Biost 518 / Biost 515
Applied Biostatistics II / Biostatistics II

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

Lecture 5:
Simple Logistic Regression Model
Simple Poisson Regression Model

January 15, 2014

1

Lecture Outline

- Regression with Binary Response
  - Risk difference: linear regression
  - Risk ratio: Poisson regression
  - Odds ratio: logistic regression
- Simple Logistic Regression Models
- Simple Poisson Regression Models

2

Regression with Binary Response

3

Binary Random Variables

- Many variables can take on only two values
  - For convenience code as 0 or 1 “indicator” variable
    - Vital status: “Dead” coded 0= alive 1= dead
    - Sex: “Female” coded 0= male 1= female
    - Intervention: “Tx” coded 0= control 1= new therapy

- Sometimes dichotomize variables
  - For scientific reasons (statistically less precise)
    - Blood pressure less than 160 mm Hg
    - PSA less than 4 ng/ml
    - Serum glucose less than 120 mg/dl

4

Bernoulli Probability Distribution

- A binary variable Y_i must have a Bernoulli probability distribution
  - A single parameter: $p = Pr(Y_i = 1)$ with $0 \leq p \leq 1$
  - We write $Y \sim B(1, p)$
  - Probability mass function: $Pr (Y_i = y) = p^{y} (1 – p)^{1-y}$

- Mean: $E[Y_i] = p$ Variance: $Var(Y_i) = p (1 – p)$
  - A "mean – variance" relationship
    - If the mean is different between two groups, the variance must also be different
  - Maximum variance of 0.25 when $p = 0.5$

- The sum of n independent Bernoulli random variables has a binomial distribution: $S_n = Y_1 + \ldots + Y_n \sim B(n, p)$
  - Mean: $E[S_n] = n p$ Variance: $Var(S_n) = n p (1 – p)$

5

Regression with Binary Response

- Conceptually, there should be no problem modeling the proportion (which is the mean of the distribution)

- However, there are several scientific and technical reasons why we do not use linear regression very often with binary response

6

Statistical Hypotheses

- Summary measures of interest for a Bernoulli random variable are pretty much limited to either
  - The proportion *p* (a mean), or
  - The odds *o = p / (1-p)*

- Contrasts used to compare the distribution of a Bernoulli random variable across subpopulations thus include
  - Difference in proportions: $p_1 - p_0$ (risk difference (RD))
  - Ratio of proportions : p_1 / p_0 (risk ratio (RR))
  - Odds ratio : $o_1 / o_0 = \frac{p_1 / (1-p_1)}{p_0 / (1-p_0)}$ (odds ratio (OR))

7

Choice of Summary Measure / Contrast

- We thus have three choices of regression with a binary response
  - RD: linear regression
  - RR: Poisson regression
  - OR: logistic regression

- In choosing among these regression models, we consider scientific issues including
  - Any desire to accentuate public health impact
  - Any desire to accentuate differences with rare events
  - Possibility of avoiding major nonlinearities and effect modification
  - Interplay of
    - How we want to eventually use results of the analysis
      - Which variable do we ideally want to condition on?
    - And the way we sampled our data
      - Were sample sizes in any subpopulations fixed by design?

Choice of Summary Measure / Contrast

- Public health impact is typically best measured by the difference in proportions ➔ perhaps prefer RD regression
  - RD, as a difference in proportions, can estimate the number of affected people in a larger population
- With rare events, the existence of an association is best demonstrated using ratios ➔ perhaps prefer RR regression
  - Though unlikely in either case, I am many more times likely to win the lottery by buying a ticket than by finding a winning ticket

9

Choice of Summary Measure / Contrast

- Greater possibility of avoiding major nonlinearities ➔ perhaps prefer OR regression
  - Based on fact that $0 \leq p \leq 1$, but $0 \leq o \leq \infty$

- In RD and RR regression, constraints on p dictate that either
  - strength of association between Y and X is constrained relative to range of possible values that X can have, or
  - the association between Y and X must be nonlinear

10

Choice of Summary Measure / Contrast

- Greater possibility of avoiding effect modification ➔ perhaps prefer OR regression
  - Based on fact that $0 \leq p \leq 1$, but $0 \leq o \leq \infty$

- Any covariate that is strongly associated with outcome *Y* must modify the effects of other moderately associated random variables on the RD or RR scale
  - We will discuss this further when we consider adjusted analyses, but the following example illustrates the general idea

11

Effect Modification Example

- The restriction on ranges for probabilities also make it likely that effect modification will often be present with proportions
- Example: 2 Yr Relapse rates by nadir PSA > 4, bone scan score (BSS) in hormonally treated prostate cancer
  - Both nadir PSA and bone scan score show strong associations with relapse
- If bone scan score < 3: A difference of 0.60
  - 40% of men with nadir PSA < 4 relapse in 24 months
  - 100% of men with nadir PSA > 4 relapse in 24 months
- If bone scan score = 3:
  - 71% of men with nadir PSA < 4 relapse in 24 months
  - Thus impossible for men with nadir PSA > 4 to have an absolute difference of 0.60 higher

12

Choice of Summary Measure / Contrast

- Interplay of
  - How we want to eventually use results of the analysis
    - Which variable do we ideally want to condition on?
  - And the way we sampled our data
    - Were sample sizes in any subpopulations fixed by design?

