Multiple logistic regression
timestamper!!!!

Multiple logistic regression

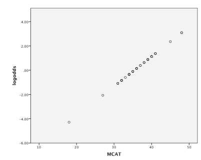
- Don’t be afraid!
- We have essentially all the tools already, from multiple linear regression and simple logistic regression

Still two forms of the model

(90% corr)

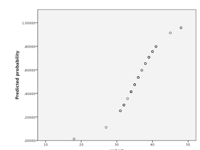
- Logit form

  $\log(\text{odds})=\beta_0 + \beta_1 X_1 + \beta_2 X_2 \ \dots + \beta_k X_k$



- Probability form

$$\pi = \frac{e\hat{}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 \ \dots + \beta_k X_k)}{1+e\hat{}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 \ \dots + \beta_k X_k)}$$



Familiar ways of assessing model

- Individual betas: Wald z-test
- Overall model test: Omnibus test which uses -2loglikelyhood.
- Compare models: -2loglikelyhood
- Same assumptions as simple logistic
  - Linearity
  - Independence:
    - No pairing or clustering of the data in space or time
  - Random:
    - Need to make sure a 'spinner model' is valid – super important!!
    - Want random sample from population or random assignment within an experiment

Medical school admissions example

- Previously we used MCAT scores to predict medical school acceptance
- We might imagine GPA has a significant effect on whether someone gets into medical school also.
- log(odds of acceptance)=$\beta_0 + \beta_1$MCAT + β_2GPA

Assumptions

- Use Box-Tidwell test on each individual predictor (52% corr)
- Do we have to scale either predictor to make it positive?

H_0: $\beta_2 = 0$

H_a: $\beta_2 \neq 0$

Since p is large (p=.278 and p=.942) for both predictors, we fail to reject the null hypothesis. This suggests that MCAT scores and GPA scores both have a linear relationship with the log(odds) of acceptance to medical school. This suggest we meet the assumption of linearity.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	GPA	-57.254	57.344	.997	1	.318	.000
	GPA by logGPA	27.885	25.683	1.179	1	.278	1.28E+12
	Constant	77.811	87.972	.782	1	.376	6.208E+33

a. Variable(s) entered on step 1: GPA, GPA * logGPA .

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 .

Assumptions: the hard ones!

- Randomness?
  - Is data randomly selected or the result of random assignment?
  - Is the spinner model accurate?
  - While med school students are not a random selection from all college students, is there a reason to think our sample is a biased selection from the population of interest – med school students?
    - Sample from representative school? Year?
  - Is there any reason to think GPA or MCAT scores would be misleading predictors?
    - MCAT a standardized test
    - GPA somewhat dependent on classes and school, but med students take mostly the same classes
  - Any hidden variables that would be causing bias?

Assumptions: the hard ones!

- Independence?
  - Any relation in time and space?
  - Is the yes/no decision subjective?
  - 55 medical school applicants from a liberal arts college in the midwest.
  - A spatial relationship, but since applying to many med schools all over, one students acceptance probably doesn't affect another's.

Now we can assess the model

- Make sure to use original variables if you had to scale any to run Box-Tidwell
- Overall model?

H_0: $\beta_1 = \beta_2 = 0$

H_a: at least one beta is not equal to zero

Since p<.001, we can reject the null hypothesis. This suggests that MCAT and GPA scores together are useful in predicting whether a student gets accepted to medical school or not.

The model correctly predicts 74.5% of the cases, doing somewhat better on predicting who gets accepted (76.7%) than who gets rejected (72.0%).

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	21.777	2	.000
	Block	21.777	2	.000
	Model	21.777	2	.000

Classification Table[a]

	Observed		Predicted: Acceptance 0	Predicted: Acceptance 1	Percentage Correct
Step 1	Acceptance	0	18	7	72.0
		1	7	23	76.7
	Overall Percentage				74.5

a. The cut value is .500

Assess the betas

H_0: $\beta_2 = 0$

H_a: $\beta_2 \neq 0$

Since p<.01, we reject the null hypothesis. This suggests that there is a significant log-linear relationship between GPA scores and the odds of whether someone gets into medical school, after accounting for their MCAT score.

