36-309/749
Experimental Design for Behavioral and Social Sciences

Nov 17, 2015

Lecture 10: Categorical Outcomes

Review

An analysis choice grid:

	Quantitative DV	Categorical DV
Quantitative IV	Regression (HLM)	Logistic regression (k=2) (generalized HLM)
One Categorical IV	ANOVA R.M. ANOVA	Chi-square test of independence
Both	Regression (ANCOVA)	Logistic Regression (k=2)
	R.M. ANCOVA (HLM)	(generalized HLM)

Tests for categorical outcomes (DVs)

- No normality or equal variance assumption. Still assume independent errors assumption.
- ***Chi-square test of independence***
  - Two categorical variables with any number of levels
  - Null hypothesis is equal probability distributions of one factor across levels of another factor, i.e., knowing X tells you nothing about Y.
- ***Logistic regression***
  - Binary categorical DV (i.e., k=2) with any IV(s).
  - Model how the probability of "success" varies with some categorical and/or continuous explanatory variable(s).
- Not covered by the above: DV has >2 levels and >1 IV or a quantitative IV.

Chi-square Test of Independence

- Setting: Categorical IV and categorical DVs (or two categorical DVs)

- Null hypothesis
  - X is independent of Y **or**
  - "distribution of Y|X=a" equals the "distribution of Y|X=b"= … **or**
  - Pr(Y=1) is equal across all levels of X (only when Y has k=2 levels)

- Examples of truth in the population:
  - Null: P(Y=1)=0.4, P(Y=2)=0.3, P(Y=3)=0.3 for **each** level of x
  - Specific alternative: P(Y=1|x=1)=0.4, P(Y=2|x=1)=0.3, P(Y=3|x=1)=0.3
    P(Y=1|x=2)=0.3, P(Y=2|x=2)=0.3, P(Y=3|x=2)=0.4
    etc.

Chi-square Test, continued

- EDA: Construct a sample ***contingency table*** (cross-tabulation) which counts the numbers of subjects for each combination of levels of variable x and variable Y. Example (100 people with 20 given each of 5 drugs):

Memory\Drug	A	B	C	D	E	All
Improved	3	5	8	10	14	40
Not	17	15	12	10	6	60

  - Percentages may be calculated by column or by row.

Chi-square Test, continued

Memory\Drug	A	B	C	D	E	All
Improved	3	5	8	10	14	40
Not	17	15	12	10	6	60

- **<u>Expected value</u>** ***under the null hypothesis***: Here, overall 40/100 improved, 40% of 20 is 8.0, so the expected cell count is 8.0 for each dose for "improved" and 12.0 for "not".

- Statistic is $X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$

Chi-square Test, continued

- Under H_0, the X^2 statistic follows an asymptotic chi-square (χ^2) distribution with (r-1)*(c-1) d.f. (assume independence across subjects; incorrect for tiny expected cell counts)
  - "Asymptotic" means "for large N". Many programs warn of possible incorrect p-values when any cell has its "expected value" (not observed value) less than 5. This warning is fairly conservative, so a few 4's, 3's, or 2's is unlikely to have much negative affect.

- SPSS (Analyze/Descriptive/Crosstab with "Statistics" set to "ChiSquare"):

Dose * Success Crosstabulation					
			Success		
			0	1	Total
Drug	A	Count	17	3	20
		Expected Count	12.0	8.0	20.0
		% within Drug	85.0%	15.0%	100.0%
	B	Count	15	5	20
		Expected Count	12.0	8.0	20.0
		% within Drug	75.0%	25.0%	100.0%
	C	Count	12	8	20
		Expected Count	12.0	8.0	20.0
		% within Drug	60.0%	40.0%	100.0%
	D	Count	10	10	20
		Expected Count	12.0	8.0	20.0
		% within Drug	50.0%	50.0%	100.0%
	E	Count	6	14	20
		Expected Count	12.0	8.0	20.0
		% within Drug	30.0%	70.0%	100.0%
Total		Count	60	40	100
		Expected Count	60.0	40.0	100.0
		% within Drug	60.0%	40.0%	100.0%

Chi-Square Test, cont.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	15.417[a]	4	.004
Likelihood Ratio	16.120	4	.003
Linear-by-Linear Association	15.036	1	.000
N of Valid Cases	100		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 8.00.

- "Pearson" is as good as the others so (pre-)choose it.
- "Contrasts": Sub-tables analyzed by χ^2 with Bonferroni corrections for post-hoc testing.