- Want to condition on exposures, but use case-control sampling with rare disease ➔ perhaps prefer OR regression
  - If we constrain sample sizes for diseased vs non-disease, analyses should in general be based on *Pr(Exposure | Disease)*
  - The one exception is when using the odds ratio, because the odds ratio based on conditioning on disease status must be equal to the odds ratio when conditioning on exposure status

13

Case-Control Studies

- Scientific interest:
  - Distribution of “effect” across groups defined by “cause”
  - E.g., how does risk of lung cancer differ by smoking behavior

- Common sampling schemes
  - Cohort study: Sample by exposure
    - Sample 1000 smokers, 1000 nonsmokers
    - Estimate risk of cancer in exposure groups

  - Case-control study: Sample by outcomes
    - Sample 1000 cancer patients, 1000 controls
    - In general: estimate prevalence of smoking in diagnosis groups
      - E.g., proportion (or odds) of smokers among people with or without cancer

14

Use of Odds Ratios

- Cohort study
  - Odds of cancer among smokers : odds of cancer among nonsmokers
- Case-control study
  - Odds of smoking among cancer : odds of smoking among noncancer
- Mathematically, the two odds ratios are the same
  - Hence, when using case-control sampling, it is valid to estimate either odd ratio
  - (But the intercept may be completely uninterpretable based on available data)

15

Example: Two Sample Studies

- Investigate association between mortality and smoking in a population of elderly adults
  - Death within 4 years of some “sentinel event”
  - Smoking behavior current at time of the “sentinel event”

- Sampling schemes that might be considered
  - Cross-sectional sampling of 4,994 subjects
  - Cohort study of 400 smokers and 1,200 nonsmokers
  - Case-control study of 300 deaths within 4 years of sentinel event and 900 controls alive 4 years after the sentinel event

16

Cross-sectional Study (N_{Tot} = 4,994)

- Valid estimates: Mortality conditioning on smoking behavior
- Valid estimates: *Smoking behavior conditioning on mortality*

		Death w/in 4 Yr		Pr (Dth = 1 | Smk)		Odds (Dth = 1 | Smk)
		0	1			
Smoking	0	**3,966**	**424**	**0.0966**		**0.1069**
	1	**533**	**71**	**0.1175**		**0.1332**
Pr (Smk = 1 | Dth)		*0.1185*	*0.1434*	**RD** **Dth | Smk: .0210** *Smk | Dth: .0250*	**RR** **Dth | Smk: 1.217** *Smk | Dth: 1.211*	
Odds (Smk = 1 | Dth)		*0.1344*	*0.1675*			**OR** **Dth | Smk: 1.246** *Smk | Dth: 1.246*

17

Cohort Study (N_{Smk}= 400; N_{NS}= 1,200)

- Valid estimates: Mortality conditioning on smoking behavior
- Valid estimates: *Smoking behavior conditioning on mortality*

		Death w/in 4 Yr		Pr (Dth = 1 | Smk)		Odds (Dth = 1 | Smk)
		0	1			
Smoking	0	**1,090**	**110**	**0.0917**		**0.1009**
	1	**358**	**42**	**0.1050**		**0.1173**
Pr (Smk = 1 | Dth)		~~*0.2472*~~	~~*0.2763*~~	**RD** **Dth | Smk: .0133** ~~*Smk | Dth: .0291*~~	**RR** **Dth | Smk: 1.145** ~~*Smk | Dth: 1.118*~~	
Odds (Smk = 1 | Dth)		~~*0.3284*~~	~~*0.3818*~~			**OR** **Dth | Smk: 1.163** ***Smk | Dth: 1.163***

18

Case-Control Study (N_{Die}= 300; N_{Surv}= 900)

- Valid estimates: Mortality conditioning on smoking behavior
- Valid estimates: *Smoking behavior conditioning on mortality*

		Death w/in 4 Yr		Pr (Dth = 1 | Smk)		Odds (Dth = 1 | Smk)
		0	1			
Smoking	0	**783**	**259**	0.2486		0.3308
	1	**117**	**41**	0.2595		0.3504
Pr (Smk = 1 | Dth)		*0.1300*	*0.1367*	**RD** Dth | Smk: .0109 *Smk | Dth: .0067*	**RR** Dth | Smk: 1.044 *Smk | Dth: 1.051*	
Odds (Smk = 1 | Dth)		*0.1494*	*0.1583*			**OR** Dth | Smk: 1.059 *Smk | Dth: 1.059*

