
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,
 $i = 1, \dots, 3$, $j = 1, \dots, 30$.

μ_i = average serum fluorescence in group i

We have the model:

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

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 $\text{Var}(\varepsilon_{ij}) = \sigma^2$. 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,

$$H_0 : \mu_1 = \mu_2 = \mu_3. \quad \text{💡}$$

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 $\hat{\theta} \rightarrow \theta$ as $n \rightarrow \infty$.
 - 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, $E(y_i | \mathbf{x}_i)$ is a linear function of the parameters β

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 *multiple 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 , $i = 1, \dots, n$

x_i =height of subject i

We have the model:

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

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 variables y_i , the salary of faculty member i , $i = 1, \dots, n$, and x_i , the years of service at UNC for faculty member i , and fit the model

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

which may be written in matrix form as

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

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

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

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

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/>

Lecture 2: Linear Algebra Review

Reading

- 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)

Why Linear Algebra?

Data for linear models can be represented as vectors and matrices

Simplification of theory and estimation of linear models, general representation

Easier to address certain problem in estimating such models

Basics of Notation

A *matrix* is a two-dimensional array of elements.

$\mathbf{A} = \{a_{ij}\}$ means \mathbf{A} is the matrix whose i, j^{th} element (or component) is the *scalar* a_{ij} , where i indexes row and j indexes column, $i = 1, \dots, r$, $j = 1, \dots, c$. In this context, a scalar s will represent a real number. Capital letters are used to represent matrices, and lowercase letters are used for vectors.

The *dimension* of \mathbf{A} is $(r \times c)$ (say “r by c”), often written $\mathbf{A}_{r \times c}$.

The entire matrix \mathbf{A} may be represented

$$\mathbf{A}_{r \times c} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1c} \\ a_{21} & \ddots & & \dots \\ \dots & & \ddots & \dots \\ a_{r1} & \dots & \dots & a_{rc} \end{bmatrix}.$$

A matrix with one column, i.e. an $(r \times 1)$ matrix, is called a *vector* or

column vector. A $(1 \times r)$ matrix is called a *row vector*. We can represent the matrix $\mathbf{A}_{r \times c}$ in terms of its columns $\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_c\}$. For example, we might represent the \mathbf{X} matrix from the one of the examples in lecture 1 as

$$\begin{aligned}\mathbf{X} &= \{1, \mathbf{height}\} \\ &= \begin{bmatrix} 1 & 61.61 \\ 1 & 59.17 \\ 1 & 54.75 \\ \vdots & \vdots \\ 1 & 62.92 \end{bmatrix}.\end{aligned}$$

Types of Matrices

A *square matrix* has the same number of rows as columns, so that $r = c$.

For $\mathbf{A}_{r \times c}$ and $r \geq c$, the *diagonal* of \mathbf{A} is $\{a_{11}, a_{22}, \dots, a_{cc}\}$.



A *symmetric matrix* is a square matrix with $a_{ij} = a_{ji}$ for all i, j . That is, the entry in row i and column j is the same as the entry in row j and column i . Only square matrices can be symmetric. The matrix

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{bmatrix}$$



is symmetric (note that we refer to symmetry about the main diagonal, which runs from the upper left to the lower right).

A square matrix is called a *diagonal matrix* if all elements off the main diagonal are zero; that is, if $i \neq j$, then $a_{ij} = 0$. We write $\text{Diag}(\mathbf{b}) = \text{Dg}(\mathbf{b})$ for a square diagonal matrix created from a vector. For example,

$$\text{Diag} \left(\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \right) = \begin{bmatrix} b_1 & 0 & 0 \\ 0 & b_2 & 0 \\ 0 & 0 & b_3 \end{bmatrix}.$$

An *identity matrix*, denoted $\mathbf{I} = \mathbf{I}_n = \mathbf{I}_{n \times n}$, is a diagonal matrix with all 1's on the main diagonal and 0's elsewhere. That is, $a_{ij} = 1$ for $i = j$ and $a_{ij} = 0$ for $i \neq j$. Thus

$$\mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Denote an $(n \times 1)$ vector of 1's by


$$\mathbf{1} = \mathbf{1}_n = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n \times 1} = \mathbf{J}_n.$$

A *zero matrix* or *null matrix*, denoted $\mathbf{0}_{r \times c}$, has $a_{ij} = 0$ for all i, j . So

$$\mathbf{0}_{2 \times 2} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

An *upper triangular matrix* \mathbf{A} has $a_{ij} = 0$ for $i > j$. That is,

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix}$$

is an upper triangular matrix. A *lower triangular matrix* has $a_{ij} = 0$ for $i < j$.

A *partitioned* matrix has elements grouped into submatrices by combinations of vertical and horizontal slicing. For example,

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \vdots & a_{13} \\ a_{21} & a_{22} & \vdots & a_{23} \\ \dots & \dots & \dots & \dots \\ a_{31} & a_{32} & \vdots & a_{33} \end{bmatrix}_{3 \times 3} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}_{3 \times 3},$$

where

$$\mathbf{A}_{11} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}_{2 \times 2}, \quad \mathbf{A}_{12} = \begin{bmatrix} a_{13} \\ a_{23} \end{bmatrix}_{2 \times 1},$$

$$\mathbf{A}_{21} = \begin{bmatrix} a_{31} & a_{32} \end{bmatrix}_{1 \times 2}, \text{ and } \mathbf{A}_{22} = \begin{bmatrix} a_{33} \end{bmatrix}_{1 \times 1}.$$

A *block diagonal matrix* has square diagonal submatrices with all

off-diagonal submatrices equal to 0. For example,

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix}$$

is a block diagonal matrix.

Matrix Operations

The *trace* of a square matrix $\mathbf{A}_{n \times n}$ is the sum of the diagonal elements of \mathbf{A} . That is, $\text{trace}(\mathbf{A}) = \sum_{i=1}^n a_{ii}$.

$$\text{trace} \left(\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \right) = a_{11} + a_{22} + a_{33} .$$

The *transpose* of a matrix \mathbf{A} , denoted \mathbf{A}' , changes the rows of \mathbf{A} into the columns of a new matrix \mathbf{A}' . If \mathbf{A} is an $(r \times c)$ matrix, its transpose \mathbf{A}' is a $(c \times r)$ matrix.

The transpose of a column vector is a row vector, i.e.

$$\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}'_{4 \times 1} = [1 \quad 2 \quad 3 \quad 4]_{1 \times 4}.$$

A symmetric matrix has $\mathbf{A}' = \mathbf{A}$.

Matrix addition of two matrices \mathbf{A} and \mathbf{B} is defined only if \mathbf{A} and \mathbf{B} have the same number of rows and the same number of columns, *matrix addition* yields $\mathbf{A}_{r \times c} + \mathbf{B}_{r \times c} = \{a_{ij} + b_{ij}\}_{r \times c}$.

Exercise:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} + \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \quad .$$

There are several types of *matrix multiplication*. We will discuss the following types:

1. scalar multiplication,
2. matrix multiplication

-
- Define *scalar* multiplication of any matrix \mathbf{A} by a scalar s as
 $s\mathbf{A} = \{sa_{ij}\}.$

Exercise: 

$$\sigma^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \quad .$$

- Two matrices \mathbf{A} and \mathbf{B} conform for the *matrix multiplication* \mathbf{AB} if the number of columns of \mathbf{A} is equal to the number of rows of \mathbf{B} . If the matrices \mathbf{A} and \mathbf{B} conform, *matrix multiplication* is defined as

$$\mathbf{A}_{r \times s} \mathbf{B}_{s \times t} = \left\{ \sum_{k=1}^s a_{ik} b_{kj} \right\} = \mathbf{C}_{r \times t}.$$

Multiplying the i^{th} row of \mathbf{A} with the j^{th} column of \mathbf{B} yields the scalar $c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \dots + a_{is}b_{sj}$:

Calculation of c_{11} :

$$\text{row} \longrightarrow \begin{bmatrix} a_{11} & \rightarrow & \rightarrow & a_{1s} \end{bmatrix}_{r \times s} \quad \begin{bmatrix} b_{11} \\ \downarrow \\ \downarrow \\ b_{s1} \end{bmatrix}_{s \times t} \quad \downarrow \text{column}$$

So $c_{11} = a_{11}b_{11} + a_{12}b_{21} + \dots + a_{1s}b_{s1}$.

Note that $\mathbf{AI} = \mathbf{A}$ and $\mathbf{IA} = \mathbf{A}$. In general, $\mathbf{AB} \neq \mathbf{BA}$.

Exercise: Let $\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{bmatrix}$.

$$\mathbf{X}'\mathbf{X} =$$

Orthogonal Matrices

An *orthogonal matrix* is a **square matrix** with $\mathbf{A}' = \mathbf{A}^{-1}$. That is, a square matrix \mathbf{A} is orthogonal if $\mathbf{A}'\mathbf{A} \stackrel{?}{=} \mathbf{I} = \mathbf{A}\mathbf{A}'$. To establish that \mathbf{A} is orthogonal, it is sufficient to show either that $\mathbf{A}'\mathbf{A} = \mathbf{I}$ or that $\mathbf{A}\mathbf{A}' = \mathbf{I}$.

- The vectors \mathbf{x} and \mathbf{y} are *orthogonal vectors* if $\mathbf{x}'\mathbf{y} = 0$.
- The vectors \mathbf{x} and \mathbf{y} are *orthonormal vectors* if \mathbf{x} and \mathbf{y} are orthogonal vectors and **are normalized:** $\mathbf{x}'\mathbf{y} = 0$, $\mathbf{x}'\mathbf{x} = 1$, and $\mathbf{y}'\mathbf{y} = 1$.
- Some authors, including Muller and Fetterman, call an orthogonal matrix a column orthonormal matrix. While this is more accurate, we will retain standard terminology and refer to a matrix with $\mathbf{A}'\mathbf{A} = \mathbf{I} = \mathbf{A}\mathbf{A}'$ as orthogonal. Thus, somewhat paradoxically, an orthogonal matrix has orthonormal columns.

Rules of Matrix Operation

Suppose \mathbf{A} and \mathbf{B} conform for the operation of interest. The following laws apply to \mathbf{A} and \mathbf{B} .

Commutative Laws

- $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$
- $a\mathbf{B} = \mathbf{B}a$

In general, $\mathbf{A}\mathbf{B} \neq \mathbf{B}\mathbf{A}$.

Distributive Laws

- $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$
- $(\mathbf{B} + \mathbf{C})\mathbf{D} = \mathbf{BD} + \mathbf{CD}$
- $a(\mathbf{B} + \mathbf{C}) = a\mathbf{B} + a\mathbf{C} = (\mathbf{B} + \mathbf{C})a$

Associative Laws

- $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$
- $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$

Transpose Operations

- $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$
- $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$

Linear Dependence and Rank

Consider the system of equations

$$5x_1 + 2x_2 = 2 \tag{1}$$

$$10x_1 + 4x_2 = 4, \quad \text{?} \tag{2}$$

or

$$\mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 = \mathbf{b},$$

where

$$\mathbf{a}_1 = \begin{pmatrix} 5 \\ 10 \end{pmatrix}, \mathbf{a}_2 = \begin{pmatrix} 2 \\ 4 \end{pmatrix}, \text{ and } \mathbf{b} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}.$$

- This system of equations has infinite solutions given by
 $x_1 = \frac{2}{5} - \frac{2}{5}x_2.$
- We do not have one unique solution because the coefficients in

the second equation are a multiple of the first and thus are *linearly dependent*.

- So equation (2) does not provide any additional information about x_1 or x_2 .
- Examining \mathbf{a}_1 and \mathbf{a}_2 , we see that $a_{11} = \frac{5}{2}a_{12}$ and $a_{21} = \frac{5}{2}a_{22}$. Thus \mathbf{a}_1 and \mathbf{a}_2 are linearly dependent vectors.

The n vectors $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$ in \mathcal{R}^n are *linearly dependent* if there exist numbers c_1, c_2, \dots, c_n (not all zero) such that

$$\mathbf{a}_1c_1 + \mathbf{a}_2c_2 + \dots + \mathbf{a}_nc_n = 0.$$

If this equation is true only when $c_1 = c_2 = \dots = c_n = 0$ then the vectors are *linearly independent*. 

More generally, suppose we have the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & 2 & 4 \\ 1 & 0 & 2 \\ 0 & 8 & 0 \end{bmatrix}.$$

Each column of \mathbf{A} may be viewed as a vector. The *column space* of \mathbf{A} , denoted $C(\mathbf{A})$, is the set of all vectors that may be written as a linear combination of the columns of \mathbf{A} . That is, $C(\mathbf{A})$ is the set of all vectors that may be written as


$$\lambda_1 \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 2 \\ 0 \\ 8 \end{bmatrix} + \lambda_3 \begin{bmatrix} 4 \\ 2 \\ 0 \end{bmatrix} = \mathbf{A} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} = \mathbf{A}\boldsymbol{\lambda},$$

for some vector $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)'$.

For example, the column space of \mathbf{J}_2 includes all vectors of the form

$$\lambda \mathbf{J}_2 = \begin{pmatrix} \lambda \\ \lambda \end{pmatrix},$$

including

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} -0.03 \\ -0.03 \end{pmatrix}, \text{ and } \begin{pmatrix} \pi \\ \pi \end{pmatrix}.$$

The columns of \mathbf{A} are *linearly dependent* if they contain redundant information. If we can find two distinct vectors, say $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$, such that $\mathbf{A}\boldsymbol{\lambda} = \mathbf{A}\boldsymbol{\gamma} = \mathbf{x}$, then the columns of \mathbf{A} are linearly dependent.

An equivalent definition, obtained by letting $\boldsymbol{\delta} = \boldsymbol{\lambda} - \boldsymbol{\gamma}$ (prove to yourself!), is that the columns of \mathbf{A} are linearly dependent if there exists a vector $\boldsymbol{\delta} \neq \mathbf{0}$ such that $\mathbf{A}\boldsymbol{\delta} = \mathbf{0}$. If the columns of \mathbf{A} are not linearly dependent, then they are *linearly independent*.

Exercise: Does the matrix


$$\mathbf{A} = \begin{bmatrix} 2 & 2 & 4 \\ 1 & 0 & 2 \\ 0 & 8 & 0 \\ 3 & 1 & 6 \end{bmatrix}$$

have linearly independent or linearly dependent columns?

Exercise: Does the matrix

$$\mathbf{B} = \begin{bmatrix} 2 & 2 & 4 \\ 1 & 0 & 2 \\ 0 & 8 & 0 \\ \text{💡} & 4 & 1 & 6 \end{bmatrix}$$

have linearly independent or linearly dependent columns?

The *rank* of a matrix \mathbf{A} is the number of linearly independent columns in \mathbf{A} . Knowledge of the matrix rank is important in determining the existence and multiplicity of solutions to a system of linear equations.

If \mathbf{A} is an $(r \times c)$ matrix with $r \geq c$, we say \mathbf{A} is *full rank* if $\text{rank}(\mathbf{A})=c$. If $\text{rank}(\mathbf{A}) < c$, then we say \mathbf{A} is less than full rank. In linear regression, the matrix of covariates \mathbf{X} must have full rank in order for our parameter estimates, $\hat{\beta}$, to be unique.

A square matrix that is less than full rank is called *singular*, while a full rank square matrix is called *nonsingular*. One method to find the rank of a matrix is by determining the number of linearly independent columns of a matrix.



Another method to find rank is by using elementary row operations to transform the matrix to a triangular matrix; once the matrix is in triangular form, we can determine the rank visually.

The three elementary row operations are

1. multiplying a row by a nonzero constant,
2. adding one row to another, and
3. exchanging two rows.

Other approaches to finding rank include using matrix decompositions or finding eigenvalues (the rank of a square matrix is equal to the number of nonzero eigenvalues).

Determinants

The *determinant* is a single number summary of a **square matrix** that gives us information about the rank of the matrix. The determinant of a square matrix \mathbf{A} is denoted $| \mathbf{A} |$ or $\det(\mathbf{A})$.

- The determinant of a diagonal or triangular matrix equals the product of the diagonal values.

Exercise:

$$\left| \begin{bmatrix} 1 & 6 & 5 \\ 0 & 2 & 4 \\ 0 & 0 & 3 \end{bmatrix} \right| = \text{ } \square$$

- For any 2×2 matrix,

$$\left| \begin{bmatrix} a & b \\ c & d \end{bmatrix} \right| = ad - bc.$$

-
- For any 3×3 matrix,

$$\begin{aligned} \left| \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \right| &= a \left| \begin{bmatrix} e & f \\ h & i \end{bmatrix} \right| - d \left| \begin{bmatrix} b & c \\ h & i \end{bmatrix} \right| + g \left| \begin{bmatrix} b & c \\ e & f \end{bmatrix} \right| \\ &= a(ei - hf) - d(bi - hc) + g(bf - ec). \end{aligned}$$

Useful properties of determinants

$|\mathbf{A}_{n \times n}| = 0 \Leftrightarrow \text{rank}(\mathbf{A}) < n$
 $\Leftrightarrow \mathbf{A}$ is less than full rank
 \Leftrightarrow the *inverse* of \mathbf{A} does not exist
 \Leftrightarrow the columns of \mathbf{A} are linearly dependent,

while

$|\mathbf{A}_{n \times n}| \neq 0 \Leftrightarrow \text{rank}(\mathbf{A}) = n$
 $\Leftrightarrow \mathbf{A}$ is full rank
 \Leftrightarrow the *inverse* of \mathbf{A} exists
 \Leftrightarrow columns of \mathbf{A} are linearly independent. 

