Stat E-150 Section #12

Neha Dhawan

dhawan@g.harvard.edu

Office hours by request for one-one help

Multiple Logistic Regression

Model:

$$\text{Log(odds) or } \log(\pi/1-\pi) = \beta_0 + \beta_1 X + \beta_2 X$$

- We say log(odds) but its actually the ln(odds)

Probability form of the model:

$$\hat{\pi} = \frac{e^{(\beta_0 + \beta_1 X + \beta_2 X)}}{1 + e^{(\beta_0 + \beta_1 X + \beta_2 X)}}$$

Example

- Suppose interested in testing birthweight and number of drinks mom had while pregnant help us predict odds survival of infants. They took the data from a small town hospital in MA which had 50 cases the past year.

Model:

$$\text{Log(odds surviving)} = \beta_0 + \beta_1 \text{Birthweight} + \beta_2 \text{Drinks}$$

$$\text{Probability of surviving } \hat{\pi} = \frac{e^{\beta 0 + \beta 1 \text{Birthweight} + \beta 2 \text{Drinks}}}{1+e^{\beta 0 + \beta 1 \text{Birthweight} + \beta 2 \text{Drinks}}}$$

Assumptions

- Linearity:
  – Box-Tidwell test (for each individual predictor)
- Randomness:
  – Random selection or random assignment?
- Independence:
  – No pairing or clustering of the data in space or time (no time/space order)

No e so don't have to worry about normality or equal variance

Linearity

- Box-Tidwell test for each predictor

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Birthweight	-.100	.050	4.060	1	.044	.905
	xbylogx	.012	.006	4.094	1	.043	1.012
	Constant	20.907	10.995	3.616	1	.057	1201354381

a. Variable(s) entered on step 1: Birthweight, xbylogx.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Drinks_1	33.842	49778.796	.000	1	.999	4.983E+14
	xbylogx	-21.916	21711.177	.000	1	.999	.000
	Constant	-11.743	76457.433	.000	1	1.000	.000

a. Variable(s) entered on step 1: Drinks_1, xbylogx.

Linearity

Birthweight:

H_0: β_2=0

H_a: $\beta_2 \neq 0$

Since the $p<0.05$, we can reject our null hypothesis. This suggests that we do not meet the assumption of linearity.

Drinks:

H_0: β_2=0

H_a: $\beta_2 \neq 0$

Since the $p>0.05$, we fail to reject our null hypothesis. This suggests that we meet the assumption of linearity.

Other assumptions

- Randomness
  - Random selection or random assignment?
  - Spinner model accurate?
  - Is there a bias? Representative sample?
- Independence
  - Is there a time-ordered relationship?
  - Is there a spatial relationship?
  - Yes/no decision?

Assessing the model

Omnibus test

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	30.498	3	.000
	Block	30.498	3	.000
	Model	30.498	3	.000

H_0: $\beta_1 = \beta_2 = 0$

H_a: at least one beta is not 0

Since the $p<0.05$, we can reject our null hypothesis. This suggests that the baby birthweight and number of drinks together are useful in predicting whether a baby will survive or not

Assessing the model

Classification table

Classification Table[a]

Observed			Predicted: Survive 0	Predicted: Survive 1	Predicted: Percentage Correct
Step 1	Survive	0	11	0	100.0
		1	0	11	100.0
	Overall Percentage				100.0

a. The cut value is .500

The model predicts 100% of the cases.

