Biost 518 / Biost 515
Applied Biostatistics II / Biostatistics II

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

Lecture 1:
Course Structure;
Regression Setting

January 5, 2015

1

The Use of Statistics to Answer Scientific Questions

Scott S. Emerson, M.D., Ph.D.

Professor of Biostatistics

University of Washington

2

General Philosophy

"Everything should be as simple as possible, but no simpler."

- A. Einstein (paraphrased)

3

Lecture Outline

- Course Structure
- Course Overview
- Regression Setting

4

Course Structure

5

Course Structure

- Instructor: Scott S. Emerson, M.D., Ph.D.
  - » Fair warning
- TAs:
  - LaNae Schaal
  - Jon Fintzi
- Time and Place:
  - Lectures: 9:30 - 10:20 am MWF HST 747
  - Data Analysis:
    - 8:30 - 9:20 am M HST 478
    - 8:30 - 9:20 am W HST 478
    - 8:30 - 9:20 am F HST 531

6

Assumed Prior Knowledge

- This course covers multiple regression
  - Linear regression
  - Logistic regression
  - Poisson regression
  - Proportional hazards regression
- Equivalent of Biost 514/517
  - Descriptive statistics (complete and censored data)
  - One sample inference:
    - Means, geometric means, proportions, medians
  - Two sample inference
    - Means, geometric means, proportions, odds, medians, hazard ratios
- Or permission of instructor

7

Recording of Lectures

- Camtasia
  - Audio and computer video on web
  - Posted approximately 24 hours after class

- No guarantees: “Mistakes happen”
  - However, lectures from 2014 are also posted on the web pages

8

Textbooks

- Optional references
  - Vittinghoff, Glidden, Shiboski, McCulloch: *Regression Methods…*
  - Kleinbaum, Kupper Muller, Nizam: *Applied Regression…*
  - Kleinbaum: *Logistic Regression…*
  - Kleinbaum: *Survival Analysis…*

9

Computer Software

- Extensively used for data analysis

- Students may use any program that will do what is required, however
  - R is free and we are starting to move toward it for that reason
    - R functions will be provided on the website
  - Historically, Stata was used in Biostat 536, 537, 540
  - R, Stata commands will be provided in lecture
  - Help will presume the use of R or Stata

10

R

- The ultimate in flexible statistical languages
  - Interactive
  - Many user-supplied functions
- Graphical functions generally very good
- Open source and free
- A collection of functions that we have written will be made available on the webpages
  - Supplementary info on web page

11

Stata

- Extremely flexible statistical package
  - Interactive
  - Excellent complement of biostatistical methods
- Graphical, report capabilities suboptimal
- Available in microcomputer lab
- Supplementary info on web page
- Syntax introduced in lectures as needed

12

Computer Software: Comments

- Designed for people who know statistics, but do not want to write basic functions
- Tries to be all things to all people
  - Much output that you will not want
  - Much output that I will recommend against