For full rank matrices that conform, $|\mathbf{AB}| = |\mathbf{A}| |\mathbf{B}|$. In addition,
 $|\mathbf{A}'| = |\mathbf{A}|$.

Positive Definite and Semidefinite Matrices

Let \mathbf{A} be an $n \times n$ symmetric matrix. \mathbf{A} is *positive definite* if and only if

- 1. $a_{ii} > 0$ for all $i = 1, \dots, n$
- 2. the determinant of every square submatrix of upper-left corner of \mathbf{A} is positive. That is,

$$\begin{array}{c} \text{ } \\ \text{ } \end{array} \quad \left| \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \right| > 0,$$

$$\left| \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{bmatrix} \right| > 0,$$

⋮

$$|\mathbf{A}| > 0.$$

The matrix \mathbf{A} is *positive semidefinite* if we replace “ > 0 ” in (1) and (2) with “ ≥ 0 ”. A matrix is called *nonnegative definite* if it is positive definite or positive semidefinite.

Covariance matrices are nonnegative definite. 

Exercise:

Let

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 1 & 1 \\ -1 & 4 & 0 & 2 \\ 1 & 0 & 1 & 3 \\ 1 & 2 & 3 & 2 \end{bmatrix}.$$

Is \mathbf{A} positive definite, positive semidefinite, or neither? 

Inverses and Generalized Inverses

Suppose we have a system of equations like

$$(\mathbf{X}_{n \times p})' (\mathbf{X}_{n \times p}) \hat{\boldsymbol{\beta}} = (\mathbf{X}_{n \times p})' \mathbf{y}_{n \times 1},$$

where are the *normal equations* for the linear model. In order to solve these equations and obtain our estimate, $\hat{\boldsymbol{\beta}}$, we would like to divide by  $\mathbf{X}'\mathbf{X}$ in some sense. Because $\mathbf{X}'\mathbf{X}$ is a matrix, this presents us some difficulty. Thus we develop the idea of a matrix inverse.

Consider the $n \times n$ matrix \mathbf{A} . If \mathbf{A} has full rank, then $\text{rank}(\mathbf{A})=n$, and there exists a unique matrix, \mathbf{A}^{-1} , called the *inverse* of \mathbf{A} , such that 

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}.$$

Some properties of inverses

1. For a scalar, $\mathbf{A}_{1 \times 1} = a$, $\mathbf{A}^{-1} = \frac{1}{a}$.
2. The inverse of a diagonal matrix is the diagonal matrix of reciprocals of the diagonal elements.

-
- 3. For conforming full rank matrices, $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.
 - 4. A symmetric matrix has a symmetric inverse.
 - 5. The inverse of the transpose is the transpose of the inverse. That is, $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$. $(\mathbf{A}')^{-1}$ is also sometimes denoted \mathbf{A}^{-T} .
 - 6. The determinant of the inverse is the inverse of the determinant. That is, $|\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|}$.
7. The inverse of the 2×2 matrix $\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ is given by
- $$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}.$$

Generalized Inverses For certain linear models (some ANOVA models for example), we will want to obtain inverses of matrices that are not full rank. For any $\mathbf{A}_{r \times c}$, there exists a *generalized inverse*, denoted $\mathbf{A}_{c \times r}^{-}$, such that

$$\mathbf{A}_{r \times c} \mathbf{A}_{c \times r}^{-} \mathbf{A}_{r \times c} = \mathbf{A}_{r \times c}.$$

The *generalized inverse* is not unique.

The *Moore-Penrose generalized inverse (MPGI)*, \mathbf{A}^{+} , is a unique type of generalized inverse satisfying the following properties:

- $\mathbf{A}\mathbf{A}^{+}\mathbf{A} = \mathbf{A}$ (definition of generalized inverse)
- $\mathbf{A}^{+}\mathbf{A}\mathbf{A}^{+} = \mathbf{A}^{+}$ (\mathbf{A} is a generalized inverse of \mathbf{A}^{+})
- $(\mathbf{A}^{+}\mathbf{A})' = \mathbf{A}^{+}\mathbf{A}$ ($\mathbf{A}^{+}\mathbf{A}$ is symmetric)
- $(\mathbf{A}\mathbf{A}^{+})' = \mathbf{A}\mathbf{A}^{+}$ ($\mathbf{A}\mathbf{A}^{+}$ is symmetric)

Facts about the MPG1

1. For $\mathbf{A}_{r \times c}$ with $r \leq c$ and \mathbf{A} of full row rank r ,

$$\mathbf{A}_{c \times r}^+ = \mathbf{A}'(\mathbf{A}\mathbf{A}')^{-1}.$$

2. For $\mathbf{A}_{r \times c}$ with $c \leq r$ and \mathbf{A} of full column rank c ,

$$\mathbf{A}_{r \times c}^+ = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'.$$

3. The Moore-Penrose generalized inverse of a less than full rank diagonal matrix is the diagonal matrix with reciprocals of the nonzero elements on the diagonal in the same locations as the nonzero elements, and zero elsewhere.

Both the Moore-Penrose generalized inverse and other generalized inverses reduce to the regular inverse when the matrix of interest is square and full rank.



Eigenvalues, Eigenvectors, and the Spectral Decomposition

Eigenanalysis is defined only for square matrices. Most interest in decomposing matrices in statistics lies with symmetric matrices, for example, covariance matrices.

Suppose \mathbf{A} is an $n \times n$ matrix (not necessarily symmetric). A *right eigenvector* of \mathbf{A} is any nonzero $n \times 1$ vector \mathbf{x} satisfying



$$\mathbf{Ax} = \lambda \mathbf{x},$$

where λ is the *eigenvalue* corresponding to \mathbf{x} . Eigenvalues and eigenvectors are also called *characteristic values* and *characteristic vectors* (*eigen* is German for *characteristic*).



Eigenvectors are not unique (prove it?), the convention is to scale the eigenvector \mathbf{x} so that $\mathbf{x}'\mathbf{x} = 1$, normalizing it to unit length.

Finding Eigenvectors and Eigenvalues

Using the definition of eigenvectors, we can find the *characteristic equation* that is used to find eigenvalues.

$$\mathbf{Ax} = \lambda \mathbf{x}$$

$$\mathbf{Ax} - \lambda \mathbf{x} = \mathbf{0}$$

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = \mathbf{0}$$

$$|\mathbf{A} - \lambda \mathbf{I}| = \mathbf{0}.$$

The last step follows because \mathbf{x} is a nonzero vector.

The characteristic equation of an $(n \times n)$ matrix equals an n^{th} degree polynomial in λ . An n^{th} degree polynomial has n roots, so a $(n \times n)$ matrix has n eigenvalues. For a (2×2) matrix,

$$\begin{aligned}
 |\mathbf{A} - \lambda\mathbf{I}| &= \left| \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0 \\
 \left| \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right| &= \left| \begin{bmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{bmatrix} \right| = 0 \\
 (a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21} &= 0 \\
 \lambda^2 - \lambda(a_{11} + a_{22}) + a_{11}a_{22} - a_{12}a_{21} &= 0.
 \end{aligned}$$

Once the eigenvalues are found, eigenvectors corresponding to these eigenvalues may be found using the equation $\mathbf{Ax} = \lambda\mathbf{x}$.

Example:

Suppose

$$\mathbf{A} = \begin{bmatrix} 5 & -3 \\ 4 & -2 \end{bmatrix}.$$

The characteristic equation is

$$\begin{aligned} |\mathbf{A} - \lambda\mathbf{I}| &= \left| \begin{bmatrix} 5 - \lambda & -3 \\ 4 & -2 - \lambda \end{bmatrix} \right| \\ &= (5 - \lambda)(-2 - \lambda) - (-3)(4) \\ &= \lambda^2 - 3\lambda + 2 \\ &= (\lambda - 2)(\lambda - 1) = 0, \end{aligned}$$

so the eigenvalues of \mathbf{A} are 2 and 1.

We find eigenvectors corresponding to these eigenvalues by solving $\mathbf{Ax} = \lambda\mathbf{x}$ for $\lambda = 1$ and $\lambda = 2$.



For $\lambda = 1$, we have

$$\begin{bmatrix} 5 & -3 \\ 4 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

which leads to the system of equations

$$5x_1 - 3x_2 = x_1$$

$$4x_1 - 2x_2 = x_2.$$

Rearranging both equations, we see $x_1 = \frac{3}{4}x_2$. So one eigenvector corresponding to the eigenvalue 1 is $(3, 4)'$.

To normalize this vector, we take

$$\begin{bmatrix} 3 & 4 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix} = 9 + 16 = 25$$

and divide by the square root of this number. So the normalized eigenvector corresponding to the eigenvalue 1 is $(\frac{3}{5}, \frac{4}{5})'$.

Similarly, we can show the normalized eigenvector corresponding to $\lambda = 2$ is $\frac{1}{\sqrt{2}}(1, 1)'$.

Some Properties of Eigenvalues and Eigenvectors

- For $\mathbf{A}_{n \times n}$, the number of distinct eigenvalues ranges from 1 to n .
- The trace of a matrix is the sum of the eigenvalues. That is,
 $\text{trace}(\mathbf{A}) = \sum_{i=1}^n \lambda_i$. 
- The determinant of a matrix is the product of its eigenvalues.
That is, $|\mathbf{A}| = \prod_{i=1}^n \lambda_i$.
- \mathbf{A} full rank $\Leftrightarrow \mathbf{A}$ has no zero eigenvalues 
- $|\mathbf{A}| = 0 \Leftrightarrow$ at least one eigenvalue is zero $\Leftrightarrow \mathbf{A}$ is not full rank
- The number of nonzero eigenvalues of \mathbf{A} is $\text{rank}(\mathbf{A})$.
- Small eigenvalues imply that there are near-linear dependencies in the columns of a matrix \mathbf{A} .
- \mathbf{A} is positive definite if $\min(\lambda_i) > 0$
- \mathbf{A} is positive semidefinite if $\min(\lambda_i) \geq 0$

Spectral Decomposition

It is sometimes easier to deal with matrices if we write them as a product of more simple matrices that have special structures. Matrix decomposition allows us to write matrices as a product of simpler matrices. We will see one such decomposition, the *spectral decomposition*, later in the course.

The *spectral decomposition* allows us to write any symmetric matrix in terms of an orthogonal matrix and a diagonal matrix of eigenvalues.

Spectral Theorem: Suppose \mathbf{A} is an $(n \times n)$ symmetric matrix. Then there exists an orthogonal (column orthonormal) matrix \mathbf{V} such that

$$\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}',$$

where $\Lambda = \text{Diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ is an $(n \times n)$ diagonal matrix of the ordered eigenvalues of \mathbf{A} so that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$. \mathbf{V} is the orthogonal matrix of eigenvectors corresponding to the eigenvalues of \mathbf{A} . The eigenvalues and eigenvectors must be in the same order.

Example: Verify that the spectral decomposition of

$$\mathbf{A} = \begin{bmatrix} 1.5 & -0.5 \\ -0.5 & 1.5 \end{bmatrix}$$

is given by $\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}'$, where


$$\mathbf{V} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}, \text{ and } \Lambda = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}.$$

Once a matrix is in spectral decomposition form, it is easy to obtain the inverse.

- For \mathbf{A} full rank, we have $\mathbf{A}^{-1} = \mathbf{V}(\Lambda)^{-1}\mathbf{V}'$. 
- For \mathbf{A} less than full rank, we have $\mathbf{A}^+ = \mathbf{V}(\Lambda)^+\mathbf{V}'$.

Thus computing the reciprocals of the eigenvalues allows us to find the inverse or Moore-Penrose generalized inverse of a symmetric matrix in this form.

We will see the spectral decomposition again in Chapter 8.

Singular Value Decomposition (SVD)

The *singular value decomposition* gives us a more accurate way to find the inverse of an ill-conditioned matrix. Both SAS and S-plus use the SVD when fitting linear models. The SVD is valid for *any* matrix, while the spectral decomposition is valid only for symmetric matrices.

For *any* matrix $\mathbf{A}_{m \times n}$ with $m \geq n$, there exist orthogonal matrices $U_{m \times m}$ and $V_{n \times n}$ along with an $(m \times n)$ matrix \mathbf{S} such that

$$\mathbf{A}_{m \times n} = \mathbf{U}_{m \times m} \mathbf{S}_{m \times n} \mathbf{V}'_{n \times n}.$$

We define

$$\mathbf{S}_{m \times n} = \begin{bmatrix} \text{Diag}(s) \\ \mathbf{0}_{(m-n) \times n} \end{bmatrix},$$



where the vector $s_{n \times 1}$ contains the n *singular values* of \mathbf{A} . The rank r of \mathbf{A} is the number of nonzero singular values of \mathbf{A} . Singular values are computed as the positive square roots of the eigenvalues of $\mathbf{A}'\mathbf{A}$.



Facts about the SVD

- $\mathbf{U}'\mathbf{U} = I_m \quad \mathbf{V}'\mathbf{V} = I_n$
-  $\mathbf{A}'\mathbf{A} = \mathbf{V}\text{Diag}(\mathbf{s}^2)\mathbf{V}'$ (note: $\text{Diag}(\mathbf{s}^2)$ here means $\text{Diag}(\{s_i^2\})$)
-  $\mathbf{A}\mathbf{A}' = \mathbf{U}\text{Diag}(\mathbf{s}^2, \mathbf{0}_{m-n})\mathbf{U}'$
- The Moore-Penrose generalized inverse of \mathbf{A} may be written

$$\begin{aligned}\mathbf{A}^+ &= \mathbf{V}\mathbf{S}^+\mathbf{U}' \\ &= \mathbf{V} [\text{Diag}(s)^+ \quad \mathbf{0}_{n \times (m-n)}] \mathbf{U}',\end{aligned}$$

where $\text{Diag}(s)^+ = \text{Diag}(\{s_1^{-1}, s_2^{-1}, \dots, s_r^{-1}, 0, \dots, 0\})$.

- SVD is also widely used in statistics where it is related to principal component analysis, and in signal processing and pattern recognition.

Random Vectors and Matrices

Suppose

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

is a vector of random variables with $E(Y_i) = \mu_i$, $\text{Var}(Y_i) = \sigma_{ii}$, and $\text{Cov}(Y_i, Y_j) = \sigma_{ij}$.

The expectation of the random vector \mathbf{Y} is defined

$$E(\mathbf{Y}) = \begin{pmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \boldsymbol{\mu}.$$

Suppose \mathbf{Z} is an $(n \times p)$ matrix of random variables. Then

$$E(\mathbf{Z}) = \begin{pmatrix} E(Z_{11}) & \dots & E(Z_{1p}) \\ \vdots & \dots & \vdots \\ E(Z_{n1}) & \dots & E(Z_{np}) \end{pmatrix}.$$

Thus the expectation of a random matrix is the matrix of the expectations.

For \mathbf{Y} an $(n \times 1)$ random vector, the *covariance matrix* of \mathbf{Y} is

$$\text{Cov}(\mathbf{Y}) = E[(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})'] \quad \square$$

$$= \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{n1} & \dots & \dots & \sigma_{nn} \end{pmatrix},$$



where $\sigma_{ij} = E[(Y_i - \mu_i)(Y_j - \mu_j)']$, $i, j = 1, \dots, n$.

Suppose $\boldsymbol{\mu} = E(\mathbf{Y})$ and covariance matrix $\boldsymbol{\Sigma} = \text{Cov}(\mathbf{Y})$. In addition, suppose $\mathbf{A}_{r \times n}$ is a matrix of constants and $\mathbf{b}_{r \times 1}$ is a vector of constants. Then

$$E(\mathbf{A}\mathbf{Y} + \mathbf{b}) = \mathbf{A}E(\mathbf{Y}) + \mathbf{b} = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$$

$$\text{Cov}(\mathbf{A}\mathbf{Y} + \mathbf{b}) = \mathbf{A}\text{Cov}(\mathbf{Y})\mathbf{A}' = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'. \quad \square$$

Let $\mathbf{W}_{r \times 1}$ be a random vector with $E(\mathbf{W}) = \boldsymbol{\gamma}$. Then

$$\text{Cov}(\mathbf{W}, \mathbf{Y}) = E [(\mathbf{W} - \boldsymbol{\gamma})(\mathbf{Y} - \boldsymbol{\mu})'] , \quad \text{Talk icon}$$

where $\text{Cov}(\mathbf{W}, \mathbf{Y})$ is an $(r \times n)$ matrix of covariances with ij^{th} element equal to $\text{Cov}(W_i, Y_j)$.

Important Distributions for Linear Models

The theory of linear models involves many statistical distributions, including the following distributions:

1. Gaussian (Normal) Distribution
2. Multivariate Normal Distribution
3. Chi-Squared Distribution
4. Non-Central Chi-Squared Distribution
5. t Distribution
6. Non-Central t Distribution
7. F Distribution
8. Non-Central F Distribution.