19

Take Home Message 1

- The corresponding valid estimates from each study design are estimating the same quantity
  - E.g., both the cross-sectional and cohort studies can be used to estimate the population *Pr[Die w/in 4 years | Smoke]*
- I created a single cohort design and a single case-control design by sampling from the cross-sectional design.
  - There is of course less precision in those derived designs, because the sample sizes are smaller
  - Furthermore, the cross-sectional design did not have all that much precision
    - The inference from the cross-sectional study estimated an odds ratio of 1.246, with a 95% CI of 0.940 to 1.63
    - The estimated OR from the cohort and case-control studies (which were 1.163 and 1.059, respectively) were consistent with that lack of precision

20

Take Home Message 2

- All study designs are estimating the odds ratio comparing the odds of death within 4 years for smokers to the odds of death within 4 years for nonsmokers
  - The cross-sectional and cohort studies can do this directly
  - The case-control study can do this indirectly from a scientific standpoint
    - But because this is true scientifically, and because the OR is mathematically the same in either direction, we can actually fit the "reverse" logistic regression model and get the same answer for the slope (though the intercept in that model is not estimating a population-based odds)
- This property is an advantage of looking at OR, because with rare events, case-control sampling is more feasible and economical

21

OR Interpretation in Case-Control Studies

- The odds ratio is easily interpreted when trying to investigate rare events
  - Odds = prob / (1 – prob)
  - Rare event: (1 – prob) is approximately 1
    - Odds is approximately the probability
    - Odds ratio is approximately the risk ratio
      - Risk ratios are easily understood
- Case-control studies are typically used when events are rare
- Note that in the previous example, the probability of death was on the order of 10% in the cross-sectional study, so the OR and the RR are only approximately equal.

22

Statistics: Mean-Variance

- There is also a technical problem with using classical linear regression
- Classical linear requires equal variances in each predictor group in order for CI and p values to be valid
- But with binary *Y*, the variance within a group depends on the mean
  - Mean: $E[Y] = p$ Variance: $Var(Y) = p(1-p)$
  - In the presence of an association between response and POI, we will definitely have heteroscedasticity
- When using the Huber-White sandwich estimate of robust standard errors, this problem not such a limitation
  - Moderate sample sizes are needed

23

Simple Logistic Regression

Inference About the Odds

24

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”

25

Simple Logistic Regression

- Modeling odds of binary response Y on predictor X

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

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

$X_i = 0$ $\log \text{odds} = \beta_0$

$X_i = x$ $\log \text{odds} = \beta_0 + \beta_1 \times x$

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

26

Interpretation as Odds

- Exponentiation of regression parameters

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

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

$X_i = 0$ $\quad \text{odds} = e^{\beta_0}$

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

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

27

Estimating Proportions

- Proportion = odds / (1 + odds)

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

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

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

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

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

28

Parameter Interpretation

- Interpretation of the logistic regression parameters based on odds

- 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: $\exp(\beta_1)$

29

Similarity to Other Regressions

- Logistic regression uses *maximum likelihood estimation* to find parameter estimates
- If a saturated model is fit, the estimated odds of event in each group will agree exactly with the sample odds
- In large samples, the regression parameter estimates are approximately normally distributed
  - P values and CI that are displayed for each parameter estimate are Wald- based estimates

$$95\%\,\text{CI}:\ (estimate) \pm (crit\ value) \times (std\ err) \qquad \hat{\beta} \ \pm \ z_{1-\alpha/2} \times \hat{se}\left(\hat{\beta}\right)$$

$$\text{Test stat}: \qquad Z = \frac{(estimate) - (null)}{(std\ err)} \qquad Z = \frac{\hat{\beta} - \beta_0}{\hat{se}\left(\hat{\beta}\right)}$$

30

Technical Details

- Unlike linear regression, there is no closed form expression to find the logistic regression parameter estimates
- Instead, computer programs use an iterative search
- This search may fail in saturated or nearly saturated models if some parameter corresponds to a group having all events or no events
  - In this setting, logistic regression parameters modeling the log odds are trying to estimate positive or negative infinity
  - The sample size is too small for the model

31

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

32

Example

- Prevalence of stroke (cerebrovascular accident- CVA) by age in subset of Cardiovascular Health Study