Testing our Betas

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Drinks	.165	.103	2.543	1	.111	0.179
	Birthweight	4.676	1.642	8.115	1	.004	1.07
	Constant	-22.373	6.454	12.017	1	.001	.000

H_0: $\beta_1 = 0$

H_a: $\beta_1 \neq 0$

Since the $p > 0.05$, we fail to reject our null hypothesis. This suggests that number of drinks doesn't significantly predict whether the baby will survive or not, **after accounting** for birthweight

Testing our Betas

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Drinks	.165	.103	2.543	1	.111	0.179
	Birthweight	4.676	1.642	8.115	1	.004	1.07
	Constant	-22.373	6.454	12.017	1	.001	.000

H_0: $\beta_2 = 0$

H_a: $\beta_2 \neq 0$

Since the $p<0.05$, we can reject our null hypothesis. This suggests that there is a significant log-linear relationship between birthweight and whether the baby will survive or not, **after accounting** for number of drinks

Interpretation:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Drinks	.165	.103	2.543	1	.111	0.179
	Birthweight	4.676	1.642	8.115	1	.004	1.07
	Constant	-22.373	6.454	12.017	1	.001	.000

Drinks:

For each additional drink, the odds of surviving decrease by a factor of 0.179, **after accounting for** birthweight

Birthweight:

For each additional gram in birthweight, the odds of surviving increase by a factor of 1.07 **after accounting for** number of drinks

Maybe there is an interaction?

- Maybe adding another variable will help the model?
- Nested likelihood ratio test!
- Drop in deviance test
- Full model vs. nested model

Nested Model: $\text{Log(odds surviving)} = \beta_0 + \beta_1\text{Birthweight} + \beta_2\text{Drinks}$

Full model(with added term): $\text{Log(odds surviving)} = \beta_0 + \beta_1\text{Birthweight} + \beta_2\text{Drinks} + \beta_3\text{Drinks*Birthweight}$

H_0: β_i =0 for all predictors in subset

H_a: β_i =/0 for at least one predictor in subset

Vs.

H_0: $\beta_{Drinks*Birthweight}$=0 **[beta(s) different between the full and nested model**

H_a: Not all betas are zero

Test statistic= -2log(nested) – -2log(full)

DF= number of betas in full model – number of betas in nested model [i.e. #betas being tested]

Comparing models

- -2 log Likelihood is a measure of how well the data fit the model
  - Unexplained variability
  - Want a smaller number

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	28.120[a]	.102	.137

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Nested

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	10.000[a]	7.50	1.500

a. Estimation terminated at iteration number 20 because maximum iterations has been reached.

Full

H_0: $\beta_3=0$

H_a: $\beta_3 \neq 0$

Test statistic= 28.120- 10.0= 18.120

DF= 1

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

p-value = 0.0001

H_0: β_3=0

H_a: $\beta_3 \neq 0$

p-value = 0.0001

Since $p<0.05$, we can reject the null and conclude that the interaction term significant predicts the odds of a baby surviving **after accounting for** drinks and birthweight.

Individual beta test?

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Birthweight	-8.657	63.857	.018	1	.000	.892
	Drinks	-9598.896	70962.139	.018	1	.892	.921
	int	-.448	4.486	.010	1	.000	.639
	Constant	30949.391	228014.825	.018	1	.892	.

a. Variable(s) entered on step 1: Birthweight, Drinks, int.

H_0: $\beta_3=0$

H_a: $\beta_3 \neq 0$

Always go with LRT if two don't agree

Since the $p<0.05$, we can reject our null hypothesis. This suggests that the interaction term significantly predicts whether the baby will survive or not, **after accounting** for birthweight and number of drinks

Centering

- Instead of starting our axis at 0 we 'center' it to the mean value

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Birthweight	22	1122	3573	2317.73	798.835
Drinks	22	0	4	1.18	1.402
Valid N (listwise)	22				

- Subtract mean from every value for each predictor to center it

Centering

- Useful when we have interactions and categorical predictors

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	cBirthweight	-8.657	63.857	.018	1	.000	. 63
	cDrinks	-9598.896	70962.139	.018	1	.892	. 21
	cint	-.448	4.486	.010	1	.000	2.13
	Constant	30949.391	228014.825	.018	1	.892	.448

a. Variable(s) entered on step 1: Birthweight, Drinks, int.

- Interpret the constant
  – For an average birthweight baby from a mom who drinks an average number of drinks, the odds of survival decrease by a factor of .448

Don’t forget about Course Evals!