An Overview of Logistic Regression

Christoph Maier
Coordinator of the Applied Research Lab

Stats For Lunch

Observed Likelihood and the Predicted Likelihood of Winning

Use SPSS to Estimate the Likelihood (Probability) of Winning

Important Fields in the Variable View Tab:

From the SPSS Output

Variables in the Equation

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

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

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

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

probability of the event 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**

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 $0.88*1.1=0.8*(1.1)^2 = 0.968$

$.8*(1.1)^{10} = 2.07$

Interpretation of Odds Ratios for Continuous Variables

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 # 2
What is r^2 for this model?

Model Summary

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

a. Estimation terminated at iteration number

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?

[redacted]

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?

[redacted]

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

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

-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	GoalsScored	1.52	.33
	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]

The coefficients (B) in Logistic

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

						95% C.I.for EXP(B)

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?

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?

Look for values of |Zresid| >3

Two games | ames»[-C,!56P5ž„ËÀ3Two g

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

MultiColinearity

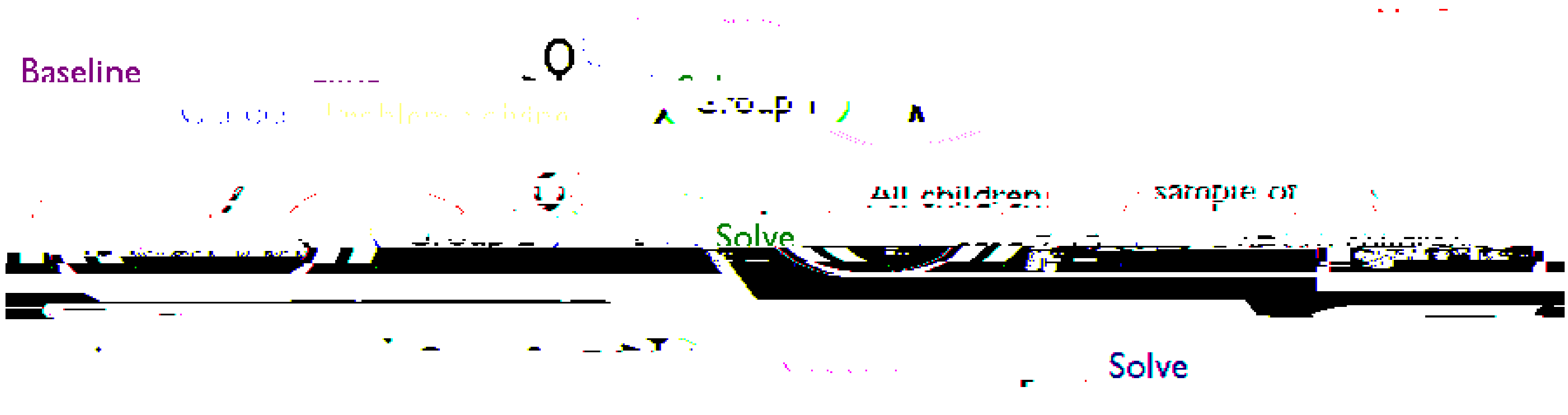
Look for values of |r| > .8 between predictors

Where r=Pearson Correlation Coefficient

Correlations

[redacted]