H_0: $\beta_1 = 0$

H_a: $\beta_1 \neq 0$

Since p>.05, we fail to reject the null hypothesis. This suggests that there is not a significant log-linear relationship between MCAT scores and the odds of whether someone gets into medical school, after accounting for their GPA score.

- What is the fitted equation?
  - Log(odds^)=−22.373 +.165MCAT +4.676GPA
- What does the constant represent?
  - The log odds of 'success' when MCAT=0, GPA=0

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	MCAT	.165	.103	2.543	1	.111	1.179
	GPA	4.676	1.642	8.115	1	.004	107.389
	Constant	-22.373	6.454	12.017	1	.001	.000

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

Would adding a variable for gender improve the model?

- Answer this question with the equivalent of the nested F-test
  - Was calling it drop-in-deviance test: to test model versus model with just constant (essentially the omnibus test)
  - More general form is the nested likelihood ratio test (LRT)
  - 75% corr

H_0: nested model (log(odds)=$\beta_0 + \beta_1$MCAT + β_2GPA)

H_a: full model (log(odds)=$\beta_0 + \beta_1$MCAT + β_2GPA + β_3Male)

- But we will write as a test of the betas that differ

H_0: β_{Male}=0

H_a: not all betas are zero

G-statistic is still difference in -2loglikellihood, with degrees of freedom equal to the number of betas tested (82% corr)

Compare nested models

- Categorical variable – don't need to check for linearity.
- Full model:

log(odds)=$\beta_0 + \beta_1$MCAT + β_2GPA + β_3Male

- Nested model:

log(odds)=$\beta_0 + \beta_1$MCAT + β_2GPA

H_0: $\beta_3 = 0$

H_a: $\beta_3 \neq 0$

Test statistic: 54.014 – 50.786 = 3.228

DF = 1

P-value: .0723

https://www.fourmilab.ch/rpkp/experiments/analysis/chiCalc.html

Calculate probability from X^2 and df

Full Model

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	50.786[a]	.365	.488

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

Nested Model

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	54.014[a]	.327	.437

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

Another way to answer this question?

- Test the beta of Male
- pvalue =.085 compared to LRT where pvalue = .0723.
- Even though same hypothesis test (81% corr)

H_0: $\beta_3 = 0$

H_a: $\beta_3 \neq 0$

- Go with the LRT test. Though here they both suggest gender is not a significant predictor after accounting for MCAT and GPA scores.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	MCAT	.181	.108	2.805	1	.094	1.198
	GPA	5.139	1.851	7.710	1	.005	170.585
	Male	-1.258	.730	2.967	1	.085	.284
	Constant	-23.985	6.969	11.846	1	.001	.000

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

Using and interpreting model

- Interpret the beta for GPA:
- For a one unit change in GPA, the odds of getting in to medical school increase by a factor of 170.6, for a given value of MCAT score and gender.
- Interpret the beta for Male:
- If you are a male, you are .284 times less likely to get into medical school than a female, for a given value of MCAT and GPA.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	MCAT	.181	.108	2.805	1	.094	1.198
	GPA	5.139	1.851	7.710	1	.005	170.585
	Male	-1.258	.730	2.967	1	.085	.284
	Constant	-23.985	6.969	11.846	1	.001	.000

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

Using and interpreting model

- What are your odds of getting into medical school as a male, with a GPA of 3.5 and a MCAT score of 44
- Log(odds^)=-23.99+.18MCAT +5.14GPA -1.26Male
- Log(odds^)=-23.99 +.18(44) +5.14(3.5) -1.26
- Log(odds^)=.66
- Odds^ = $e^{.66}$=1.93
- What is the probability of getting into medical school as a male with a GPA of 3.5 and a MCAT score of 44?
- $\hat{\pi} = \dfrac{e^{\wedge} 23.99+.18MCAT +5.14GPA -1.26Male}{1+e^{\wedge}23.99+.18MCAT +5.14GPA -1.26Male}$
- Just figured out what e^24.99+.18MCAT +5.14GPA -1.26Male is: 1.93
- π^ = 1.93/1+1.93 = .659 = 65.9% chance

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	MCAT	.181	.108	2.805	1	.094	1.198
	GPA	5.139	1.851	7.710	1	.005	170.585
	Male	-1.258	.730	2.967	1	.085	.284
	Constant	-23.985	6.969	11.846	1	.001	.000

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

Prediction and confidence intervals?