We will now discuss many of these distributions in detail.

Gaussian (Normal) Distribution

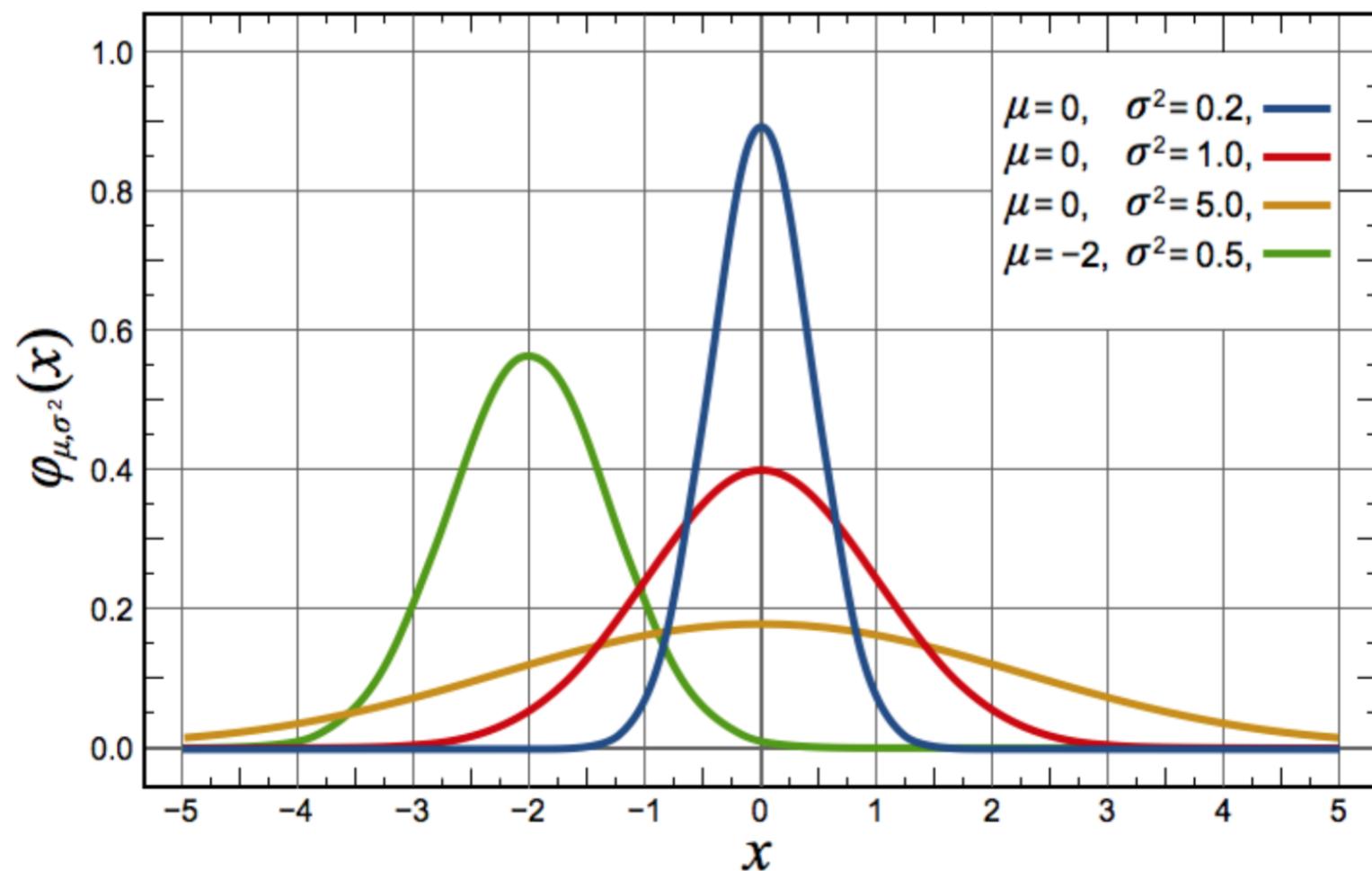
A random variable X has a Gaussian (normal) distribution with mean μ and variance σ^2 , written $X \sim N(\mu, \sigma^2)$, if X has density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}.$$

When $\mu = 0$ and $\sigma^2 = 1$, then we say $X \sim N(0, 1)$ has a *standard normal distribution*.

If $X \sim N(\mu, \sigma^2)$, then $\frac{x-\mu}{\sigma} \sim N(0, 1)$. For the standard normal distribution, 95% of the probability mass falls between -1.96 and 1.96 (roughly 2 standard deviations of the mean). Much of hypothesis testing is based on this fact.

The normal distribution is symmetric about its mean.



Multivariate Normal Distribution

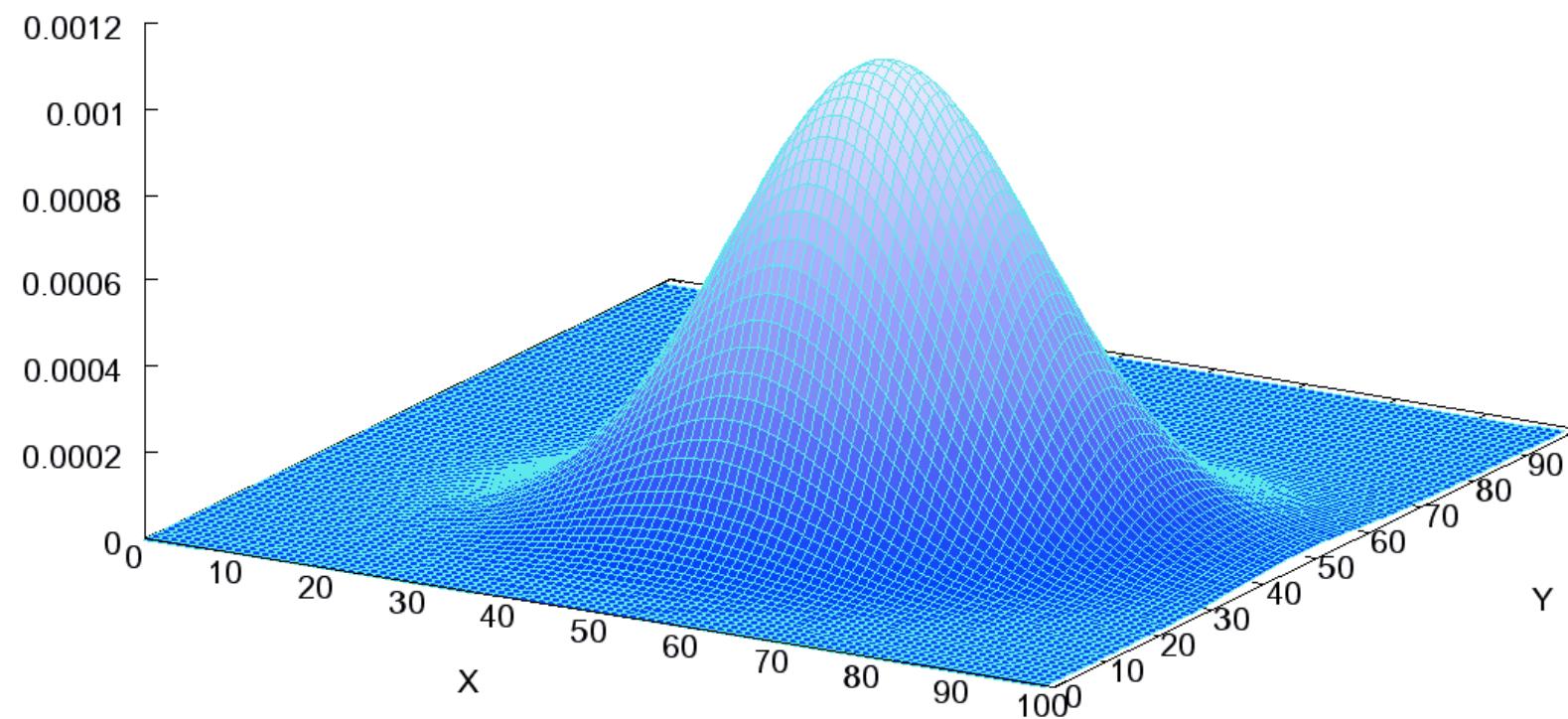
Suppose $X = (X_1, \dots, X_n)'$. Then X has an n dimensional multivariate normal distribution with mean μ and covariance matrix Σ if X has density

$$f(x) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (x - \mu)' \Sigma^{-1} (x - \mu) \right\}.$$

We write this $X \sim N_n(\mu, \Sigma)$.



This definition requires Σ to be positive definite.



Facts about the multivariate normal distribution

1. A linear transformation of a multivariate normal distribution yields another multivariate normal distribution. Suppose $X \sim N_n(\mu, \Sigma)$. For $A_{r \times n}$ a matrix of constants and $b_{r \times 1}$ a vector of constants, then $Y = AX + b$ has the multivariate normal distribution given by $Y \sim N_r(A\mu + b, A\Sigma A')$.
2. A linear combination of independent multivariate normal distributions is a multivariate normal distribution. Suppose X_1, \dots, X_k are independent with $X_i \sim N_n(\mu_i, \Sigma_i)$, $i = 1, \dots, k$. Suppose a_1, \dots, a_k are scalars and define 

$$Y = a_1X_1 + \dots + a_kX_k.$$

Then $Y \sim N(\mu^*, \Sigma^*)$, where $\mu^* = \sum_{i=1}^k a_i\mu_i$ and $\Sigma^* = \sum_{i=1}^k a_i^2\Sigma_i$. 

3. Marginal distributions of a multivariate normal distribution are multivariate normal distributions. Suppose $X \sim N_n(\mu, \Sigma)$.

Partition X into $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ where X_1 is $r \times 1$ and X_2 is $(n - r) \times 1$. Partition μ as $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ where μ_1 is $r \times 1$ and μ_2 is $(n - r) \times 1$. Similarly, partition Σ as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where Σ_{11} is $r \times r$, Σ_{12} is $r \times (n - r)$, $\Sigma_{21} = \Sigma'_{12}$ is $(n - r) \times r$, and Σ_{22} is $(n - r) \times (n - r)$. Then the marginal distribution of X_1 is given by $X_1 \sim N_r(\mu_1, \Sigma_{11})$, and the marginal distribution of X_2 is given by $X_2 \sim N_{(n-r)}(\mu_2, \Sigma_{22})$.

4. Conditional distributions of multivariate normal distributions are multivariate normal distributions. Suppose that $X \sim N_n(\mu, \Sigma)$. 

Using the same partition as above, then we have

$$\begin{aligned} X_1 \mid X_2 &= x_2 \\ &\sim N_r(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma^*), \end{aligned}$$

where $\Sigma^* = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

Chi-Squared Distribution

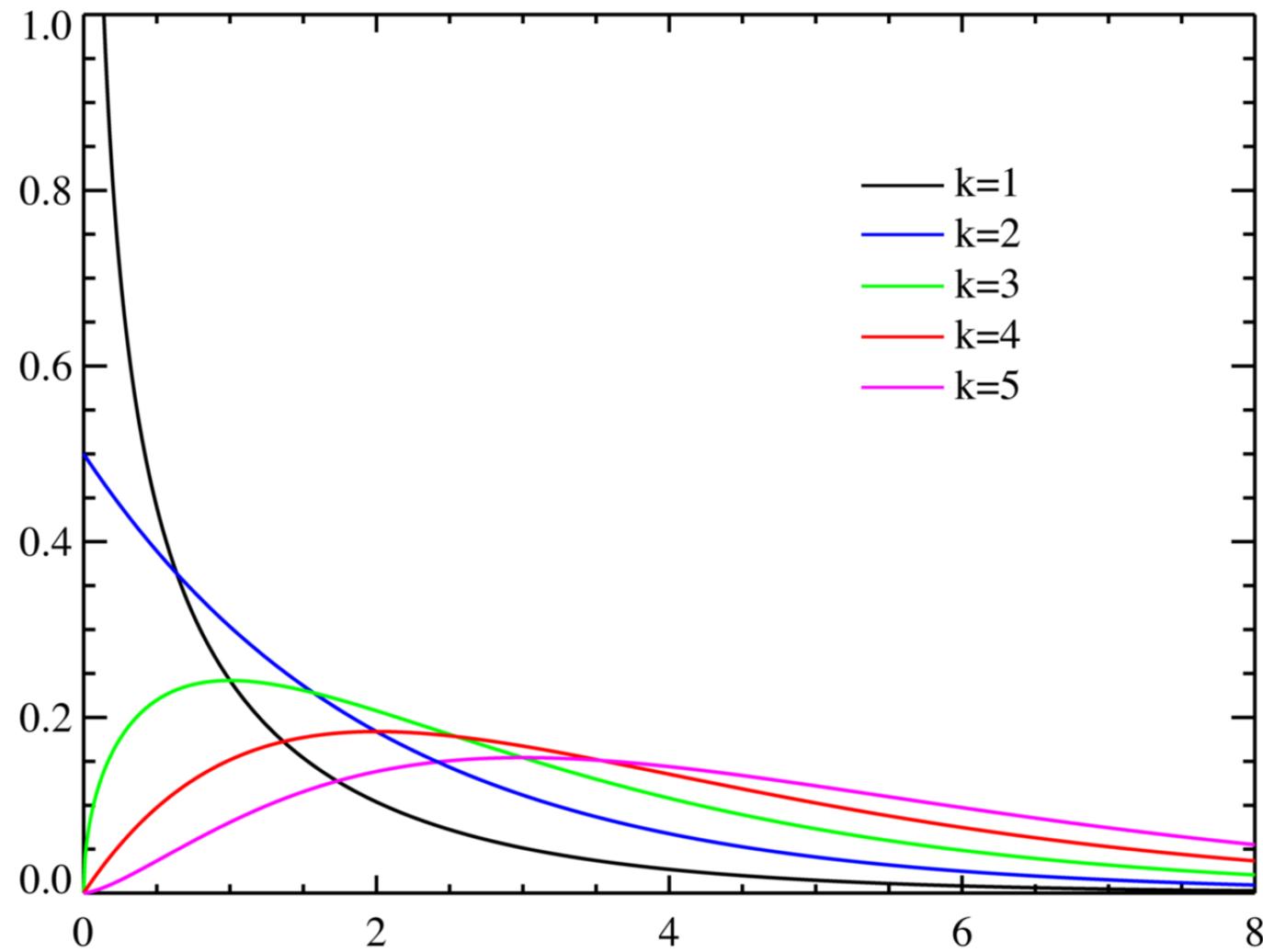
A random variable X has a central chi-squared distribution with n degrees of freedom, written $X \sim \chi^2(n)$, if the density of X is given by

$$f(x) = \left(\frac{1}{\Gamma(\frac{n}{2})} \right) \left(\frac{1}{2} \right)^{\frac{n}{2}} x^{\frac{n}{2}-1} \exp \left\{ -\frac{x}{2} \right\},$$

where $\Gamma(a)$ is the complete gamma function, given by

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx.$$

The chi-squared distribution is asymmetric and restricted to positive numbers. Its degrees of freedom determine the mean and variance of the distribution.



The chi-squared distribution is related to the normal distribution. If the random variable $Z \sim N(0, 1)$, then $Z^2 \sim \chi^2(1)$. In addition, if Z_1, Z_2, \dots, Z_n are independent, identically distributed $N(0, 1)$ random variables, then $W = \sum_{i=1}^n Z_i^2$ has a chi-squared distribution with n degrees of freedom; that is, $W \sim \chi^2(n)$.

The mean of a $\chi^2(n)$ distribution is n , and its variance is $2n$.

If $Z \sim N(\mu, 1)$, then Z^2 follows a non-central chi-squared distribution with 1 degrees of freedom and non-centrality parameter μ^2 .

The chi-squared distribution is used widely in the analysis of categorical data.

