Statistics

March 10, 2010

Odds ratio and Logistic regressions

Jobayer Hossain Ph.D. & Tim Bunnell Ph.D.

Nemours Bioinformatics Core Facility

Class Objectives -- You will Learn:

- Odds and odds ratio of an event
- Logit and Logistic regression
- Multiple logistic regression
- Multinomial and ordinal logistic regressions
- Calculating odds ratio and modeling logistic regression using statistical package SPSS

Odds of an event

- The **odds** in favor of an event (e.g. occurrence of a disease) is the ratio of the probabilities of the event occurring to that of not occurring, i.e., $p/(1-p)$
  - where p is the probability of the event occurring
- If probability of an event is 0.6 (i.e., it is observed in 60% of the cases) then the probability of its not occurring is (1 - 0.6) = 0.4
- The odds in favor of the event occurring is thus 0.6/0.4 = 1.5
- The greater the odds of an event, the greater it's probability

Odds of an event

	Response Variable		
Predictor Variable	Cancer	No Cancer	Total
Smokers	a	b	a+b
Non- Smokers	c	d	c+d
Total	a+c	b+d	N=a+b+c+d

The proportion of smokers with cancer: $p = a/(a+b)$
(this is the *likelihood* of smokers having cancer)

The proportion of smokers without cancer: $1\text{-}p = b/(a+b)$
(this is the likelihood of smokers not having cancer)

The odds of smokers having cancer: $p / (1 - p) = (a/(a+b))/(b/(a+b)) = a/b$

Odds of an event

	Response Variable		
Predictor Variable	Cancer	No Cancer	Total
Smokers	a	b	a+b
Non-Smokers	c	d	c+d
Total	a+c	b+d	N=a+b+c+d

The proportion of non-smokers with cancer: $p = c/(c+d)$

The proportion of non-smokers without cancer: $1\text{-}p = d/(d+d)$

The odds of cancer among non-smokers: $(c/(c+d)) / (d/(c+d)) = c/d$

Odds ratio of an event

- The ***odds ratio*** of an event is the ratio of the odds of the event occurring in one group to the odds of it occurring in another group.
- Let p_1 be the probability of an event in group 1 and p_2 be the probability of the same event in group 2. Then the odds ratio (OR) of the event in these two groups is:

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

- The odds ratio compares the likelihood of an event between two groups using relative odds of that event (e.g. disease occurrence) in two groups
- The odds ratio is a measure of effect size

Odds ratio of an event

- In the previous example,

	Cancer	No Cancer	Column Total
Smokers	a	b	a+b
Non- Smokers	c	d	c+d
Row Total	a+c	b+d	N=a+b+c+d

The odds of cancer among smokers is a/b

The odds of cancer among non-smokers is c/d

So, the odds ratio of cancer among smokers vs non-smokers $= \dfrac{\frac{a}{b}}{\frac{c}{d}} = \dfrac{ad}{bc}$

Odds ratio of an event

Estimating odds ratio in a 2x2 table

	Cancer	No Cancer	
Smokers	120	80	200
Non- Smokers	60	140	200
	180	220	N=400

The odds of cancer in smokers is 120/80 = 1.5 and the odds of cancer in non-smokers is 60/140 = 0.4286 and the ratio of two odds is,

$$OR = \frac{120/80}{60/140} = \frac{120\text{x}140}{60\text{x}80} = 3.5$$

Interpretation: There is a 3.5 fold greater odds of cancer for smokers than for non-smokers (in this sample)

Odds ratio of an event

- The odds ratio must be greater than or equal to zero.
- As the odds of the first group approaches to zero, the odds ratio approaches to zero.
- As the odds of the second group approaches to zero, the odds ratio approaches to positive infinity
- An odds ratio of 1 indicates that the condition or event under study is equally likely in both groups. In our example, that would mean no association between cancer and smoking was observed.

Odds ratio of an event

- An odds ratio greater than 1 indicates that the condition or event is more likely in the first group. In the previous example, an odds ratio of 2 means that odds of cancer is 2 times more likely in smokers compared to non-smokers
- An odds ratio less than 1 indicates that the condition or event is less likely in the first group. In our example, an odds ratio of 0.8 would mean that the odds of cancer is 20% (i.e., 1 - 0.8) less likely in smokers compared to non-smokers

Odds ratio of an event : SPSS demonstration

- Analyze <- Descriptive statistics ->Crosstabs -> enter response variable in column and other group variable in row -> select statistics - > risk then click ok

Logit

- The logit of a number *P* between 0 and 1 is log(*P*/*1*-*P*). It is defined by logit(*P*)
- If *P* is the probability of an event, then *P*/*1*-*P* is odds of that event and logit(*P*) is the log(*odds*) = log(*P*/*1*-*P*).
- The difference between the logits of two probabilities is the log of the odds ratio (O*R).*
- log (OR) = $\log\left(\frac{p_1/(1-p_1)}{p_2/(1-p_2)}\right)=\log\left(\frac{p_1}{1-p_1}\right)-\log\left(\frac{p_2}{1-p_2}\right)=\log it(p_1)-\log it(p_2)$
- The logit scale is linear and functions much like a z-score scale.

