

Biost 518

Applied Biostatistics II

.....

Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics
University of Washington

Lecture 5: Review of Simple Regression II

January 12, 2007

1

© 2002, 2003, 2005 Scott S. Emerson, M.D., Ph.D.

Lecture Outline

-
- General Regression Setting
 - Inference on Means
 - Inference about Geometric Means
 - Inference about Odds
 - Inference about Rates
 - Inference about Hazards

2

Simple Logistic Regression

.....

Inference About the Odds

3

Logistic Regression

-
- Binary response variable
 - Allows continuous (or multiple) grouping variables
 - But is OK with binary grouping variable also
 - Compares odds of response across groups
 - “Odds ratio”

4

Why not Linear Regression?

- Many misconceptions about the advantages and disadvantages of analyzing the odds
- Reasons that I consider valid
 - Scientific basis
 - Use of odds ratios in case-control studies
 - Plausibility of linear trends and no effect modifiers
 - Statistical basis
 - Mean variance relationship (if not using robust SE)

5

Simple Logistic Regression

- Modeling odds of binary response Y on predictor X

Distribution $\Pr(Y_i = 1) = p_i$

Model $\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \times X_i$

$X_i = 0$ log odds = β_0

$X_i = x$ log odds = $\beta_0 + \beta_1 \times x$

$X_i = x+1$ log odds = $\beta_0 + \beta_1 \times x + \beta_1$

6

Interpretation as Odds

- Exponentiation of regression parameters

Distribution $\Pr(Y_i = 1) = p_i$

Model $\left(\frac{p_i}{1-p_i}\right) = e^{\beta_0} \times e^{\beta_1 \times X_i}$

$X_i = 0$ odds = e^{β_0}

$X_i = x$ odds = $e^{\beta_0} \times e^{\beta_1 \times x}$

$X_i = x+1$ odds = $e^{\beta_0} \times e^{\beta_1 \times x} \times e^{\beta_1}$

7

Estimating Proportions

- Proportion = odds / (1 + odds)

Distribution $\Pr(Y_i = 1) = p_i$

Model $p_i = \frac{e^{\beta_0} \times e^{\beta_1 \times X_i}}{1 + e^{\beta_0} \times e^{\beta_1 \times X_i}}$

$X_i = 0$ $p_i = e^{\beta_0} / (1 + e^{\beta_0})$

$X_i = x$ $p_i = \frac{e^{\beta_0} \times e^{\beta_1 \times x}}{1 + e^{\beta_0} \times e^{\beta_1 \times x}}$

$X_i = x+1$ $p_i = \frac{e^{\beta_0} \times e^{\beta_1 \times x} \times e^{\beta_1}}{1 + e^{\beta_0} \times e^{\beta_1 \times x} \times e^{\beta_1}}$

8

Simple Logistic Regression

.....

- Interpretation of the model
 - Odds when predictor is 0
 - Found by exponentiation of the intercept from the logistic regression: $\exp(\beta_0)$
 - Odds ratio between groups differing in the value of the predictor by 1 unit
 - Found by exponentiation of the slope from the logistic regression: $\exp(\beta_1)$

9

Stata

.....

- `"logit respvar predvar, [robust]`
 - Provides regression parameter estimates and inference on the log odds scale
 - Intercept, slope with SE, CI, P values
- `"logistic respvar predvar, [robust]`
 - Provides regression parameter estimates and inference on the odds ratio scale
 - Only slope with SE, CI, P values

10

Example

.....

- Prevalence of stroke (cerebrovascular accident- CVA) by age in subset of Cardiovascular Health Study
 - Response variable is CVA
 - Binary variable: 0= no history of prior stroke, 1= prior history of stroke
 - Predictor variable is Age
 - Continuous predictor

11

Odds Ratios using "logistic"

.....

```
.logistic cva age, robust
Logistic regression   Number of obs   =          735
                    LR chi2(1)         =           2.52
                    Prob > chi2        =          0.1127
                    Log likelihood    = -240.98969
                    Pseudo R2         =           0.0051
```

<u>cva</u>	<u> Odds Ratio</u>	<u>StdErr</u>	<u>z</u>	<u>P> z </u>	<u>[95% Conf Int]</u>
age	1.034	.0219	1.59	0.113	.992 1.078

12

Example: Interpretation

.....

