Advanced Methods in Biostatistics I

Lecture 1

Martin Lindquist

August 29, 2017

Advanced Methods in Biostatistics

- *Advanced Methods in Biostatistics* is a four term sequence.
- The first two terms (751-2) focus on linear models.
- The last two terms (753-4) focus on generalized linear models, mixed models, and longitudinal data analysis.

- Together these topics are arguably some of the most important techniques used in statistics.

Linear Models

- Linear models are the cornerstone of statistical methodology and one of its most studied branches, both in theory and application.
- They allow us to model the relationship between a response variable and one or more explanatory variables, and determine which of these variables are important.
- They are relatively easy to work with, and are applicable to many real-life settings.
- Properly constructed, they encompass many standard statistical techniques, including t-tests, linear regression, analysis of variance, and analysis of covariance.

Notation

- Suppose we want to model the relationship between a response variable and one or more explanatory variables.

- Let y_i be the response variable and $(x_{i1}, x_{i2}, \dots x_{ip})$ a set of explanatory variables for observation $i = 1, \dots n$.

- Note that the response variable can also be referred to as the dependent variable or the output.
- The explanatory variables can also be referred to as the covariates, predictors, factors, independent variables, or inputs, depending on the setting.

Notation

- In a linear model, the response variable is real valued.
- For now we assume that the explanatory variables are known constants. They can be:
  - Quantitative variables or transformations of quantitative variables;
  - Basis expansions (i.e., $\mathbf{x}_1$, $\mathbf{x}_2 = \mathbf{x}_1^2$, $\mathbf{x}_3 = \mathbf{x}_1^3$, etc);
  - ‘Dummy variables’ that code the levels of a categorical variable (i.e., $x = 0$ for male, $x = 1$ for female).
  - Interactions between variables (i.e., $\mathbf{x}_3 = \mathbf{x}_1\mathbf{x}_2$).

Linear Models

- In a linear model the relationship between the response variable and explanatory variables can be expressed as:

  $$y_i = \mu(x_{i1}, x_{i2}, \ldots x_{ip}) + \epsilon_i$$

  where $\mu(x_{i1}, x_{i2}, \ldots x_{ip})$ is some function of the explanatory variables and ϵ a random variable.

Linear Models - Deterministic part

- The function $\mu(.)$ is the deterministic part of the model.
- Its form is generally assumed to be known, but it contains unknown parameters that need to be estimated.
- The term *linear* implies that $\mu(.)$ is a linear function of some unknown parameters.
- Examples:
  - $\mu(x_{i1}, x_{i2}) = \beta_1 x_{i1} + \beta_2 x_{i2}$
  - $\mu(x_{i1}, x_{i2}) = \beta_1 x_{i1} + \beta_2 x_{i2}^2$

Linear Models - Stochastic part

- The error term ϵ is random.
- Throughout the course we will make a number of additional assumptions about the mean, variance, and distribution of the error, but for now we leave this information unspecified.

Goals

There are multiple goals when working with linear models, including:

- Finding the best ‘fit’ between the explanatory variables and the response. The standard procedure for this is the method of least squares, but other methods exist.
- Obtaining good parameter estimates that allow us to understand the relationship between variables.
- Making predictions that allow us to determine what response we can expect to observe under a new set of experimental conditions.
- Performing inference.

Example - Simple linear regression

- Does blood pressure (y) level depend upon age (x)?
- Model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \text{for } \; i = 1, \ldots n$$

or

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Example - Polynomial regression

- Is the relationship quadratic?
- Model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \qquad \text{for} \;\; i = 1, \ldots n$$

or

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Example - Multiple linear regression

- Does blood pressure (y) level depend upon age (x_1), weight (x_2), and BMI (x_3)?
- Model:

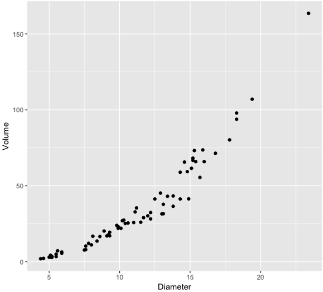
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i \qquad \text{for} \quad i = 1, \ldots n$$

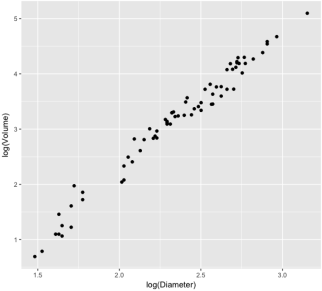
or

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Example - Data transformations

- Is the volume of a tree (*y*) related to its diameter (*x*)?
- We can often use a linear model after transforming either or both the explanatory and response variables.
- Note the errors must be additive on the transformed scale.