Binary Outcome: Logistic Regression

- Purpose: With the ***two*** outcome categories labeled as "success" and "failure", model how the chance of success depends on the explanatory variable(s).

- Model "**log odds of success**":

  $\log[\Pr(S)/\Pr(F)] = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$

  (As usual, factors with k>2 need to be coded.)

  - The change in the chance of success due to "x going up by 1" can be expressed as an addition on the log odds scale, a multiplication on the odds scale, but in no general way on the probability scale.

Mathematics of Logistic Regression

- Define: S = Success F = Failure
- Probability of success: P(S) or Pr(S)
  - $0 \le P(S) \le 1$ $P(F) = 1 - P(S)$
  - 0.5 is the "middle"
  - Cannot keep adding, so $P(S)=\beta_0+\beta_1 x$ fails
- Odds: $Odds(S) = P(S) / P(F) = P(S) / [1 - P(S)]$
  - $0 \le Odds(S) < \infty$
  - 1 is the "middle"
  - Interpretation of odds values:
    - odds=3=3/1 means 3 succeed for every 1 who fails
    - odds=0.21 ≈ 2/10 = 1/5 means 1 succeeds for every 5 who fail
  - Can keep multiplying, so
    $Odds(S|X=x+1) = \gamma\, Odds(S|X=x)$ does work

Logistic Regression Math, cont.

- Log odds: Logit(S) = ln(Odds(S)) = $\log_e$(Odds(S))
  - $-\infty \leq \text{logit}(S) \leq \infty$
  - Zero is the "middle"
  - Can keep adding, so logit(S|x)= $\beta_0 + \beta_x x$ is OK.

- Needed formulas
  - Odds(S) = $e^{\text{logit}(S)}$ = exp(logit(S))
    - Exp on many calculators is actually "*10^"
    - Use "Inverse" of "ln" on most calculators
  - P(S) = Odd(S) / (1 + Odds(S))

Logistic Regression Math, cont.

- E.g. p = 0.2, 0.5, 0.75; odds = 0.25, 1, 3;
  log(odds) = -1.30, 0, 1.10

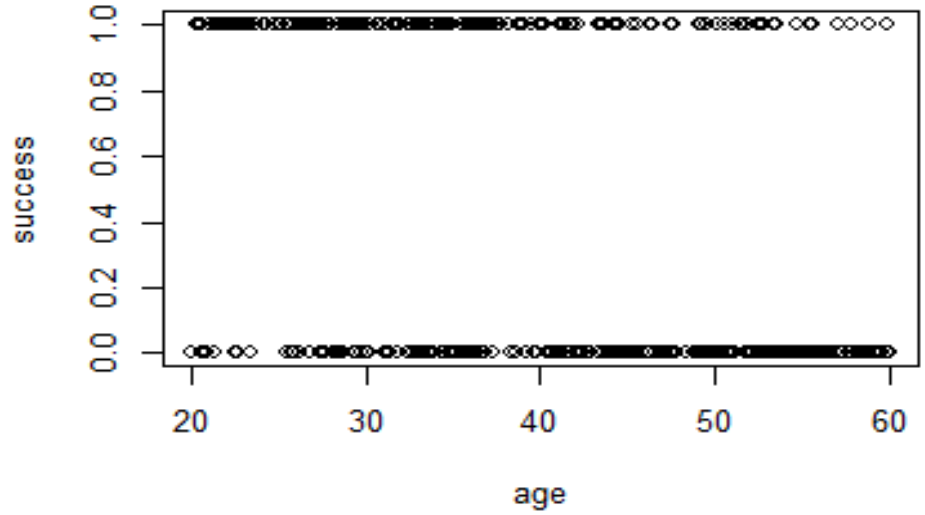
- Adding, e.g., 1.10 to log odds(S) is the same as multiplying odds(S) by exp(1.10)=3.00.