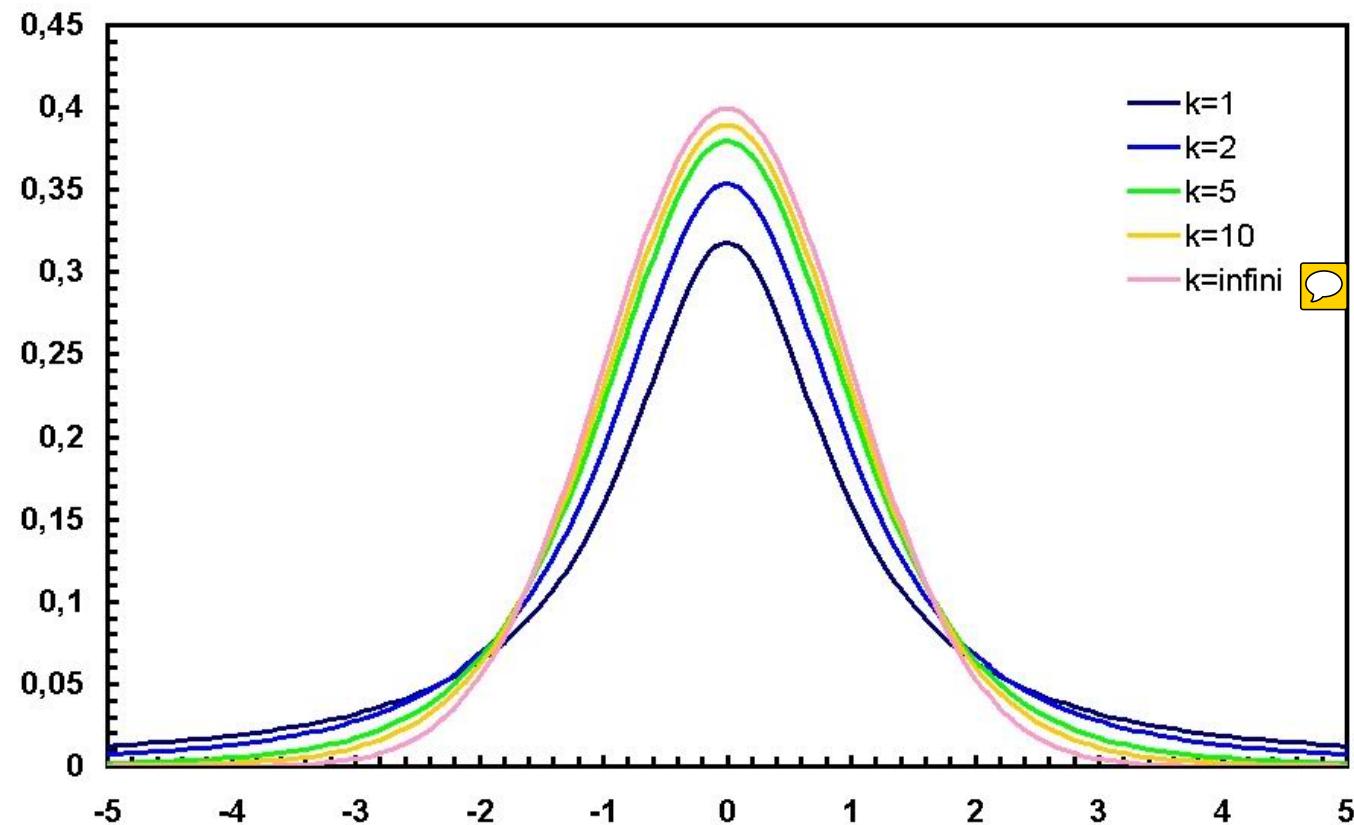
Student's t Distribution

Suppose $X \sim N(0, 1)$, $Y \sim \chi^2(n)$, with X and Y independent. The random variable

$$T = \frac{X}{\sqrt{\frac{Y}{n}}}$$



has a t distribution with n degrees of freedom. We write this as $T \sim t(n)$. Like the standard normal distribution, the t distribution is symmetric about 0.



If $X \sim N(\mu, 1)$, then T has a non-central t-distribution with n degrees of freedom and non-centrality parameter μ .

The *degrees of freedom* n determines the amount of variability in the distribution. As the number of degrees of freedom increases, the variability of the t distribution decreases. In fact, as the number of degrees of freedom gets large, the t distribution approximates the standard normal distribution. With smaller degrees of freedom, the t distribution resembles a normal distribution with fatter tails.

A $t(1)$ distribution, which has 1 degree of freedom, is not well-behaved and is called a *Cauchy distribution*.

Fisher's F Distribution

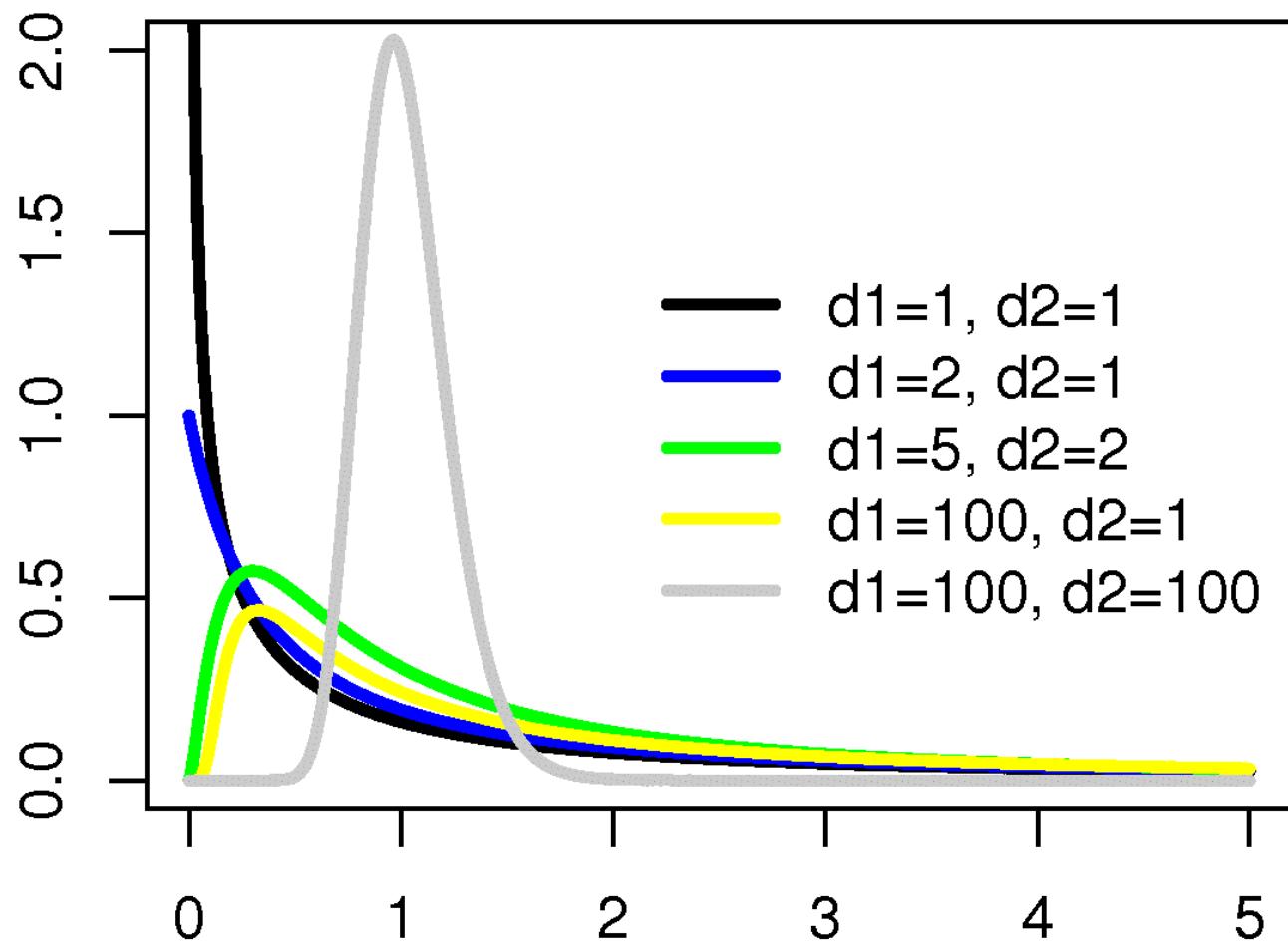
Suppose $X_1 \sim \chi^2(n_1)$ and $X_2 \sim \chi^2(n_2)$, and X_1 and X_2 are independent. The random variable

$$F = \frac{\left(\frac{X_1}{n_1}\right)}{\left(\frac{X_2}{n_2}\right)}$$

has a central F distribution with (n_1, n_2) degrees of freedom. We write this $F \sim F(n_1, n_2)$.

We call n_1 the *numerator* degrees of freedom and n_2 the *denominator* degrees of freedom. The central F distribution is used in hypothesis tests of nested linear models. If $T \sim t(\nu)$, then $T^2 \sim F(1, \nu)$. The F distribution is asymmetric and restricted to positive numbers.

If $X_1 \sim \chi^2(n_1; \mu^2)$, then F follows a non-central F distribution.



Maximum Likelihood Estimation

Maximum likelihood estimates (MLE's) have excellent large-sample properties and are applicable in a wide variety of situations. Examples of maximum likelihood estimates include the following:

- The sample average \bar{X} of a group of independent and identically normally distributed observations X_1, \dots, X_n is a MLE.
- Parameter estimates in a linear regression model fit to normally distributed data are maximum likelihood estimates.
- Parameter estimates in a logistic regression model are maximum likelihood estimates.
- The estimate $s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$ of the variance of a group of independent and identically normally distributed observations is *not* a MLE. (The MLE of the variance is $(\frac{n-1}{n}) s^2$.)



Finding Maximum Likelihood Estimates

Let $L(\mathbf{Y} \mid \boldsymbol{\theta})$ denote the *likelihood function* for $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ from some population described by the parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_p)$.

The maximum likelihood estimate of $\boldsymbol{\theta}$ is given by the estimator $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_p)$ for which


$$L(\mathbf{Y} \mid \hat{\boldsymbol{\theta}}) > L(\mathbf{Y} \mid \boldsymbol{\theta}^*),$$

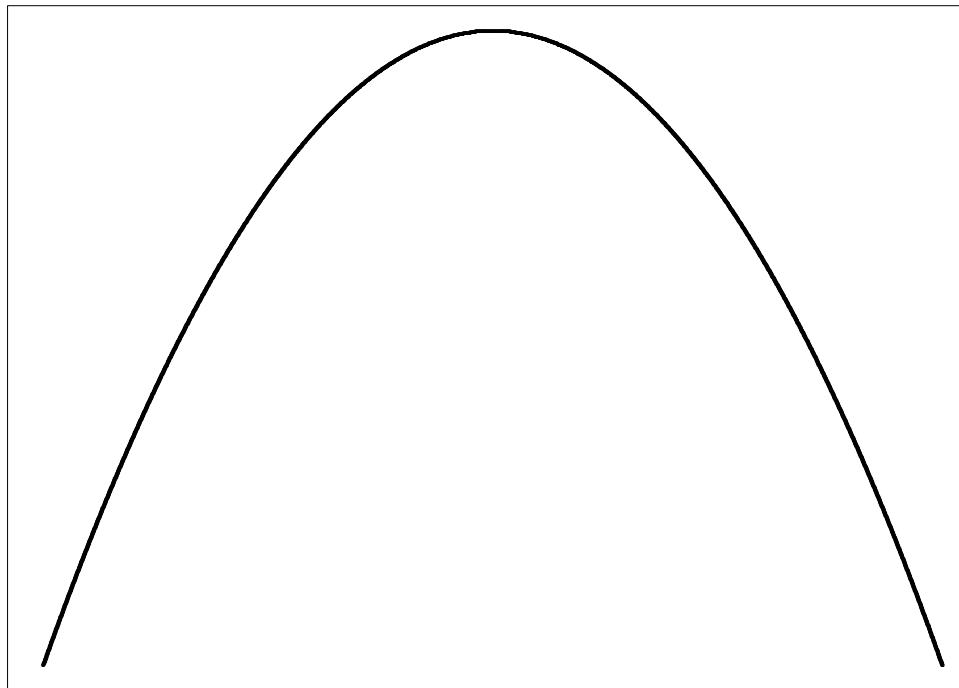
where $\boldsymbol{\theta}^*$ is any other estimate of $\boldsymbol{\theta}$.

Thus the maximum likelihood estimate is the most “probable” or “likely” for the data.

Maximizing the likelihood $L(\mathbf{Y} | \boldsymbol{\theta})$ is equivalent to maximizing the natural logarithm $\ln(L(\mathbf{Y} | \boldsymbol{\theta})) = \ell(\mathbf{Y} | \boldsymbol{\theta})$, called the log-likelihood.

The maximum likelihood estimates are typically found as the solutions of the p equations obtained by setting the p partial derivatives of $\ell(\mathbf{Y} | \boldsymbol{\theta})$ with respect to each θ_j , $j = 1, \dots, p$, equal to zero.

Why do we solve the derivatives for zero? The derivative gives us the slope of the likelihood (or log-likelihood), and when the slope is zero, we know that we are at either a local minimum or local maximum. (The second derivative is negative for a maximum and positive for a minimum.)



When closed form expressions for maximum likelihood estimates do not exist, computer algorithms may be used to solve for the estimates.

Example:

Let Y_i , $i = 1, \dots, n$ be i.i.d. normal random variables with mean μ and variance σ^2 , so $Y_i \sim N(\mu, \sigma^2)$, $i = 1, \dots, n$.

The density of Y_i is given by

$$f(Y_i | \mu, \sigma^2) = (2\pi)^{-\frac{1}{2}} (\sigma^2)^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2\sigma^2} (Y_i - \mu)^2 \right\}.$$

Find the maximum likelihood estimates of μ and σ^2 .

Hypothesis Testing and Interval Estimation with MLE's

It can be shown (based on large-sample properties of MLE's) that

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_1)}}$$

is approximately $N(0, 1)$ when the sample size is large.

A test of $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$ can be based on the Z statistic $\frac{\hat{\beta}_1}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_1)}}$, which has approximately a $N(0, 1)$ distribution under H_0 . This test is called a *Wald test*.

By a similar argument, an approximate $100(1 - \alpha)\%$ large-sample confidence interval for β_1 takes the form

$$\hat{\beta}_1 \pm Z_{1 - \frac{\alpha}{2}} \sqrt{\widehat{\text{Var}}(\hat{\beta}_1)},$$

where $Pr(Z > Z_{1-\frac{\alpha}{2}}) = \frac{\alpha}{2}$ when $Z \sim N(0, 1)$.

NOTE: Testing σ^2 is more complicated, and the Wald test for the hypothesis $\sigma^2 = 0$ is not recommended because the value $\sigma^2 = 0$ is on the boundary of the parameter space for σ^2 .

Next: Simple Linear Regression

Reading Assignment:

- Weisberg Chapter 2: Simple Linear Regression

Lecture 3: Simple Linear Regression

Reading

- Weisberg, Chapter 2: “Simple Linear Regression” (Required)

We will consider the case in which we observe a single response and one covariate. (For most of the course, we assume that the covariates are fixed and known.)

This lecture will serve as a gentle introduction to the matrix-based material utilized in the rest of this course, and provide some intuition and motivations for later concepts.

Model Notation

- Roman letters (a, b, c, \dots, x, y, z) represent constants and random variables.
 - Letters at beginning of the alphabet, (a, b, c, \dots), often represent constants.
 - Letters in the middle of the alphabet, (\dots, i, j, k, \dots), often represent indices.
 - Letters at the end of the alphabet, (\dots, x, y, z), often represent random variables.
- Greek letters ($\alpha, \beta, \gamma, \dots, \chi, \psi, \omega$) represent parameters to be estimated.
- Lower-case symbols in bold type ($\mathbf{x}, \mathbf{y}, \boldsymbol{\beta}, \dots$) represent vectors.
- Upper-case symbols in bold type ($\mathbf{X}, \mathbf{A}, \boldsymbol{\Sigma}, \dots$) represent matrices.

Simple Linear Regression

Regression analysis is used to explore the nature of the relationship between a response variable and one or more covariates. In simple linear regression, we focus on the case when we consider only a single covariate.

Hypothesis: State-specific melanoma mortality rates are related to the latitude of the state.

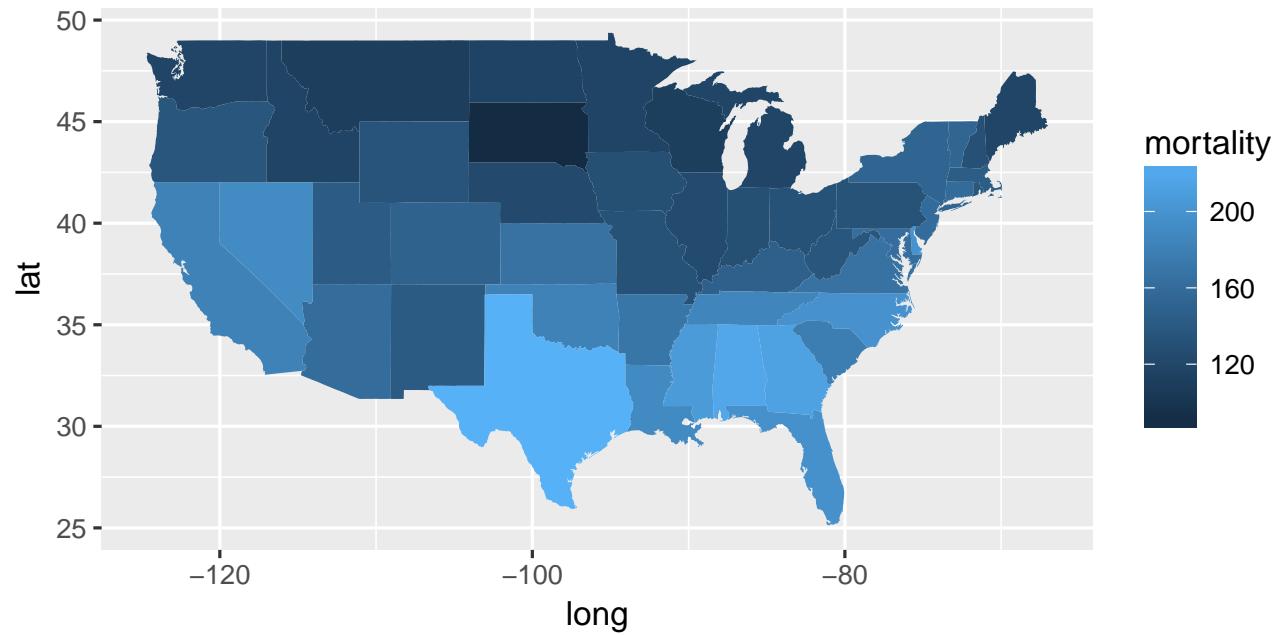
Define:

y_i =annual melanoma mortality in state i , $i = 1, \dots, 50$

x_i =latitude of the center of the state (in degrees) 

Our observations are $(x_1, y_1), \dots, (x_{50}, y_{50})$, where the x_i 's are known fixed values (*predictors*, *covariates*, or *independent variables*), and the y_i 's are random *response* variables (*dependent variables*).