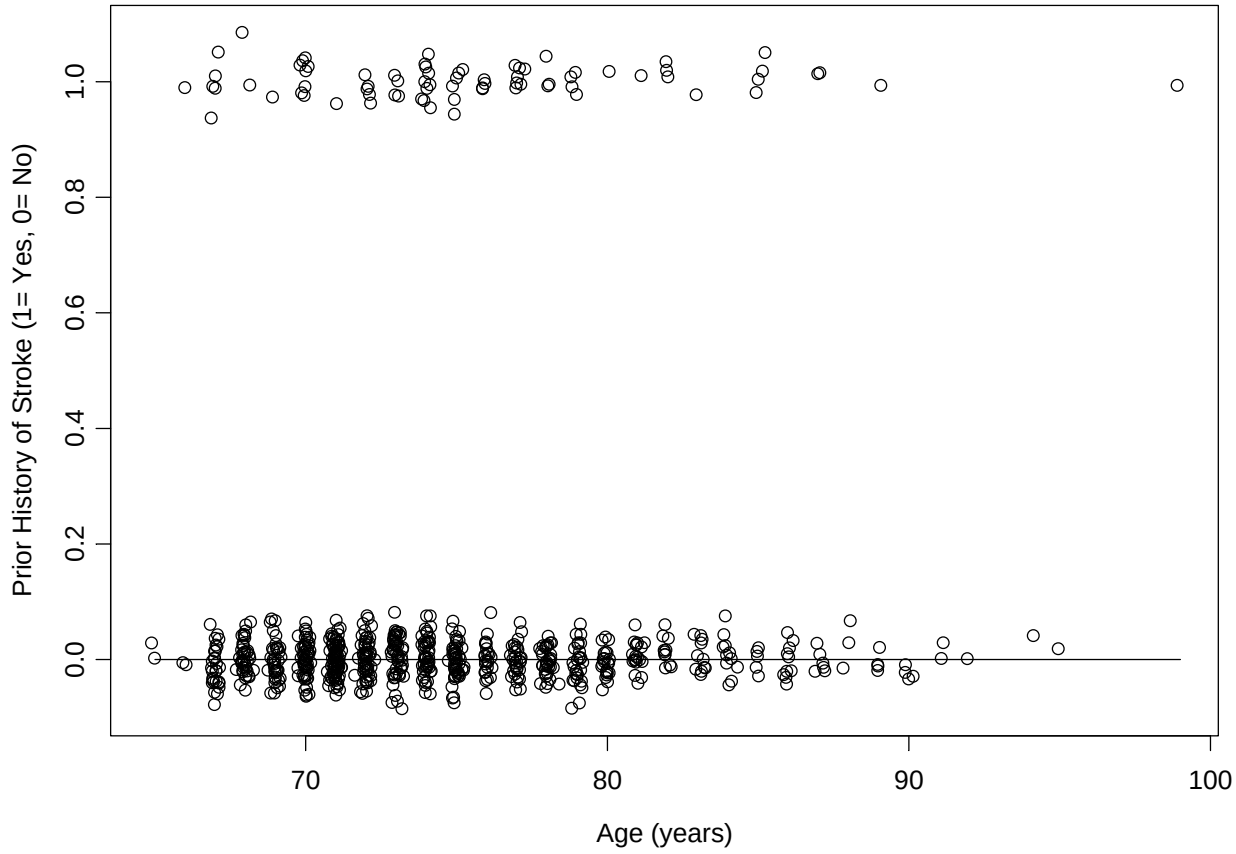
- Response variable is CVA
  - Binary variable: 0= no history of prior stroke, 1= prior history of stroke

- Predictor variable is Age
  - Continuous predictor

33

Lowess Smooth of CVA vs Age

- Scatter plot is pretty useless



34

Characterization of Plot

- Clearly the scatterplot (even with superimposed smooth) is pretty useless with a binary response

- (Note that we are estimating proportions– not odds– with this plot, so we can not even judge linearity for logistic regression)

35

Example: Regression Model

- Answer question by assessing linear trends in log odds of stroke by age
- Estimate best fitting line to log odds of CVA within age groups

$$\log odds\left(CVA \middle| Age \right) = \beta_0 + \beta_1 \times Age$$

- An association will exist if the slope (β_1) is nonzero
  - In that case, the odds (and probability) of CVA will be different across different age groups

36

Parameter Estimates

```
. logit cva age
(iteration info deleted)
                           Number of obs   =        735
                           LR chi2(1)      =       2.45
                           Prob > chi2     =     0.1175
                           Log likelihood  = -240.98969
                           Pseudo R2       =     0.0051

------------------------------------------------------------------------------
         cva |      Coef   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |      .0336   .0210     1.59   0.111    -.0077    .0748
       _cons |     -4.69    1.591    -2.95   0.003    -7.810   -1.572
------------------------------------------------------------------------------
```