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

Example # 3
SPSS Dummy Variables

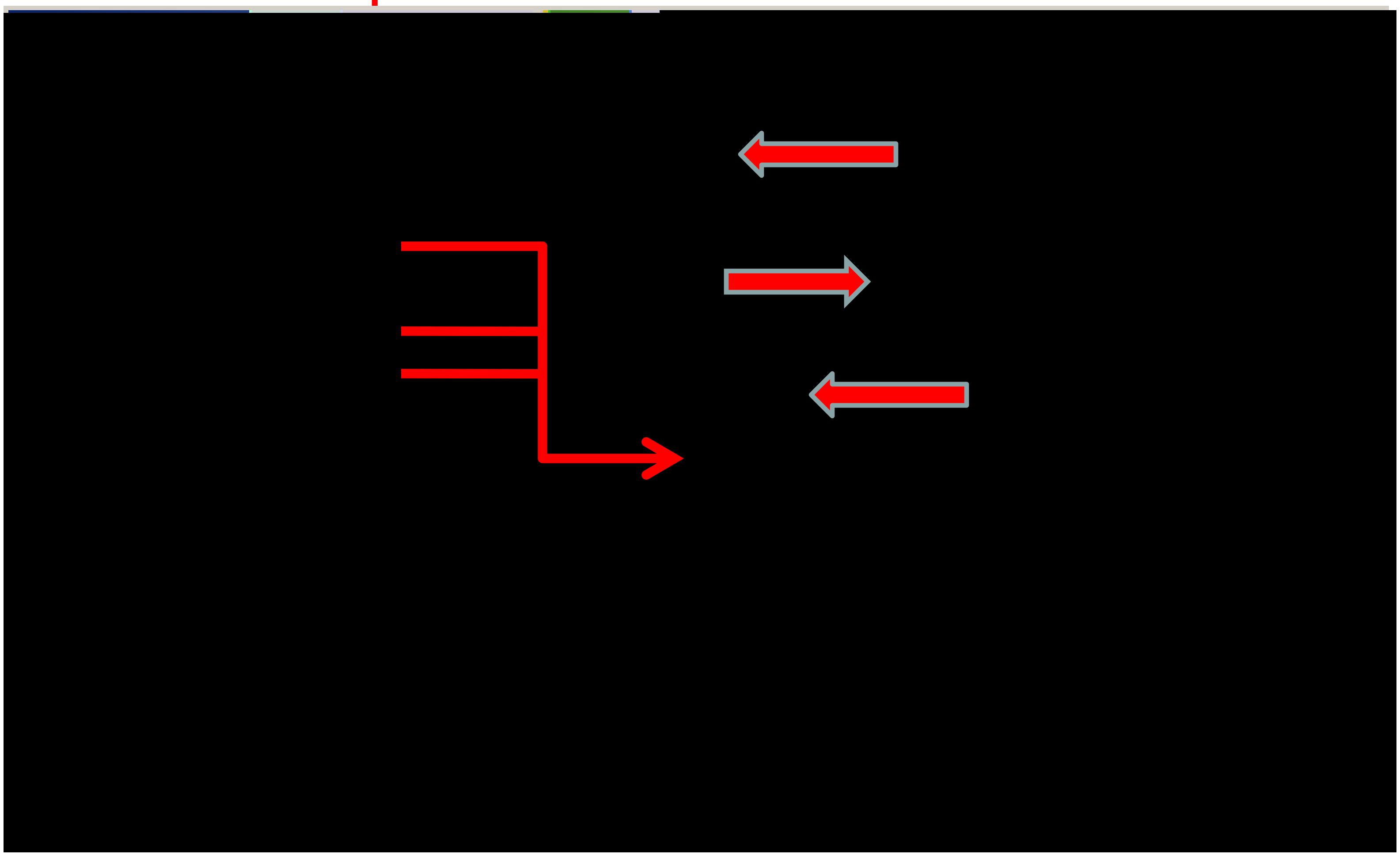
Gender 1=Female 2=Male

→ reference category: Male
first dummy: Female 0=No 1=Yes

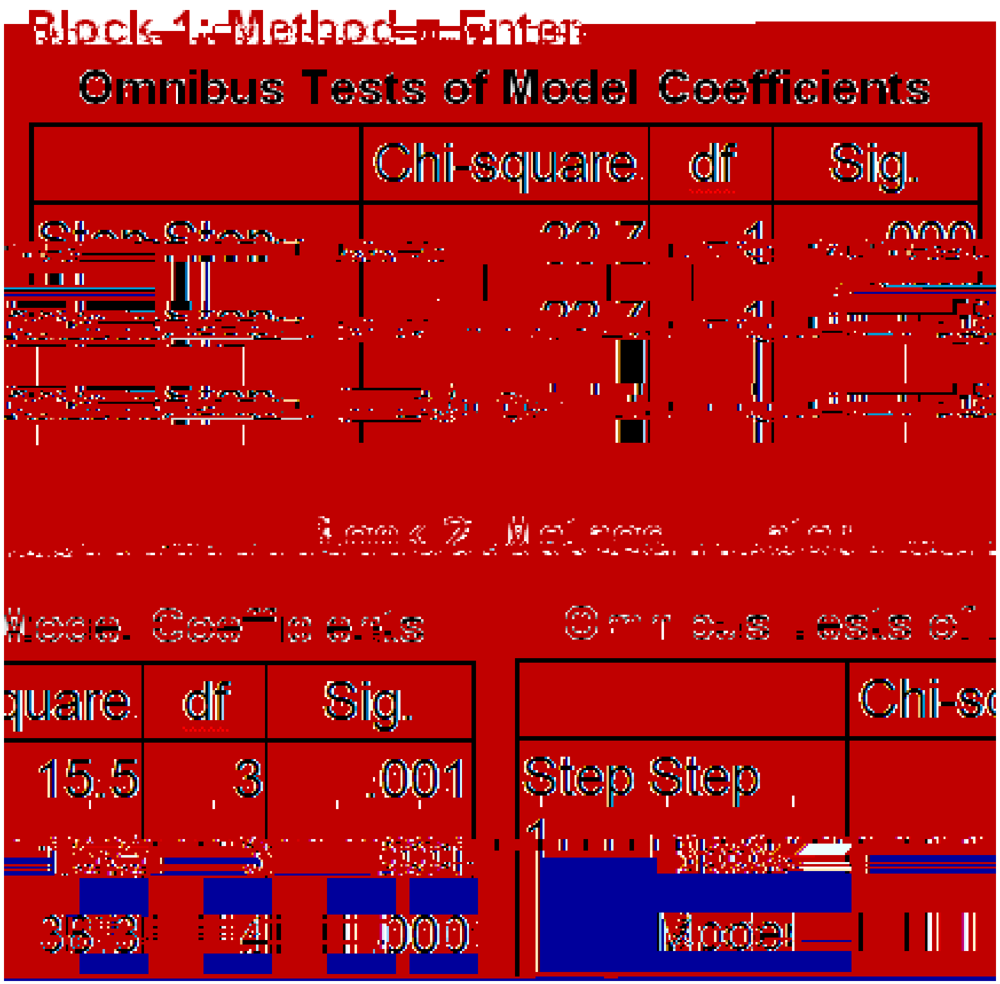
Strategy 1=No strategy (control)
2=Strategy A
3=Strategy B