Example - Data transformations

- Model:

$$log(y_i) = \beta_0 + \beta_1 log(x_i) + \epsilon_i \qquad \text{for} \;\; i = 1, \ldots n$$

or

$$\begin{pmatrix} log(y_1) \\ log(y_2) \\ \vdots \\ log(y_n) \end{pmatrix} = \begin{pmatrix} 1 & log(x_1) \\ 1 & log(x_2) \\ \vdots & \vdots \\ 1 & log(x_n) \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Example - One-way analysis of variance

- Does cholesterol level (y) differ between subjects in a drug group and a control group?
- Model:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad \text{for } \; i = 1, 2; \;\; j = 1, \ldots J$$

or

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1J} \\ y_{21} \\ \vdots \\ y_{2J} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1J} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2J} \end{pmatrix}$$

General Linear Model

The general linear model in matrix form:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{10} & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ x_{20} & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n0} & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

General Linear Model

Equivalent shorthand form:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon.$$

Here $\mathbf{y}$ $(n \times 1)$ is the response vector, $\mathbf{X}$ $(n \times p)$ is the design (or model or regression) matrix, β $(p \times 1)$ is the vector of regression coefficients, ε $(n \times 1)$ is the error vector.

General Linear Model

- Usually $x_{i0} = 1$ for all i, i.e., there is an intercept term β_0 in the model.
- We refer to $x_{i0}, x_{i1}, \ldots, x_{i,p-1}$ as the predictor variables or regressor variables. They are known constants.
- The model is linear in the unknown regression coefficients $\beta_0, \beta_1, \ldots, \beta_{p-1}$, i.e.,

  $$\mathbf{y} = \sum_{j=0}^{p-1} \beta_j \mathbf{x}_j + \varepsilon,$$

  where $\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})'$.

Least Squares Estimation

- How do we estimate the unknown parameters β?
- There are many ways to do this, but here we will focus on least-squares estimation.

Least Squares Estimation

Definition

An estimate $\hat{\beta}$ is a least-squares estimate of β if it minimizes $||\mathbf{y} - \mathbf{X}\beta||^2$ over all β.

Course outline - 751

- In the first part of the class we make no explicit assumptions about the error term and explore least-squares estimation from a matrix algebra perspective.
- Later we will add assumptions about the moments of the error, allowing us to prove optimality of the least-squares solution and partition the variance.
- Finally, after adding distributional assumptions we will discuss inference.

Course outline - 752

- In 752 we will continue discussing inference in the linear model framework.
- We will also discuss linear mixed models, generalized least squares, and penalized least squares.

Software

- R is a programming language and software environment for statistical computing and graphics that is highly extensible.
- Linear models can be fit in R using using the `lm` command.
- We will discuss R and the `lm` command throughout the course.

R code

- We illustrate simple linear regression using the `mtcars` data set that is directly available in R. The data was extracted from the 1974 Motor Trend US magazine, and consists of gas consumption (`mpg`) and 10 other aspects of automobile design and performance for a total of 32 cars.

R code

A simple linear regression using gas consumption (`mpg`) as the response variable and weight (`wt`) as the explanatory variable.

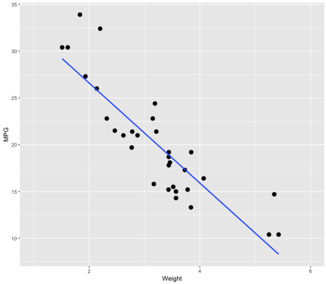
```
> fit <- lm(mpg ~ wt, data=mtcars)
> fit

Call:
lm(formula = mpg ~ wt, data = mtcars)

Coefficients:
(Intercept)           wt
     37.285       -5.344
```

R code

A scatter plot of gas consumption (`mpg`) and weight (`wt`) with the regression line superimposed.



R code

```
> summary(fit)

Call:
lm(formula = mpg ~ wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5432 -2.3647 -0.1252  1.4096  6.8727

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt           -5.3445     0.5591  -9.559 1.29e-10 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```