“From logistic regression analysis, we estimate that for each year difference in age, the odds of stroke is 3.4% higher in the older group, though this estimate is not statistically significant ($P = .113$). A 95% CI suggests that this observation is not unusual if a group that is one year older might have odds of stroke that was anywhere from 0.8% lower or 7.8% higher than the younger group.”

13

Logistic Regression and χ^2 Test

.....

- Logistic regression with a binary predictor (two groups) corresponds to familiar chi squared test
 - Three possible statistics from logistic regression
 - Wald: The test based on the estimate and SE
 - Score: Corresponds to chi squared test, but not given in Stata output
 - Likelihood ratio test: Can be obtained using post-regression commands in Stata (next quarter)

14

Simple Poisson Regression

.....

Inference About Rates

15

Count Data

.....

- Sometimes a random variable measures the number of events occurring over some region of space and interval of time
 - E.g.,
 - Number of polyps recurring in a patient's colon during a 3 year interval between colonoscopies
 - Number of actinic keratoses developing over a three month period on a patient's left arm
 - Number of pulmonary exacerbations experienced by a cystic fibrosis patient during a year

16

Event Rates

.....

- When a response variable measures counts over space and time, we most often summarize the response across patients by considering the event rate
 - Event rate = expected number of events per unit of space-time
 - The rate is thus a mean count
 - In most statistical problems, we know the interval of time and volume of space sampled

17

Poisson Probability Model

.....

- Frequently: Assume counts are Poisson
 - The Poisson distribution can be derived from the following assumptions
 - The expected number of events occurring in an interval of time is proportional to the size of the interval
 - The probability that two events occur in an infinitesimally small interval of space-time is 0
 - The number of events occurring in separate intervals of space-time are independent
 - (Assumption of a constant rate with independence over separate intervals is pretty strong)

18

Poisson Distribution

.....

- Counts the events occurring at a constant rate λ in a specified time (and space) t
 - Independent intervals of time and space
 - Probability distribution has parameter $\lambda > 0$
 - For $k=0, 1, 2, 3, 4, \dots$

$$\Pr(Y = k) = \frac{e^{-\lambda t} (\lambda t)^k}{k!}$$

- Mean $E(Y) = \lambda t$; variance $Var(Y) = \lambda t$
- Poisson approx to Binomial for low p

19

Regression with Counts

.....

- When the response variable represents counts of some event, we typically model the (log) rate using Poisson regression
 - Compares rates of response per space-time (person-years) across groups
 - “Rate ratio”

20

Why not Linear Regression?

- Primarily statistical:
 - The rate is in fact a mean
 - For Poisson Y measured over time t and having event rate λ
 - $E(Y) = \lambda t$
 - $\text{Var}(Y) = \lambda t$
 - But
 - Want to account for different areas or length of time for measurement
 - Need to account for mean-variance relationship (if not using robust SE)

21

Why a Multiplicative Model?

- In Poisson regression, we tend to use a log link when modeling the event rate
 - Thus we are assuming a multiplicative model
 - “Multiplicative model” = comparisons between groups based on ratios
 - “Additive model” = comparisons between groups based on differences
 - Technical statistical properties:
 - Log rate is the “canonical parameter” for the Poisson

22

Poisson Regression

- Response variable is count of event over space-time (often person-years)
- “Offset” variable specifies space-time
- Allows continuous (or multiple) grouping variables
 - But is OK with binary grouping variable also
- “Offset” variable specifies space-time

23

Simple Poisson Regression

- Modeling rate of count response Y on predictor X

$$\text{Distn} \quad Y_i \sim P(\lambda_i t_i) \Rightarrow \Pr(Y_i = k | T_i = t_i) = \frac{e^{-\lambda_i t_i} (\lambda_i t_i)^k}{k!}$$

$$\text{Model} \quad E(Y_i | T_i, X_i) = \log(\lambda_i T_i) = \log(T_i) + \beta_0 + \beta_1 \times X_i$$

$$X_i = 0 \quad \log \lambda_i = \beta_0$$

$$X_i = x \quad \log \lambda_i = \beta_0 + \beta_1 \times x$$

$$X_i = x + 1 \quad \log \lambda_i = \beta_0 + \beta_1 \times x + \beta_1$$

24

Interpretation as Rates

.....

