Biostatistics 515: Biostatistics II

Introduction to Regression

January 6, 2004

Course Structure

Instructor: Elizabeth Brown

Meeting times and locations:

Lectures: 9:00 - 10:20 am, Tues. and Thurs.

Quiz Section: 1:30-2:20 pm, Thurs.

Texts:

Kleinbaum et al: Applied Regression Analysis

Harrell: Regression Modeling

Software

We will make extensive use of statistical computing tools in this course. All examples and answer keys will be given in R. However, you are free to use your favorite statistical computing package to complete your assignments.

Evaluations

Weekly assignments	**30%**
Class discussion	**5%**
Midterm	**20%**
Project	**25%**
Final Exam	**20%**

Before 515

- **Estimates, confidence intervals and tests for one sample problems, e.g.:**
  - Estimating mean, median, variance
  - Construction and interpretation of CIs
  - Testing whether mean is different from some hypothesized value
- **Estimates, confidence intervals and tests for two sample problems, e.g.:**
  - Testing equality of means across groups
  - Non-parametric tests for differences in distributions of continuous variables between two groups

In 515

- **Extend these methods to the case where there are more than two groups, e.g.:**
  - Viral load levels across 4 treatment groups
  - FEV across (conceptually) an infinite number of height groups
- **Adjusting for a third (or more) variable(s) to obtain**
  - Inference about effect modification
  - Inference about independent effects of variables
  - Greater precision for prediction

Two Variable Setting

Many statistical problems can be regarded as considering the association between two variables

Response variable (outcome, dependent variable)

Grouping variable (predictor, independent variable)

The scientific question is addressed by comparing 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

Correspondence to Number of Samples

In introductory statistics courses, there is a tendency to characterize problems according to the number of samples and whether the samples are independent

The correspondence between that nomenclature and the two variable setting is based on the type of variable used as the grouping variable

- Constant: One sample problem
- Binary: Two sample problem
- Categorical: *k* sample problem (e.g., ANOVA)
- Continuous: Infinite sample problem
  - Regression

Infinite Sample Problem

When the grouping variable is continuous, there are conceptually an infinite number of groups

E.g., when investigating the blood pressure across age groups

- If measured with enough precision, no two people have exactly the same age

It is, of course, rare that we would have an infinite number of groups in our sample

- (and possibly not even in our population)

It is common to have 1 (or fewer) subjects in a particular group in our sample

Regression Methods

In Biost 515, we explore methods to solve 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

Regression Methods

The primary statistical methods that we will use are those referred to as regression

As in one and two sample problems, we will focus on the parameter compared across groups

- Means → Linear regression
- Odds → Logistic regression
- Rates → Poisson regression
- Hazards → Proportional Hazards regression
- Quantiles → Parametric survival regression

Regression vs Two Sample Methods

A very convenient feature of the regression methods is that when used with a binary grouping variable they reduce to the corresponding two variable methods

Logistic regression with a binary predictor

- Chi square test of odds ratios (score test)

Linear regression with a binary predictor

- t test with equal variance
- (approx t test with unequal variance when using "robust" standard errors)

Proportional hazards regression with a binary predictor

- Logrank test

Linear Regression Setting Example

Association between blood pressure and age

Scientific question:

- Does aging affect blood pressure?

Statistical question:

- Does the distribution of blood pressure differ across age groups?
  - Acknowledges variability of response
  - Acknowledges uncertainty of cause and effect
    - (Differences could be related to calendar time instead of age)

Linear Regression Setting Example

Association between blood pressure and age (cont.)

Definition of variables

- Response: Systolic blood pressure
  - continuous
- Predictor of interest (grouping): Age
  - continuous
    - an infinite number of ages are possible
    - we probably will not sample every one of them

Linear Regression Setting Example

Association between blood pressure and age (cont.)

Answering the question is possible if we try to assess linear trends in, say, average SBP by age

Estimate best fitting line to average SBP within age groups

$$E\left(SBP \mid Age\right) = \beta_0 + \beta_1 \times Age$$

An association will exist if the slope (β_1) is nonzero

- In that case, the average SBP will be different across different age groups

Linear Regression Setting Example

Association between blood pressure and age (cont.)

The regression model thus produces something similar to "a rule of thumb"

E.g., "Normal SBP is 100 plus age"

$$E\left(SBP \mid Age\right) = 100 + 1 \times Age$$

Regression: Necessary Ingredients

Response variable

The distribution of this variable will be compared across the groups

Notation:

- It is extremely common to use *Y* to denote the response variable when discussing general methods
- The predictor variable is often denoted by *X*.

