Biost 536 / Epi 536
Categorical Data Analysis in Epidemiology

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

Lecture 12:
Regression Based Inference

November 12, 2013

1

Lecture Outline

- Scientific Questions Addressed With Regression
  - Clustering of Cases
  - Clustering of Variables
  - Estimation of Summary Measures
  - Associations Between Variables
  - Prediction
- Alternative Classification
  - Unsupervised exploration
  - Supervised prediction
  - Variable importance
- Validity of Statistical Inference

2

Statistical Questions: Classification

- Clustering of observations
- Clustering of variables
- Comparing distributions
- Quantification of distributions
- Prediction of individual observations

3

1. Cluster Analysis

- Focus is on identifying similar groups of observations
- Divide a population into subgroups based on patterns of similar measurements
- All variables treated symmetrically
  - No delineation between outcomes and groups
- “Unsupervised learning”

4

1. Cluster Analysis: Examples

- Classic example: Identifying species of irises
- Clinical patterns of diabetes at diagnosis
  - Age at diagnosis; weight loss; insulin levels; autoantibodies
- Clinical patterns of weight gain by age
  - Rapid weight gain vs steady weight vs cycling
- Identifying individuals with similar patterns of gene expression
  - Subtypes of cancer
- Identifying individuals with similar patterns of pharmacokinetics
- Identifying individuals within social networks
  - Spread of epidemics

5

1. Cluster Analysis: Categorical Data

- The goal of cluster analysis is usually the identification of "latent" categorical variables
  - We sometimes attach labels (e.g., "JODM" vs "AODM")
  - (When specifically interested in identifying the cluster to which a given subject belongs, we may use terms like "classification")
- When using only binary data to find the clusters, we are pretty much limited to talking about "concordance" and "discordance"
  - Sometimes useful to think about clusters defined by
    - "AND" (must have both indicators), or
    - "OR" (must have at least one of the indicators)
- When using mixed binary and continuous data, we tend to think more about means and correlations

6

1. Cluster Analysis: Methods Overview

- Cluster membership: Many options
  - Each observation belongs to exactly one cluster
  - Each observation has a probability of belonging to each cluster
  - Some observations may belong to no cluster ("outliers")
  - Hierarchical clustering: defines "distance" between all pairs
    - User can choose level of detail, and hence number of clusters
- Known or unknown number of clusters
- Clustering criteria (partial list)
  - Denseness of data
  - Distance from center of cluster
  - Similarity of covariance among variables
  - Shape of distributions (mixture models)

7

1. Cluster Analysis: Methods Overview

- Usually "unsupervised" clustering using
  - Measured covariates: Univariate, multivariate
  - Derived variables:
    - Contrasts across measurements (BMI, ankle-arm index, PSA acceleration)
    - Averages across measurements (area under curve (AUC))
- Inference about clustering is generally difficult
  - How should we measure the reproducibility of assigning individuals to a cluster?
  - Repeated experiments may end up with very different clusters
    - Labeling of clusters is itself a random variable
- Typically, cluster analysis is used only in the earliest exploratory analyses for a given area of investigation

8

2. Clustering Variables

- Identifying hidden variables indicating groups that tend to have similar measurements of some outcome
  - A "latent variable"
- Predictors that imprecisely measure some abstract quality
- Desire to find patterns in predictors that more precisely reflect the abstract quality

9

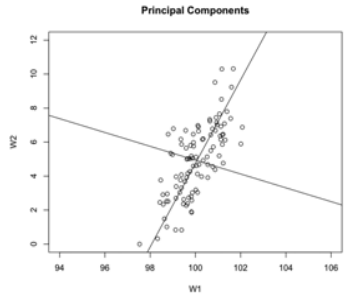
2. Clustering Variables: Examples

- "Personality types" from psychological testing
- Measures of disease severity (e.g., liver function tests, drug toxicity)
- Biochemical pathways
  - Co-expression of genes