- Exponentiation of parameters

$$\text{Distn } Y_i \sim P(\lambda_i t_i) \Rightarrow \Pr(Y_i = k | T_i = t_i) = \frac{e^{-\lambda_i t_i} (\lambda_i t_i)^k}{k!}$$

$$\text{Model } E(Y_i | T_i, X_i) = \log(\lambda_i T_i) = \log(T_i) + \beta_0 + \beta_1 \times X_i$$

$$X_i = 0 \quad \lambda_i = e^{\beta_0}$$

$$X_i = x \quad \lambda_i = e^{\beta_0} \times e^{\beta_1 \times x}$$

$$X_i = x + 1 \quad \lambda_i = e^{\beta_0} \times e^{\beta_1 \times x} \times e^{\beta_1}$$

25

Simple Poisson Regression

.....

- Interpretation of the model
 - Rate when predictor is 0
 - Found by exponentiation of the intercept from the Poisson regression: $\exp(\beta_0)$
 - Rate ratio between groups differing in the value of the predictor by 1 unit
 - Found by exponentiation of the slope from the Poisson regression: $\exp(\beta_1)$

26

Example: Setting

.....

- Chemosensitizers for cancer chemotherapy
 - In vitro evaluation of the ability of some drugs to potentiate the cytotoxic effects of doxorubicin
 - Cells cultured in the laboratory are exposed to doxorubicin at several concentrations with and without chemosensitizers
 - This example: Only the control group

27

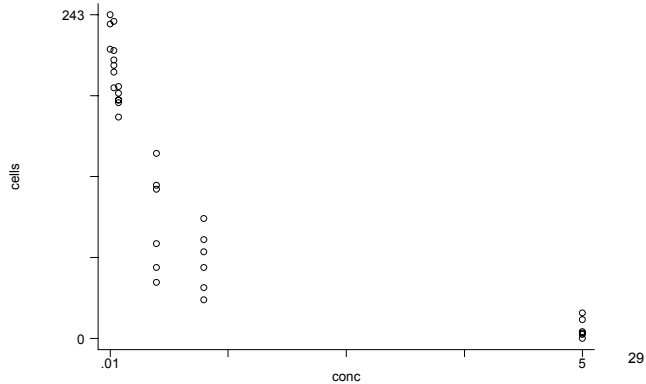
Example: Variables

.....

- Response:
 - Number of surviving cell colonies
 - Each presumably arising from a single cell
- Offset:
 - Default value of 1
 - Same volume of culture used for all samples
- Predictor:
 - Concentration of doxorubicin

28

Scatterplot Cells vs Dox Conc

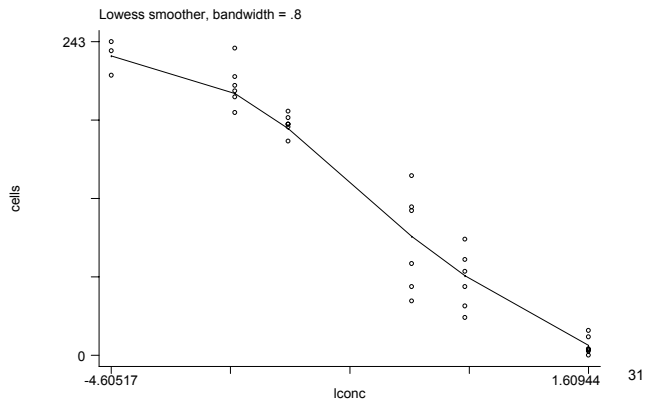


Characterization of Scatterplot

- Characterization of scatterplot
 - Doxorubicin concentration was sampled on log scale
 - This sampling scheme was used because it was known that proportion of cells killed is more or less linear in log concentration
 - Michaelis-Menten kinetics: Actually S shaped in log concentration, but well approximated linearly over a range of doses

30

Scatterplot: Cells vs log (Conc)



Characterization of Scatterplot

- Outliers:
 - None obvious
- First order trend:
 - Decreasing cell survival with increasing log concentration
- Second order trend:
 - Hint of S-shaped curve, but counts fairly well approximated by straight line
- Within group variability:
 - Decreasing variance for lower group means (note smaller sample size in first group)

