Lecture 1: Introduction and Overview

Linear models are used to study how a quantitative response variable depends on one or more explanatory variables. The model is called *linear* because it assumes a linear relationship,

$$E(y|x) = \beta_0 + \beta_1 x,$$

between a response y and predictor(s) x.

WHY USE LINEAR MODELS?

- *A Scientific Question:* Suppose an obstetrician is interested in knowing whether babies of women who smoke during pregnancy weigh less than babies of women who do not smoke. (Low birth weight is associated with increased morbidity and mortality of infants.) The obstetrician has a dataset containing information about all the deliveries at County General Hospital. For each delivery, you know the birth weight of the infant and whether or not the mother smoked.
- *Your Initial Analysis:* To answer the scientific question of interest, you conduct a two sample t-test to compare birth weights of infants with smoking mothers to the birth weights of infants with nonsmoking mothers.

  Question: How does the t-test answer the question of interest?

- *Physician Reaction:* The obstetrician was thrilled with the results of your analysis (clearly, the p-value was < 0.05).
- *Physician Reconsideration:* On the way out of your office, the physician comments that teenaged mothers are more likely to have low birth weight infants than older mothers and are also more likely to smoke. In addition, the obstetrician comments that perhaps the number of cigarettes smoked per day is related to the birth weight as well. Do more cigarettes lead to lower birth weight? What about family income or maternal stress? Alcohol use or cocaine use? Smoking marijuana? What about women who stop smoking or reduce smoking levels after learning about the pregnancy?

- *Secondary Analysis:* You realize that the t-test does not control for these additional factors that may be related to birth weight and may influence the relationship between cigarette smoking and birth weight. Accurate estimation of the effect of cigarette smoking on birthweight will depend on how smoking, as well as other variables like maternal age and drug use, are related to birthweight and each other. You need to use a *model* that incorporates the various exposures and potential confounders into your analysis; such *multiple regression* models are extremely important in observational (i.e., non-randomized) studies.

Linear models are one of the world's most popular statistical tools. Understanding the theory of linear models is also a foundation for understanding more complex models used in statistics (including generalized linear models, longitudinal and multivariate models, and survival models).

Connection to Previous Coursework

In BIOS662, we considered the basics of sampling, one- and two-sample parametric and nonparametric inference, and simple methods for the analysis of data from more than two groups (ANOVA and simple linear regression).

In BIOS663, we will discuss the general linear model in detail.

- We begin by discussing standard results for least-squares model fitting and basic procedures of inference (testing simple and complex hypotheses, constructing confidence intervals and regions, and making predictions).
- Then, we move to aspects of model checking, including residual analysis and detection and treatment of outliers.
- Next, we discuss basic procedures for inference in polynomial models

and consider basic smoothing techniques.

- We devote considerable time to model-building, variable selection, and model validation.
- We conclude by considering the generalized linear model (logistic and Poisson regression in particular) and methods for correlated response data.

Example 1: One-Way ANOVA

Analysis of Variance (ANOVA) involves comparing random samples from several populations. One-way ANOVA is an extension of the 2-sample t-test to three or more samples.

Hypothesis: NO_2 exposure leads to damaged lung tissue in mice.

Lung damage was measured by percent serum fluorescence, where higher readings indicate greater damage. Investigators exposed 30 mice to a high dose of NO_2, 30 mice to a low dose, and selected 30 mice as controls. (Nitrogen dioxide is found in tobacco smoke and can also be produced by kerosene heaters and unvented gas stoves.)

Define:

$y_{ij}=$ serum fluorescence of the j^{th} mouse in the i^{th} dose group, $i = 1, \ldots, 3$, $j = 1, \ldots, 30$.

$\mu_i=$average serum fluorescence in group i

We have the model:

$y_{ij} = \mu_i + \varepsilon_{ij}$, $i = 1, \ldots, 3$, $j = 1, \ldots, 30$,

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)^T$ is the vector of parameters to be estimated. We assume that the ε_{ij}'s are *i.i.d.* from some distribution with $E(\varepsilon_{ij}) = 0$ and Var$(\varepsilon_{ij}) = \sigma^2$. This model is often called a *means* model.

We may wish to use this model to do the following.

- Estimate the μ_i's and σ^2.
- Test hypotheses about the μ_i's, for example,

  $$H_0 : \mu_1 = \mu_2 = \mu_3.$$

  What does this hypothesis mean in terms of the subject matter?
- Decide which group mean is the largest, smallest, etc.

Review of Terminology and Basic Concepts

Scales of Measurement

Scales of measurement help to determine what type of analysis or model should be used for the data.