37

Interpretation of Stata Output

- Regression model for CVA on age
- Intercept is labeled by "_cons"
  - Estimated intercept: -4.69
- Slope is labeled by variable name: "age"
  - Estimated slope: 0.0336
- Estimated linear relationship:
  - log odds CVA by age group given by

$$\text{log odds } CVA = -4.69 + 0.0336 \times Age_i$$

38

Interpretation of Intercept

$$\log \text{odds } CVA = -4.69 + 0.0336 \times Age_i$$

- Estimated log odds CVA for newborns is -4.69
  - Odds of CVA for newborns is $e^{-4.69} = 0.0092$
  - Probability of CVA for newborns
    - Use prob = odds / (1+odds): .0092 / (1+.0092)= .0091
- Pretty ridiculous to try to estimate
  - We never sampled anyone less than 67
  - In this problem, the intercept is just a tool in fitting the model

39

Interpretation of Slope

$$\log \text{odds } CVA = -4.69 + 0.0336 \times Age_i$$

- Estimated difference in log odds CVA for two groups differing by one year in age is 0.0336, with older group tending to higher log odds
  - Odds Ratio: $e^{0.0336} = 1.034$
  - For 5 year age difference: $e^{5 \times 0.0336} = 1.034^5 = 1.183$

- (If a straight line relationship is not true, we interpret the slope as an average difference in log odds CVA per one year difference in age)

40

Stata: "logit" versus "logistic"

- We are rarely interested in the intercept by itself
  - We do have to use it when estimating odds of an event in a single group

- Given that we are rarely interested in the intercept, we might as well use the "`logistic`" command
  - It will provide inference for the odds ratio
  - We don't have to exponentiate the slope estimate

41

Odds Ratios using "logistic"

```
.logistic cva age
Logistic regression   Number of obs   =        735
                      LR chi2(1)      =       2.45
                      Prob > chi2     =     0.1175
                      Log likelihood  = -240.98969
                      Pseudo R2       =     0.0051

------------------------------------------------------------------------------
         cva | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.034    .0218     1.59   0.111      .992       1.078
------------------------------------------------------------------------------
```

42

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."

43

Comments on Interpretation

- I express this as a difference between group odds rather than a change with aging
  - We did not do a longitudinal study
- To the extent that the true group log odds have a linear relationship, this interpretation applies exactly
- If the true relationship is nonlinear
  - The slope estimates the “first order trend” for the sampled age distribution
  - We should not regard the estimates of individual group probabilities / odds as accurate

44

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 (covered with adjusting for covariates)

45

Signal and Noise

- Note that the Signal and Noise idea does not apply so well here
- We do not tend to quantify an “error distribution” with logistic regression

46

Simple Poisson Regression

Inference About Rates

47

Count Data

- Sometimes a random variable measures the number of events occurring over some region of space and interval of time
- 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

48

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

49

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

50

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 $k = 0, 1, 2, 3, 4, \ldots$

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

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

51

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”

52

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
    - $E(Y) = \lambda t$
    - $Var(Y) = \lambda t$
- 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)

53

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

54

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

55

Simple Poisson Regression

- Modeling rate of count response Y on predictor X

Distn $\quad Y_i \sim \mathrm{P}(\lambda_i t_i) \Rightarrow \quad \Pr(Y_i = k \mid T_i = t_i) = \dfrac{e^{-\lambda_i t_i}(\lambda_i t_i)^k}{k!}$

Model $\quad E(Y_i \mid T_i, X_i) = \log(\lambda_i T_i) = \log(T_i) + \beta_0 + \beta_1 \times X_i$

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

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

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

56

Interpretation as Rates

- Exponentiation of parameters

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

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

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

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

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

57

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($\beta 1$)

58

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]`
- By default, Stata reports estimates on the log mean and log mean ratio scale
  - Specifying the option `irr` will cause Stata to suppress output of the intercept and to report "incidence rate ratios"

59

Similarity to Other Regressions

- Poisson regression uses *maximum likelihood estimation* to find parameter estimates
- If a saturated model is fit, the estimated mean in each group will agree exactly with the sample mean
- In large samples, the regression parameter estimates are approximately normally distributed
  - P values and CI that are displayed for each parameter estimate are Wald- based estimates

$$95\%\,\text{CI}:\ (estimate) \pm (crit\ value) \times (std\ err) \qquad \hat{\beta} \ \pm \ z_{1-\alpha/2} \times \hat{se}\left(\hat{\beta}\right)$$

$$\text{Test stat}: \qquad Z = \frac{(estimate) - (null)}{(std\ err)} \qquad Z = \frac{\hat{\beta} - \beta_0}{\hat{se}\left(\hat{\beta}\right)}$$

60

Technical Details

- Unlike linear regression, there is no closed form expression to find the Poisson regression parameter estimates
- Instead, computer programs use an iterative search
- This search may fail in saturated or nearly saturated models if some parameter corresponds to a group having no events
  - In this setting, Poisson regression parameters modeling the log mean are trying to estimate negative infinity
  - The sample size is too small for the model