Logit

- Logit is a continuous score in the range -∝ to ∝
- p = 0.50, then logit = 0.0

  p = 0.70, then logit = 0.85

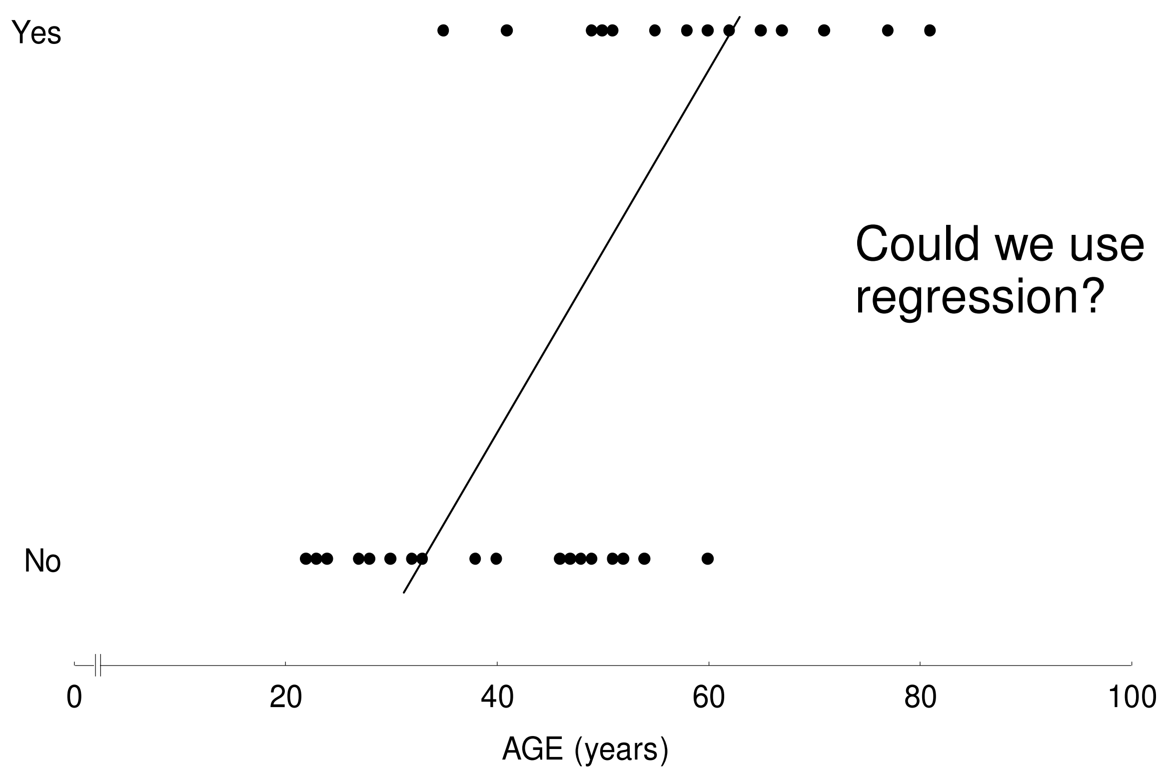
  p = 0.30, then logit = -0.85
- The standard deviation of logit is sqrt(1/a + 1/b + 1/c + 1/d).

Logistic Regression

- Recall: For a categorical variable, we focus on number or proportion for each category.
- Proportion of a category simply says about how likely to happen that category
- Suppose, y is a variable that represent occurrence or not occurrence of cancer (two categories only).
- And y=1 indicates occurrence and y=0 indicates (not occurrence).
- Let, p=likelihood of the event (y=1), so 1-p = likelihood of the event (y=0).
- We want to relate p or (1-p) i.e. likelihood of happening or not happening, instead the response y itself, with an independent variable