10

2. Clustering Variables: Methods

- Most methods derive from "principal components" which looks at linear combinations of variables
  - Finds the linear combinations that have greatest variability among individuals
  - Has roots in continuous data (and multivariate normal)



11

2. Clustering Variables: Methods

- In "factor analysis" we often consider the linear combinations that have the greatest variability and then try to give a name to that quantity based on the variables that cluster together
  - We are most often interested in some sort of "data reduction"
  - Which variables contribute the most to any given "factor"?
- Inference is difficult
  - Contributions of each variable to a single "principal axis" can vary
  - Order of the "principal axes" can vary
  - Hence, difficult to describe the variation from experiment to experiment
- Typically, such clustering of variables is considered during early stages of scientific investigation in a specific area

12

1-2. Clustering Observations & Variables

- With "big data" problems, there has been a lot of interest in analyses that merge questions of my type 1 and 2.
- Simultaneously consider the clustering of observations and the clustering of variables

13

Example: Netflix

- Combination of clustering cases and variables
- Measure movie preferences on all customers
- Make recommendations based on
  - Similarity of movies?
    - The movies cluster: *Rocky* is like *Pretty Baby*
  - Similarity of customers?
    - Customers who like *Rambo* also like *The Three Stooges*
  - (Does it matter to Netflix which is which?)

14

Example: Genomics/Proteomics

- Combination of clustering cases and variables
- Measure expression of 10,000 genes on (usually small) number of patients
- Identify genes that tend to act the same way across patients
  - Pathways?
- Identify groups of patients that tend to have the same patterns of gene expression
  - Subtypes of disease?
  - Similar behaviors that lead to the disease?
  - Similar behaviors in response to the disease?

15

3. Quantifying Distributions

- Focus is on distributions of measurements within a population
- Scientific questions about tendencies for specific measurements within a population
  - Point estimates of summary measures
  - Interval estimates of summary measures
    - Quantifying uncertainty
  - Decisions about hypothesized values
- May desire estimates within subgroups
  - E.g., estimates by sex, age, race

16

3. Quantifying Distributions: Example

- Sample of patients newly diagnosed with stage II breast cancer
  - Follow for survival time (may be censored)
- Scientific question: Some summary measure for distribution
  - Survival probability at a fixed point in time
  - Median survival time (or some other quantile)
  - Mean survival time (restricted mean due to censoring)
  - Average hazard
- Statistical analysis
  - Best estimate of the median survival (K-M?)
  - Quantify uncertainty in that estimate
  - Compare to some clinically important time range (e.g., 10 years)

17

3. Quantifying Distributions: Categorical

- With categorical data, there are limited summary measures that we might be interested in
- Binary data: Proportions (probability) or odds
- Poisson data: Event rate parameter
- Time to event data: (Average) hazard over age, time
  - Directly standardized rates

18

4. Comparing Distributions

- Comparing distributions of measurements across populations
- 4a. Identifying groups that have different distributions of some measurement
- 4b. Quantifying differences in the distribution of some measurement across predefined groups (effects or associations)
- 4c. Quantifying differences in effects across subgroups (interactions or effect modification)

19

4a. Identifying Groups

- Identifying groups that have different distributions of some measurement
- Focus is on some particular outcome measurement
- Identify groups based on other measurements
  - E.g., exploring dose response relationships
  - E.g., quantifying distributions within subgroups
  - E.g, stepwise regression models to identify most promising predictors
- (cf: Cluster analysis where all measurements are treated symmetrically)

20

4a. Identifying Groups: Example 1

- Many reports in the scientific literature report increased risk of CVD in subjects with higher serum cholesterol
- Our analysis of the Cardiovascular Health Study data suggests a linear trend among 65-100 yo in which subjects with higher serum cholesterol trend toward higher probability of survival
- We can consider more flexible models to differentiate between
  - Data that is not at all consistent with those prior studies, and
  - A curvilinear relationship that might give different trends according to the distribution of cholesterol levels (and other factors)

21

4a. Identifying Groups: Example 1

- Statistical Tasks:
- Cohort study of elderly
  - Measure serum cholesterol
  - Measure mortality over time
- Statistical analysis
  - Selected regression models of varying flexibility
    - Polynomials, dummy variables, splines
    - (Use p values to rank interest in particular curves?)