13

Guiding Principles

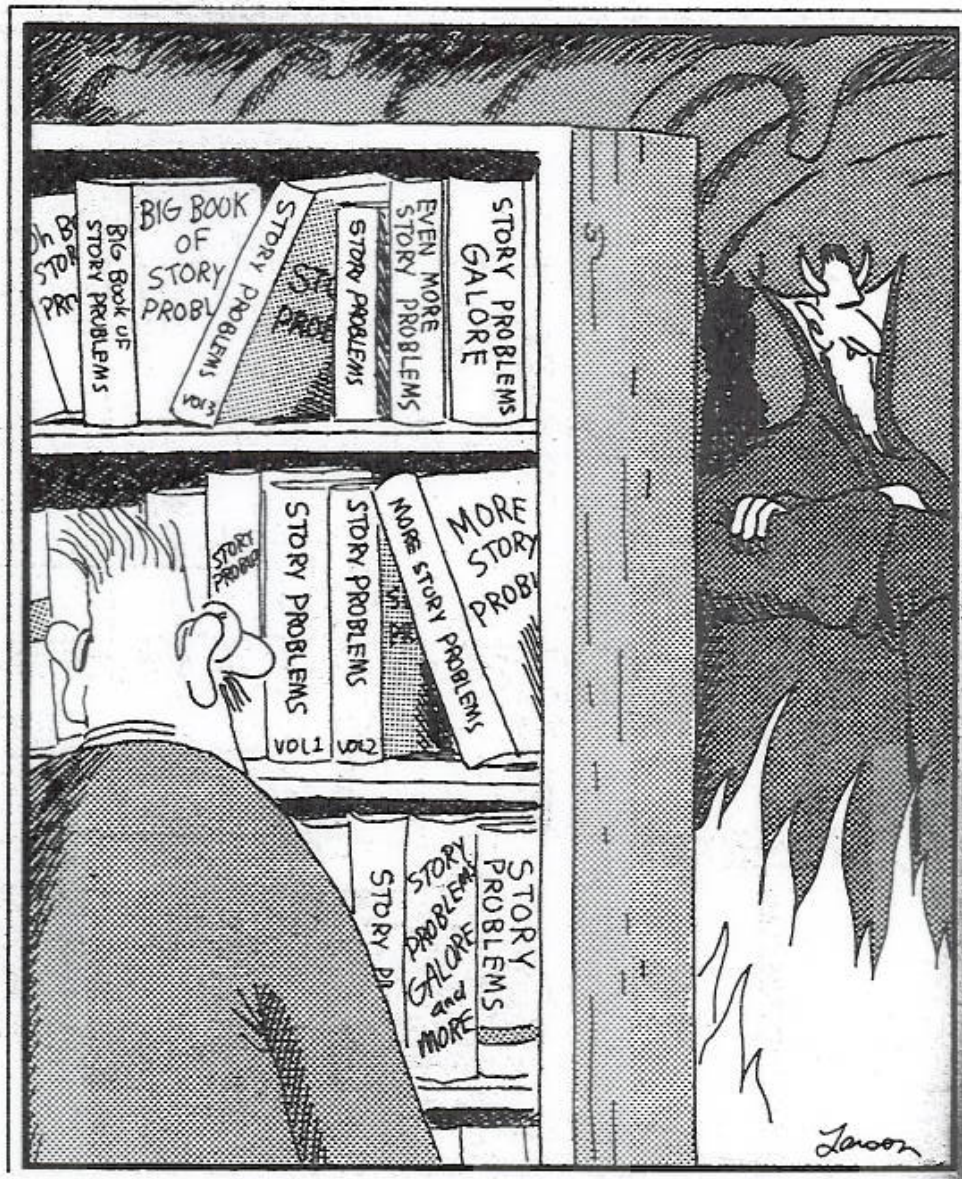
- This is a course in biostatistics, not statistical software
- I will tell you how you can get the statistics I teach you to use
  - There are often multiple ways
  - I tend to teach one of them
- I will not explain every number that appears on the printout

14

Written Homeworks

- Homeworks approximately every week: analysis of real data
  - Questions directed toward specific analyses
    - But questions will still be stated in as scientific terms (as opposed to statistical) terms as possible
  - Work handed in is expected to be organized scientifically
    - I expect nicely formatted tables, figures
    - Unedited Stata / R code is totally unacceptable
- Homeworks must be submitted:
  - Electronically
  - On-time (exceptions only in the most dire circumstances)
  - Anonymized
    - You must use ID codes supplied for each homework
    - It is your responsibility to ensure no indication of your identity as document author, etc.

15



Hell's library

16

Errors to Avoid

- Unedited Stata / R output is TOTALLY unacceptable
- Any assignments that are handed in should be only your work
- Electronically submitted homeworks should be anonymized
  - Use ID code provided to you
  - Remove identifying information from your files
  - **Name file appropriately**
- Submission of homeworks and grades must be on time.

