The first three steps in a logistic regression analysis with examples in IBM SPSS.

Steve Simon

P.Mean Consulting

www.pmean.com

2. Why do I offer this webinar for free?

I offer free statistics webinars partly for fun and partly to build up goodwill for my consulting business,

- www.pmean.com/consult.html.

Also see my Facebook and LinkedIn pages

- www.facebook.com/pmean
- www.linkedin.com/in/pmean

I provide a free newsletter about Statistics, The Monthly Mean.

- www.pmean.com/news
- www.facebook.com/group.php?gid=302778306676

3. Abstract

- Abstract: This training class will give you a general introduction in how to use SPSS software to compute logistic regression models. Logistic regression models provide a good way to examine how various factors influence a binary outcome. There are three steps in a typical logistic regression analysis: First, fit a crude model. Second, fit an adjusted model. Third, examine the predicted probabilities. These steps may not be appropriate for every logistic regression analysis, but they do serve as a general guideline. In this presentation, you will see these steps applied to data from a breast feeding study, using SPSS software.

4. Objectives

- Objectives: In this class, you will learn how to:
  - compute and interpret simple odds ratios;
  - relate the output of a logistic regression model to these odds ratios; and
  - examine the assumptions behind your logistic model.

5. Sources

Much of the material for this webinar comes from:

- Stats #04: Using SPSS to Develop a Logistic Regression Model
  - www.childrens-mercy.org/stats/training/hand04.asp

6. Pop quiz #1

The logistic regression model can accommodate all the following settings, except:

1. A categorical outcome variable
2. A categorical predictor variable
3. A continuous outcome variable
4. A continuous predictor variable
5. Multiple predictor variables.
6. Don’t know/not sure

7. Pop quiz #2

In a logistic regression model, the slope represents the:

1. baseline risk in the control group
2. change in the log odds
3. change in the probability
4. odds ratio
5. relative risk
6. don’t know/not sure

8. Definitions

Categorical data is data that consist of only small number of values, each corresponding to a specific category value or label. Ask yourself whether you can state out loud all the possible values of your data without taking a breath. If you can, you have a pretty good indication that your data are categorical.

Continuous data is data that consist of a large number of values, with no particular category label attached to any particular data value. Ask yourself if your data can conceptually take on any value inside some interval. If it can, you have a good indication that your data are continuous.

9. Interpreting regression coefficients

The logistic regression model is useful when the outcome variable is categorical. The simplest and most commonly used form of logistic regression is binary logistic regression, used when the outcome variable is binary. The logistic regression model can accommodate either categorical or continuous predictor variables. It can also handle multiple predictor variables.

10. Titanic data set

I will be using a data example that shows information about mortality for the 1,313 passengers of the Titanic. This data set comes courtesy of OzDASL and the Encyclopedia Titanica.

Variable list (5 variables)

- name (name of the passenger)
- pclass (passenger class)
- age (age in years)
- sex (male of female)
- survived (1=yes, 0=no)

11. What are odds?

- While probability represents the ratio of successes to successes plus failures, odds represent the ratio of successes to failures.
- During the flu season, you might see ten patients in a day. One would have the flu and the other nine would have something else. So the probability of the flu in your patient pool would be one out of ten. The odds would be one to nine.

12. What are odds?

- It's easy to convert a probability into an odds. Simply take the probability and divide it by one minus the probability. Here's a formula.
  - Odds = Prob / (1-Prob)

13. What are odds?

If you know the odds in favor of an event, the probability is just the odds divided by one plus the odds. Here's a formula.

- Prob = Odds / (1+Odds)

- You should get comfortable with converting probabilities to odds and vice versa. Both are useful depending on the situation.

14. What are odds?

Example: If both of your parents have an Aa genotype, then the probability that you will have an AA genotype is .25. The odds would be 1/3, which can also be expressed as one to three.

If both of your parents are Aa, then the probability that you will be Aa is .50. In this case, the odds would be 1. We will sometimes refer to this as even odds or one to one odds.

15. What are odds?

- When the probability of an event is larger than 50%, then the odds will be larger than 1. When both of your parents are Aa, the probability that you will have at least one A gene is .75. This means that the odds are 3, which we can also express as 3 to 1 in favor of inheriting that gene.
- Let's convert that odds back into a probability. An odds of 3 would imply that
  - Prob = 3 / (1+3) = 0.75.