- A *nominal* variable is for mutually exclusive, but not ordered, categories, e.g. blood type or gender; also called *categorical*
- An *ordinal* variable is one where the order matters but not the difference between values, e.g. AP Basketball poll (also called *ranked data*) or pain severity scale (none, minor, moderate, severe)
- An *interval* variable is a measurement where the difference between two values is meaningful. The difference between a temperature of 100 degrees and 90 degrees is the same difference as between 90 degrees and 80 degrees. Examples include temperature in degrees Centigrade.

- A *ratio* variable has all the properties of an interval variable, and also has a clear definition of 0. When the variable equals 0, there is none of that variable. Variables like height, weight, enzyme activity are ratio variables. Temperature, expressed in F or C, is not a ratio variable. A temperature of 0 on either of those scales does not mean 'no temperature'. However, temperature in degrees Kelvin is a ratio variable, as 0 degrees Kelvin really does mean 'no temperature'. When working with ratio variables, but not interval variables, you can look at the ratio of two measurements. A weight of 4 grams is twice a weight of 2 grams, because weight is a ratio variable. A temperature of 100 degrees C is not twice as hot as 50 degrees C, because temperature C is not a ratio variable.

Interval and ratio variables are types of *continuous* variables. Ratio variables typically have error variance proportional to the size of the measurement, so transformations are often used to stabilize the variance.

Types of Variables

- *Response* or *Dependent Variable*: variable that is to be described in terms of other variables
- *Predictor* or *Independent Variable* or *Covariate*: used (perhaps with other independent variables) to describe a response variable
- *Control* or *Nuisance Variables*: may affect relationships but of no real interest in current study.
  - *Confounder*: a third factor that can lead to an observed association (or lack of association) due in fact to mixing of effects between the dependent variable, independent variable, and the confounding variable. For example, let the dependent variable be development of lung cancer, the independent variable be smoking, and the potential confounder be a genetic mutation. This particular genetic mutation may increase the likelihood of

lung cancer, as well as the likelihood to be addictive to smoking.

Sets of Interest

- *Population:* any set of interest. A *parameter* describes a property of a population.
- *Sample:* any subset of a population. A *statistic* describes a property of a sample.

Statistical Activities

- *Parameter Estimation:* means of providing a value (the "estimator") thought to be a "good" approximation of a parameter.

  What is "good"?

  - An estimate $\hat{\theta}$ of a parameter θ is *consistent* if $\hat{\theta} \to \theta$ as $n \to \infty$.
  - An estimate $\hat{\theta}$ of a parameter θ is *unbiased* if $E(\hat{\theta}) = \theta$.
  - *Maximum likelihood* estimates are asymptotically unbiased and consistent.

- *Inference:* Statistical inference or statistical induction comprises the use of statistics and random sampling to make inferences concerning some unknown aspect of a population. It is distinguished from descriptive statistics. Two schools of statistical inference are frequency probability and Bayesian inference.

Purposes of Statistical Analysis

- *Confirmatory:* Relies on prespecified variables, model, and hypothesis test of interest. Generally, when pharmaceutical companies evaluate drugs in an effort to obtain FDA (U.S. Food and Drug Administration) approval, they conduct confirmatory analyses. (Does the new drug lead to a better patient outcome than the standard treatment?)

- *Exploratory:* Seeks to find patterns or explain variation in data; "fishing expedition." A major epidemiologic study, for example, might involve a primary confirmatory-type analysis for which a study is designed (Do infections during pregnancy lead to dangerously early births?) and several exploratory analyses (Do factors such as diet, exercise, socioeconomic status, education, drug use, etc. also affect the length of pregnancy?).

In an exploratory analysis, p-values generally must be interpreted with caution due to multiple testing of the data.

General Linear Model (GLM)

General: Applicable to wide variety of problems of estimation and testing

Linear: Regression function, $E(y_i \mid \mathbf{x}_i)$ is a linear function of the parameters $\boldsymbol{\beta}$

Model: Describes the relationship between the one response and one or more predictors.

Muller & Fetterman terminology, also known as "Multiple Regression" in ALR. A simple *model function* is given by

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

with corresponding *regression function*

$$E(y_i|x_i) = \mu_i = \beta_0 + \beta_1 x_i.$$

The GLM is a *univariate* model, which means that we consider only one response variable. For this reason, the GLM is sometimes called the GLUM (general linear univariate model, or Simple Linear Regression in ALR). We may have one predictor or many predictors, in which case we may call our model a *multivariable* model, or refer to the regression problem as *mutiple regression*.

Multivariate models involve more than one response (researchers outside biostatistics sometimes use the word *multivariate* instead of *multivariable* to describe models with one response and many predictors).