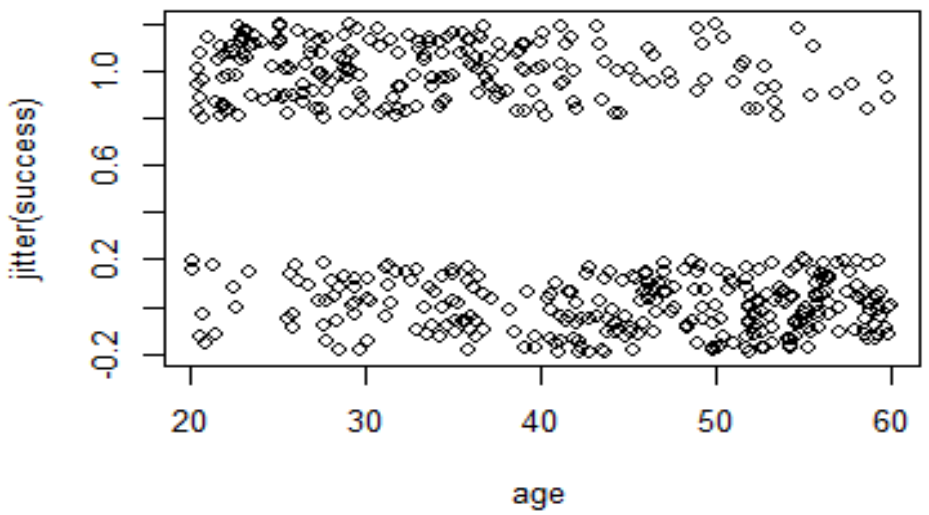
- Deeper math: $e^{a+b} = e^a e^b$ $e^{cd} = (e^c)^d$
  ($e^a + e^b$ does not simplify)
  - E.g., log odds(S) = $\beta_0 + \beta_1 x$ ➔
    odd(S) = exp($\beta_0 + \beta_1 x$) = $e^{\beta 0 + \beta 1 X}$
    = $e^{\beta 0}$ $e^{\beta 1 x}$ = $e^{\beta 0}$ $(e^{\beta 1})^x$

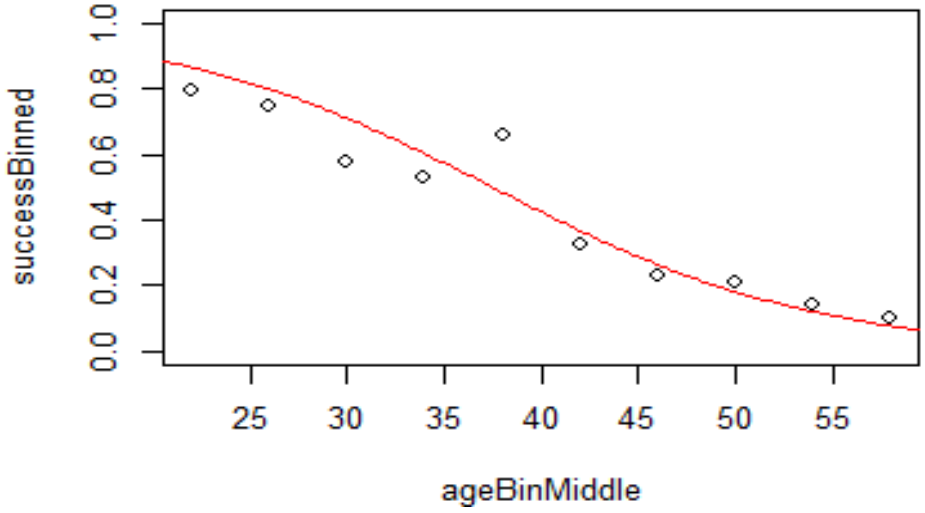
EDA

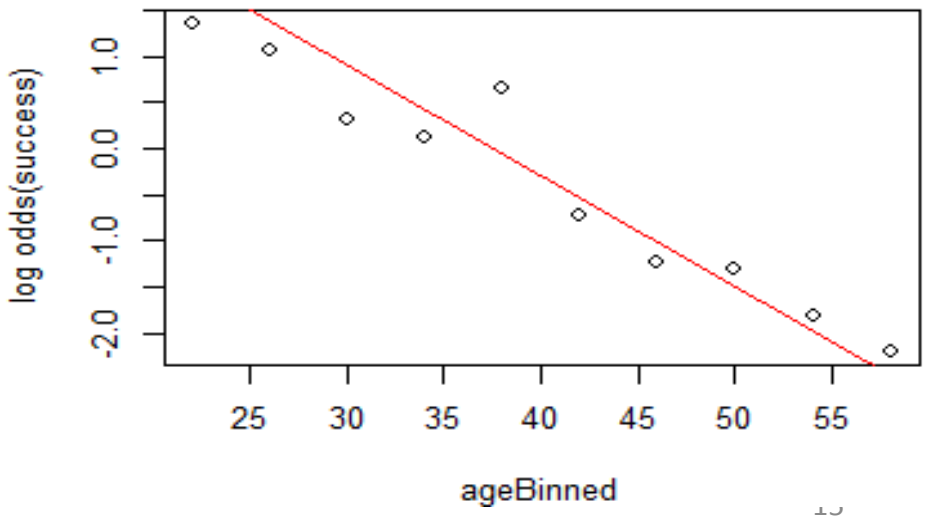
- EDA is problematic because there are only two values of the DV.
- For categorical IVs, a crosstabulation should be done. (Some plots of the crosstab are available in some programs, but are only slightly helpful.)
- For quantitative IVs, the best EDA is to “cut” the IV into bins, compute the fraction of success in each bin, and plot fraction(S) vs., e.g., the mean or middle IV value in the bin.
- This should be “S” shaped on a linear scale and a straight line on a log odds scale.