61

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

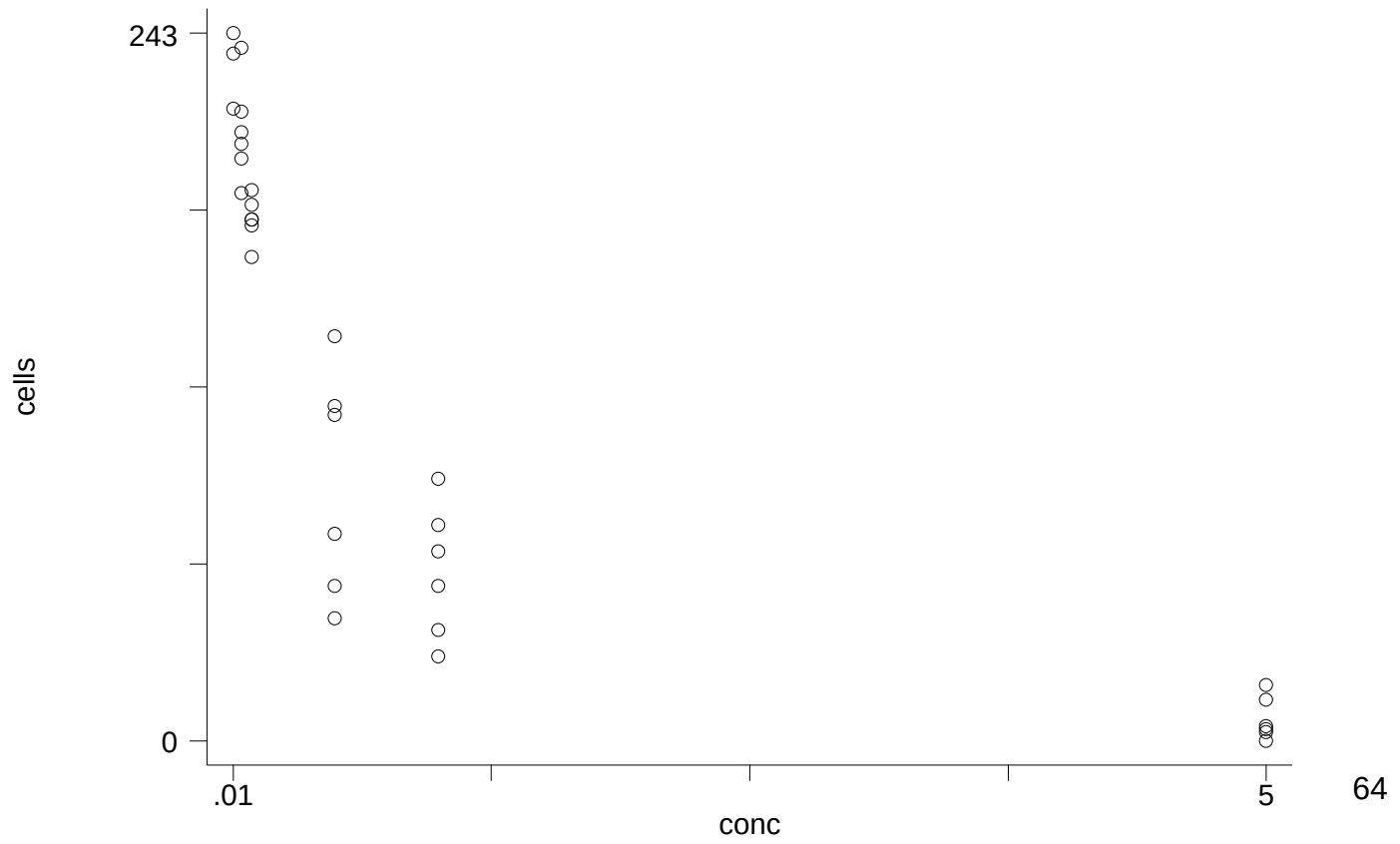
62

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

63

Scatterplot Cells vs Dox Conc

