51.5	2	.000
51.5	2	.000
51.5	2	.000

$X^2(2) = 51.5$ $p < .0005$
so yes, at least one of these predictors does help predict the likelihood of winning the game.

Overall test or omnibus test of the model

- Compares -2Log likelihood of the intercept only model vs.
  -2LL of the model with these two predictors.
- Smaller -2LL means that the model fits better.
- The difference follows a chi-square distribution with degrees of freedom = number of predictors

Question # 2
What is r^2 for this model?

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	61.378[a]	.466	.624

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Cox & Snell underestimates R^2

So using Nagelkerke, the model as a whole explains 62.4% of the variability in outcomes of the game.

Question # 3
How well does the model predict wins and losses?

Classification Table[a]

Observed			Predicted: Won game no	Predicted: Won game yes	Predicted: Percentage Correct
Step 1	Won game	no	31	6	83.8
		yes	8	37	82.2
	Overall Percentage				82.9

a. The cut value is .500

Predict a win if likelihood > .5 (default)

The Penguins lost 31+6=37 of their games. The model correctly predicted a loss in 31 (83.8%) of those games (specificity).

The Penguins won 8+37=45 of their games. The model correctly predicted a win in 37 (82.2%) of those games (sensitivity).

Question # 4
Are the individual predictors statistically significant?

Variables in the Equation

		B	S.E.	Wald	df	Sig.
Step 1[a]	GoalsScored	1.52	.33	21.5	1	.000
	HomeGame	.87	.65	1.78	1	.182
	Constant	-4.8	1.08	19.3	1	.000

a. Variable(s) entered on step 1: GoalsScored, HomeGame.

GoalsScored
$X^2(1)=21.5$
$p<.0005$
significant

HomeGame
$X^2(1)=1.78$
$p=.182$
Not significant

Wald's test also has a Chi-square distribution

Warning: This test can under some circumstances tend to declare that statistically significant variables are not statistically significant.

Question # 5
Equation for Predicting likelihood of winning?

Variables in the Equation

		B	S.E.
Step 1[a]	GoalsScored	1.52	.33
	HomeGame	.87	.65
	Constant	-4.8	1.08

a. Variable(s) entered on step 1: GoalsScored, HomeGame.

The coefficients (B) in Logistic regression are called "Logits", because they are the natural log of the odds ratio.

$$P(winning) = \frac{1}{1 + e^{-(b_0 + b_1 \, NumGoals + b_2 \, HomeGame)}}$$

$$= \frac{1}{1 + e^{-(-4.8 + 1.52 \, NumGoals + .87 \, HomeGame)}}$$

Question # 6
What is the effect of GoalsScored?

Variables in the Equation

		B	S.E.	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
Step 1[a]	GoalsScored	1.52	.33	.000	4.6	2.4	8.7
	HomeGame	.87	.65	.182	2.4	.66	8.6
	Constant	-4.8	1.08	.000	.009		

a. Variable(s) entered on step 1: GoalsScored, HomeGame.

Use odds ratio = Exp(B)

The odds of winning the game increases by a factor of 4.6 for every additional goal scored! (more than quadruples)

95% confident that the odds of winning the game increases by a factor of between 2.4 and 8.7 for every additional goal scored.

Question # 7
What is the effect of HomeGame?

Variables in the Equation

		B	S.E.	Sig.	Exp(B)	95% C.I.for EXP(B)	
						Lower	Upper
Step 1[a]	GoalsScored	1.52	.33	.000	4.6	2.4	8.7
	HomeGame	.87	.65	.182	2.4	.66	8.6
	Constant	-4.8	1.08	.000	.009		

a. Variable(s) entered on step 1: GoalsScored, HomeGame.

The odds of winning a home game is 2.4 times the odds of winning an away game.

95% confident that the odds of winning a home game is between 0.66 and 8.6 times the odds of winning an away game.   Note that 1 falls in the interval [0.66, 8.6]

Question # 8
Which predictor is the most important predictor of winning a game?

Variables in the Equation

		B	S.E.	Sig.	Exp(B)	95% C.I.for EXP(B)	
						Lower	Upper
Step 1[a]	GoalsScored	1.52	.33	.000	4.6	2.4	8.7
	HomeGame	.87	.65	.182	2.4	.66	8.6
	Constant	-4.8	1.08	.000	.009		

a. Variable(s) entered on step 1: GoalsScored, HomeGame.