Example 2: Wingspan and Height Relationship

Hypothesis: Wingspan (the distance from one outstretched fingertip to the other) is equal to one's height.

y_i=wingspan of subject i, $i = 1, \ldots, n$

x_i=height of subject i

We have the model:

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, \ldots, n$.

Hypotheses to test: $H_0 : \beta_1 = 1$ and $\beta_0 = 0$.

Example 3: UNC Faculty Salaries

Suppose we wish to model the relationship between faculty salary and years of service at UNC. We define the varables y_i, the salary of faculty member i, $i = 1, \ldots, n$, and x_i, the years of service at UNC for faculty member i, and fit the model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

which may be written in matrix form as

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

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

Now suppose we are interested in the relationship between rank (instructor, assistant professor, associate professor, or full professor) and faculty salary. We define three indicator variables, x_1, x_2, x_3, as follows:

$$x_1 = \begin{cases} 1 & \text{assistant professor} \\ 0 & \text{otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{associate professor} \\ 0 & \text{otherwise} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{full professor} \\ 0 & \text{otherwise.} \end{cases}$$

In this coding scheme, instructors have $(x_1, x_2, x_3) = (0, 0, 0)$ (reference group), assistant professors have $(x_1, x_2, x_3) = (1, 0, 0)$, associate professors have $(x_1, x_2, x_3) = (0, 1, 0)$, and full professors have

$(x_1, x_2, x_3) = (0, 0, 1)$. An ANOVA model is given by $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1,n_{\text{assistant}}} \\ y_{21} \\ \vdots \\ y_{2,n_{\text{associate}}} \\ y_{31} \\ \vdots \\ y_{3,n_{\text{full}}} \\ y_{41} \\ \vdots \\ y_{4,n_{\text{instructor}}} \end{pmatrix}_{n\times 1} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \end{pmatrix}_{n\times 4} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1,n_{\text{assistant}}} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2,n_{\text{associate}}} \\ \varepsilon_{31} \\ \vdots \\ \varepsilon_{3,n_{\text{full}}} \\ \varepsilon_{41} \\ \vdots \\ \varepsilon_{4,n_{\text{instructor}}} \end{pmatrix}_{n\times 1}.$$

Next, suppose that we are interested in evaluating both rank and years at UNC in one regression model. An ANCOVA model is given by

$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1,n_{\text{asst}}} \\ y_{21} \\ \vdots \\ y_{2,n_{\text{assoc}}} \\ y_{31} \\ \vdots \\ y_{3,n_{\text{full}}} \\ y_{41} \\ \vdots \\ y_{4,n_{\text{inst}}} \end{pmatrix}_{n\times 1} = \begin{pmatrix} 1 & 1 & 0 & 0 & x_{11} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & x_{1,n_{\text{asst}}} \\ 1 & 0 & 1 & 0 & x_{21} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 & x_{2,n_{\text{assoc}}} \\ 1 & 0 & 0 & 1 & x_{31} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 & x_{3,n_{\text{full}}} \\ 1 & 0 & 0 & 0 & x_{41} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & x_{4,n_{\text{inst}}} \end{pmatrix}_{n\times 5} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1,n_{\text{asst}}} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2,n_{\text{assoc}}} \\ \varepsilon_{31} \\ \vdots \\ \varepsilon_{3,n_{\text{full}}} \\ \varepsilon_{41} \\ \vdots \\ \varepsilon_{4,n_{\text{inst}}} \end{pmatrix}_{n\times 1}.$$

Finally, we may consider an interaction between rank and years of service in the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & x_{11} & x_{11} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & x_{1,n_{\text{asst}}} & x_{1,n_{\text{asst}}} & 0 & 0 \\ 1 & 0 & 1 & 0 & x_{21} & 0 & x_{21} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 & x_{2,n_{\text{assoc}}} & 0 & x_{2,n_{\text{assoc}}} & 0 \\ 1 & 0 & 0 & 1 & x_{31} & 0 & 0 & x_{31} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 & x_{3,n_{\text{full}}} & 0 & 0 & x_{3,n_{\text{full}}} \\ 1 & 0 & 0 & 0 & x_{41} & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & x_{4,n_{\text{inst}}} & 0 & 0 & 0 \end{pmatrix}_{n \times 8}$$

This is *full model in every cell* by Muller and Fetterman.

Next: Linear Algebra Review

Reading Assignment

- Weisberg, Appendix A.6-7: "A Brief Introduction to Matrices and Vectors"
- Muller and Fetterman, Appendix A: "Matrix Algebra Useful for Linear Models"
- Namboodiri: "Matrix Algebra: An Introduction" (Optional)

To do

- Download and install R: `https://cloud.r-project.org/`
- Order and install SAS: `https://software.sites.unc.edu/software/`