Week10: Assessing logistic regression models

Time stamper

Odds clarification

- Mega millions (1 to 259 million)
- Odds of getting stuck by lightening (1 in 3000)
- Odds of getting married by age 40 (6.14 to 1)
- But what does it mean?

Odds clarification

- Dice example: odds of rolling a 6 with 1 die?
  - 1 to 5
  - If you roll the die a bunch of times, you would expect in general, that for every 1 time you roll a six, there will be 5 times you don't
  - So the odds are written as the chance of event A to the chance of not A
  - If the two events were equal, the odds would be 1 to 1
  - So when the odds of getting struck by lightening are 1 to 3000, it means for every time someone gets struck by lightening, there are 3000 times they don't.
  - When the odds of getting married by age 40 are 6.14 to 1, it means for about every 6 times someone over 40 gets married, 1 doesn't

Odds ratio

- Another way to compare two things. But now instead of comparing event A to not event A, you are comparing two separate events/conditions.
  - If there was no relationship between the two events, then the OR would be 1 – the odds of 'success', being pain free, would be the same for both treatments
  - If greater than one, then suggests chances of 'success' greater with treatment, less then one, suggests chances of 'success' greater without treatment.
- Odds for a sample is same as dividing the number of successes by the number of failures (works out to be the same as $\pi/1-\pi$)
  - Odds pain free with TMS= 39/61=.693 Odds pain free with sham= 22/78=.282
- How do the odds of being pain free with TMS compare to that of sham?
  - Odds ratio (OR) = .639/.282 =2.27
- Interpret: **The odds of being pain free were 2.27 times higher with TMS than with the sham.**

	TMS	Placebo	Total
Pain free	39	22	61
Not pain free	61	78	139
total	100	100	

Predicting medical school acceptance from MCAT score

- Odds ratio, or Exp(B) – what to interpret to talk about relationship between predictor (MCAT) and response (success of getting into med school – log odds of succes)
  - For each additional point on your MCAT, your odds of being accepted to medical school increase by a factor of 1.279.
- If no relationship, OR=1.
- values greater than 1, increases your odds of 'success', less than one decreases your odds
  - Success is whatever is coded as 1 for response variable
- No 'on average' – beauty of the 'spinner model' underlying logistic regression. Deals with error in a different way.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	MCAT	.246	.089	7.573	1	.006	1.279
	Constant	-8.712	3.237	7.246	1	.007	.000

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

What happened to the ε?

- Simple linear regression:

$Y=\beta_0 + \beta_1 X + \varepsilon$

- Where the randomness in the model was in the error term
  - errors were independent, normal, and had constant variance
- Logistic regression:

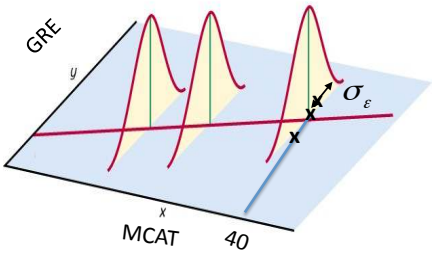
$\log(odds)=\beta_0 + \beta_1 X$

- No ε
- It's the beauty of the 'spinner model'/Bernoulli distribution

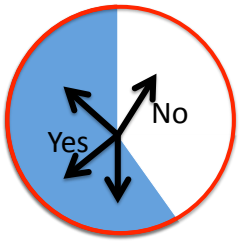
What happened to the ε?

Linear regression

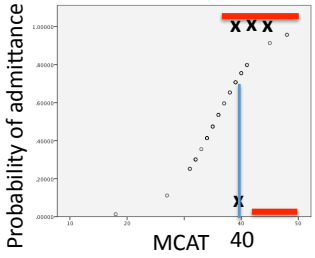
MCAT 40 -> predicted Y =75



Logistic regression



MCAT 40 -> predicted probability of 60



That's why we lose the ε and the 'average'

- Simple linear regression:

$Y=\beta_0 + \beta_1 X + \varepsilon$

- For a 1 unit increase in X
- Y increases/decreases on average by β_1
- Logistic regression:

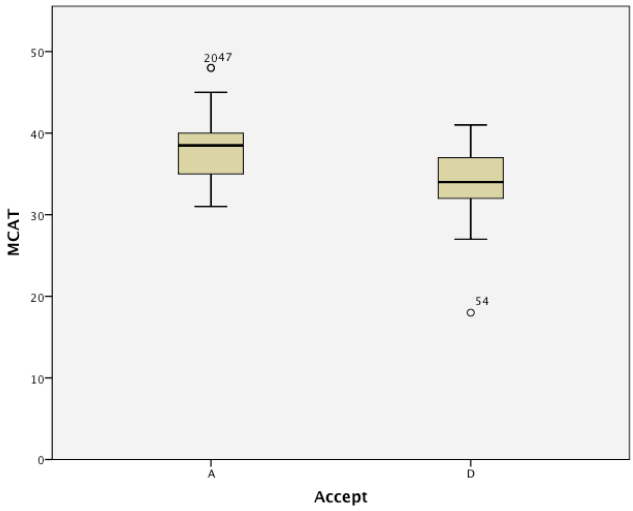
$\log(odds)=\beta_0 + \beta_1 X$

- For a 1 unit increase in X
- the odds of success increase/decrease by a factor of $e^{\beta 1}$
- Both are in units of predictor variable