Regression: Necessary Ingredients

Summary measure (parameter) of distribution to model

The appropriate choice of parameter will depend upon scientific relevance (and type of variable)

- Binary response: odds of event
- Count response: rate of event
- Continuous response: mean or median
- Censored response: hazard (instantaneous probability of event) or median

Notation:

- θ will denote an arbitrary parameter
- p, μ, $\lambda(t)$ will denote proportions, mean, hazard

Regression: Necessary Ingredients

Transformation of parameter to be modeled linearly

For statistical stability, we often consider best fitting lines to a transformation of the parameter

- Notation: $g(\theta)$

Common transformations:

- Mean: none
- Geometric mean: log
- Median: log
- Odds: log
- Rates: log
- Hazard: log

Regression: Necessary Ingredients

Regression model

We typically consider a "linear predictor function" that is linear in the modeled predictors

$$g(\theta) = \beta_0 + \beta_1 \times X$$

In later lectures we will discuss the interpretation of the "regression parameters" for specific types of regression models

- intercept β_0
- slope β_1

Uses of Regression Models

Regression models can be used to answer the most commonly encountered statistical questions

Prediction

- Estimating a future observation of response *Y*
- Often we use the mean or median

Quantifying distributions

- Describing the distribution of response *Y* within groups by estimating the parameter θ

Comparing distributions across groups

- Distributions differ across groups if the regression slope parameter β_1 is nonzero

$$g(\theta) = \beta_0 + \beta_1 \times X$$

Uses of Regression Models

General comments

Regression models are a useful tool for estimating general trends in the distribution of response across groups defined by the predictor

- The assumption of straight line relationships in the modeled (transformed) parameter need not hold exactly for these purposes

Interpolation to unobserved groups is less risky than extrapolation outside the range of predictors

Simple Regression

Modeling the distribution of some response variable across groups defined by a single predictor

Examples

- Comparing mean DSST across age groups
- Comparing geometric mean FEV across height groups
- Comparing the odds of being in remission at 24 months across nadir PSA groups
- Comparing instantaneous risk of death across age groups

Simple Regression

Modeling distribution of response

In regression, we consider that the distribution of the response variable might be different in each group defined by the predictor

Each subject in the sample might be in a different group

- There might not be any two subjects who have the same value for the predictor variable

Simple Regression

Distribution of response for an individual

By the "distribution of response" for a given individual we mean the distribution of response across individuals who have the same value of the grouping variable as that individual

Thus, when speaking of the mean, median, etc. response for an individual, we usually really mean the mean, median, etc. for a population who has the same value of the predictor of that individual

Simple Regression

General notation for variables and parameter

Y_i Response measured on the *i*th subject

X_i Value of the predictor for the *i*th subject

θ_i Parameter of distribution of Y_i

The parameter might be the mean, geometric mean, odds, rate, instantaneous risk of an event (hazard), etc.

Simple Regression

General notation for simple regression model

$$g(\theta_i) = \beta_0 + \beta_1 \times X_i$$

$g(\ )$ "link" function used for modeling

β_0 "Intercept"

β_1 "Slope (for predictor X)"

The link function is usually either the identity function (so no link) or log ( )

Simple Regression

The regression model allows us to make inference about groups that have few (if any) subjects

We "borrow information" from other groups in order to estimate the value of a parameter for each group

- "Borrowing" is usually based on estimation of a line on some scale (the scale of the link)
- (Two points determine a line)

Even when we do not think a line represents the true relationship between parameters across groups, we can estimate "average rate of change"

Simple Regression

Interpretation for simple regression parameters with no link function (identity link)

E.g., linear regression modeling the mean

No link : $\theta_i = \beta_0 + \beta_1 \times X_i$

β_0 Value of parameter in group with $X = 0$

β_1 Difference of parameters between groups which differ by 1 unit in value of X

Simple Regression

Interpretation for simple regression parameters with no link function (identity link)

E.g., logistic regression modeling the odds

Log link : $\log(\theta_i) = \beta_0 + \beta_1 \times X_i$

e^{β_0} Value of parameter in group with $X = 0$

e^{β_1} Ratio of parameters between groups which differ by 1 unit in value of X

Summary

So far, we've discussed general linear models and deterministic forms for modeling parameters of interest. Starting in the next lecture, we will discuss

- Probabilistic models for the linear models (the random component)
- Estimation of the coefficients in the linear models
- Hypothesis testing for the coefficients