32

Stata Commands

- Same form as for other regression models
 - Exception:
 - If the observed counts are measured over different amounts of time or space, we must specify the length of “exposure”
 - `poisson respvar predvar, exposure(tm) [robust]`
 - Exposure can also be given as the “offset”, which is just the log of the exposure time
 - `poisson respvar predvar, offset(logtm) [robust]`

33

Estimation of Regression Model

```
. poisson cells lconc  
(Iteration information omitted)
```

```
Number of obs   =       282  
LR chi2(1)      =   14724.65  
Prob > chi2     =     0.0000  
Pseudo R2     =     0.6242
```

cells	Coef.	StErr.	z	P> z	[95% CI]	
lconc	-.366	.003	-115	0.000	-.372	-.360
_cons	3.75	.011	329	0.000	3.72	3.77

34

Interpretation of Stata Output

$$\log \text{ rate} = 3.75 - 0.366 \times \text{lconc}_i$$

- Regression model for cells on log concentration
 - Intercept is labeled by “_cons”
 - Estimated intercept: 3.75
 - Slope is labeled by variable name: “lconc”
 - Estimated slope: -0.366

35

Interpretation of Intercept

$$\log \text{ rate} = 3.75 - 0.366 \times \text{lconc}_i$$

- Estimated count rate for lconc 0 is found by exponentiation: $\exp(3.75) = 42.5$
 - lconc= 0 corresponds to a concentration of 1.0
 - This was the highest concentration sampled
 - In this problem, the intercept is of interest if the linear relationship between log concentration and log rate is correct

36

Interpretation of Slope

.....

$$\log \text{ rate} = 3.75 - 0.366 \times \text{lconc}_i$$

- Estimated ratio of rates for two groups differing by 1 in log concentration is found by exponentiation slope: $\exp(-0.366) = 0.694$
 - Group one log unit higher has survival rate only 0.694 as large (69.4% as large)
 - 1 log unit = 2.718 times higher concentration
 - 10 fold increase in concentration tends to cause a survival rate only $10^{-0.3660} = 0.431$ as large
 - 56.9% decrease in survival rate

37

Role of Linearity

.....

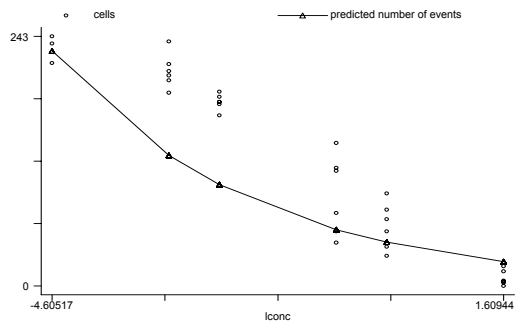
- We have to be careful in interpreting this model if the linear relationship does not hold
 - Scatterplot suggested linear relationship between cell counts and log concentration was reasonable
 - But we modeled the log rate versus log concentration

38

Fitted Regression Model

.....

```
. predict fcells  
. graph cells fcells lconc, s(oT) c(.1)
```



Simple Proportional Hazards Regression

.....

Inference About Hazards

40

Right Censored Data

- A special type of missing data: the exact value is not always known
 - Some measurements are known exactly
 - Some measurements are only known to exceed some specified value (perhaps different for each subject)
- Typically represented by two variables
 - An observation time: Time to event or censoring, whichever came first
 - An indicator of event: Tells us which were observed events

41

Statistical Methods

- In the presence of censored data, the “usual” descriptive statistics are not appropriate
 - Sample mean, sample median, simple proportions, sample standard deviation should not be used
 - Proper descriptives should be based on Kaplan-Meier estimates
- Similarly, special inferential procedures are needed with censored data

42

Notation

Unobserved :

True times to event : $\{T_1^0, T_2^0, \dots, T_n^0\}$

Censoring Times : $\{C_1, C_2, \dots, C_n\}$

Observed data :

Observation Times : $T_i = \min(T_i^0, C_i)$

Event indicators : $D_i = \begin{cases} 1 & \text{if } T_i = T_i^0 \\ 0 & \text{otherwise} \end{cases}$

43

Proportional Hazards Model

- Instantaneous rate of failure at each time among subjects who have not failed
 - Proportional hazards assumes that the ratio of these instantaneous failure rates is constant in time between two groups
 - Proportional hazards (Cox) regression treats the survival distribution within a group semiparametrically
 - A semi-parametric model: The hazard ratio is the parameter, there is no intercept

44

Borrowing Information

.....