22

4a. Identifying Groups: Example 2

- Chromosomal abnormalities associated with ovarian cancer
- Cytogenetic analysis of dividing cells identifies regions of the chromosomes with defects
  - Cancer is caused by some defects, and cancer causes other defects
  - Approximately 370 identifiable regions
- Which of the regions are the most promising to explore in more focused studies?

23

4a. Identifying Groups: Example 2

- Statistical Tasks:
  - Sample of cancer tissues
    - Measure type of cancer (ovarian, melanoma, etc.)
    - Measure chromosomal defects
  - Statistical analysis
    - Stepwise regression models of chromosomal abnormalities predicting cancer type
      - (Use p values to rank interest in particular regions?)

24

4a. Identifying Groups: Example 3

- Risk factors for diabetes
  - Variables most associated with diabetes risk may give clues about etiology and eventual prevention

25

4a. Identifying Groups: Example 3

- Statistical Tasks
  - Sample subjects to measure risk factors and disease prevalence
    - Cohort study
    - Case-control study
  - Statistical analysis
    - Stepwise model building
      - (Rank most interesting variables by p value?)

26

4a. Identifying Groups: Comments

- A key aspect that I have considered in questions of my type 4a is that of credibility of P values and inference
- The examples that I gave here were quite exploratory in nature
  - The questions related more to screening for the most interesting hypotheses to later confirm
  - It is very difficult to control type 1 errors and power when engaging in “model building”

27

4b. Detecting Associations

- Associations between variables – distributions of one variable differ across groups defined by another
  - Existence of differences
  - Direction of tendency of effect
  - First, second order relationships in a summary measure
  - Characterization of dose-response in a summary measure

28

Definition of an Association

- The distributions of two variables are not independent
  - Independence: Equivalent definitions
    - Probability of outcome and exposure is product of
      - Overall probability of outcome, and
      - Overall probability of exposure
    - Distribution of exposure is EXACTLY the same across ALL outcome categories
    - Distribution of outcome is EXACTLY the same across ALL exposure categories

29

Summary Measures

- Generally we consider some summary measure of the distribution
  - For instance, when we use the mean, we show an association by showing either
    - Mean outcome differs across exposure groups
    - Mean exposure differs across outcome groups

30

Justification

- This works, because if two distributions are the same, ALL summary measures should be the same
  - If some summary measure is different, then we know the distributions are different
- HOWEVER: This means that it is easier to prove an association, than to prove no association

31

Example: Detecting Association

- Effect of blood cholesterol levels on risk of heart attacks
  - Understanding etiology of heart attacks may lead to prevention and/or treatment strategies

32

4b. Detecting Association: Example 1

- Statistical tasks
  - Measure risk factors, MIs on sample
    - Cohort or case-control sample
  - Statistical analysis
    - Regression model (possibly adjusted as pre-specified)
      - Cohort: Incidence of MIs across cholesterol levels
      - Case-control: Cholesterol levels across MI status
      - (Comparison can be at many levels of detail)
    - Quantify estimates, precision, confidence in decisions

33

4b. Detecting Associations: Example 2

- Many reports in the scientific literature report increased risk of CVD in subjects with higher serum cholesterol
- Based on our prior understanding that the elderly with poor liver function will tend to have low cholesterol, we want to investigate whether the association between mortality and serum cholesterol is nonlinear
- We consider a function that models cholesterol as a quadratic

34

4a. Identifying Groups: Example 1

- Statistical Tasks:
- Cohort study of elderly
  - Measure serum cholesterol
  - Measure mortality over time
- Statistical analysis
  - We fit a quadratic curve (possibly with other pre-specified adjustment)
  - We use the p value for the squared term to test our hypothesis

35

4c. Detecting Effect Modification

- Quantifying differences in effects across subgroups (interactions or effect modification)
  - Existence of interaction
  - Direction of interaction (synergy, antagonism)
  - Quantification of exact relationship of interaction