Can not just compare the odds ratios since they are dependent on the magnitude of the unit.

One strategy: standardize the units.

Goals Scored: M=3.22 SD=1.785

HomeGame: M=.5 SD=.503

Which predictor is the most important predictor of winning a game?

Goals Scored:

M=3.22 SD=1.785 OR=1.52 $OR^{SD} = 1.52^{3.22} = 3.85$

HomeGame:

M=0.5 SD=.503 OR=2.4 $OR^{SD} = 2.4^{.503} = 1.55$

Which factor is a more important predictor?

GoalsScored: odds increases by a factor of 3.85 when GoalsScored increases by 1 SD. ☺ more important

HomeGame: odds increases by a factor of 1.55 when HomeGame is increased by 1 SD.

Question # 9
Are there any outliers?

Casewise List[b]

Case	Selected Status[a]	Observed			Temporary Variable	
		Won game	Predicted	Predicted Group	Resid	ZResid
35	S	y**	.038	n	.962	5.0
62	S	y**	.086	n	.914	3.3

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2.0 are listed.

Look for values of |Zresid| >3

Two games– won both but model predicts a loss

35 They won this away game by a score of 1-0.

#62 They won this home game by a score of 1-0.

Note: Good to look at values of Cook's D > 1

And |Leverage values| > 3(number of predictors+1)/n

Question # 10
Does the data meet the conditions for using Logistic Regression

MultiColinearity

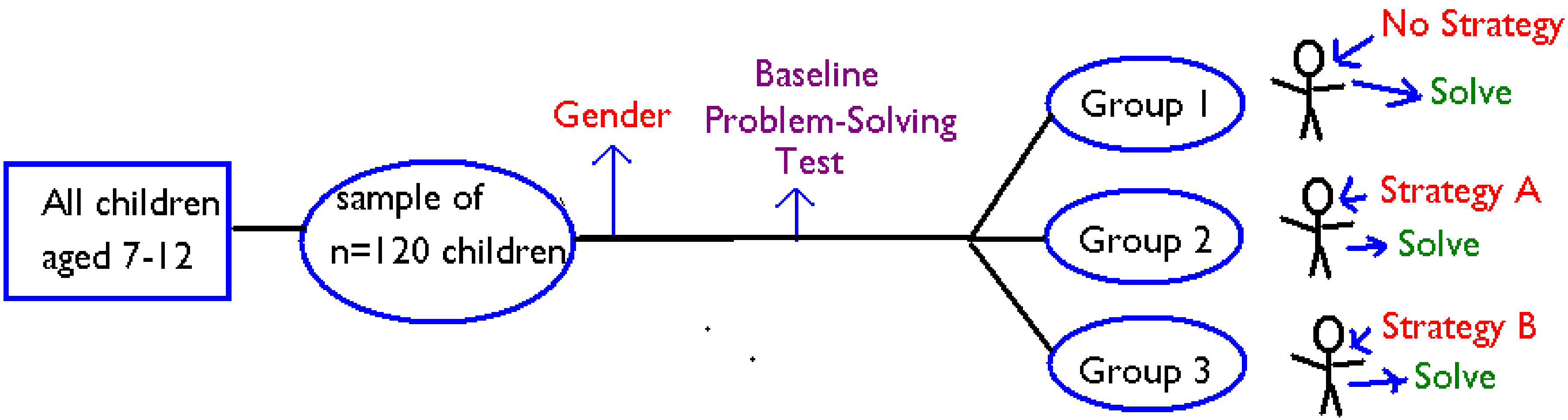
Look for values of |r| > .8 between predictors

Where r=Pearson Correlation Coefficient

Correlations

		Goals Scored	Won game	Home game
Pearson Correlation	Goals Scored	1	.665**	.027
	Won game	.665**	1	.123
	Home game	.027	.123	1

Example # 3



Variables

- Pretest Scale Control Variable
- Gender Nominal Independent Variable
- Strategy Nominal Independent Variable
- Solve Nominal Dependent Variable

Example # 3
How the SPSS Variables were coded

- Gender   1=Female 2=Male
- Pretest   scale of 0 to 100 points
- Strategy 1=No strategy (control)
  2=Strategy A
  3=Strategy B
- Solve   0=No, not correctly solved
  1=yes, correctly solved

Example # 3
SPSS Dummy Variables

- Gender 1=Female 2=Male

  → reference category: Male

  first dummy: Female 0=No 1=Yes

- Strategy 1=No strategy (control)