- Use other groups to make estimates in groups with sparse data
 - Borrows information across predictor groups
 - E.g., 67 and 69 year olds would provide some relevant information about 68 year olds
 - Borrows information over time
 - Relative risk of an event at each time is presumed to be the same under Proportional Hazards

45

Simple PH Regression Model

.....

- “Baseline” hazard function is unspecified
 - Similar to an intercept

Model $\log(\lambda(t | X_i)) = \log(\lambda_{i_0}(t)) + \beta_1 \times X_i$

$X_i = 0$ log hazard at $t = \log(\lambda_0(t))$

$X_i = x$ log hazard at $t = \log(\lambda_0(t)) + \beta_1 \times x$

$X_i = x+1$ log hazard at $t = \log(\lambda_0(t)) + \beta_1 \times x + \beta_1$

46

Model on Hazard scale

.....

- Exponentiating parameters

Model $\lambda(t | X_i) = \lambda_0(t) \times e^{\beta_1 \times X_i}$

$X_i = 0$ hazard at $t = \lambda_0(t)$

$X_i = x$ hazard at $t = \lambda_0(t) \times e^{\beta_1 \times x}$

$X_i = x+1$ hazard at $t = \lambda_0(t) \times e^{\beta_1 \times x} \times e^{\beta_1}$

47

Interpretation of the Model

.....

- No intercept
 - Generally do not look at baseline hazard
 - But can be estimated
- Slope parameter
 - Hazard ratio between groups differing in the value of the predictor by 1 unit
 - Found by exponentiation of the slope from the proportional hazards regression: $\exp(\beta_1)$

48

Stata

- `"stcox obsvar eventvar, [robust]"`
 - Provides regression parameter estimates and inference on the hazard ratio scale
 - Only slope with SE, CI, P values

49

Example

- Prognostic value of nadir PSA relative to time in remission
 - PSA data set: 50 men who received hormonal treatment for advanced prostate cancer
 - Followed at least 24 months for clinical progression, but exact time of follow-up varies
 - Nadir PSA: lowest level of serum prostate specific antigen achieved post treatment

50

Estimation of Regression Model

```
. stset obstime relapse
. stcox nadir
Cox regression -- Breslow method for ties
No. of subj =      50      No. of obs =      50
No. fail    =      36
Time at risk = 1423

Wald chi2(1) = 16.79
Log likelihood = -113.3      Prob > chi2 = 0.0008

|_____|_____|_____|_____|_____|_____|
|          Robust
|_____|_____|_____|_____|_____|_____|
t | HzRat StdErr  z    P>|z|    [95% Conf Int]
nadir | 1.016  .0038  4.10  0.000  1.008  1.023
```

51

Interpretation of Stata Output

- Scientific interpretation of the slope

$$\text{Hazard ratio} = 1.015^{\Delta \text{nadir}}$$

- Estimated hazard ratio for two groups differing by 1 in nadir PSA is found by exponentiation slope (Stata only reports the hazard ratio):
 - Group one unit higher has instantaneous event rate 1.015 times higher (1.5% higher)
 - Group 10 units higher has instantaneous event rate $1.015^{10} = 1.162$ times higher (16.2% higher)

52

Statistical Validity of Inference

.....

- Inference (CI, P vals) about associations requires three general assumptions
 - Assumptions about approximate normal distribution for parameter estimates
 - Assumptions about independence of observations
 - Assumptions about variance of observations within groups

53

Normally Distributed Estimates

.....

- Assumptions about approximate normal distribution for parameter estimates
 - Classically or Robust SE:
 - Large sample sizes
 - Definition of “large” depends on underlying probability distribution

54

Independence / Dependence

.....

- Assumptions about independence of observations for linear regression
 - Classically:
 - All observations are independent
 - Robust standard error estimates:
 - Allow correlated observations within identified clusters

55

Within Group Variance

.....