17

Peer Grading

- Keys to the homeworks will be generally be available on the web pages after the deadline for submission
  - My answers will typically go beyond what I expected you to do
    - Extra information will be identified in special fonts
    - You are responsible for any new information that I provide in the homework keys, even if that information is not otherwise presented in class
    - Annotated Stata / R output will often be included
- Each student is expected to use the key to grade another student's paper
  - Double blind: both submitted homework and comments should be anonymous
  - Appeals of grades are decided by instructor

18

Quizzes and Lecture Discussion

- Periodically a part of lecture will be used to take a brief quiz on the subject matter
  - Basic knowledge / judgment
- Students will hand in written answers that will be graded as part of “Quizzes and discussion”
  - There is no way to make up a missed quiz
  - You are allowed to miss one, no questions asked
- We will then discuss the reasoning that should be used to answer the quiz questions and current homework assignments
  - Participation in the discussion is required

19

Discussion Section

- Data analysis to answer scientific questions
- You will be given a scientific question and a data set which was collected to try to answer that question
  - Setting is more realistic than that which is given on written homeworks
- We will discuss the approach to the whole problem
- Often nothing to hand in, but participation in discussion is required
  - I will often call on students at random
    - It is okay to be wrong, but not okay to be unprepared or inattentive
  - You must inform me if you are attending a different discussion section

20

Grading

- 20% Homeworks and peer grading (approx 9)
- 10% Quizzes and discussion
- 25% One Midterm (in class, closed book)
- 20% Data Analysis and Report (group project, except SAP)
- 25% Final Exam (in class, closed book)

21

Course Web Pages

- Address: www.emersonstatistics.com/b518/
- Content
  - Syllabus (you are responsible for knowing this information)
  - Lecture handouts
  - Recordings of lectures (and discussions if I think it worthwhile)
  - Homework assignments and keys
  - Datasets
  - Supplemental materials not discussed in class
    - Handouts, noteworthy emails

22

Supplementary Materials

- Reporting associations
- Use of logarithms in statistical analyses
- Approach to analysis of a data set
- The general regression model
- Special topics
  - The importance of the normality assumption
  - The Wilcoxon rank sum statistic
  - Adjustment for baseline in RCT
  - Missing data in RCT

23

Course Structure

- Biost 517
  - One response variable; one grouping variable
    - In 2013, binary grouping variable
  - One-, two-, K-sample description and inference
- Biost 518
  - Simple regression
    - Linear, logistic, Poisson, proportional hazards
  - Adjustment for confounding, precision, effect modification
  - Stratified description and inference
  - Multivariable regression
    - Linear, logistic, Poisson, proportional hazards

24

Biost 518 Topics

- Review: Two variable problem
  - Means, geometric means, proportions, odds, hazard ratios
- Simple regression
  - Linear, logistic, Poisson, proportional hazards regression
- Confounding, precision, effect modification
- Stratified analyses

25

Biost 518 Topics

- Multiple regression
  - Models, interpretation of parameters
  - Modeling associations
  - Interactions
  - Time varying covariates; clustered data
  - Prediction
  - Missing data
  - Diagnostics
  - Exploratory models

26

Overview of Setting

Scientific Method

27

Purpose of Statistics

- Statistics is about science
  - (Science in the broadest sense of the word)

- Science is about proving things to people
  - (The validity of any proof rests solely on the willingness of the audience to believe it)

28

First Stage of Scientific Investigation

- Hypothesis generation
- Observation
- Measurement of existing populations
- Disadvantages:
  - Confounding
  - Limited ability to establish cause and effect

29

Further Stages of Scientific Investigation

- Refinement and confirmation of hypotheses
- Experiment: Intervention
- Elements of experiment
  - Overall goal
  - Specific aims (hypotheses)
  - Materials and methods
  - Collection of data
  - Analysis
  - Interpretation; Refinement of hypotheses

30

Do You Need Statistics?

- Two question test (Both must be YES)
- In a deterministic world, do YOU know how to answer your question?
  - Is the question answerable in the real world?
  - How do you use a number to answer the scientific question?
- In a world subject to variation, do YOU know how you would answer your question if you had the entire population?