Logistic Regression EDA









Formal Logistic Regression Analysis

- "Regression / Binary Logistic" in SPSS
- Technically, this is a general***ized*** linear model with link function "logistic".
- Think of predicting log odds of success from the usual right hand side:
  $\eta_x = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$.
  - β_0 is the log odds of success for "baseline" subjects, i.e., when all x's are zero
  - $\exp(\beta_0)$ is the odds of success for baseline
  - $\exp(\beta_0) / (1 + \exp(\beta_0))$ is the baseline prob(S)
  - β_1 is the (<u>additive</u>) change in the <u>log odds</u> of success when x_1 goes up by one, holding all other x's constant
  - $\exp(\beta_1)$ is the <u>multiplicative</u> change in the <u>odds</u> of success when x_1 goes up by one, holding all other x's constant
  - Nothing in general can be said about the effect of a change in x on the <u>probability</u> scale.

Example: Donner Party

- Scientific hypothesis: women are better able to survive harsh conditions than men (children excluded from the analysis)

- Statistical model:
  - LogOdds(survival | age, gender) = $\beta_0 + \beta_{age}Age + \beta_{female}Female$
  - $H_{0G}: \beta_{female}=0$ $H_{0A}: \beta_{age}=0$

- Meaning of the parameters
  - β_{female} is the (additive) change in the log odds of success when comparing a female to a male of any age.
  - $\exp(\beta_{female})$ is the multiplicative change in the odds of success when comparing a female to a male of any age.

  - β_{age} is the (additive) change in the log odds of success when comparing a person to another of the same sex who is 1 year younger.
  - $\exp(\beta_{age})$ is the multiplicative change in the odds of success when comparing a person to another of the same sex who is 1 year younger.

Donner, cont.

Dependent Variable Encoding

Original Value	Internal Value
Died	0
Survived	1

Categorical Variables Codings

			Parameter coding
		Frequency	(1)
female	male	30	.000
	female	15	1.000

Classification Table(a)

	Observed		Predicted		
			survived		Percentage Correct
			died	survived	Died
Step 1	Survived	Died	23	2	92.0
		Survived	8	12	60.0
	Overall Percentage				77.8

Donner, cont.

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Age	-.078	.037	4.399	1	.036	.925
female(1)	1.597	.756	4.470	1	.034	4.940
Constant	1.633	1.110	2.164	1	.141	5.120

- β_0: (Useless) extrapolation to baseline
  - Baseline is newborn males
  - Estimated log odds of survival is 1.633
  - Estimated odds(S) is exp(1.633) = 5.12
  - Estimated p(S) is 5.12/(1+5.12) = 0.84 (84%)

Donner, cont.

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Age	-.078	.037	4.399	1	.036	.925
female(1)	1.597	.756	4.470	1	.034	4.940
Constant	1.633	1.110	2.164	1	.141	5.120

- β_{female}: Female "slope" parameter
  - Describes females compared to males at the same age
  - Estimated log odds of survival is 1.60 higher (adding) for females than males of the same age
  - Estimated odds of survival 4.94 times (multiplying) as high for females than males of the same age
    - E.g., odds of survival for 25 y/o males is exp(1.633-0.078(25)) = 0.73 (0.73 survive for every 1 who dies; or 3S for 4D)
    - Odds of survival for 25 y/o females is 0.73*4.94=3.61 (3.61 survive for 1 who dies; or 14S for 4D)

Donner, cont.