  2=Strategy A

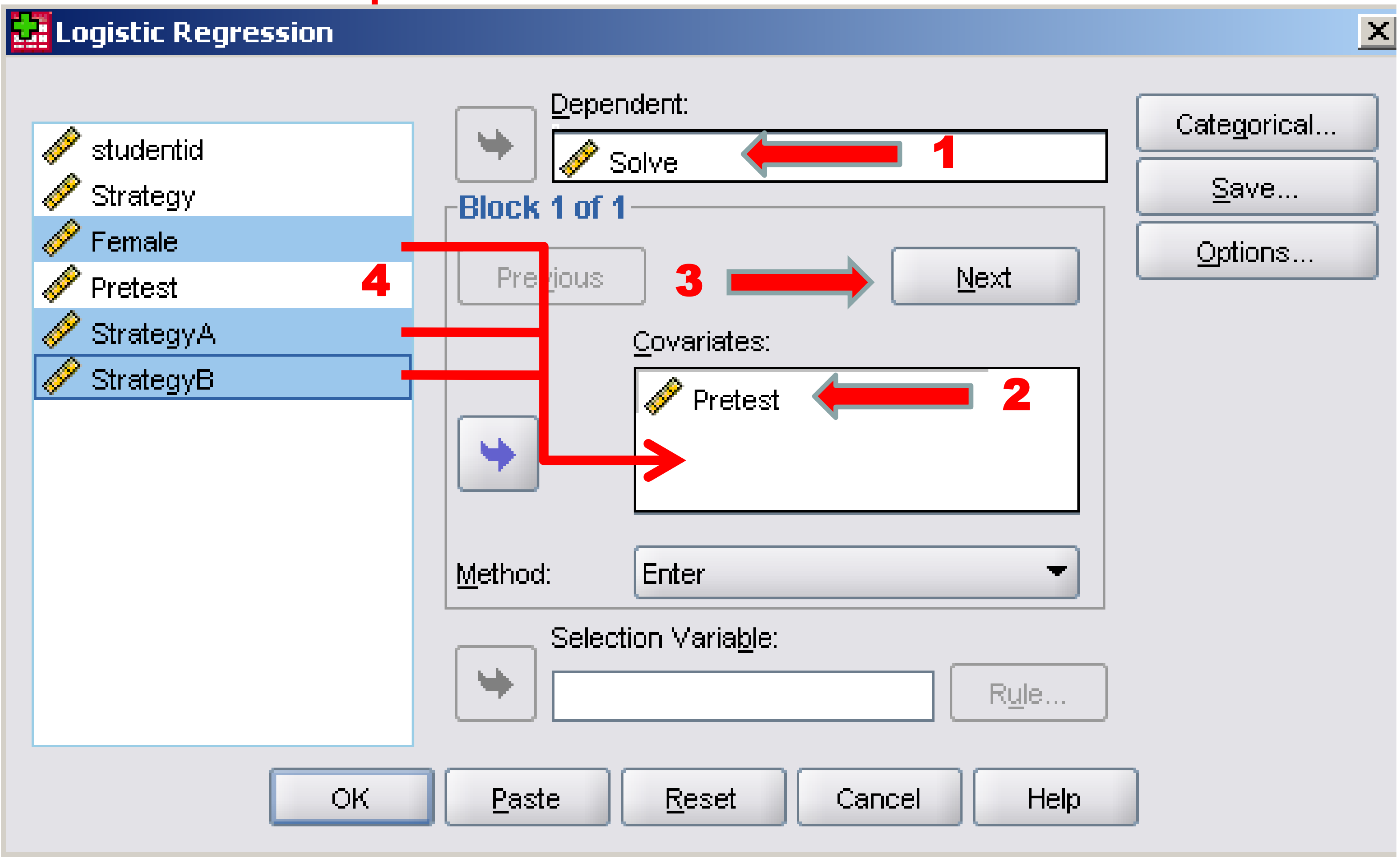
  3=Strategy B

  → reference category: control

  first dummy: StrategyA 0=no 1=yes

  second dummy: StrategyB 0=no 1=yes

Hierarchical Logical Regression in SPSS
Use two blocks: control variables in the first block and predictors in the second block



SPSS Screen
Analyze → Regression → Logistic

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	22.7	1	.000
	Block	22.7	1	.000
	Model	22.7	1	.000

Block 1
Effect of the control variables (pretest score)

Block 2: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	15.5	3	.001
	Block	15.5	3	.001
	Model	38.3	4	.000

Block 2
Effect of the Predictors (female, Strategy A, Strategy B) after adjusting for control variables

How to contact the ARL?