36

Example: Effect Modification

- Identifying whether effect of cholesterol on heart attacks differs by sex
  - Comparing association between blood cholesterol level and incidence of heart attacks between sexes
    - Quantify association in men
    - Quantify association in women
    - Compare measures of association

37

Approach Common to #3 & #4

- Inference based on some summary measure of a distribution
  - #3: Estimate the summary measure
  - #4: Compare and contrast summary measures
- In answering each scientific question, statistics typically provides four numbers
  - Best estimate
    - "Best" can be defined by frequentist or Bayesian criteria
  - Interval describing precision
    - Confidence interval or Bayesian credible interval
  - Quantification of belief in some hypothesis
    - P value or Bayesian posterior probability

38

Example: Detecting Association

- Association between sex and prevalence of MI in elderly population
  - 59 of 366 males have had MI: 16.1%
  - 32 of 367 females have had MI: 8.7%
  - Association measured by difference
    - Best estimate: Prevalence 7.4% higher in males
    - Interval estimate: Between 2.7% and 12.2%
      - (95% confidence interval)
    - Strength of evidence: P value = 0.002
      - If there were no real difference, the observed data is pretty unlikely: Probability of this data is 0.002

39

5. Prediction

- Focus is on individual measurements
  - Point prediction:
    - Best single estimate for the measurement that would be obtained on a future individual
      - Continuous measurements
      - Binary measurements (discrimination)
  - Interval prediction:
    - Range of measurements that might reasonably be observed for a future individual

40

5. Prediction

- Focus is on individual measurements
- Point prediction:
  - Best single estimate for the measurement that would be obtained on a future individual
    - Continuous measurements
    - Binary measurements (discrimination)
- Interval prediction:
  - Range of measurements that might reasonably be observed for a future individual

41

Example: Continuous Prediction

- Creatinine clearance
  - Creatinine
    - Breakdown product of creatine
    - Removed by the kidneys by filtration
      - Little secretion, reabsorption
  - Measure of renal function
    - Amount of creatinine cleared by the kidneys in 24 hours

42

Example: Continuous Prediction

- Problem:
  - Need to collect urine output (and blood creatinine) for 24 hours
- Goal:
  - Find blood, urine measures that can be obtained instantly, yet still provide an accurate estimate of a patient's creatinine clearance

43

Example: Continuous Prediction

- Statistical Tasks:
  - Training sample
    - Measure true creatinine clearance
    - Measure sex, age, weight, height, creatinine
  - Statistical analysis
    - Regression model that uses other variables to predict creatinine clearance
    - Quantify accuracy of predictive model
      - (Mean squared error?)

44

Example: Discrimination

- Diagnosis of prostate cancer
  - Use other measurements to predict whether a particular patient might have prostate cancer
    - Demographic: Age, race, (sex)
    - Clinical: Symptoms
    - Biological: Prostate specific antigen (PSA)
  - Goal is a diagnosis for each patient

45

Example: Discrimination

- Statistical Tasks:
  - Training sample
    - "Gold standard" diagnosis
    - Measure age, race, PSA
  - Statistical analysis
    - Regression model that uses other variables to predict prostate cancer diagnosis
    - Quantify accuracy of predictive model
      - (ROC curve analysis?)

46

Example: Interval Prediction

- Determining normal range for PSA
  - Identify the range of PSA values that would be expected in the 95% most typical healthy males
  - Age, race specific values

47

Example: Interval Prediction

- Statistical Tasks:
  - Training sample
    - Measure age, race, PSA
  - Statistical analysis
    - Regression model that uses other variables to define prediction interval
      - (Mean plus/minus 2 SD?)
      - (Confidence interval for quantiles?)
    - Quantify accuracy of predictive model
      - (Coverage probabilities?)