<i>State</i>	<i>Melanoma Mortality Rate</i> (Deaths/Million)	<i>Latitude</i> (in degrees)
Alabama	219	33.00
Alaska	220	63.25
Florida	197	28.00
Hawaii	330	19.96
Maine	117	45.20
Minnesota	116	46.00
Mississippi	207	32.80
North Dakota	115	47.50
North Carolina	199	35.50
Tennessee	186	36.00
Vermont	153	44.00



How we do determine the relationship between melanoma mortality
and latitude?

Simple Linear Regression

Let us write the simple linear regression (SLR) model as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \dots, n \quad (1)$$

where

- n is the number of observations or sampling units
- y_1, \dots, y_n are the random responses,
- x_1, \dots, x_n are the fixed and known covariates
- $\varepsilon_1, \dots, \varepsilon_n$ is a vector of unobserved random errors, s.t.
 $\varepsilon_i \stackrel{i.i.d}{\sim} N(0, \sigma^2)$ 
- β_0, β_1 , and σ^2 are parameters (fixed and to be estimated from the data)

The index i corresponds to subjects or sampling units.

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

where

- $E[\varepsilon_i | x_i] = E[\varepsilon_i] = 0$
- On right side, only ε_i is random
- (y_i, x_i) forms the data for observation i

Model Assumptions

- Homogeneity 
- Independence 
- Linearity
- Existence 
- Gaussian Errors

Violations of these assumptions lead to problems with the application, estimation, and interpretation of linear models.

We also may write the simple linear regression (SLR) model by taking expectations of both sides

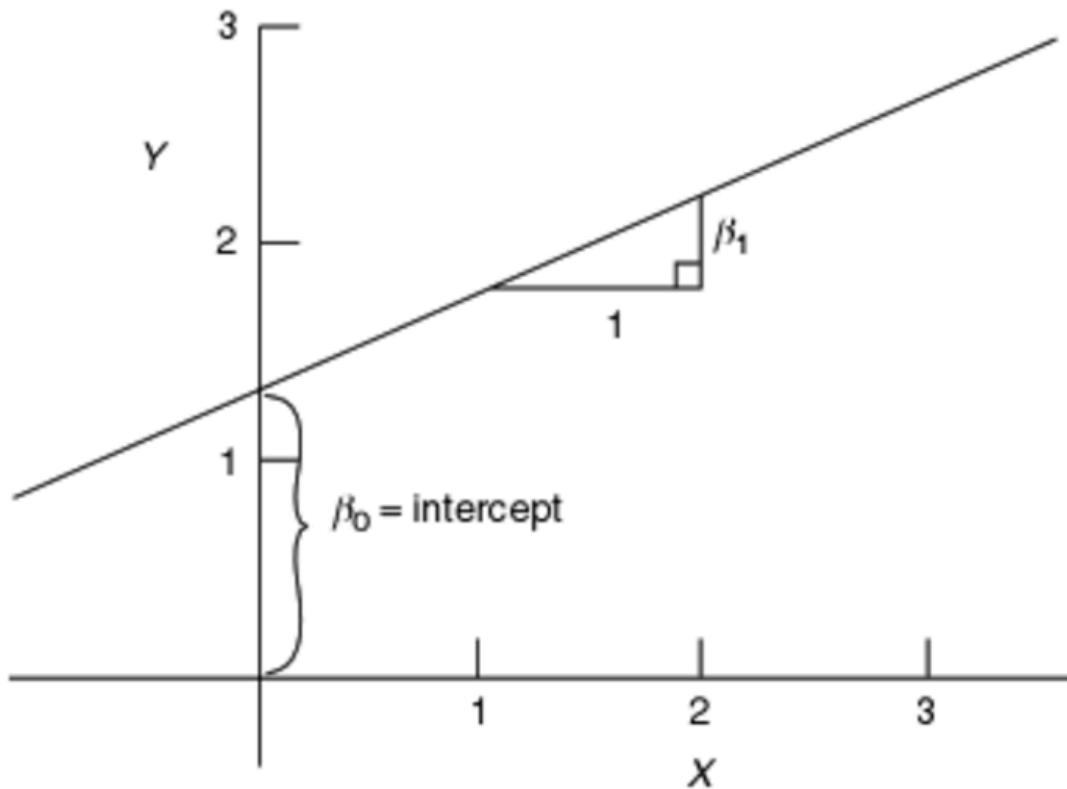
$$E[y_i|x_i] = E[\beta_0 + \beta_1 x_i + \epsilon_i|x_i]$$

$$= \beta_0 + \beta_1 x_i$$

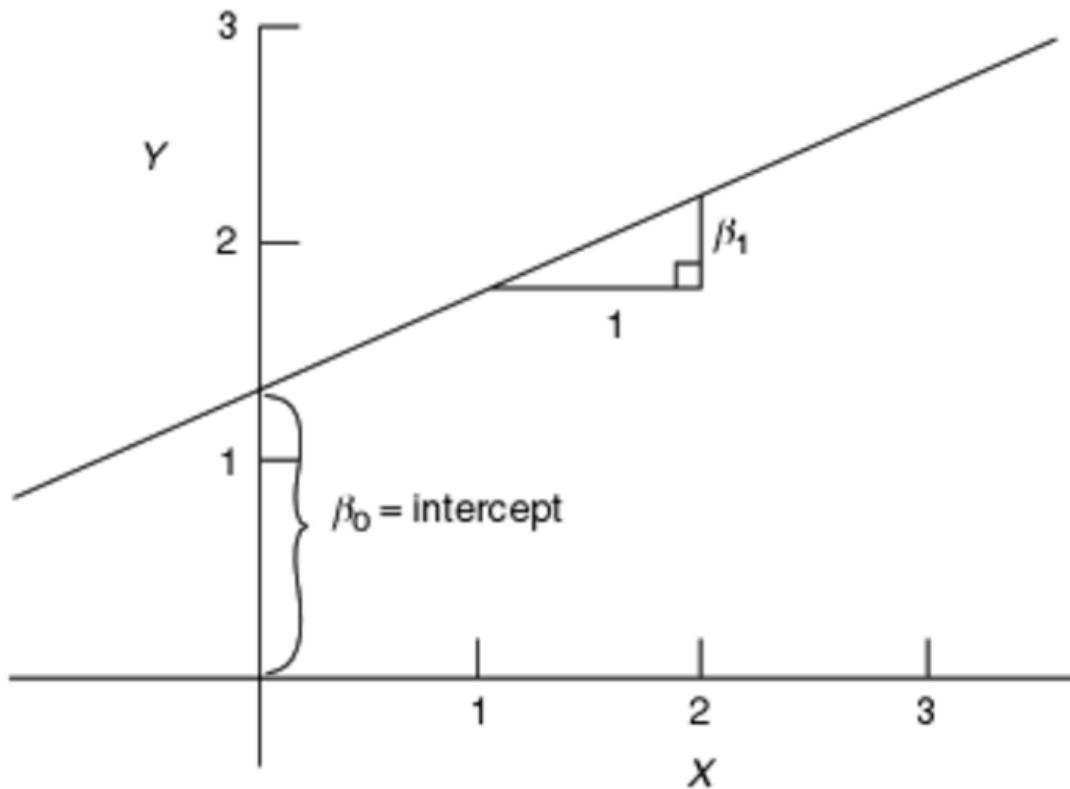
$$Var[y_i|x_i] = Var[\beta_0 + \beta_1 x_i + \epsilon_i|x_i]$$

$$= \sigma^2$$

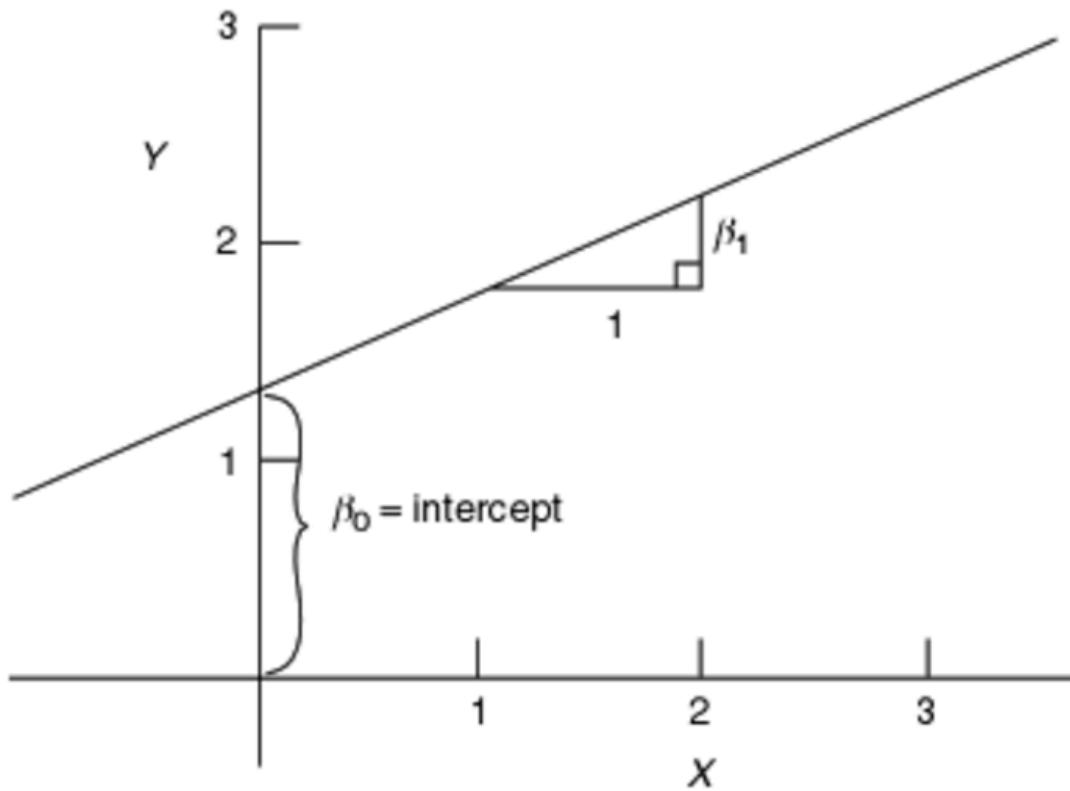
Based on the assumptions of the SLR, how did we arrive at this?
What does this say about the distribution of y_i ? We can plot this line
on the next slide



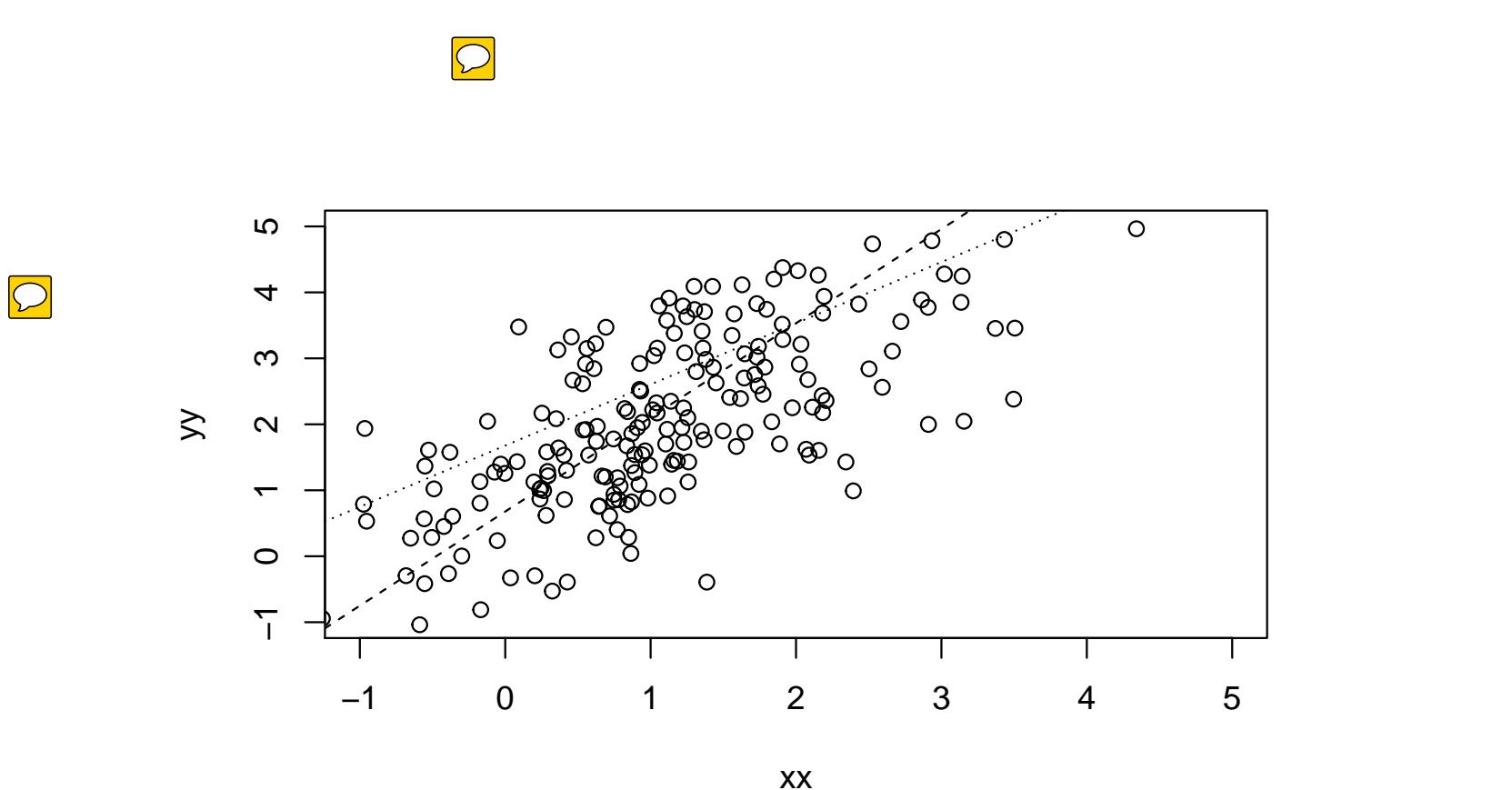
How we interpret the parameters β_0 , β_1 , and σ^2 in this model?



- β_0 : intercept, $E[y_i|x_i = 0]$
- β_1 : slope, change in $E[y_i|x_i]$ when $x_i \uparrow$ by 1 unit
- σ^2 : Variance of y_i about $E[y_i|x_i]$



Lets say that we knew β_0 , β_1 , and σ^2 , and had covariates x_1, \dots, x_n . Then, we could generate new data y_1, \dots, y_n using eq. 1.



However, in reality we typically only have observed data points $(x_1, y_1), \dots, (x_n, y_n)$ and do not know β_0 , β_1 , and σ^2 , and do not observe $\varepsilon_1, \dots, \varepsilon_n$. We estimate these quantities given the observed data. How to choose the best line?

Going back to the original example, we have the model:

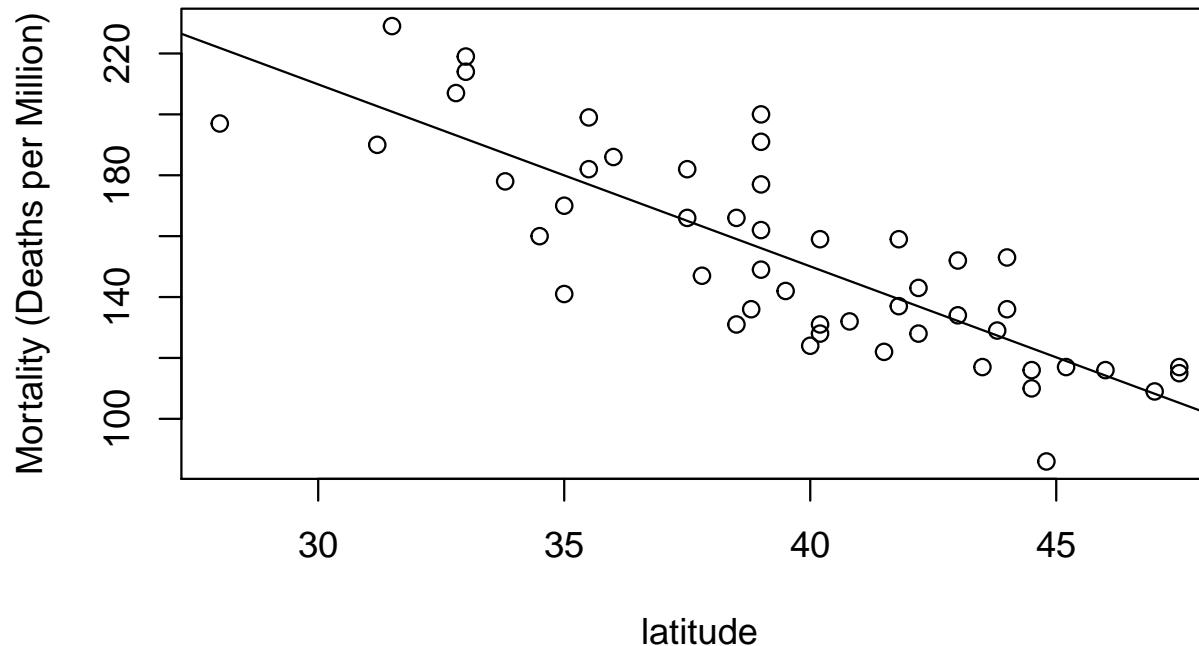
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, 50,$$

where (β_0, β_1) are *parameters* to be estimated. The ε_i 's represent random errors, which account for the fact that the response will vary even for states with the same latitudes (see, for example, MS and AL). We may wish to use this model to do the following.