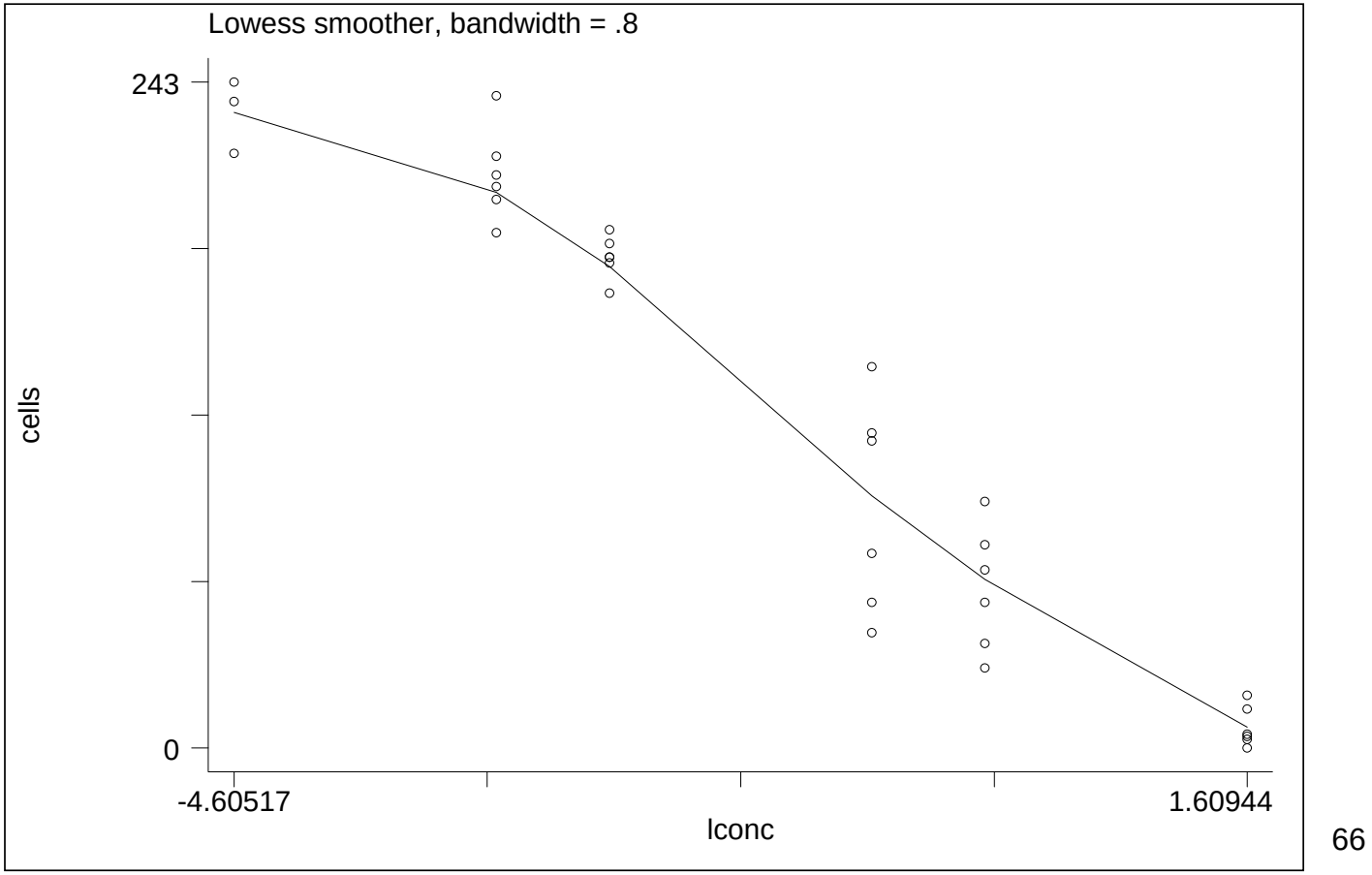


Characterization of Scatterplot

- 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

65

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)