- They are possible, available code for R, but still an area of research.

Another example

- The National Snow and Ice Data Center collects data on the ice melt ponds in the Canadian artic. Environmental engineers at the University of Colorado are using these data to study how climate impacts he sea ice. Data for 504 randomly selected ice melt ponds from Canada were collected. One thing they are interested in is what type of ice forms in the long-term, which can be either multiyear or landfast ice. The pond characteristics that are hypothesized to impact ice type are depth, broadband surface albedo, and visible surface albedo. They used a logistic model to investigate this, where y =1 if landfast ice, 0 if multiyear.

Assumptions

- How many test for linearity?
  - 3, since all three variables are quantitative.
- Do we need to scale? Look closely at the data or look at descriptive statistics (Analyze -> descriptive statistics -> descriptive)
- No negatives, but have 0 for depth, so that needs to be scaled

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
depth	504	.00	.86	.2642	.16251
broadband-alb	504	.03	.75	.2103	.09802
visible-alb	504	.05	.77	.3602	.13561
Valid N (listwise)	504				

Randomness and independence

- Random selection of ponds
- All in similar region of Canada, but that is the population we are interested in
- No reason to think decision about whether landfast ice or not is biased
- Should be all good.

Assess overall model

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$

H_a: at least one beta is not equal to zero

Since $p<.001$, we can reject the null hypothesis. This suggests that depth, broadbandalb and visablealb together are useful in predicting the odds of whether pond ice is landfast or not.

The model correctly predicts 70.7% of the cases, and does a little better at predicting when ice is not landfast (79.1% vs 61.2%).

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	70.456	3	.000
	Block	70.456	3	.000
	Model	70.456	3	.000

Classification Table[a]

Observed			Predicted		
			landfast		Percentage Correct
			.00	1.00	
Step 1	landfast	.00	174	46	79.1
		1.00	76	120	61.2
	Overall Percentage				70.7

a. The cut value is .500

Assess the individual betas

H_0: $\beta_i = 0$
H_a: $\beta_i \neq 0$
Since p<.001, we reject the null hypothesis for depth, broadbandalb, and visiblealb . This suggests that there is a significant log-linear relationship between each predictor and the odds of whether pond ice is landfast or not, after accounting for the other predictors.

- Why is the Exp(B) so large and small?

Think back to the descriptive statistics, max values were under 1, so when we say a 1 unit increase in X, this isn't all that meaningful of a thing to say, and can lead to very extreme values for Exp(B).

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
	depth	4.128	.809	26.021	1	.000	62.065
1[a]	broadbandalb	47.123	7.477	39.724	1	.000	2.919E+20
	visiblealb	-31.144	4.730	43.356	1	.000	.000
	Constant	.296	.434	.466	1	.495	1.345

iable(s) entered on step 1: depth, broadbandalb, visiblealb.

Assess the individual betas

Depth: For each additional one unit increase in ice depth, the odds of being landfast increase by a factor of 62, for a given value of broadband and visible albedo.

Broadband: For each additional one unit increase in boradband surface albedo, the ice is 2.919×10^{20} times as likely to be landfast, after accounting for depth and visible albedo.