48

Comment About Prediction

- Prediction often requires more understanding about the problem than does detecting associations
  - Relies more heavily on assumptions
- For me to consider a problem to be purely a prediction problem, interest must lie solely in the predicted value, and not in the way that value was obtained
  - E.g., in weather prediction, we might just want to know the weather tomorrow
    - We won't be trying to impress upon our audience the way it should be predicted
  - I do not think this is very often the case

49

Alternative Categorization of Questions

- Exploratory
  - Unsupervised: All variables treated identically
    - Cluster analysis
    - Factor analysis
  - Supervised: A response variable is pre-specified
    - (Stepwise) Model building
    - Classification and regression trees; Machine learning
- Confirmatory
  - Variable importance: (Regression) Analyses with pre-specified
    - Analysis model
    - Variable included as predictors and form of the variables
    - The specific target of inference
  - Predictive models
    - Where target is the prediction and accuracy is evaluated through cross-validation or out-of-sample predictions

50

Regression Based Inference

51

Two Variable Setting

- Many statistical problems consider the association between two variables
  - Response variable
    - (outcome, dependent variable)
  - Grouping variable
    - (predictor, independent variable)

52

Addressing Scientific Question

- Compare the distribution of the response variable across groups that are defined by the grouping variable
  - Within each group, the value of the grouping variable is constant

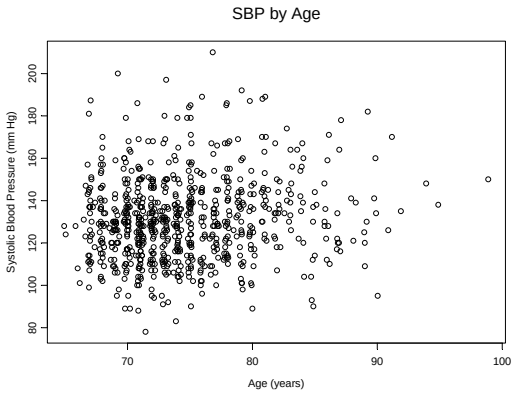
53

Intro Course Classification

- Characterize statistical analyses by
  - Number of samples (groups), and
  - Whether subjects in groups are independent
- Correspondence with two variable setting
  - By characterization of grouping variable
    - Constant: One sample problem
    - Binary: Two sample problem
    - Categorical: k sample problem (e.g., ANOVA)
    - Continuous: Infinite sample problem
      - Regression

54

Example: SBP and Age



55

Regression Methods

- Regression extends one and two sample statistics (e.g., the t test) to the infinite sample problem
  - While we don't really ever have (or care) about an infinite number of samples, it is easiest to use models that would allow that in order to handle
    - Continuous predictors of interest
    - Adjustment for other variables

56

Regression vs Two Samples

- When used with a binary grouping variable common regression models reduce to the corresponding two variable methods
  - Linear regression with a binary predictor
    - Classical: t test with equal variance
    - Robust SE: t test with unequal variance (approx)
  - Logistic regression with a binary predictor
    - Score test: Chi squared test for association
  - Cox regression with a binary predictor
    - Score test: Logrank test

57

Guiding Principle

"Everything is regression."

- Scott Emerson

58

Uses of Regression

- Two major uses of regression
- Borrow information to address "sparse data" in some groups
  - E.g., 68 and 70 year olds provide information about 69 year olds
  - Question: How far away do you want to go?
- Provide a statistical "contrast" to compare distribution of response across groups
  - Think of a "slope" as an average comparison of summary measures per unit difference in the grouping variable

59

Regression Inference

- Estimates
  - Slope: (average) contrasts across groups
  - Fitted values: estimated summary measure in a group
- Standard errors
- Confidence intervals
- P values testing for
  - Intercept of zero (who cares?)
  - Slope of zero (test for linear trend in summary measures)

60

Robust Standard Errors

- I have recommended the use of robust standard errors
- Relaxes assumptions about variance of data within groups
- Allows tests of weak null hypotheses
  - Statements about equality of summary measures rather than equality of entire distributions
- Later: Allows regression with correlated data