Location: G10 Donna D Putt Hall

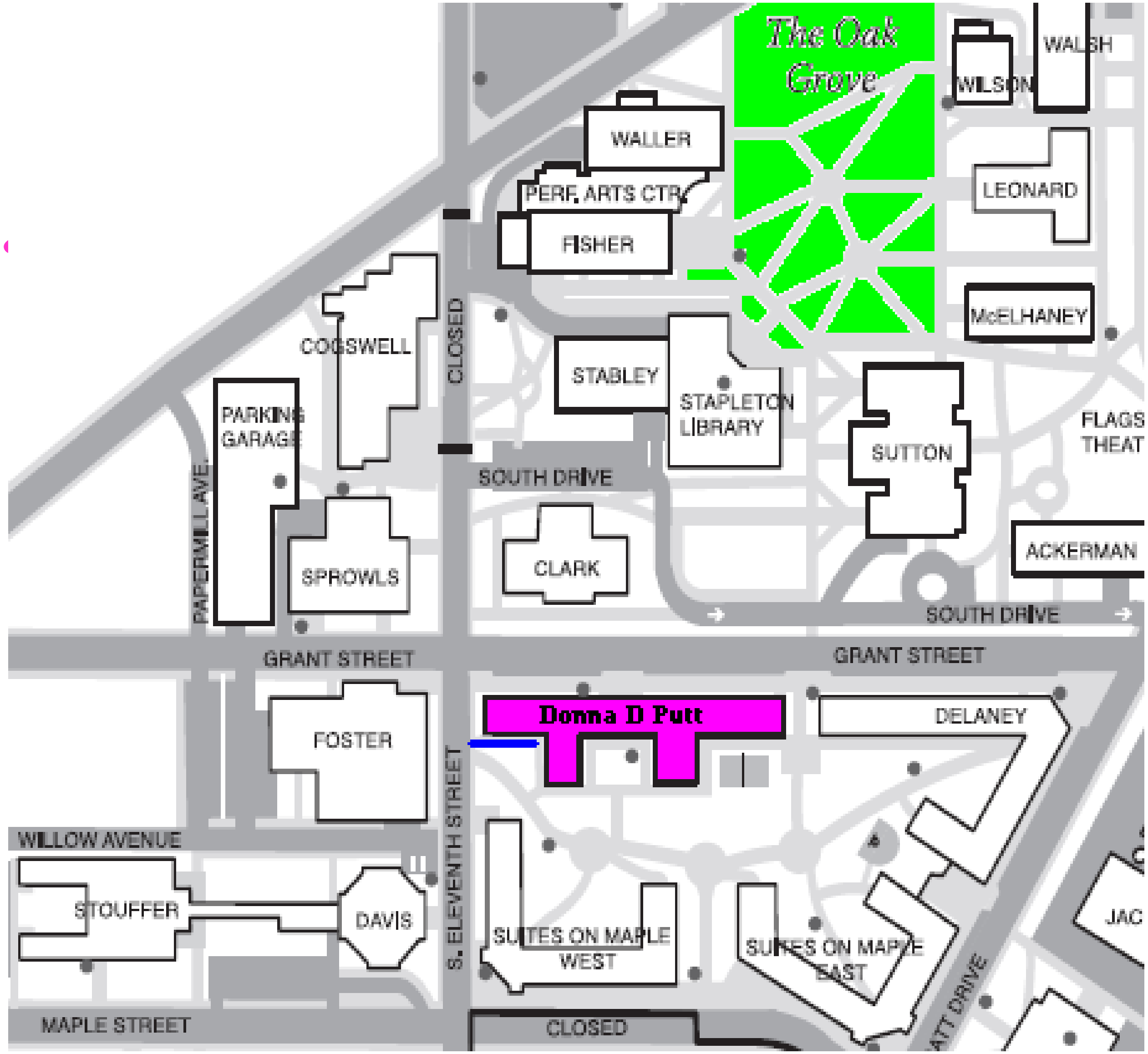
Hours: Monday through Friday 8:00– 4:00
(Fall, Spring, Summer I, and Summer II)