Visible: For each additional one unit increase in visible surface albedo, the odds of being landfast decrease by a factor of 2.982×10^{-14}, for a given value of broadband albedo and depth.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
depth	504	.00	.86	.2642	.16251
broadband-alb	504	.03	.75	.2103	.09802
visible-alb	504	.05	.77	.3602	.13561
Valid N (listwise)	504				

		B	S.E.	Wald	df	Sig.	Exp(B)
	depth	4.128	.809	26.021	1	.000	62.065
1[a]	broadbandalb	47.123	7.477	39.724	1	.000	2.919E+20
	visiblealb	-31.144	4.730	43.356	1	.000	.000
	Constant	.296	.434	.466	1	.495	1.345

iable(s) entered on step 1: depth, broadbandalb, visiblealb.

To make more reasonable:

- Scale the variables, so a one unit increase is meaningful.
- Multiple each variable by 100

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	70.456	3	.000
	Block	70.456	3	.000
	Model	70.456	3	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	504.857[a]	.156	.208

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

Classification Table[a]

			Predicted		
			landfast		Percentage Correct
Observed			.00	1.00	
Step 1	landfast	.00	174	46	79.1
		1.00	76	120	61.2
	Overall Percentage				70.7

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	depth100	.041	.008	26.021	1	.000	1.042
	bradband100	.471	.075	39.724	1	.000	1.602
	visible100	-.311	.047	43.356	1	.000	.732
	Constant	.296	.434	.466	1	.495	1.345

a. Variable(s) entered on step 1: depth100, bradband100, visible100.

Original model

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	70.456	3	.000
	Block	70.456	3	.000
	Model	70.456	3	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	504.857[a]	.156	.208

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

Classification Table[a]

			Predicted		
			landfast		Percentage Correct
Observed			.00	1.00	
Step 1	landfast	.00	174	46	79.1
		1.00	76	120	61.2
	Overall Percentage				70.7

a. The cut value is .500

Other terms?

- The researchers have reason to believe that the predictors variables might interact with each other. For instance, the depth of ice might have a different effect on whether ice is landfast or not for high values of visiblealb than for low. If they want to test all interactions, what model do they want to test?

$\text{Log(odds)} = \beta_0 + \beta_1\text{Depth} + \beta_2\text{Broadbandalb} + \beta_3\text{Visiblealb} + \beta_4\text{DepthBroadbandalb} + \beta_5\text{DepthVisiblealb} + \beta_6\text{BroadbandalbVisiblealb}$

- What would be a good way to test whether these higher order terms are needed as a group?
- The nested likelihood ratio test!

Likelihood ratio test

- Full model:

$\text{Log(odds)} = \beta_0 + \beta_1\text{Depth} + \beta_2\text{Broadbandalb} + \beta_3\text{Visiblealb} + \beta_4\text{DepthBroadbandalb} + \beta_5\text{DepthVisiblealb} + \beta_6\text{BroadbandalbVisiblealb}$

- Reduced model:

$\text{Log(odds)} = \beta_0 + \beta_1\text{Depth} + \beta_2\text{Broadbandalb} + \beta_3\text{Visiblealb}$

- Hypothesis want to test:

H_0: $\beta_4 = \beta_5 = \beta_6 = 0$

H_a: at least one beta is not zero

Likelihood ratio test

H_0: $\beta_4 = \beta_5 = \beta_6 = 0$
H_a: at least one beta is not zero
test-statistic, G-statistic :
504.857-472.664=32.193
Df:
3
Pvalue:
p<.001, reject the null.
Suggests that the interaction terms together are useful in predicting the odds of whether pond ice is landfast or not, after accounting for the first order effects.

Reduced model

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	504.857[a]	.156	.208

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

Full model

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	472.664[a]	.219	.292

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

Agree with betas?

- Seems to! One of the interaction terms is significant (H_0: β_6=0, H_a: $\beta_6 \neq 0$, p<.001), which is consistent

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	depth100	-.030	.030	.991	1	.319	.970
	bradband100	.106	.125	.712	1	.399	1.111
	visible100	-.397	.077	26.390	1	.000	.672
	bradband100 by depth100	.005	.004	1.629	1	.202	1.005
	depth100 by visible100	-.001	.002	.065	1	.798	.999
	bradband100 by visible100	.006	.001	21.763	1	.000	1.006
	Constant	6.097	1.444	17.830	1	.000	444.411

a. Variable(s) entered on step 1: depth100, bradband100, visible100, bradband100 * depth100 , depth100 * visible100 , bradband100 * visible100 .