31

Statistical Tasks

- Understand overall goal
- Refine specific aims (stat hypotheses)
- Materials and methods: Study design
- Collection of data: Advise on QC
- Analysis
  - Describe sample (materials and methods)
  - Analyses to address specific aims
- Interpretation

32

Statistical Questions

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

33

Scientific Hypotheses

- Usual statement: (Cause and Effect)

The intervention (exposure) when given to the target population will tend to result in outcome measurements that are

higher than,	an absolute standard, or
lower than, or	measurements in a comparison group
about the same as	

34

Statistical Hypotheses

- Better statement: (Associations)

An exposed population will tend to have in outcome measurements that are

{ higher than, / lower than, or / about the same as } { an absolute standard, or / measurements in a comparison group }

35

Refining Scientific Hypotheses

- Statistical hypotheses precisely define
  - the intervention (or risk factor)
  - the outcome
    - advise on precision of measurement
  - the target population(s)
    - covariates
  - “tend to” (the standards for comparison)
    - summary measures
    - contrasts (difference or ratio)
    - relevance of absolute or relative standards

36

Statistical Role of Variables

- Statistical hypotheses involve
  - "Response" or "Outcome"
    - Can be either the "effect" or the "cause"
  - "Grouping Variable(s)"
    - Primary scientific question
      - Predictor of interest
      - Effect Modifiers
    - Adjustment for covariates
      - Confounders
      - Precision variables

37

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 the same across all outcome categories

  – Distribution of outcome is the same across all exposure categories

38

Detecting Associations

- It takes an infinite sample size to show exact equality of distributions
  - Hence, we do not try to show no association
- Instead, we detect associations by showing that two variables are not independent
  - Thus, we show that two distributions are different
- Most often we show that some summary measure of the distributions is different
  - This works, because if two distributions are the same, ALL summary measures should be the same
  - Hence, if some summary measure is different, then we know the distributions are different

39

Hierarchy of Null Hypotheses

- Strong Null
  - Distribution of response identical in all groups
- Intermediate Null
  - Summary measure identical in all groups
    - Summary measures on a flat line
- Weak Null
  - No linear trend in summary measure across groups
    - On average, summary measures on a flat line

40

Impact of Study Design

- To establish an association
  - Cohort studies must examine whether
    - $\Pr(O \mid C = c_1) \neq \Pr(O \mid C = c_2)$
  - Case-control studies must examine whether
    - $\Pr(C \mid O = o_1) \neq \Pr(C \mid O = o_2)$
  - Cross sectional studies can examine either of the above, as well as whether
    - $\Pr(O, C) \neq \Pr(O) \times \Pr(C)$

41

Impact of Study Design on Analysis

- We are ultimately interested in
  - Detecting associations (hypothesis testing)
  - Quantifying associations (estimate difference in means, etc.)

- For testing, we can often ignore study design when deciding "response" vs "predictor"
  - P values will generally be similar
- For quantifying magnitude of association, we will most often need to "condition" on the variable with constrained sample sizes
  - Cohort design: condition on "exposure"
  - Case-control design: condition on "disease"
  - (Notable exception: With odds ratios (logistic regression), we can do either)

42

Univariate Summary Measures

- Many times, statistical hypotheses are stated in terms of summary measures for the distribution within groups
  - Means (arithmetic, geometric, harmonic, …)
  - Medians (or other quantiles)
  - Proportion exceeding some threshold
  - Odds of exceeding some threshold
  - Time averaged hazard function (instantaneous risk)
  - …

43

Comparisons Across Groups

- Comparisons across groups then use differences or ratios
  - Difference / ratio of means (arithmetic, geometric, …)
  - Difference / ratio of proportion exceeding some threshold
  - Difference / ratio of medians (or other quantiles)
  - Ratio of odds of exceeding some threshold
  - Ratio of hazard (averaged across time?)
  - …

44

Based on Type of Data

- Correspondence to relevance of descriptive statistics
- Binary or dichotomous:
  - mean (proportion); odds
- Nominal (unordered categories):
  - frequencies; odds
- Ordinal (ordered categories):
  - median (quantiles); odds; ? mean
- Quantitative (addition makes sense):
  - mean; median; proportion > c; hazards, …