67

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
------------------------------------------------------------------------------
```

68

Interpretation of Stata Output

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

- 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

69

Interpretation of Intercept

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

- 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

70

Interpretation of Slope

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

- 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
- 95% CI for 10 fold increase in conc: 56.3% to 57.5% decrease
  - $10^{-0.372} = 0.425$, $10^{-0.360} = 0.437$

71

Example: Interpretation

- "From Poisson regression analysis, we estimate that for each 10 fold increase in concentration of doxorubicin, the probability of cell survival decreases by 56.9%, highly statistically significant observation (P < 0.0001). A 95% CI suggests that this observation is not unusual if a cell culture exposed to a 10-fold higher concentration of doxorubicin might have a cell survival rate that was decreased anywhere from 56.3% to 57.5% compared to the lower concentration group."

72

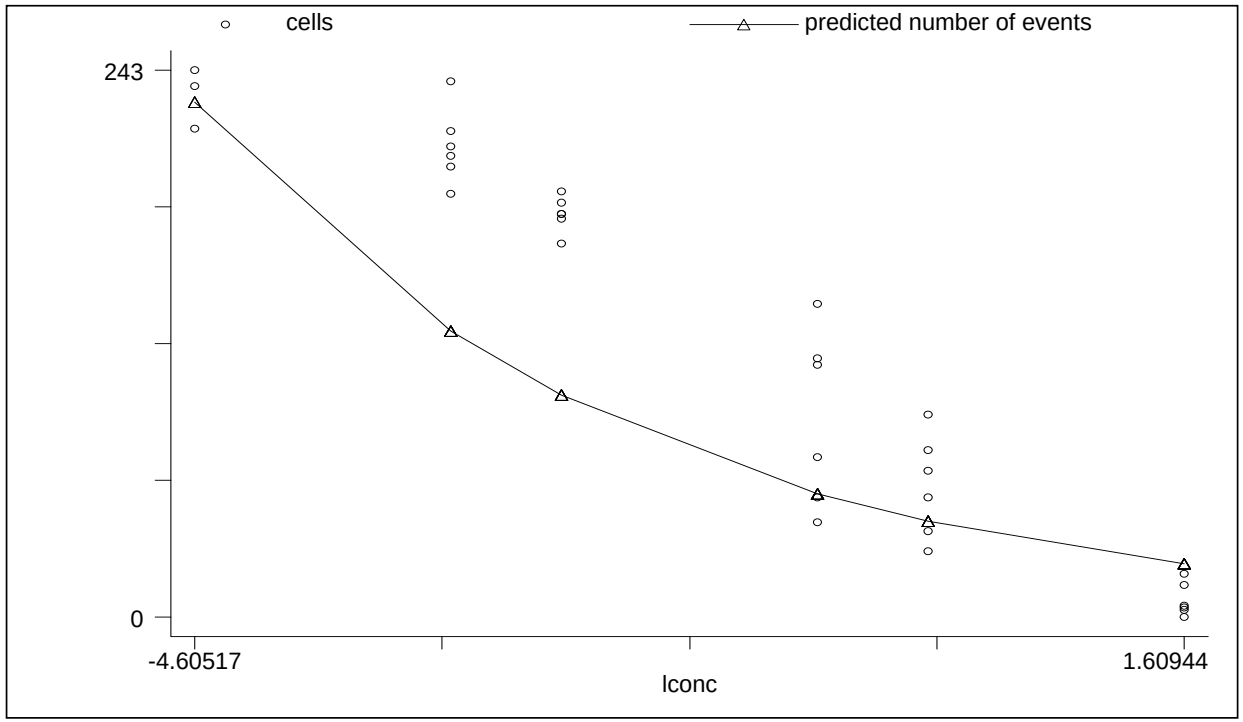
Role of Linearity

- 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

73

Fitted Regression Model

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



74

Role of Linearity

- The true goal of this experiment was to estimate the concentration at which 50% of cells might be expected to die
  - The LC_{50}
- There significance of the slope was taken for granted
- The lack of linearity means that we cannot trust the Poisson regression with a linear term to provide the estimate we want
  - We are trying to make estimate of each group's mean
- In real life, I fit a nonlinear curve to this data
  - That curve was based on the S-shaped curve that Michaelis-Menten kinetics would suggest

75

Disease Incidence Rates

- Poisson regression is frequently used to investigate disease incidence rates
  - Incidence rate ratios are then the target of inference
- For a specified combination of covariates, data might consist of
  - Total person-years of observation while at risk
  - Number of events observed within that time period
- Often such data is grouped into age intervals, with a particular individual contributing person-years to multiple strata
  - This can be used as a “piecewise exponential” survival analysis

76

Wider Use of Poisson Regression

- More generally, Poisson regression can be used to model means on a multiplicative scale in a wide variety of settings
  - Exponentiated slope estimates are mean ratios
  - (cf: linear regression on log *Y* as geometric mean ratios)
- Because classical Poisson regression presumes a particular mean-variance relationship, robust SE should usually be used to remove that assumption
- With Bernoulli data having *low* event rates, the Poisson approximation to the Binomial justifies the use of Poisson regression, and the mean-variance relationship is correct
  - Bernoulli: Var (Y) = p (1 – p)
  - Poisson: Var (Y) = p

77

Example: CVA vs Age

- We can compare logistic regression to Poisson regression on analyses investigating an association between cerebrovascular accidents (CVA – stroke) and age
  - Logistic regression has odds ratio (OR) as target of inference
  - Poisson regression has risk ratio (RR) as target of inference
- In this data, 75 of 735 subjects have had a history of CVA
  - Overall estimated probability is 0.1136
  - Overall estimated odds is 0.1136 / (1 - 0.1136) = 01282
  - This incidence is not so low as to regard that OR and RR will be exactly the same
  - I use robust SE in the Poisson regression model

78

Example: CVA vs Age

```
. poisson cva age, robust irr
Poisson regression                                Number of obs   =        735
                                                  Wald chi2(1)    =       2.61
                                                  Prob > chi2     =     0.1064
Log pseudolikelihood = -245.08706                 Pseudo R2       =     0.0044

             |               Robust
         cva |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Age          |   1.030208   .0189904     1.61   0.106     .9936513    1.068109
```

```
. logistic cva age
Logistic regression                               Number of obs   =        735
                                                  LR chi2(1)      =       2.45
                                                  Prob > chi2     =     0.1175
Log likelihood = -240.98969                       Pseudo R2       =     0.0051

         cva | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.034134   .0217632     1.59   0.111     .9923468    1.077681
```