16. What are odds?

Suppose the odds against winning a contest were eight to one. We need to re-express as odds in favor of the event, and then apply the formula. The odds in favor would be one to eight or 0.125. Then we would compute the probability as

- Prob = 0.125 / (1+0.125) = 0.111

17. What are odds?

Notice that in this example, the probability (0.125) and the odds (0.111) did not differ too much. This pattern tends to hold for rare events. In other words, if a probability is small, then the odds will be close to the probability. On the other hand, when the probability is large, the odds will be quite different. Just compare the values of 0.75 and 3 in the example above.

18. Alternate models for probability

- Let's consider an artificial data example where we collect data on the gestational age of infants (GA), which is a continuous variable, and the probability that these infants will be breast feeding at discharge from the hospital (BF), which is a binary variable. We expect an increasing trend in the probability of BF as GA increases. Premature infants are usually sicker and they have to stay in the hospital longer. Both of these present obstacles to BF.

19. Alternate models for probability

- Linear model for probability.
  - prob BF = 4 + 2*GA
- This means that each unit increase in GA would add 2 percentage points to the probability of BF.

GA	prob BF
28	60%
29	62%
30	64%
31	66%
32	68%
33	70%
34	72%

20. Alternate models for probability

- This linear model can sometimes produce “bad” probabilities.
  - prob BF = 4 + 3*GA
- The linear model is still sometimes used when you know that all probabilities are between about 20% and 80%.

GA	prob BF
28	88%
29	91%
30	94%
31	97%
32	100%
33	103%
34	106%

21. Alternate models for probability

- Multiplicative model for probability:
  - Each unit increase in GA leads to a 3-fold increase in risk.
- This model can also produce impossible probabilities. Still, it can be used when you know that all the probabilities are small (e.g., less than 20%).

GA	prob BF
28	0.01%
29	0.03%
30	0.09%
31	0.27%
32	0.81%
33	2.43%
34	7.29%

22. Alternate models for probability

- Multiplicative model for odds:
  - Each unit increase in GA leads to a 3-fold increase in odds.
- This model will always produce reasonable odds and thus, reasonable probabilities.

GA	odds BF
28	27 to 1 against (.037)
29	9 to 1 against (.111)
30	3 to 1 against (.333)
31	1 to 1 (1)
32	3 to 1 in favor (3)
33	9 to 1 in favor (9)
34	27 to 1 in favor (27)

23. Alternate models for probability

- The multiplicative odds model becomes a linear model in the log odds.

GA	odds BF	log odds
28	27 to 1 against (.037)	-3.30
29	9 to 1 against (.111)	-2.20
30	3 to 1 against (.333)	-1.10
31	1 to 1 (1)	0.00
32	3 to 1 in favor (3)	1.10
33	9 to 1 in favor (9)	2.20
34	27 to 1 in favor (27)	3.30