61

Simple Linear Regression

62

Interpretation

- Interpretation of "regression parameters"
  - Intercept β_0: Mean Y for a group with X=0
    - Quite often not of scientific interest
      - Often outside range of data, sometimes impossible
  - Slope β_1: Difference in mean Y across groups differing in X by 1 unit
    - Usually measures association between Y and X

$$E(Y \mid X) = \beta_0 + \beta_1 \times X$$

63

Derivation of Interpretation

- Simple linear regression of response Y on predictor X
  - Mean for an arbitrary group derived from model
  - Interpretation of parameters by considering special cases

Model	$E[Y_i \mid X_i] = \beta_0 + \beta_1 \times X_i$
$X_i = 0$	$E[Y_i \mid X_i = 0] = \beta_0$
$X_i = x$	$E[Y_i \mid X_i = x] = \beta_0 + \beta_1 \times x$
$X_i = x+1$	$E[Y_i \mid X_i = x+1] = \beta_0 + \beta_1 \times x + \beta_1$

64

Example: Mental Function by Age

- Cardiovascular Health Study
  - A cohort of ~5,000 elderly subjects in four communities followed with annual visits
    - A subset of 735 subjects all followed for at least 4 years
  - Mental function measured at baseline by Digit Symbol Substitution Test (DSST)
  - Question: How is performance on DSST associated with mortality

65

Statistical Validity of Inference

- Inference (CI, P vals) about associations requires three general assumptions
- Approximate normal distribution for estimates
  - Normal data or large N
- Assumptions about independence of observations
  - Independence or identified clusters
- Assumptions about variance of observations within groups
  - Robust SE: relaxes requirement for equal variance

66

Prediction of Group Means

- Additional assumption about adequacy of linear model for prediction of group means with linear regression
- Classically OR robust standard error estimates:
  - The mean response in groups is linear in the modeled predictor
    - (We can model transformations of the measured predictor)

67

Prediction Intervals

- Inference (prediction intervals) about individual observations in specific groups has still another assumption
  - Assumption about distribution of errors within each group
    - Normally distributed errors

68

Ex: Classical Standard Errors

```
. regress deadin5 dsst

      Source |       SS       df       MS              Number of obs =     713
-------------+------------------------------           F(  1,   711) =   36.21
       Model |   4.6080903     1   4.6080903           Prob > F      =  0.0000
    Residual |  90.4830738   711  .127261707           R-squared     =  0.0485
-------------+------------------------------           Adj R-squared =  0.0471
       Total |  95.0911641   712  .133555006           Root MSE      =  .35674

------------------------------------------------------------------------------
     deadin5 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dsst |  -.0063776   .0010599    -6.02   0.000    -.0084584   -.0042968
       _cons |   .4210135   .0456276     9.23   0.000     .3314325    .5105945
------------------------------------------------------------------------------
```

69

Ex: Robust Standard Errors

```
. regress deadin5 dsst, robust

Linear regression                               Number of obs =     713
                                                F(  1,   711) =   31.62
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0485
                                                Root MSE      =  .35674

------------------------------------------------------------------------------
             |               Robust
     deadin5 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dsst |  -.0063776   .0011342    -5.62   0.000    -.0086044   -.0041508
       _cons |   .4210135   .0529401     7.95   0.000     .3170758    .5249511
------------------------------------------------------------------------------
```

70

Interpretation of Intercept

$$E\left[deadin5_i \mid dsst_i\right] = 0.421 - 0.00638 \times dsst_i$$

- Estimated 5 year mortality for DSST of 0 is 42%

71

Interpretation of Slope

$$E\left[deadin5_i \mid dsst_i\right] = 0.421 - 0.00638 \times dsst_i$$

- Estimated difference in 5 year mortality for two groups differing by one point in DSST is -0.00638, with higher scoring group averaging a lower mortality rate
  - For 10 point DSST difference: 10 x -0.00638 = - .0638
  - For 100 point DSST difference: - 0.638
- This suggests that a straight line relationship is not true
  - Somebody scoring perfectly would be estimated to have negative mortality
- If a straight line relationship is not true, we interpret the slope as an average difference in mortality per one year difference in DSST)