45

Descriptive Statistics

- Description of a sample
- Identification of measurement or data entry errors
- Characterize materials and methods
- Validity of analysis methods
  - Assess scientific and statistical assumptions
- (Straightforward estimates of effects-- inference)
- Hypothesis generation (inference-- estimation)

46

Descriptive Statistics

	Binary	Unordered	Ordered		
		Nominal	Categ	Quant	Cens
Distribution					
Frequency	OK	OK	OK	OK	
Cum Freq	boring		OK	OK	KM
Mode	boring	Sample	Sample	Density	
Quantiles	boring		OK	OK	KM
Dichotomize Prop / Odds	OK	OK	OK	OK	KM
Means					
Arithmetic	Prop		***	OK	(?KM)
Geometric				OK	
Std Dev	boring			OK	
Others				OK	47

Joint Summary Measures

- Other times groups are compared using a summary measure for the joint distribution
  - Median difference / ratio of paired observations
  - Probability that a randomly chosen measurement from one population might exceed that from the other
  - …

48

Commonly Used Parameters

	Binary	Unordered	Ordered		
		Nominal	Categ	Quant	Cens
Entire Distribution	OK	OK	OK	OK	OK
Proportion	OK	Dichotomize	Dichotomize	Dichotomize	Dichotomize
Odds	OK	Dich	Dichotomize, Prop Odds	Dichotomize	Dichotomize
Median			(OK)	OK	OK
Means					
Arithmetic	Prop		(OK)	OK	
Geometric				OK	
Hazard			(OK)	OK	OK
Pr (Y > X)			(OK)	OK	OK

49

Criteria for Summary Measure

- In order of importance
  - Scientifically (clinically) relevant
    - Also reflects current state of knowledge
  - Is likely to vary across levels of the factor of interest
    - Ability to detect variety of changes
  - Statistical precision
    - Only relevant if all other things are equal

50

Science vs Statistics

- Scientific summary measures
  - Summarize distributions of meaningful measurements
  - Contrasts across populations
    - E.g., a slope

- Statistical measures
  - How precisely we estimate a scientific measure
    - E.g., a P value, correlation

51

Inference

- Generalizations from sample to population
- Estimation
  - Point estimates
  - Interval estimates
- Decision analysis (testing)
  - Quantifying strength of evidence

52

Take Home Message

- Ideal: Always give 4 numbers
  - Point estimate
  - Confidence interval (lower, upper bound)
  - P value
- Even when not significant: Give 4 numbers
  - Arguably, it is more important to provide them when not significant
- Be forewarned
  - My job is to improve science into the future
  - If your advisor or the journals you publish in are wrong on this point, I want to make sure you know they are wrong

53

Biost 517

- We described tests (and sometimes CI) for comparing parameters across groups
- Not all are implemented in statistical software, though with a little work they can be obtained in most software packages
- There are some tests which technically could be applied in certain situations, but it is not very often seen (or recognized)
  - (I have denoted these cases with ?)

54

Two Independent Samples

	Binary	Unordered	Ordered		
		Nominal	Categ	Quant	Cens
Entire Distn	Chi Sq	Chi Sq	Chi Sq	Kol-Sm	Modif Kol-Sm
Diff in Proportion	Chi Sq	Chi Sq	Chi Sq	Chi Sq	KM
Odds Ratio	Chi Sq; Fish Ex	Chi Sq; Fish Ex	Chi Sq; Fish Ex; Prop Odds	Chi Sq; Fish Ex	KM

55

Two Independent Samples

	Binary	Unordrd	Ordered		
		Nominal	Categ	Quant	Cens
Diff in Medians			?(Bstrap)	Bstrap	?(Bstrp)
Median Difference			?(Sign)	?(Sign)	
Ratio of Medians					