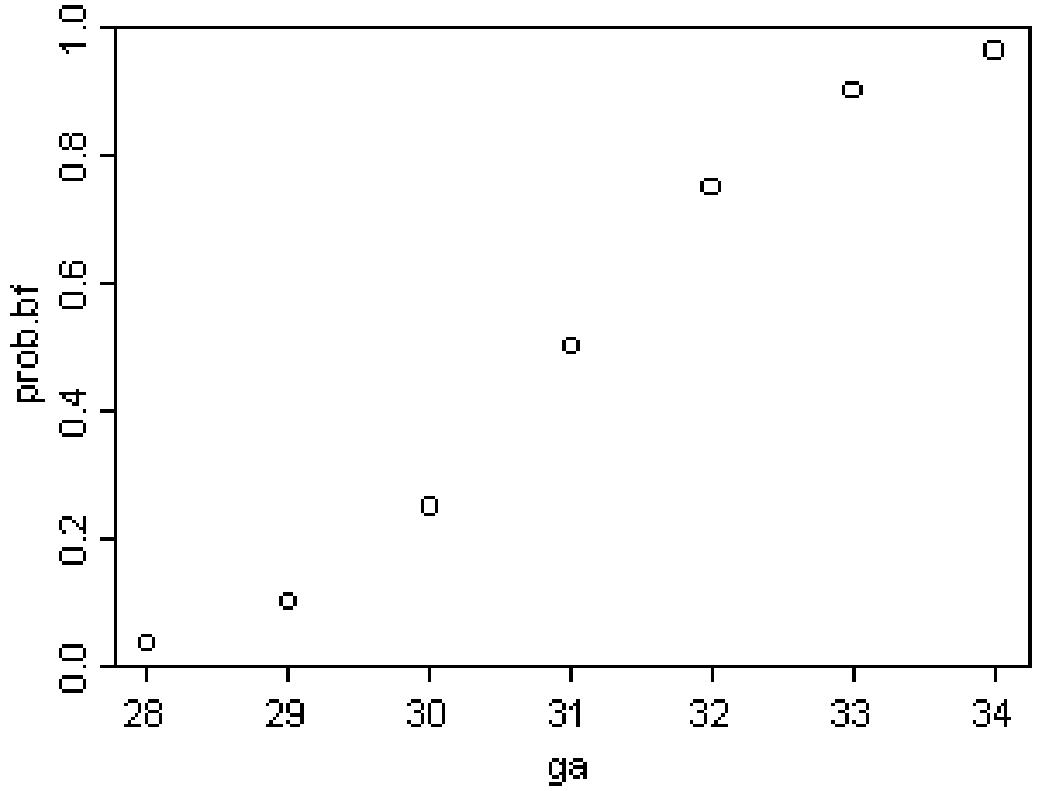
24. Alternate models for probability

- The probabilities in this model show an interesting pattern.

GA	odds BF	prob BF
28	27 to 1 against (.037)	3.6%
29	9 to 1 against (.111)	10.0%
30	3 to 1 against (.333)	25.0%
31	1 to 1 (1)	50.0%
32	3 to 1 in favor (3)	75.0%
33	9 to 1 in favor (9)	90.0%
34	27 to 1 in favor (27)	96.4%

25. Alternate models for probability

- The graph shows a classic S-shaped curve.



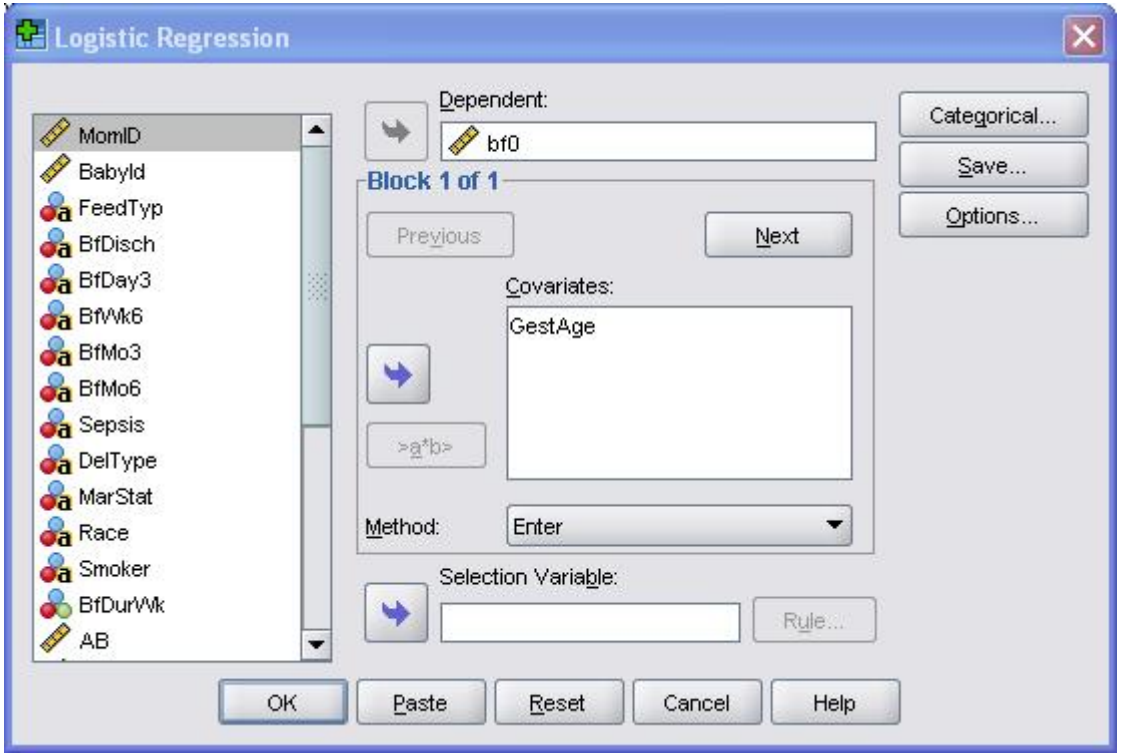
26. Some real data

- All the data up to now has been conceptual. Here is some real data. I've simplified the data set by removing some of the extreme gestational ages.