- Estimate β_0 , β_1 , and σ^2 .
- Test hypotheses about β_1 or obtain confidence limits for β_1 , relating the statistical results back to the scientific question concerning the relationship between latitude and melanoma mortality.
- Predict a future y at a given x .

Fitting the data, we obtain the estimates $\hat{\beta}_0 = 389.189$ and $\hat{\beta}_1 = -5.978$.

We can use these estimates to draw the estimated regression line.



How do we interpret our parameter estimates?

Ordinary Least Squares (OLS) Estimation

- Many possible values of β_0, β_1
- Many different possible lines
- OLS is an approach to obtain estimates of these parameters, $\hat{\beta}_0, \hat{\beta}_1$, in addition to $\hat{\sigma}^2$

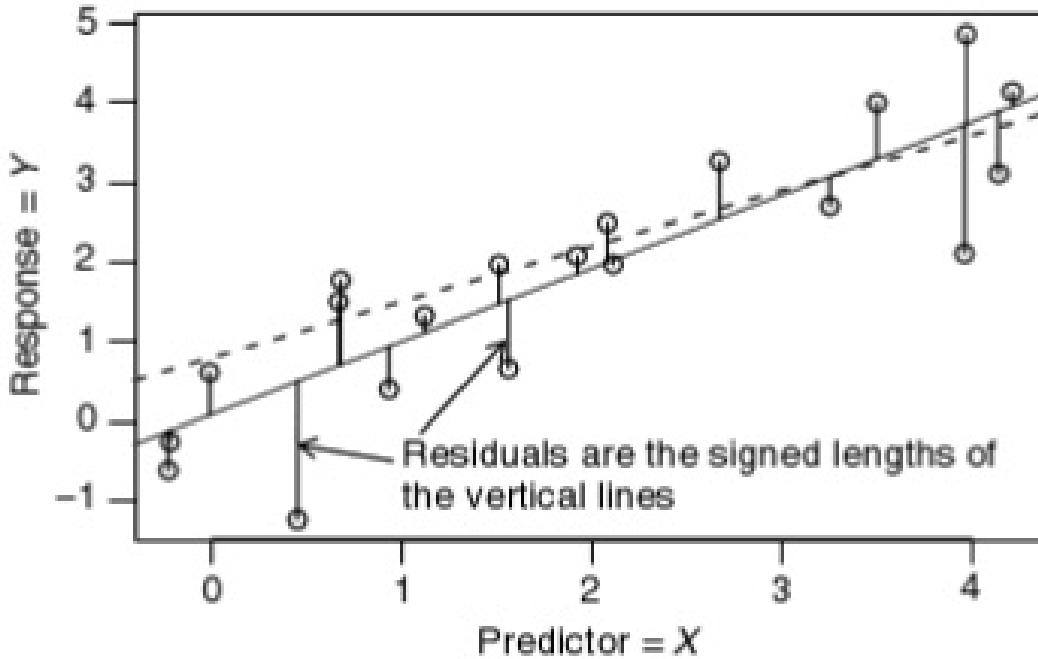
With these estimated values we may obtain two values


$$\hat{y}_i = \hat{E}[y_i|x_i] = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, \dots, n$$

where \hat{y}_i is the predicted value of y_i given $\hat{\beta}_0$, $\hat{\beta}_1$, x_i , and

$$\hat{\varepsilon}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), i = 1, \dots, n$$

where $\hat{\varepsilon}_i$ is called the residual for the i th observation.



Intuition behind OLS - graphical illustration (Weisberg pg. 25). Solid line is the fitted line, dashed line is the true line (why are these different?) that the data was generated from. But how do we select the fitted line, and how do we know it is "best" given the observed data?

Least Squares Criterion The values for $(\hat{\beta}_0, \hat{\beta}_1)$ are obtained by minimizing the RSS with respect to (β_0, β_1) . That is, we minimize

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Define the Residual Sum of Squares (RSS), also known as the Sums of Squares for Error (SSE), as the following

$$\begin{aligned} RSS(\hat{\beta}_0, \hat{\beta}_1) &= \sum_{i=1}^n [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \\ &= S_{YY} - \hat{\beta}_1^2 S_{XX} \end{aligned}$$

where $S_{XX} = \sum_{i=1}^n (x_i - \bar{x})^2$ and $S_{YY} = \sum_{i=1}^n (y_i - \bar{y})^2$

But how do we minimize this $RSS(\beta_0, \beta_1)$? Discussed in Weisberg A.3, we will briefly go over approach here.



Intuitively, $\hat{\sigma}^2$ can be estimated by averaging $\hat{\varepsilon}_i^2$. An unbiased estimate for $\hat{\sigma}^2$ is given by

$$\hat{\sigma}^2 = \frac{RSS}{n - 2}$$

This quantity is also referred to as the Residual Mean Square, or the Mean Square of the Errors (MSE). RSS may also be expressed as

$$RSS = \sum_{i=1}^n [\hat{\varepsilon}_i]^2$$

Why is this the case? Note that RSS is determined solely on $\hat{\beta}_0$ and $\hat{\beta}_1$.

Assuming that the errors are drawn from a normal distribution, then the Residual Mean Square divided by σ^2 will be distributed Chi-squared with $n - 2$ degrees of freedom. Then, we have that

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n-2} \chi^2(n-2)$$

so that

$$\begin{aligned} E[\hat{\sigma}^2] &= \frac{\sigma^2}{n-2} E[\chi^2(n-2)] \\ &= \frac{\sigma^2}{n-2} (n-2) \\ &= \sigma^2 \end{aligned}$$



→ $\hat{\sigma}^2$ is unbiased. This is in contrast to the MLE of σ^2 .

Properties of Least Squares Estimates We can see that $\hat{\beta}_0$ and $\hat{\beta}_1$ depend on the random ε_i 's (why?). If the errors are mean 0, then

$$E[\hat{\beta}_0] = \beta_0$$

$$E[\hat{\beta}_1] = \beta_1$$



In other words, $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased. Notice that we do not need the assumption of normality for this result to hold. The normal assumption of the errors impacts hypothesis testing.

$$E[\hat{\beta}_0] = \beta_0$$

$$E[\hat{\beta}_1] = \beta_1$$

Can you show how these estimators are unbiased?

Now lets look at the variances of teh estimates

$$Var[\hat{\beta}_0] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right)$$

$$Var[\hat{\beta}_1] = \frac{\sigma^2}{S_{XX}}$$

How do we obtain this result?

We can also derive the forms of the covariances and correlation of the estimates

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \frac{\bar{x}}{S_{XX}}$$

$$\rho(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\bar{x}}{\sqrt{\frac{S_{XX}}{n} + \bar{x}^2}}$$



What do these results imply? How do we obtain this result?

The Gauss-Markov theorem provides an optimality result for ols estimates. Among all estimates that are linear combinations of the y_1, \dots, y_n and unbiased, the ols estimates have the smallest variance. These estimates are called the best linear unbiased estimates , or blue. If one believes the model assumptions and is interested in using linear unbiased estimates, the ols estimates are the ones to use.

The means and variances, and covariances of the estimated regression coefficients do not require a distributional assumption concerning the errors. Since the estimates are linear combinations of the y_1, \dots, y_n , and hence linear combinations of the errors, the central limit theorem shows that the coefficient estimates will be approximately normally distributed if the sample size is large enough.

For smaller samples, if the errors are i.i.d, then the regression estimates $(\hat{\beta}_0, \hat{\beta}_1)$ will have a joint normal distribution with means, variances, and covariances as given before. When the errors are normally distributed, the OLS estimates can be justified using a completely different argument, since they are then also maximum likelihood estimates.

Estimated Variances

$$\widehat{Var}[\hat{\beta}_0] = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right)$$

$$\widehat{Var}[\hat{\beta}_1] = \frac{\hat{\sigma}^2}{S_{XX}}$$

$$se(\hat{\beta}_j) = \sqrt{\widehat{Var}[\hat{\beta}_j]}, j = 1, 2$$

Standard deviation - variability of an observable random variable like the response y_i or an unobservable random variance like the errors ε_i .

Standard error - square root of the **estimated** variance of a statistic like $\hat{\beta}_1$.

Confidence Intervals and t-tests Confidence intervals result in interval estimates, while tests provide methodology for making decisions concerning the value of a parameter or fitted value.

When the errors are $\stackrel{i.i.d}{\sim} N(0, \sigma^2)$, parameter estimates, fitted values, and predictions will be normally distributed. Why is this the case?

Confidence intervals and tests can be based on a t-distribution, which is the appropriate distribution with normal estimates but using $\hat{\sigma}^2$ to estimate the unknown variance σ^2 . Suppose we let $t(\alpha/2, d)$ be the value that cuts off $\alpha/2 \times 100\%$ in the upper tail of the t-distribution with d df.

-
- For example, the standard error of the intercept is
- se($\hat{\beta}_0$) = $\hat{\sigma}(1/n + \bar{x}^2/S_{XX})$. Hence, a $(1 - \alpha) \times 100\%$ CI is the set of points β_0 in the interval

$$\hat{\beta}_0 - t(\alpha/2, n - 2)se(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + t(\alpha/2, n - 2)se(\hat{\beta}_0).$$

$1 - \alpha$ percent of such intervals will include the true value.

A hypothesis test of

$$H_0 : \beta_0 = \beta_0^*, \beta_1 \text{ arbitrary},$$

$$H_A : \beta_0 \neq \beta_0^*, \beta_1 \text{ arbitrary}$$

is obtained by computing the t-statistic $t = \frac{\hat{\beta}_0 - \beta_0^*}{se(\hat{\beta}_0)}$ and referring this ratio to the t-distribution with $df = n - 2$, the number of df in the estimate of σ^2 .

From our Melanoma example, testing

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

would be addressing what question in particular?

Prediction If we found a new state and latitude value, x_{51} , can we predict what the mortality would be, given the fitted model? That is, we would like to know what y_{51} would be given x_{51} , assuming the existing data is relevant to the new data. A point prediction of y_{51} , say \tilde{y}_{51} , is simply 

$$\tilde{y}_{51} = \hat{\beta}_0 + \hat{\beta}_1 x_{51}$$

where \tilde{y}_{51} predicts the unobserved y_{51} .



Assuming the model is correct, then the true value of y_{51} is

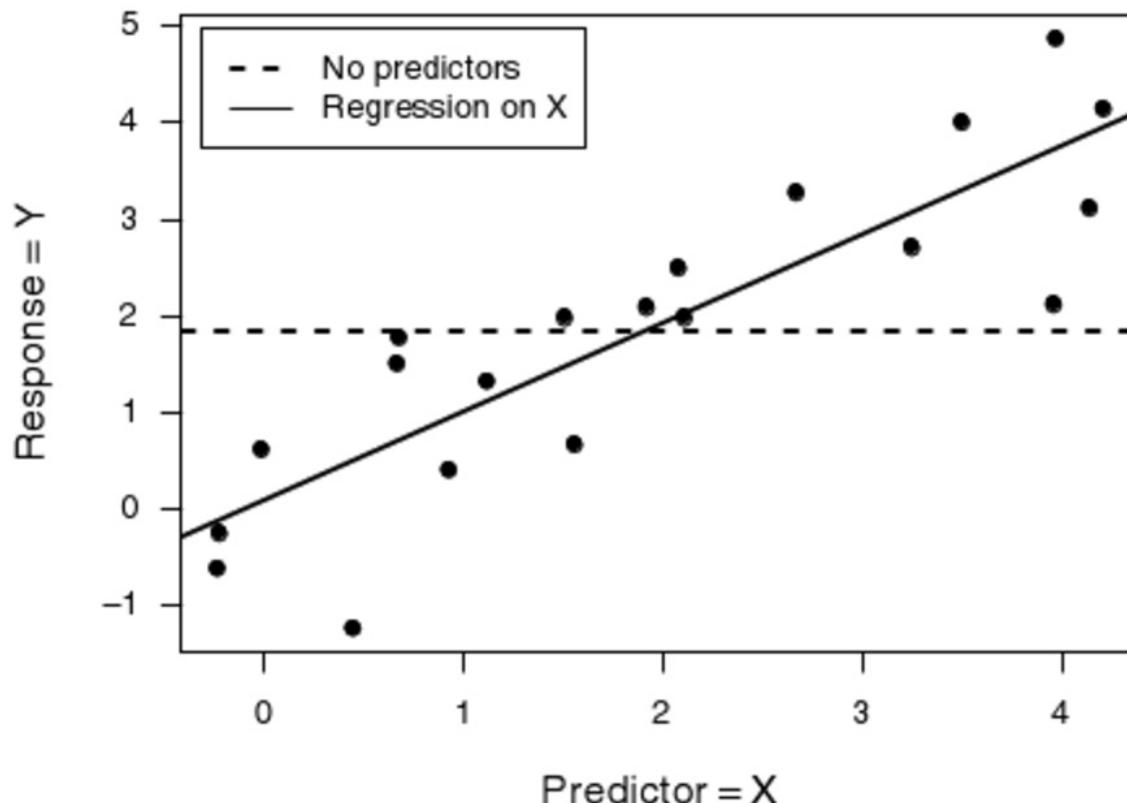
$$\tilde{y}_{51} = \beta_0 + \beta_1 x_{51} + \varepsilon_{51}$$

where ε_{51} is the random error attached to the future value. Since \tilde{y}_{51} is based on estimated values of β_0 and β_1 , the prediction error variability will have a second component that arises from the uncertainty in the estimates of the coefficients. We can show that

$$Var(\tilde{y}_{51}|x_{51}) = \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_{51} - \bar{x})^2}{S_{XX}} \right).$$

What does this imply?

Coefficient of Determination R^2

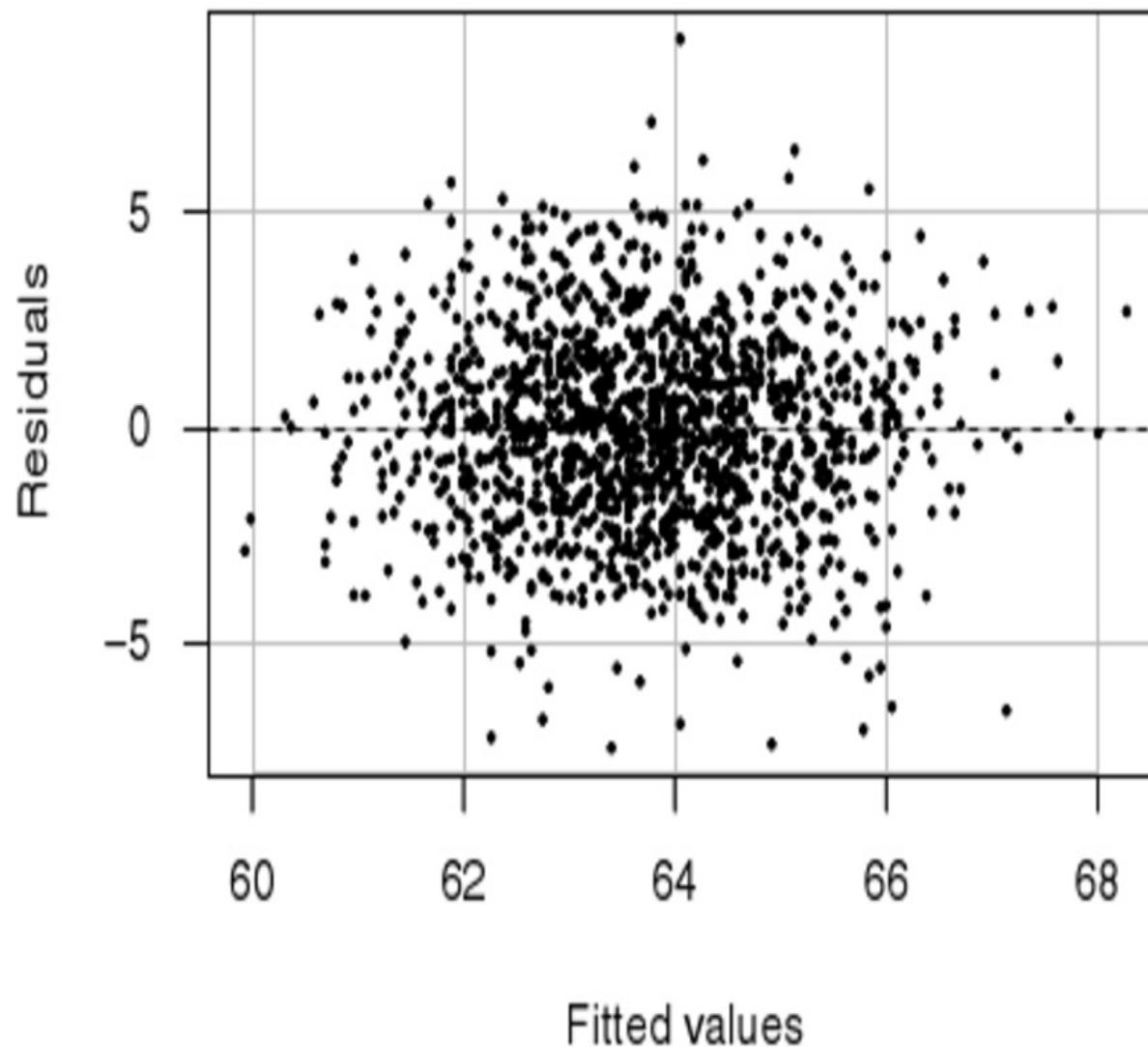


Ignoring all covariates, the best prediction of the response would be the total sum of squares $SSY = \sum_{i=1}^n (y_i - \bar{y})^2$, the observed total variation of the response, ignoring any and all predictors.