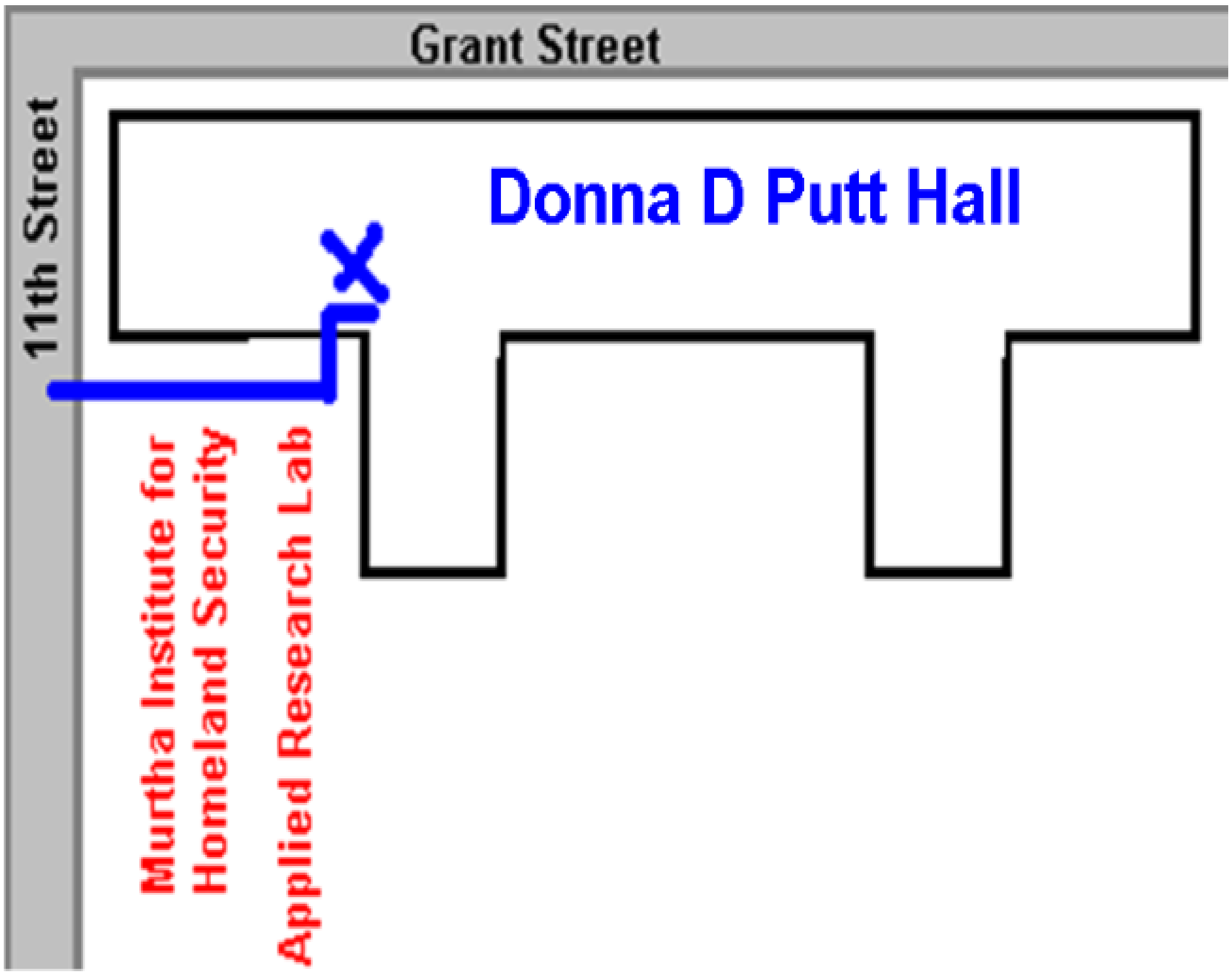
Phone: (724) 357- 4530

Web page: www.iup.edu/arl

Email: iup-arl@iup.edu

Where we are located?




Grant Street
11th Street
Donna D Putt Hall
Murtha Institute for
Homeland Security
Applied Research Lab

Personnel 2009-2010

Coordinator:

Christoph Maier

Graduate Consultants

Steven Brewer	Criminology
Ben Jarrett	Mathematics
Chad Nease	Mathematics
Danielle Smyre	Educational Psychology
Beth Watson	Psychology