Assumptions

- Linearity:
  - log(odds) is linearly related to predictor variables
- Independence:
  - No pairing or clustering of the data in space or time
- Random:
  - Want random sample from population as usual
  - Need to make sure a 'spinner model' is valid – super important!!
- Because no ε
  - No normality assumption 90% correct
  - No constant variance assumption:
    - Does not apply: In fact, variability in Y is largest when x is near 1/2, and lowest when it is near 0 or 1.

Looking at your data



Linearity

- Not as 'straight' forward as before
- Binary predictor
  - automatic!
- Quantitative
  - Book suggests logit plots: difficult to make with SPSS
  - Box-Tidwell test
    - Suggested by Hosmer and Lemeshow (1989)
    - Add the term of the form x*log(x)
      - where log means the natural log and x is a predictor
    - If 2nd order term is significant, suggests NOT linear
      - Need to make a logit plot to figure out exactly how nonlinear
    - If nonsignificant, take it out and continue!
    - Its not overly sensitive to small deviations from linearity
  - **BUT you must make sure all values of X are positive!**

Med school example

- **IMPORANT NOTE: all values of X must be positive!!!!**
  - **You can ensure this by adding a constant to all values of X, large enough to make all numbers positive, and greater than 0**
  - **Use this new X for the test, then use original for the actual model!!!!**
- Put 1st order term of MCAT and 2nd order term of MCAT x log(MCAT) into logistic model
- Box-Tidwell

H_0: $\beta_2 = 0$

H_a: $\beta_2 \neq 0$

Since p is large (p=.942), we fail to reject the null hypothesis. This suggests that relationship between log(odds) of acceptance to medical school and MCAT scores is linear.

- Implications?
- So we would take out the interaction term and go back to our original model

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	MCAT	.615	5.062	.015	1	.903	1.850
	MCAT by logMCAT	-.081	1.103	.005	1	.942	.923
	Constant	-11.609	39.912	.085	1	.771	.000

a. Variable(s) entered on step 1: MCAT, MCAT * logMCAT .

How was the data collected?

- Randomness
  - Randomness super important, because statistical tests and intervals are based on the probability model (spinner model)
    - If not valid, can't trust tests and intervals
- Book gives lots of examples to demonstrate the subtleties.
- Even if fail randomness assumption, model can still be useful description of a relationship and sometimes still used for prediction (83% correct)

Is the spinner model valid?

- Usually if you have random assignment, like in an experiment this is satisfied
- TMS study example
  - Because people randomly assigned to treatment or placebo, you know that the difference in pain is due to the treatment.
    - Not to some other systematic difference, or grouping

Is the spinner model valid?

- Medical school example
  - Who gets into med school is not decided by random assignment to yes/no
  - People who apply to med school is not a random sample
    - But probably not systematically different from our population of interest – people applying to med school
  - No reason to think that for our sample MCAT scores are systematically varied in some unusual way

Spinner model valid?

- Golf putting example
- Predict whether if someone makes a putt or not is related to the distance of the putt.
  - Outcome not random, but where the ball goes depends on so many things, like wind, grass height, incline….. That we can probably think of it as following probability model.

How was the data collected?

- Independence
- If fail randomness, don't need to think about independence
- Something can be random but not independent
  - If put ticket numbers into a hat, mix them up, and take one out
  - Choice is random, but once you draw a number, you won't be able to draw it again, so not independent.
    - If you replace number after draw, then independent
- Helps to ask yourself about time, space, and the yes/no decision
  - Are the results from a time-ordered process?
    - Usually not good…
  - Do the observational units have a spatial relationship?
    - Students from the same school
  - Is the yes/no decision based on subjective judgment?
    - If so, might might introduce dependence

Measures for investigating significance and usefulness of model

- Linear regression
  - Test of betas (t-tests)
  - Overall model test (ANOVA)
  - Quantities: std error of the model, and R^2/adjR^2
- Logistic regression
  - Test of betas (wald test: z-statistic instead of t-stat)
    - So its NOT the same distribution for the test statistic (55% correct)

    $z=\hat{\beta}/SE_{\hat{\beta}}$

Beta test

H_0: $\beta_1 = 0$
H_a: $\beta_1 \neq 0$

Since p<.01, we reject the null hypothesis.

This suggests that there is a log-linear relationship between MCAT scores and whether someone gets accepted to medical school or not.

For every additional point on the MCAT, the odds of being accepted to medical increase by a factor of 1.279.