The unexplained variation is given by RSS , the sum of squared deviations from the fitted line. We define $SSreg = S\text{SY} - RSS$. If both sides are divided by S_{YY} , then we have $\frac{SSreg}{S_{YY}} = 1 - \frac{RSS}{S_{YY}} = R^2$. R^2 , the coefficient of determination, may be thought of the proportion of the total variability explained by the model, or $1 -$ the unexplained variability. This is a scale-free one-number summary of the strength of the relationship between x_i and y_i . In later lectures we will go more into this summary.

Residuals

Plots of residuals versus other quantities are used to find failures of assumptions. The most common plot, especially useful in simple regression, is the plot of residuals versus the fitted values. A null plot would indicate no failure of assumptions. Curvature might indicate that the fitted mean function is inappropriate. Residuals that seem to increase or decrease in average magnitude with the fitted values might indicate nonconstant residual variance. A few relatively large residuals may be indicative of outliers, cases for which the model is somehow inappropriate.



In a later lecture, we will learn how to use residuals to check model assumptions.

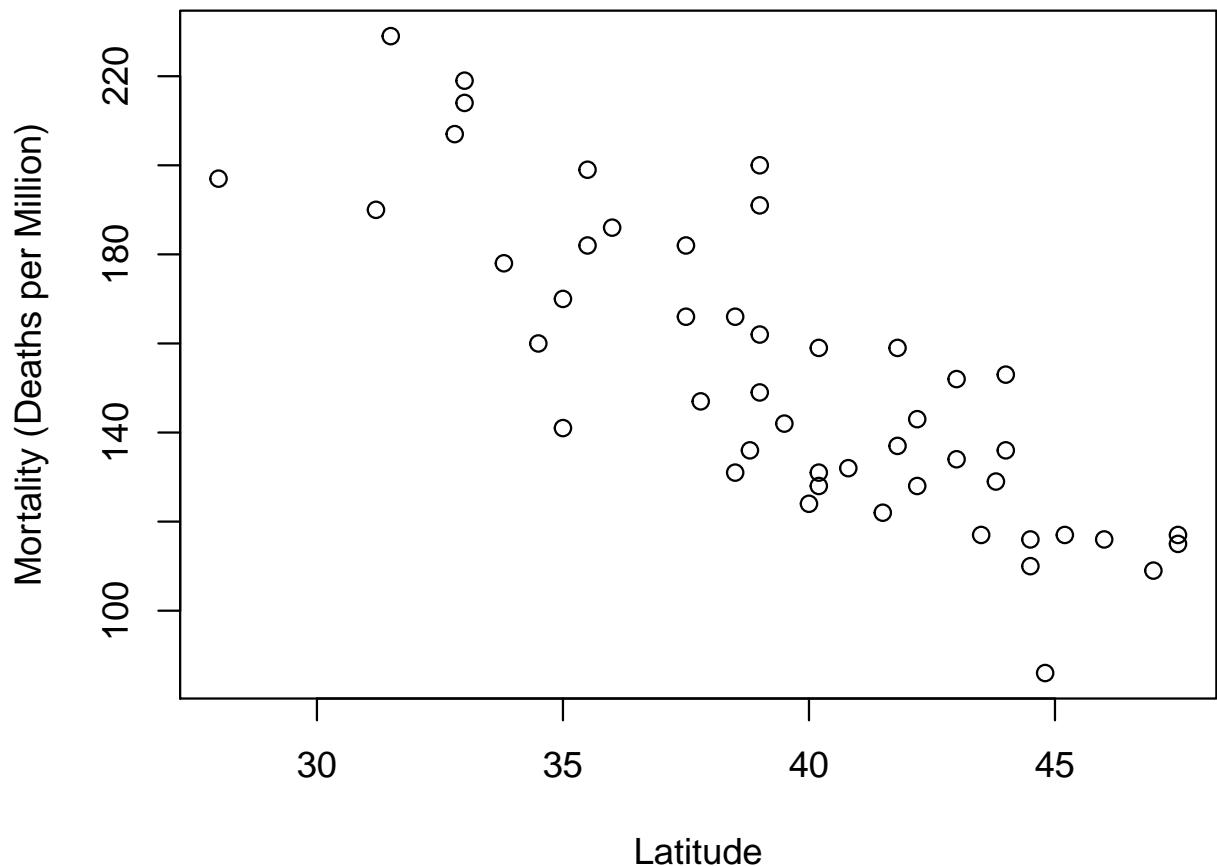
SLR in R and SAS

Lets take a peek at the data for the contiguous US

	mortality	latitude	longitude	ocean
Alabama	219	33.0	87.0	yes
Arizona	160	34.5	112.0	no
Arkansas	170	35.0	92.5	no
California	182	37.5	119.5	yes
Colorado	149	39.0	105.5	no
Connecticut	159	41.8	72.8	yes

Lets plot the data

```
> plot(USmelanoma[,2], USmelanoma[,1],  
+       ylab = "Mortality (Deaths per Million)",  
+       xlab = "Latitude"  
+ )
```



We can fit the linear regression model with the `lm` function

```
> out = lm(mortality ~ latitude, data = USmelanoma)
> out = lm(USmelanoma[,1] ~ USmelanoma[,2])
```

The summary function prints basic information about the fit

```
> summary(out)
```

Call:

```
lm(formula = USmelanoma[, 1] ~ USmelanoma[, 2])
```

Residuals:

Min	1Q	Median	3Q	Max
-38.485	-12.823	1.272	12.192	44.381

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	388.312	23.767	16.338	< 2e-16 ***
USmelanoma[, 2]	-5.966	0.597	-9.994	4.15e-13 ***

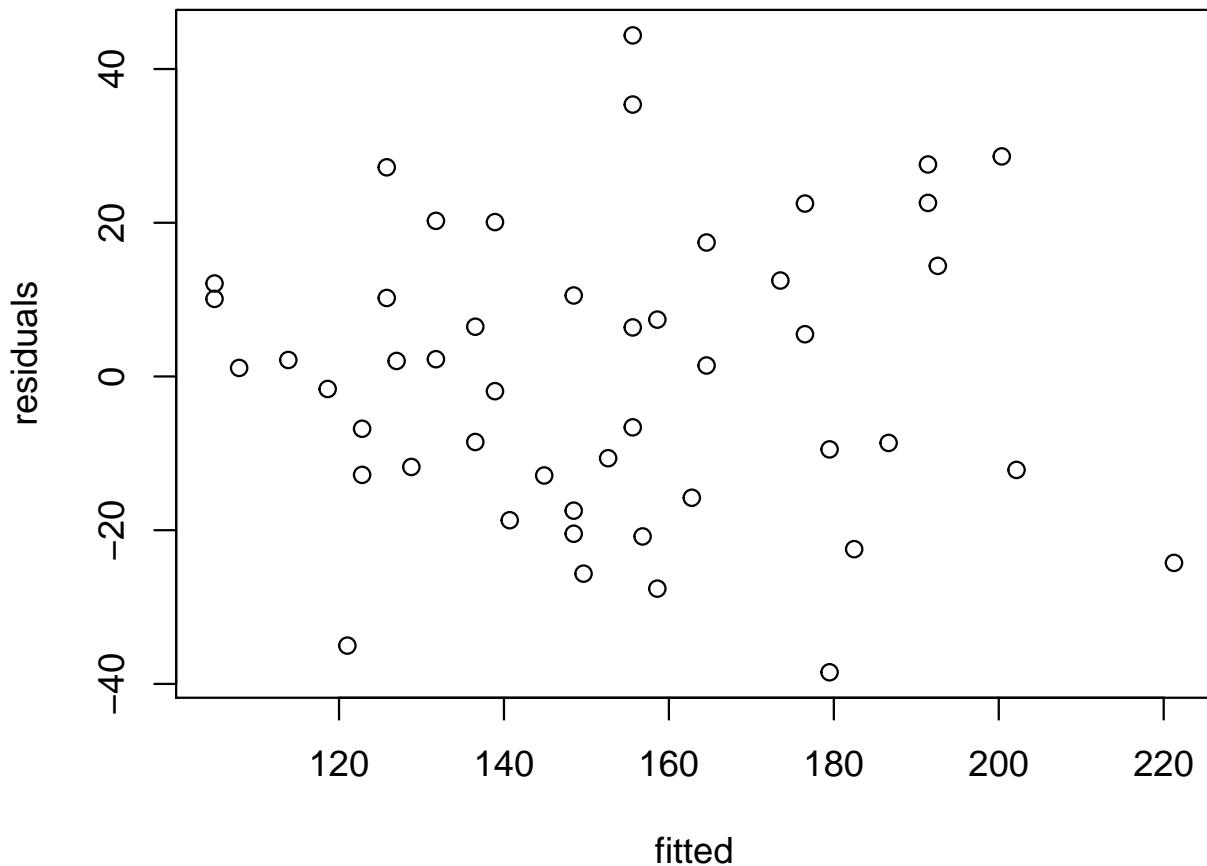
Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’
	0.1 ‘ ’			1

Residual standard error: 19.07 on 46 degrees of freedom
Multiple R-squared: 0.6847, Adjusted R-squared: 0.6778
F-statistic: 99.89 on 1 and 46 DF, p-value: 4.145e-13

We can also extract values based on the model fit

```
> residuals = out$residuals  
> fitted = out$fitted.values  
> coefficients = out$coefficients
```

```
> plot(fitted, residuals)
```



Now, lets use what we learned so far to check the results

```
> y = USmelanoma$mortality  
> x = USmelanoma$latitude  
> n = length(y)  
> ybar = mean(y)  
> xbar = mean(x)  
> S_{YY} = sum( (y - ybar)^2 )  
> S_{XX} = sum( (x - xbar)^2 )  
> S_{XY} = sum( (y - ybar)*(x - xbar) )
```

Now, lets use what we learned so far to check the results

```
> coefficients
```

```
(Intercept) USmelanoma[, 2]  
388.311830      -5.966476
```

```
> S_{XY}/S_{XX} # beta_1 hat
```

```
[1] -5.966476
```

```
> ybar - (S_{XY}/S_{XX})*xbar # beta_0 hat
```

```
[1] 388.3118
```

```
> RSS = sum( (y - fitted)^2 )
> s2_hat = RSS/(n - 2) # sigma_squared hat
> s2_hat

[1] 363.5953

> sqrt(s2_hat*(1/n + xbar^2/S_{XX})) # se of beta_0 hat
[1] 23.76711

> sqrt(s2_hat/S_{XX}) # se of beta_1 hat
[1] 0.5969898

> round(vcov(out),3) # covariance matrix of beta_0 hat, beta_1 hat
                               (Intercept) USmelanoma[, 2]
(Intercept)           564.875          -14.093
USmelanoma[, 2]      -14.093           0.356
```

From our Melanoma example, testing

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

$$t = \frac{-5.966 - 0}{0.597} = -9.994$$

```
> t = -9.994
> p.value = 2*(1 - pt(abs(t), n-2))
> p.value
[1] 4.147793e-13
```

SAS Proc Reg

```
proc reg data = USmelanoma;  
    model mortality = latitude;  
run;
```

The SAS System

The REG Procedure

Model: MODEL1

Dependent Variable: mortality

Number of Observations Read	49
Number of Observations Used	49

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	36464	36464	99.80	<.0001
Error	47	17173	365.38436		
Corrected Total	48	53637			

Root MSE	19.11503	R-Square	0.6798
Dependent Mean	152.87755	Adj R-Sq	0.6730
Coeff Var	12.50349		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	389.18935	23.81232	16.34	<.0001
latitude	1	-5.97764	0.59837	-9.99	<.0001

Mini Problems to do from Weisberg

- 2.1.3
- 2.11

Lecture 5: General Linear Model: Estimation and Testing

Reading Assignment:

- Muller and Fetterman Chapter 2
- Weisberg Chapter 3

We will now consider the general case in which we observe a single response and one or more covariates.

We write the general form of the linear model as

$$\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times 1} + \boldsymbol{\varepsilon}_{n \times 1},$$

where

- $\mathbf{y}_{n \times 1} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ contains the observed responses,

- $\mathbf{X}_{n \times p} = (\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{p-1}) = \begin{bmatrix} x_{1,0} & x_{1,1} & \cdots & x_{1,p-1} \\ x_{2,0} & x_{2,1} & \cdots & x_{2,p-1} \\ \vdots & & & \vdots \\ x_{n,0} & x_{n,1} & \cdots & x_{n,p-1} \end{bmatrix}$ is a matrix of fixed and known covariates,

-
- $\beta_{p \times 1} = (\beta_0, \beta_1, \dots, \beta_{p-1})' = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$ is a vector of parameters to be estimated, and

- $\varepsilon_{n \times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$ is a vector of unobserved random errors.

-
- The rows of \mathbf{y} , \mathbf{X} , and $\boldsymbol{\varepsilon}$ correspond to subjects or sampling units.
 - Columns of \mathbf{X} and the corresponding rows of $\boldsymbol{\beta}$ correspond to predictors. Often, the first column of \mathbf{X} , denoted \mathbf{x}_0 , corresponds to an intercept variable and takes the value 1 for all subjects so that we have $\mathbf{x}_0 = \mathbf{J} = \mathbf{1}$. (We will assume this to be the case unless we state otherwise.)

We refer to the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ as the matrix representation of the linear model. All linear models may be written in this form.

Alternatively, the model is written in scalar notation as

$$y_i = \sum_{j=0}^{p-1} x_{ij}\beta_j + \varepsilon_i, \quad i = 1, \dots, n.$$

The row of \mathbf{X} corresponding to subject i is denoted $\text{row}_i(\mathbf{X})$ or \mathbf{x}'_i (which may lead to some confusion with the columns of \mathbf{X} so that the precise meaning may depend on the context).

Example: Social Setting, Family Planning, and Birth Rate

Rodriguez (2002) considers factors related to the decline in the crude birth rate (CBR, the number of births per thousand population) between 1965 and 1975 for 20 Latin American and Caribbean countries. Covariates of interest include a social setting index (SOCIAL) and a family planning index (FAMPLAN). The social setting index is a function of literacy, school enrollment, life expectancy, infant mortality, percent of males aged 15-64 in the non-agricultural labor force, gross national product per capita, and percent of population living in urban areas. Higher social setting scores represent higher socio-economic levels. The family planning index is a function of availability of contraceptive methods, official government family planning policies, and structure of family planning programs in the country. Values of 20 or more indicate strong efforts in family planning, and values of 10-19 represent moderate efforts.

A few representative observations in the birth rate dataset are presented below.

<i>Country</i>	<i>SOCIAL</i>	<i>FAMPLAN</i>	<i>CBR</i>
Brazil	74	0	10
Costa Rica	84	21	29
Haiti	35	3	0
Mexico	83	4	9
Trinidad-Tobago	84	15	29

This model is written in the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ as follows:

$$\mathbf{y}_{20 \times 1} = \mathbf{X}_{20 \times 3} \boldsymbol{\beta}_{3 \times 1} + \boldsymbol{\varepsilon}_{20 \times 1}$$

$$\begin{bmatrix} 10 \\ 29 \\ 0 \\ 9 \\ 29 \\ \vdots \end{bmatrix}_{20 \times 1} = \begin{bmatrix} 1 & 74 & 0 \\ 1 & 84 & 21 \\ 1 & 35 & 3 \\ 1 & 83 & 4 \\ 1 & 84 & 15 \\ \vdots & \vdots & \vdots \end{bmatrix}_{20 \times 3} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}_{3 \times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \vdots \end{bmatrix}_{20 \times 1}$$

-
- β_0 is called the intercept, which is the expected decline in birth rate when the social setting and family planning indices take the value zero.
 - Some investigators prefer to center all predictors so that the mean of each predictor is zero. Why?
 - β_1 is the slope for social setting. It is interpreted as the expected increase in CBR decline for a one unit increase in the social setting index.
 - β_2 is the slope for family planning. It is interpreted as the expected increase in CBR percent decline for a one unit increase in the family planning index.
 - Each element of ϵ represents the distance between a country's observed percent CBR decline and the population regression line.