GA	Actual prob BF
28	2/6 = 33.3%
29	2/5 = 40.0%
30	7/9 = 77.8%
31	7/9 = 77.8%
32	16/20 = 80.0%
33	14/15 = 93.3%

27. Some real data

- Here's the dialog box in SPSS.



28. Some real data

- The estimated logistic regression model is
  - log odds = -16.72 + 0.577*GA

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	GestAge	.577	.198	8.517	1	.004	1.781
	Constant	-16.720	6.063	7.605	1	.006	.000

29. Some real data

- Let's examine these calculations for GA = 30. The predicted log odds would be
  - log odds = -16.72 + 0.577*30 = 0.59
- Convert from log odds to odds by exponentiating.
  - odds = exp(0.59) = 1.80
- And finally, convert from odds back into probability.
  - prob = 1.80 / (1+1.80) = 0.643
- The predicted probability of 64.3% is reasonably close to the true probability (77.8%).

30. Some real data

- Here is a table showing the results for all the values of gestational age.

GA	Predicted log odds	Predicted odds BF	Predicted prob BF
28	-0.57	0.57	36.2%
29	0.01	1.01	50.3%
30	0.59	1.80	64.3%
31	1.16	3.20	76.2%
32	1.74	5.70	85.1%
33	2.32	10.15	91.0%

31. Some real data

You might also want to take note of the predicted odds. Notice that the ratio of any odds to the odds in the next row is 1.78. For example,

- 3.20/1.80 = 1.78
- 5.70/3.20 = 1.78
- 10.15/5.70 = 1.78

It's not a coincidence that you get the same value when you exponentiate the slope term in the log odds equation.

- exp(0.59) = 1.78

32. Some real data

This is a general property of the logistic model. The slope term in a logistic regression model represents the log of the odds ratio representing the increase (decrease) in risk as the independent variable increases by one unit.

33. Categorical predictor variables

You treat categorical variables in much the same way as you would in a linear regression model. Let's start with some data that listed survival outcomes on the Titanic. That ship was struck by an iceberg and 863 passengers died out of a total of 1,313. This happened during an era where there was a strong belief in "women and children" first.

34. Categorical predictor variables

Here is the output from crosstabs.

sex * survived Crosstabulation

			survived		Total
			No	Yes	
sex	female	Count	154	308	462
		% within sex	33.3%	66.7%	100.0%
	male	Count	709	142	851
		% within sex	83.3%	16.7%	100.0%
Total		Count	863	450	1313
		% within sex	65.7%	34.3%	100.0%

35. Categorical predictor variables

You can see this in the crosstabulation shown above. Among females, the odds of dying were 2-1 against, because the number of survivors (308) was twice as big as the number who died (154). Among males, the odds of dying were almost 5 to 1 in favor (actually 4.993 to 1), since the number who survived (142) was about one-fifth the number who died (709).

36. Categorical predictor variables

Here is some additional output from SPSS.

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for sex (female / male)	.100	.077	.130
For cohort survived = No	.400	.350	.457
For cohort survived = Yes	3.995	3.393	4.704
N of Valid Cases	1313		

37. Categorical predictor variables

Here is the output from a logistic regression model in SPSS.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	SexMale	-2.301	.135	291.069	1	.000	.100
	Constant	.693	.099	49.327	1	.000	2.000

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

38. Conclusion

The logistic regression model is useful when the outcome variable is categorical. The logistic regression model can accommodate either categorical or continuous predictor variables. It can also handle multiple predictor variables. The slope in a logistic regression model is related to the odds ratio.

39. Repeat of pop quiz #1

The logistic regression model can accommodate all the following settings, except:

1. A categorical outcome variable
2. A categorical predictor variable
3. A continuous outcome variable
4. A continuous predictor variable
5. Multiple predictor variables.
6. Don’t know/not sure

40. Repeat of pop quiz #2

In a logistic regression model, the slope represents the:

1. baseline risk in the control group
2. change in the log odds
3. change in the probability
4. odds ratio
5. relative risk
6. don’t know/not sure