- Assumptions about variance of response within groups for proportional hazards regression
 - Classically:
 - Mean variance relationship for binary data
 - Proportional hazards considers odds of event at every time
 - Need proportional hazards and linearity of predictor
 - Robust standard error estimates:
 - Allow unequal variances across groups
 - (Do not need proportional hazards or linearity)

56

Linearity of Model

- Assumption about adequacy of linear model for prediction of group odds of response with logistic regression
 - The log hazard ratio across groups is linear in the modeled predictor
 - (We can model transformations of the measured predictor)

57

Example: Interpretation

“From proportional hazards regression analysis, we estimate that for each 1 ng/ml unit difference in nadir PSA, the risk of relapse is 1.6% higher in the group with the higher nadir. This estimate is highly statistically significant ($P < .001$). A 95% CI suggests that this observation is not unusual if a group that has a 1 ng/ml higher nadir might have risk of relapse that was anywhere from 0.8% higher to 2.3% higher than the group with the lower nadir.”

58

Log Transformed NadirPSA

- Based on prior experience
 - A constant difference in PSA would not be expected to confer same increase in risk
 - Comparing 4 ng/ml to 10 ng/ml is not the same as comparing 104 ng/ml to 110 ng/ml
 - A multiplicative effect on risk might be better
 - Same increase in risk for each doubling of nadir
 - Use log transformed nadir PSA

59

Estimation of Regression Model

```
. generate lnadir = log(nadir)
. stcox lnadir, robust
Cox regression -- Breslow method for ties
No. of subj   =      50      No. of obs   =      50
No. fail      =      36
Time at risk  =     1423

LR chi2(1)    =     34.04
Log likelihood = -107.3   Prob > chi2   =  0.0000
```

t	HzRat	StdErr	z	P> z	[95% Conf Int]
lnadir	1.54	.113	5.83	0.000	1.33 1.77

60

Interpretation of Parameters

.....

- Hazard ratio is 1.54 for an e-fold difference in nadir PSA
 - $e = 2.7183$
- I can more easily understand doubling, tripling, 5-fold, 10-fold increases
 - For doubling: HR : $1.54^{\log(2)} = 1.35$

61

PH Regression and Logrank Test

.....

- Proportional hazards regression with a binary predictor (two groups) corresponds to the logrank test
 - Three possible statistics from proportional hazards regression
 - Wald: The test based on the estimate and SE
 - Score: Corresponds to logrank test, but not given in Stata output
 - Likelihood ratio test: Can be obtained using post-regression commands in Stata (next quarter)

62

Interpretation of Slopes

.....

63

“Additive Models”

.....

- Identity link function
 - Means: linear regression

$$\theta_X = \beta_0 + \beta_1 \times X$$

64

“Additive Models”: Slope

- Interpretation of slope:

β_1 : (Average) Difference in summary measure between groups per 1 unit difference in X

$\Delta \times \beta_1$: (Average) Difference in summary measure between groups per Δ unit difference in X

$$\theta_X = \beta_0 + \beta_1 \times X$$

65

“Additive Models”: $\log(X)$

- Slope with log transformed predictor

$\log(k) \times \beta_1$: (Average) Difference in summary measure between groups per k -fold difference in X

$$\theta_X = \beta_0 + \beta_1 \times \log(X)$$

66

“Multiplicative Models”

- Log link function

– Geom means: linear regression on logs

– Odds: logistic regression

– Hazards: proportional hazard regression

– Means: Poisson regression

– Medians: accel failure time regression

$$\log(\theta_X) = \beta_0 + \beta_1 \times X$$

67

“Multiplicative Models”: Slope

- Interpretation of slope:

e^{β_1} : (Average) Ratio of summary measure between groups per 1 unit difference in X

$e^{\Delta \times \beta_1} = (e^{\beta_1})^\Delta$: (Average) Ratio of summary measure between groups per Δ unit difference in X

$$\log(\theta_X) = \beta_0 + \beta_1 \times X$$

68

“Multiplicative Models”: $\log(X)$

- Slope with log transformed predictor

$e^{\log(k) \times \beta_1} = k^{\beta_1} = (e^{\beta_1})^{\log(k)}$: (Average) Ratio of summary measure between groups per k -fold difference in X

$$\log(\theta_X) = \beta_0 + \beta_1 \times \log(X)$$