- Confidence interval:
  - We can be 95% confident that the true increase in the odds of acceptance is between 1.073 and 1.524.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	MCAT	.246	.089	7.573	1	.006	1.279	1.073	1.524
	Constant	-8.712	3.237	7.246	1	.007	.000		

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

Assessing the model

- Linear regression
  - Overall model test (ANOVA)
    - Sum of squares: tested how much error the model accounted for
- Logistic regression
  - Overall model test: uses a different method
  - 94% got that there is a similar test

Method of maximum likelihood

- Instead of reducing sum squared error, it tries to minimize -2logL or -2logLikelyhood
  - Called deviance
  - Behaves similar to the residual sum of squares in regression
- Can compare nested models by looking at the change in deviance
  - How we can determine if overall model is significant
  - Compare the model with just the constant to the full model using -2logL
    - Follows a chi-square distribution (X^2), instead of an F
- We want -2loglikelyhood to be LOWER, since trying to reduce 80% correct

ANOVA now becomes Omnibus test

- Look at Chi-square (book calls G, or drop in deviance)
- Values for step, block, and model are all the same because we didn't use stepwise regression or blocking
  - df: 1 for each predictor in the model
- While the hypothesis is the same as for the Beta (when just one predictor), pvalues may not always agree, in that case, go with the -2log(L) or -2LogLikelyhood

H_0: $\beta_1 = 0$
H_a: $\beta_1 \neq 0$
Since p<.01, we reject the null. This suggests that MCAT scores are useful in predicting whether someone is admitted to medical school.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	11.094	1	.001
	Block	11.094	1	.001
	Model	11.094	1	.001

Assessing the model

- Linear regression
  - Quantities: std error of the model, and R^2/adjR^2
- Logistic regression
  - Quantities?

R^2 doesn't mean the same thing

- Don't really use in logistic because just approximations
- -2LogLikelihood will become important in multiple logistic regression
  - As we saw before, can be used to compare nested models, but doesn't have any meaning on its own.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	64.697[a]	.183	.244

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Other output

- Tells you how well it is classifying
- Getting it right 60% of time when not admitted, 66.7% correct when they are accepted
- False alarms?
- Misses?

Classification Table[a]

	Observed		Predicted		
			Acceptance		Percentage
			0	1	Correct
Step 1	Acceptance	0	15	10 (FA)	60.0
		1	10 (Miss)	20	66.7
	Overall Percentage				63.6

a. The cut value is .500

Summary

- Assumptions
  - Randomness, independence: how the data were collected
  - Linearity: Box-Tidwell test
  - No normality or equal variance like in linear regression
- Logistic regression
  - Test of betas (wald test: z-statistic instead of t-stat)
  - Overall model test (Omnibus test instead of ANOVA)
  - Quantities?

Complete example

- Collected election data from 2008. Use percentage of adults with at least a college education to predict whether Obama won a majority of the votes in the state

Assumptions

- Linearity:
  - log(odds) is linearly related to predictor variables
  - Make sure all values positive
- Independence:
  - No pairing or clustering of the data
  - ask yourself about time, space, and the yes/no decision
- Random:
  - Need to make sure a 'spinner model' is valid – super important!!
  - Want random sample from population

Assumptions

- Linearity: Box-Tidwell test

All values are positive because percentage

H_0: $\beta_2 = 0$

H_a: $\beta_2 \neq 0$

Since p>.05 we fail to reject the null hypothesis. This suggests we meet the assumption of linearity

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	BA	-1.574	5.055	.097	1	.756	.207
	BA by BAlog	.457	1.192	.147	1	.701	1.580
	Constant	2.281	30.390	.006	1	.940	9.786

a. Variable(s) entered on step 1: BA, BA * BAlog .

Assumptions

- Randomness
  - Sample is all states in the US, but only want to make inference about the states in the US
  - Model can still be informative
- Independence
  - Related in space – north vs south, coasts vs middle of country
  - But at some level, each person's vote is independent, so maybe not the worst.

Output

- Test the betas

H_0: $\beta_1 = 0$

H_a: $\beta_1 \neq 0$

Since p<.01, we reject the null hypothesis. This suggests that there is a significant relationship between percentage of people with a bachelor's degree and the log odds of whether Obama won or not.

Or: a significant log-linear relationship between percentage of people with a bachelor's degree and the odds of Obama winning or or not.

- Interpret the beta

For each additional percentage point in BA, the odds of Obama winning increase by a factor of 1.449.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	BA	.371	.114	10.518	1	.001	1.449
	Constant	-9.467	2.966	10.189	1	.001	.000

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

Output

- Assess the model

H_0: $\beta_1 = 0$

H_a: $\beta_1 \neq 0$

Since p<.001, we can reject the null hypothesis. This suggests that the percentage of people with BA's is useful in predicting whether Obama won or not.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	20.048	1	.000
	Block	20.048	1	.000
	Model	20.048	1	.000