Agree with betas?

- What is the fitted equation?
- Log(odds^)=6.097 -.03Depth100 +.106 Broadband100 -.397Visible100 +.005DepthBroadband100 -.001DepthVisible100 +.006BroadbandVisible100
- Interpreting betas with interaction terms complicated just like in regular regression. Now slope for lower order terms only relevant when other variables are zero. So ice is .97 times as likely to be landfast for each additional unit increase in depth, only when Broadband and Visible are zero.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	depth100	-.030	.030	.991	1	.319	.970
	bradband100	.106	.125	.712	1	.399	1.111
	visible100	-.397	.077	26.390	1	.000	.672
	bradband100 by depth100	.005	.004	1.629	1	.202	1.005
	depth100 by visible100	-.001	.002	.065	1	.798	.999
	bradband100 by visible100	.006	.001	21.763	1	.000	1.006
	Constant	6.097	1.444	17.830	1	.000	444.411

a. Variable(s) entered on step 1: depth100, bradband100, visible100, bradband100 * depth100 , depth100 * visible100 , bradband100 * visible100 .

Zero often makes no sense

- Idea of centering
- Change what is the zero point for a dataset.
- Lets go back to MCAT example
- To center MCAT, subtract mean (36.27) from every value of MCAT
- To center GPA, subtract mean (3.5533) from every value of GPA

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
MCAT	55	18	48	36.27	4.817
GPA	55	2.72	3.97	3.5533	.28645
Valid N (listwise)	55				

Centered model

- Log(odds^)=−22.373 + .165MCAT + 4.676GPA
- Centered: Log(odds^)=.211 + .165MCAT + 4.676GPA
- In centered model, interpret constant:
  - The odds of success are 1.234 times greater than failure, if you have an average MCAT and GPA score.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	MCAT	.165	.103	2.543	1	.111	1.179
	GPA	4.676	1.642	8.115	1	.004	107.389
	Constant	-22.373	6.454	12.017	1	.001	.000

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

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Cmcat	.165	.103	2.543	1	.111	1.179
	Cgpa	4.676	1.642	8.115	1	.004	107.389
	Constant	.211	.335	.394	1	.530	1.234

a. Variable(s) entered on step 1: Cmcat, Cgpa.

Centering useful

- Particularly when have an interaction, or have categorical predictors
- Interaction: Log(odds^)=.187 +.185Cmcat +4.804Cgpa +.315CgpaCmcat
  - For each additional point on the mcat, you are 1.203 times as likely to get into medical school, when you have an average GPA.
  - Can say that because 'average GPA' means you set Cgpa to 0.
- Categorical (sex =0 for male): Log(odds^)=-.421 +.181Cmcat + 5.139Cgpa + 1.258Sex
  - Odds of getting into medical school for men is .656, when Cmcat and Cgpa =0, or when someone has an average mcat and gpa.
  - The odds for women with an average GPA and MCAT score is $e^{-.421+1.258}$ = 2.31
  - Just to verify, if take ratio of odds 2.31/.656 = 3.5 -> Odds ratio for Sex!!

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Cmcat	.185	.110	2.807	1	.094	1.203
	Cgpa	4.804	1.667	8.308	1	.004	122.004
	Cgpa by Cmcat	.315	.286	1.213	1	.271	1.371
	Constant	.187	.344	.295	1	.587	1.206

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Cmcat	.181	.108	2.805	1	.094	1.198
	Cgpa	5.139	1.851	7.710	1	.005	170.585
	Sex(1)	1.258	.730	2.967	1	.085	3.518
	Constant	-.421	.499	.711	1	.399	.656

a. Variable(s) entered on step 1: Cmcat, Cgpa, Sex.