Logistic Regression

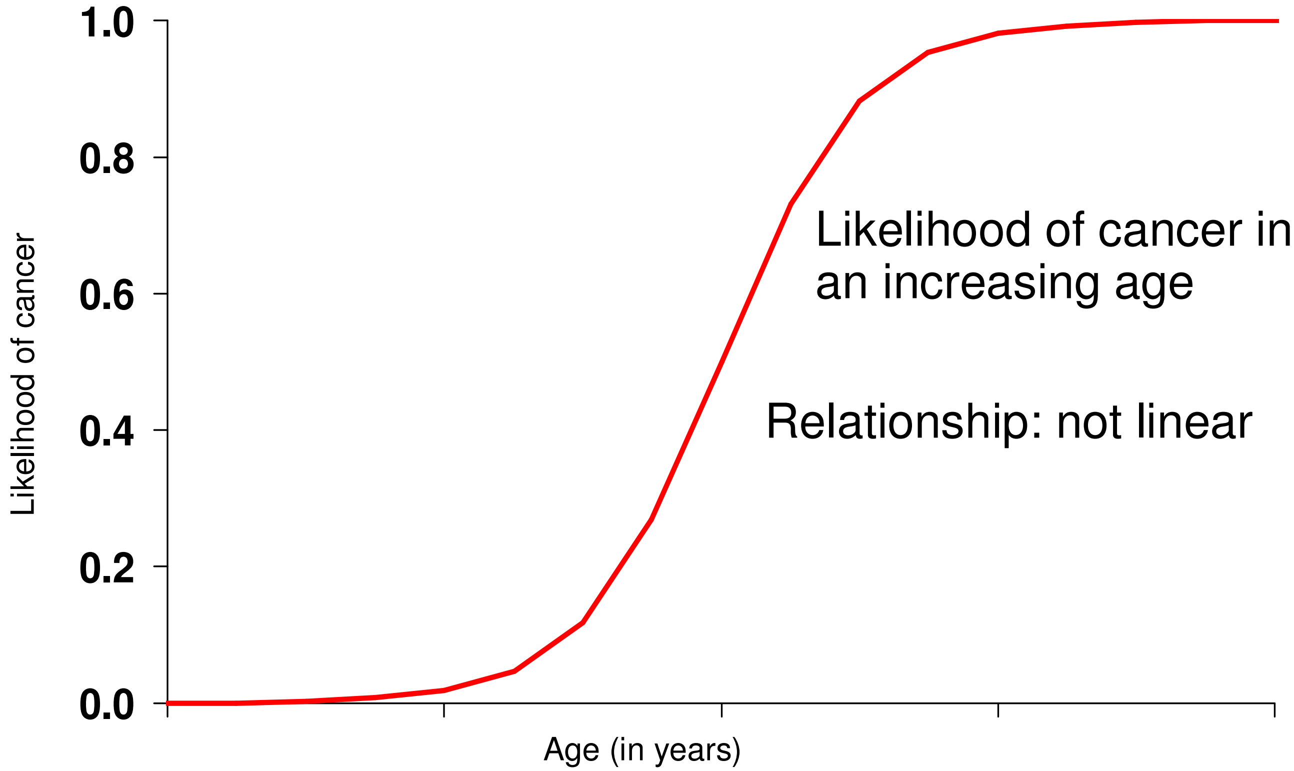
Plot of two variables: An outcome variable (say cancer happening/not happening) and an independent variable (say age)



Could we use a linear regression?