56

Two Independent Samples

	Binary	Unordrd	Ordered		
		Nominal	Categ	Quant	Cens
(Diff in) Arithmetic Means (of Diff)	Chi Sq		t test (eq,uneq vrnc)	t test (eq,uneq vrnc)	?(Restr Mean)
(Ratio of) Geometric Means (Ratio)				t test (eq,uneq vrnc) on logs	

57

Two Independent Samples

	Binary	Unordered	Ordered		
		Nominal	Categ	Quant	Cens
Hazard Ratio				Logrank	Logrank
Pr (Y > X)			Wilcox Rnk Sum	Wilcox Rnk Sum	Modif Wilcox
???			?(Wilcox Sgn Rnk)	?(Wilcox Sgn Rnk)	

58

Two Matched Samples

	Binary	Unordered	Ordered		
		Nominal	Categ	Quant	Cens
Entire Distn	McNemar (Sign); Paired t test				
Diff in Proportion	McNemar (Sign); Paired t test	McNemar (Sign); Paired t test	McNemar (Sign); Paired t test	McNemar (Sign); Paired t test	
Odds Ratio	McNemar (Sign); Paired t test	McNemar (Sign)	McNemar (Sign); Paired t test	McNemar (Sign); Paired t test	

59

Two Matched Samples

	Binary	Unordrd	Ordered		
		Nominal	Categ	Quant	Cens
Diff in Medians			?(Bstrap)	Bstrap	
Median Difference			Sign	Sign	
Ratio of Medians			?(Bstrap)	Bstrap	

60

Two Matched Samples

	Binary	Unordrd	Ordered		
		Nominal	Categ	Quant	Cens
(Diff in) Arithmetic Means (of Diff)	McNemar (Sign); Paired t test		Paired t test	Paired t test	
(Ratio of) Geometric Means (Ratio)				Paired t test on logs	

61

Two Matched Samples

	Binary	Unordered	Ordered		
		Nominal	Categ	Quant	Cens
Hazard Ratio				Logrank	
Pr (Y > X)			Sign	Sign	
???			Wilcox Sgn Rnk	Wilcox Sgn Rnk	

62

Regression Methods

- In Biost 518, we extend these methods to the case of the “infinite sample” problem
- Borrowing information in presence of sparse data
- Contrasts across multiple groups
  - Continuous grouping variables
  - Adjustment for covariates

63

Infinite Samples

- 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
  - Compare groups differing in age by 1 year
    - 3 vs 4; 8 vs 9; 11 vs 12 …
  - Figure that comparisons across groups that differ by *k* years will be *k*-fold higher
  - Average all those estimates

- Adjustment for other variables
  - Compare males to females among 30 yo, among 31 yo, …
  - Average all those estimates

64

Regression Methods

	Binary	Unordered	Ordered		
		Nominal	Categ	Quant	Cens
Entire Distn	Logist				
Diff in Prop	(Linear)	(Linear)	(Linear)	(Linear)	
Odds Ratio	Logist	Logist	Logist; Prop Odds	Logist	

65

Regression Methods

	Binary	Unordered	Ordered		
		Nominal	Categ	Quant	Cens
Diff in Medians					
Median Diffrence					
Ratio of Medians				Param Surv (AFT)	Param Surv (AFT)

66

Regression Methods

	Binary	Unordered	Ordered		
		Nominal	Categ	Quant	Cens
(Diff in) Arith Means (of Diff)	(Linear)		Linear	Linear	
Ratio of Arith Means	Poisson		Poisson	Poisson	
(Ratio of) Geometric Means (Ratio)				Linear on logs	

67

Regression Methods

	Binary	Unordered	Ordered		
		Nominal	Categ	Quant	Cens
Hazard Ratio				Prop Hazard	Prop Hazard
Pr (Y > X)					
???					

68

"Everything is Regression"

- The most commonly used two sample tests are special cases of regression

- Regression with a binary predictor
  - Linear ➔ t test
  - Logistic ➔ chi square (score test)
  - Poisson ➔ two sample test of Poisson rates
  - Proportional hazards ➔ logrank (score test)

69