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Age	-.078	.037	4.399	1	.036	.925
female(1)	1.597	.756	4.470	1	.034	4.940
Constant	1.633	1.110	2.164	1	.141	5.120

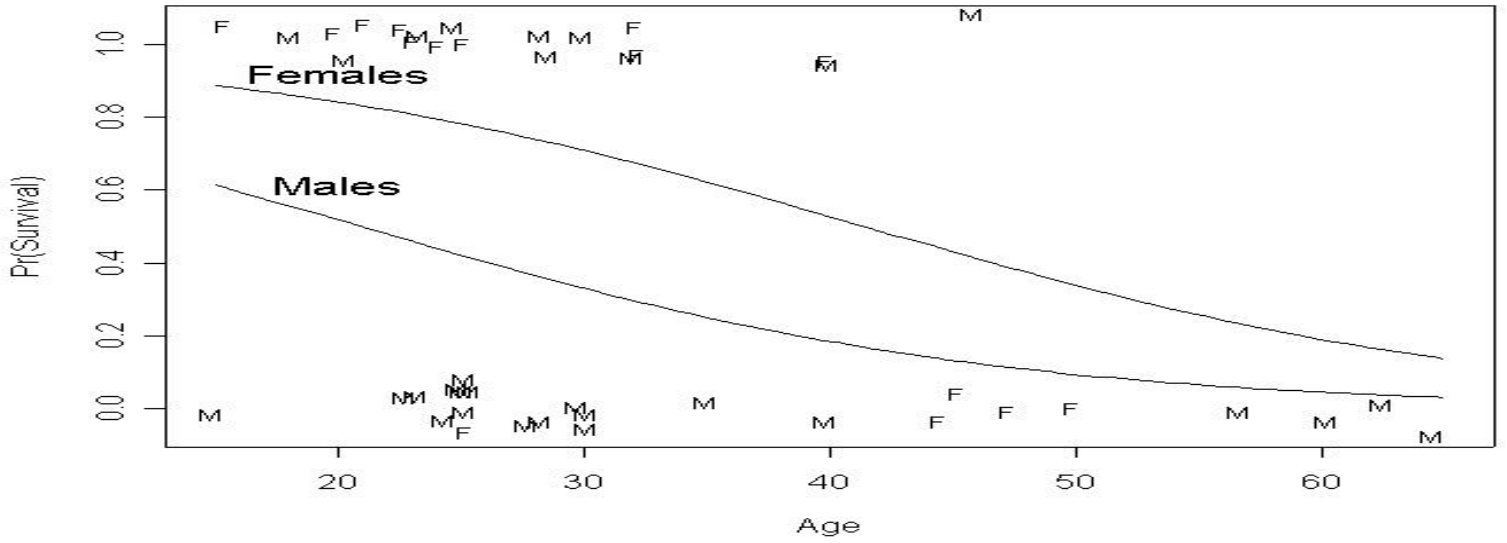
- β_{Age}: Age slope parameter
  - Describes any age compared to 1 year younger for the same sex
  - Estimated log odds of survival is 0.078 lower (subtracting) for any age than one year younger for the same sex
  - Estimated odds of survival 0.925 times (multiplying) as high (i.e., lower) for one year younger of the same sex
  - It is OK (better) to make "change in x" more meaningful:
    - 10 year increase in age lowers log odds of survival by 0.78.
    - 10 year increase in age multiplies odds of survival by $e^{0.78}$=0.46 times. (Or every 8.9 years older results in half the odds of survival.)

Model checking for Donner

- Hosmer-Lemeshow goodness of fit test: among other problems, it can detect ($p \leq 0.05$) the need for a transformation of IVs to corrected for non-linearity on the log odds scale.

Step	Chi-square	df	Sig.
1	9.320	7	.230

- Graphical summary of the model:



Logistic Regression Assumptions

- Binomial outcome
- Logistic relationship for p_x=Pr(Y=1|x) vs. each quantitative x
- Error variance(Y|x)= p_x *(1- p_x)
- X is “fixed” (measured with no or little error)
- Independent errors

Categorical DV Summary

- Chi-square test of independence
  - DV with *c* levels and IV (or second DV) with *r* levels
  - H_0: Pr(Y=i | X=j) = Pr(Y=i |X=j') for j,j' for each i (and vice versa)
- Logistic regression
  - Binary DV (one of the k=2 Y values called "success")
  - Log odds(S) = $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$
  - X's can include any factors, covariates and/or interactions
- DV with >2 levels and quantitative IV or more than one IV: ???