72

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"

73

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)

74

Simple Logistic Regression

- Modeling odds of binary response Y on predictor X

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

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

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

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

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

75

Interpretation as Odds

- Exponentiation of regression parameters

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

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

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

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

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

76

Estimating Proportions

- Proportion = odds / (1 + odds)

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

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

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

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

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

77

Simple Logistic Regression

- Interpretation of the model
  - Odds when predictor is 0
    - Found by exponentiation of the intercept from the logistic regression: exp(β0)
  - 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(β1)

78

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

79

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

80

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

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

81

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."

82

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)

83

Simple Poisson Regression

Inference About Rates

84

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

85

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

86

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)

87

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 $k = 0, 1, 2, 3, 4, \ldots$
  - Mean $E(Y) = \lambda t$; variance $Var(Y) = \lambda t$
- Poisson approx to Binomial for low p

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

88

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"

89

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)

90

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

91

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

92

Simple Poisson Regression

- Modeling rate of count response Y on predictor X

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

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

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

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

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

93

Interpretation as Rates

- Exponentiation of parameters

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

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

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

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

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

94

Simple Poisson Regression

- Interpretation of the model
  - Rate when predictor is 0
    - Found by exponentiation of the intercept from the Poisson regression: exp(β0)
  - 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(β1)

95

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

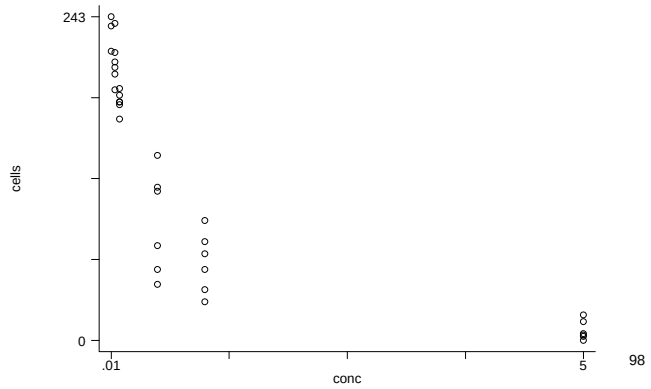
96

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

97

Scatterplot Cells vs Dox Conc



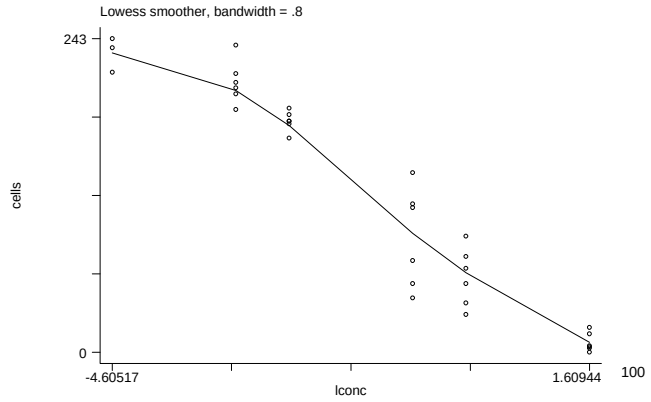
98

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

99

Scatterplot: Cells vs log (Conc)



100

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)

101

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]

102

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

103

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

104

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

105

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

106

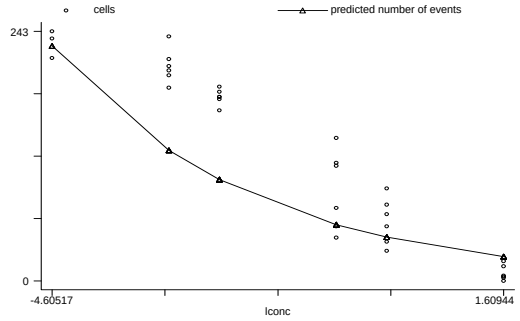
Role of Linearity

- 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

107

Fitted Regression Model

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



108