Logistic Regression

Plot of proportions for different ages

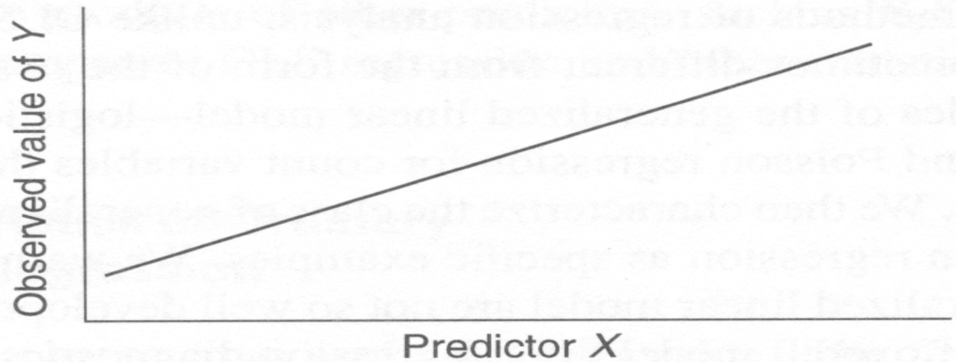


Likelihood of cancer increases with an increasing age

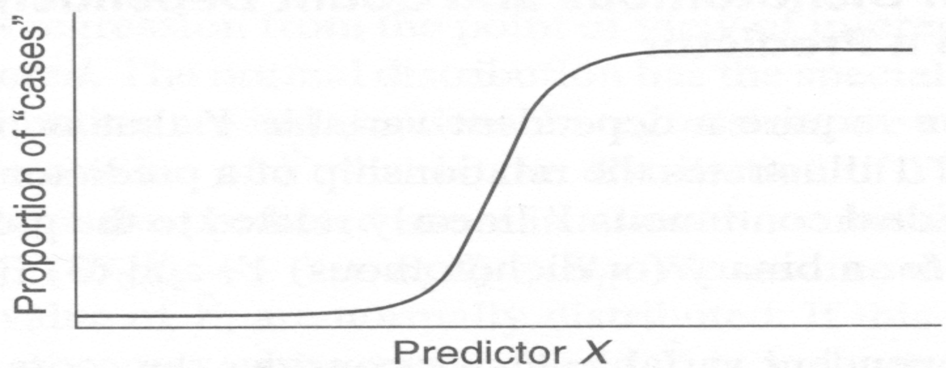
Relationship: not linear

Logistic Regression

(A) For a continuous outcome variable Y, the numerical value of Y at each value of X.



(B) For a binary outcome variable, the proportion of individuals who are "cases" (exhibit a particular outcome property) at each value of X.



Logistic Regression

- Recall: Simple linear regression:
- $y = b_0 + b_1x$, where y is a continuous quantitative outcome variable, x is a quantitative/categorical variable.
- Like y, logit is a quantitative variable, and we can replace y by logit of p where p is the likelihood of an event
- That is, $\log(p/(1-p)) = \log(\text{odds}) = b_0 + b_1x$, which is simple linear regression between $\log(p/(1-p)) = \log(\text{odds})$ and the independent variable x (say age).
- Association patterns with log(odds) are the same as the patterns with odds itself.

Logistic Regression

- log(p/1-p) = log (odds of cancer) = $b_0 + b_1$*age
- Interpretation of b_1: change of log odds of cancer for 1 year change of age.
- Let us consider two persons of ages 55 and 56, then,
- log (odds of cancer at age 55) = $b_0 + b_1$*55
- log (odds of cancer at age 56) = $b_0 + b_1$*56
- The difference, log(odds at 56) – log (55) = b_1

$$b_1 = \log\left(\frac{\text{odds at 56}}{\text{odds at 55}}\right) = \log(\text{odds ratio})$$

$$\text{Odds ratio} = \exp(\text{b}_1)$$

Logistic Regression

- $b_1=0$ (or equivalently odds ratio $\exp(b_1)=1$), indicates no association of log(Odds of cancer) with the variable age (X).
- $b_1>0$, indicates a positive association of log(Odds of cancer) with the variable age (X).
- $b_1<0$, indicates a negative association of log(Odds of cancer) with the variable age (X).
- If 95% confidence interval (CI) of b1 does not contains 0 (the null hypothesis), it indicates that the independent variable has an significant influence on the response variable at 5% level of significance.
- b_0 is the intercept

Logistic Regression

- $\exp(b_1) = 1$, No association of response with predictor. For categorical predictor, an event is equally likely in both reference as well as comparative group.
- $\exp(b_1) > 1$, indicates that an event is more likely to the comparative group compare to the reference group.
- $\exp(b_1) < 1$, indicates that an event is less likely to the comparative group compare to the reference group.
- If 95% CI of odds ratio contains 1, it indicates that likelihood of an event occurrence in two groups are not significantly different at 5% level of significance.

Logistic Regression

- Outcome (response) variable is binary
- Independent variable (predictor) can be either categorical or quantitative
- Relationship of outcome variable and predictor (s) is not linear

Logistic Regression: SPSS demonstration

- Analyze -> Regression -> Binary Logistic -> Select dependent variable to the dependent box and select independent variable to the covariate box-> Click on categorical variable to identify categorical independent variable and select reference category and then select change and other output options.

Multiple Logistic Regression

- More than one independent variables in the model i.e.
- $\log(p/1\text{-}p) = b_0 + bx_1 + bx_2$
- Interpretation: the same as it is for the simple logistic regression
- Response variable is binary

Multinomial and Ordinal Logistic Regressions

- Multinomial: The response (outcome) variable is multicategorical (e.g. race- Caucasian, African American, Hispanic, Asian etc).
- Ordinal: Categories of the response (outcome) variable can be ranked or order (e.g. disease condition: mild, moderate, and severe)

Thank you