→ 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

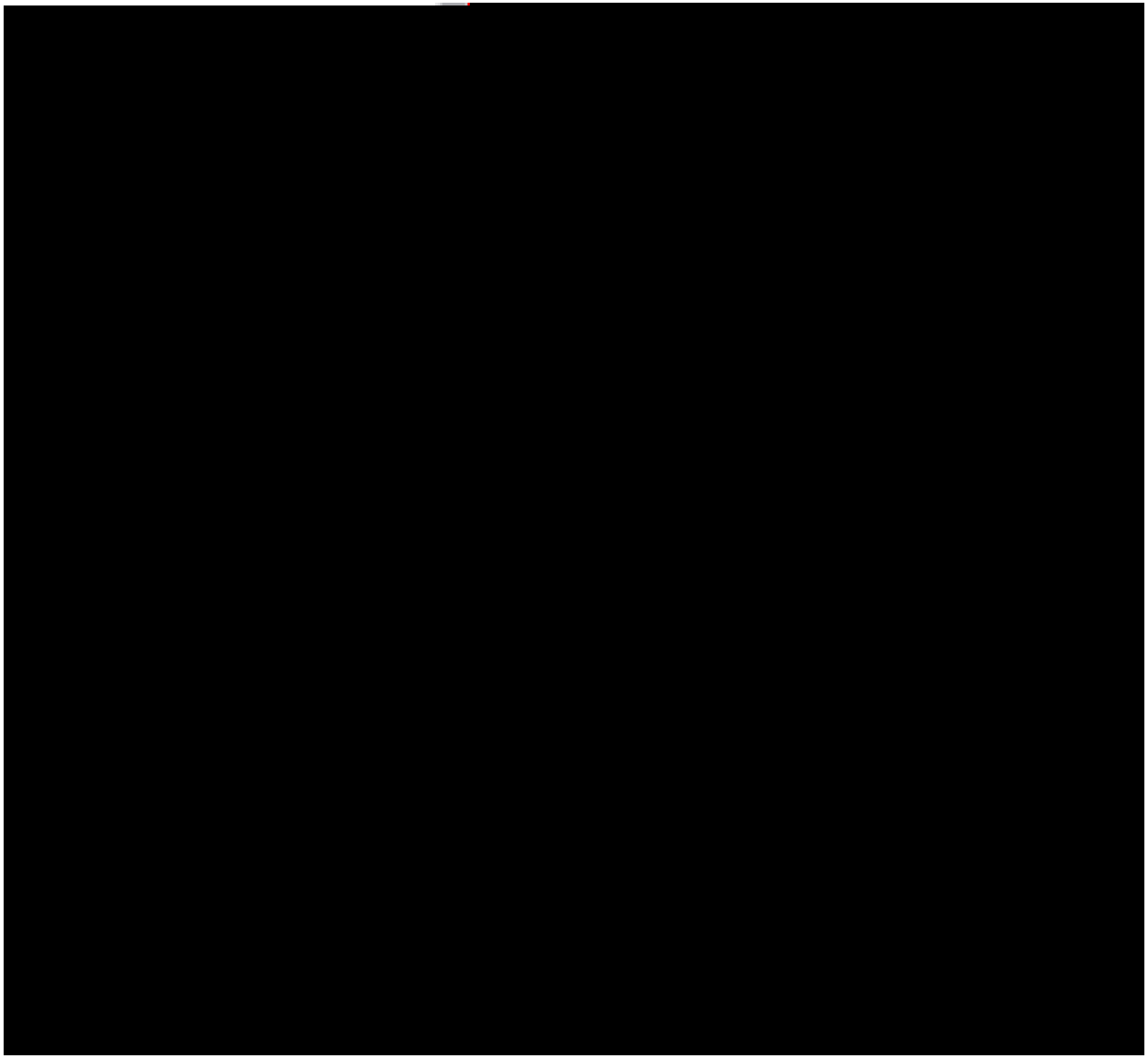


SPSS Screen
Analyze → Regression → Logistic



Block 1
Effect of the
control variables
(pretest score)

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




Grant Street

Coordinator:

Christoph Maier

Graduate Consultants

Steven Brewer	Criminology
Ben Jarrett	Mathematics
Chad Nease	Mathematics