Least Squares Estimation

The **least squares** estimate of β , denoted $\hat{\beta}$, satisfies

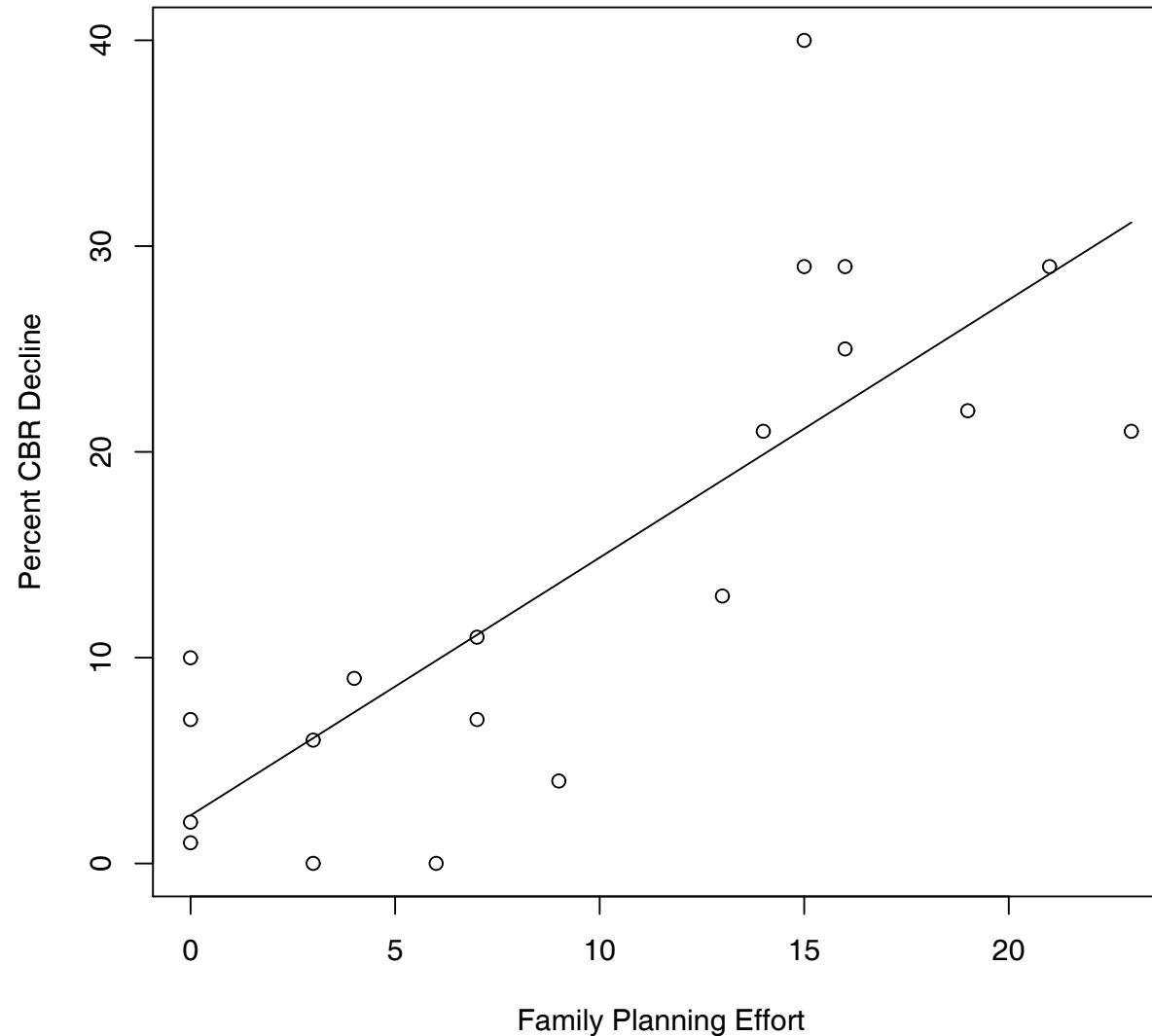
$$\begin{aligned}(y - X\hat{\beta})'(y - X\hat{\beta}) &= \min_{\beta} (y - X\beta)'(y - X\beta) \\ &= \min_{\beta} \| y - X\beta \|^2 \\ &= \min_{\beta} \sum_{i=1}^n (y_i - \mathbf{x}'_i \beta)^2,\end{aligned}$$

where \mathbf{x}'_i indicates the i -th row of matrix X . The least squares estimate of β minimizes the squared Euclidean distance between y and its mean $\mu = X\beta$.

It can be shown that when we make the additional assumption that $y_i \sim N(\mathbf{x}'_i \beta, \sigma^2)$, the least squares estimates $\hat{\beta}$ are also *maximum likelihood* estimates. Maximum likelihood estimates have several desirable properties that you will learn more about in BIOS 661.

Below is a plot of the observed percent CBR decline versus family planning effort along with the least squares regression line. Consider two steps in fitting this line. First, we calculate the mean percent CBR decline, \bar{y} . Then, we pivot a line around $(\bar{x}, \bar{y}) = (9.55, 14.3)$, where \bar{x} is the mean family planning effort, until we minimize the sum of squared deviations around the line.

CBR Decline by Family Planning Effort



The least squares estimator, $\hat{\beta}$, has several good properties.

- First, if the linear model assumption holds, then the least squares estimator is *unbiased*; that is, $E(\hat{\beta}) = \beta$. 
- Next, we will show later that if the observations are uncorrelated and have constant variance σ^2 , then the variance-covariance matrix of the least squares estimator $\hat{\beta}$ is $\text{Cov}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. This estimator is *efficient* in the sense that it has the smallest variance in the class of all unbiased estimators that are linear functions of the data. We call this type of estimator a BLUE (Best Linear Unbiased Estimator). 

-
- If we add the assumption that the error vector is normally distributed, then the least squares estimator is the “best” estimator among *all* unbiased estimators. (We define a “best” estimate to be an unbiased estimate with minimum variance. Other criteria for “best” estimates do exist but will not be addressed further here.)
 - Later, we will also show that the sampling distribution of the least squares estimator $\hat{\beta}$ in *large* samples is approximately multivariate normal with the previously given covariance; that is,
$$\hat{\beta} \sim N_p(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}).$$
 In the intercept-only model, this means that $\hat{\beta} = \bar{y} \sim N(\mu, \frac{\sigma^2}{n})$ in large samples. (This is true even if your errors are not exactly normal but are uncorrelated with constant variance. If they are exactly normal, then the sampling distribution of $\hat{\beta}$ is exactly normal, and you don’t need to worry about attaining a certain sample size in order for asymptotic theory to hold.)

Least squares estimation requires a variety of assumptions:

Existence Assumption

Assume ε_i has finite first and second moments. That is, we observe values of random variables with **finite variance**. In practice, considering a finite number of subjects ensures that this assumption holds.

Linearity Assumption

We assume the expected values (means) of the response are linear functions of the parameters. That is, we assume

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta},$$

or equivalently, that

$$E(\mathbf{y}) - \mathbf{X}\boldsymbol{\beta} = \mathbf{0} \Leftrightarrow \mathbf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}.$$

The linearity assumption refers to the relationship between the response and parameters. Consider several examples.

-
1. $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ is linear in parameters, predictors, and error
 2. $y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$ is linear in parameters and error
 3. $y_i = \beta_0 + x_i^{\beta_1} + \varepsilon_i$ is linear in error
 4. $y_i = \beta_0 \exp(-\beta_1 x_i) + \varepsilon_i$ is nonlinear in parameters
 5. $y_i = \beta_0 [\exp(-\beta_1 x_i)] \varepsilon_i \Leftrightarrow \log(y_i) = \log(\beta_0) - \beta_1 x_i + \ln(\varepsilon_i)$
 \Leftrightarrow (5b) $y_i^* = \gamma_0 + \gamma_1 x_i + \delta_i$, which is linear in parameters, predictor, and error
 6. $y_i = (\beta_0 + \beta_1 x_i) \varepsilon_i$ is nonlinear in parameters 

Of these examples, only 1, 2, and 5b meet the linear model assumption of linearity. If linearity does not hold, then we should not attempt to fit linear models, and estimates from linear models will not have the nice properties discussed previously. Nonlinear models, such as the exponential growth or decay model given by $\mathbf{y} = \beta_1 e^{\beta_2 \mathbf{x}} + \varepsilon$, will not be covered in this class.

Independence Assumption

Each element of ϵ is statistically independent of every other.

Equivalently, each element of y is statistically independent of every other, conditional on \mathbf{X} . This generally is fairly obvious to check.

For example, if a set of twins or a parent and child are included in data, the independence assumption will be violated. If positively correlated observations tend to have positively correlated predictor values, standard linear regression will yield anti-conservative tests of associations (i.e., your p-value will be too small). This is the type of error we expect with correlation in family data or longitudinal data.

We will discuss appropriate models for correlated data later in BIOS 663.

Homogeneity Assumption

We assume each element of ε has the same variance σ^2 . Equivalently, $\text{Cov}(\varepsilon) = E(\varepsilon\varepsilon') = \sigma^2\mathbf{I}$. It is possible to check this assumption, and we will spend a good deal of time later in the course discussing ways to check homogeneity of variances (which is sometimes called  **homoscedasticity**). *Discussion of Homogeneity:*

- σ_i^2 is the variance of the error for subject i .
- $\sigma_i^2 = \sigma_i^2(y_i | x_{i1}, x_{i2}, \dots, x_{ip})$ is the variance of the response, conditional on the value of the covariates \mathbf{x}_i for subject i .
- Homogeneity of variances means that $\sigma_i^2 = \sigma^2$ for all subjects i .
- Homogeneity also means that the variance about the regression function $E(\mathbf{y} | \mathbf{X})$ is constant.

Error Covariance Matrix

Recall that we assume errors have mean zero. So $E(\boldsymbol{\varepsilon}) = \mathbf{0}$. Thus

$$\begin{aligned}\text{Cov}(\boldsymbol{\varepsilon}) &= E[(\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon}))(\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon}))'] \\ &= E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') - \boldsymbol{\varepsilon}E(\boldsymbol{\varepsilon})' - E(\boldsymbol{\varepsilon})\boldsymbol{\varepsilon}' + E(E(\boldsymbol{\varepsilon})E(\boldsymbol{\varepsilon})') \\ &= E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') - \mathbf{0} - \mathbf{0} + \mathbf{0} = \mathbf{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')_{\mathbf{n} \times \mathbf{n}} \\ &= \begin{bmatrix} E(\varepsilon_1^2) & E(\varepsilon_1\varepsilon_2) & \dots & E(\varepsilon_1\varepsilon_n) \\ E(\varepsilon_2\varepsilon_1) & E(\varepsilon_2^2) & & \vdots \\ \vdots & & & \vdots \\ E(\varepsilon_n\varepsilon_1) & \dots & \dots & E(\varepsilon_n^2) \end{bmatrix}.\end{aligned}$$

If the covariance $E(\varepsilon_i\varepsilon_j) = 0$, then the correlation

$$\frac{E(\varepsilon_i\varepsilon_j)}{\sqrt{E(\varepsilon_i^2)}\sqrt{E(\varepsilon_j^2)}} = 0,$$

and vice versa. Although independence of two random observations

always implies zero correlation (and zero covariance), the converse is not true. That is, unless we have Gaussian random variables, zero covariance does *not* imply independence.

The independence assumption means that the error covariance matrix is a diagonal matrix. In addition, because the covariance of a variable with itself equals its variance, we have

$$\begin{aligned}\text{Cov}(\boldsymbol{\varepsilon}) &= E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') \\ &= \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & & & \vdots \\ 0 & \dots & \dots & \sigma_n^2 \end{bmatrix}.\end{aligned}$$

The homogeneity assumption implies that $\sigma_i^2 = \sigma^2$ so that

$$\begin{aligned}\text{Cov}(\varepsilon) &= E(\varepsilon\varepsilon') \\ &= \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & & & \vdots \\ 0 & \dots & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}_n.\end{aligned}$$

For example, if the CBR decline is more variable for countries with low family planning effort than for countries with high family planning effort, conditional on other covariates of interest, the homogeneity of variances assumption would not hold.

Gaussian Errors Assumption

The Gaussian errors assumption is that $\varepsilon_i \sim N(0, \sigma_i^2)$. Assuming homogeneity of variances as well, we have that $\varepsilon_i \sim N(0, \sigma^2)$. This implies that $y_i \sim N(\mathbf{x}'_i \boldsymbol{\beta}, \sigma^2)$ so that

$$f(y_i | \mathbf{x}_i, \boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \left(\frac{y_i - \mathbf{x}'_i \boldsymbol{\beta}}{\sigma}\right)^2\right).$$

The normality assumption is not needed for validity of least squares estimation. However, adding the assumption of Gaussian errors means that the least squares estimate of $\boldsymbol{\beta}$ is also a maximum likelihood estimate and a minimum variance unbiased estimate. The assumption of Gaussian errors also allows simple construction of exact small-sample hypothesis tests.

We combine all five assumptions as follows:

$$y_i \sim N(\mathbf{x}'_i \boldsymbol{\beta}, \sigma^2),$$



with $y_i \perp y_j$ for $i \neq j$. Muller and Fetterman use the mnemonic *HILE Gauss* to describe the five assumptions.

Under the five assumptions (HILE Gauss), the $\{\varepsilon_i\}$ are i.i.d.(independent and identically distributed). However, the $\{y_i\}$ in general are not i.i.d. as $E(y_i) \neq E(y_j)$ because $\mathbf{x}'_i \neq \mathbf{x}'_j$ for all $i \neq j$. The $\{y_i\}$ are independent Gaussian random variables with equal variances, but they do not necessarily have equal expected values.

Defining the Normal Equations

For $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, there are two distinct sets of parameters:

- $\boldsymbol{\beta}$ with p elements, and
- σ^2 with 1 element.

We almost always assume $n \gg p$ so that the sample size \gg the number of parameters.

Once the parameters $\boldsymbol{\beta}$ have been estimated, one may compute $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, the *predicted values* of the outcome. The *residuals* measure the accuracy of the prediction and are given by

$$\mathbf{y} - \hat{\mathbf{y}} = \hat{\boldsymbol{\varepsilon}} \neq \boldsymbol{\varepsilon}.$$

The error, ε_i , is unobserved, while the residual, $\hat{\varepsilon}_i$, is observed (and an estimate of ε_i).

We seek estimates $\hat{\beta}$, $\hat{\sigma}^2$ that are optimal in some sense. One criterion is least squares, which computes $\hat{\beta}$, $\hat{\sigma}^2$ that minimize the average squared distance from observed outcomes to predicted outcomes.

Total squared error of prediction, called the residual SS or the sum of squares for error (SSE), is

$$\begin{aligned} SSE &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2 \\ &= \hat{\varepsilon}' \hat{\varepsilon} = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) \\ &= (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \\ &= \mathbf{y}'\mathbf{y} - 2\beta'\mathbf{X}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta. \end{aligned}$$

To find $\hat{\beta}$ that minimizes SSE take derivatives:


$$\frac{\partial SSE}{\partial \beta} = \mathbf{0} - 2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\beta$$

Set $\partial SSE / \partial \beta = \mathbf{0}_{p \times 1}$ to create a system of p simultaneous equations:

$$\mathbf{0} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\beta}.$$

Rearranging terms yields the *normal equations*:

$$(\mathbf{X}'\mathbf{X})_{p \times p} \hat{\beta}_{p \times 1} = \mathbf{X}'_{p \times n} \mathbf{y}_{n \times 1}.$$

A $\hat{\beta}$ that solves the normal equations yields predicted values $\hat{\mathbf{y}}$ and residuals $\hat{\epsilon}$ such that $\hat{\mathbf{y}}'\hat{\epsilon} = 0$; that is, the predicted values and residuals are orthogonal (normal to each other). 

Solving Normal Equations if \mathbf{X} is Full Rank

If $\mathbf{X}'\mathbf{X}$ is full rank (i.e., $\text{rank}(\mathbf{X}) = p$ with $n > p$), then $\mathbf{X}'\mathbf{X}$ has a unique inverse, $(\mathbf{X}'\mathbf{X})^{-1}$. Then we solve the normal equations:

$$\begin{aligned} (\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} &= \mathbf{X}'\mathbf{y} \\ (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \end{aligned}$$

Facts about the Least Squares Estimate:

- $\hat{\boldsymbol{\beta}}$ is the unique least squares estimate.
- $\hat{\boldsymbol{\beta}}$ is best linear unbiased estimate (BLUE).
- With Gaussian errors, $\hat{\boldsymbol{\beta}}$ is the MLE and minimum variance unbiased estimator.

Solving Normal Equations for \mathbf{X} Less Than Full Rank

If \mathbf{X} is less than full rank, say $\text{rank}(\mathbf{X}) = r < p$, then $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist.

Solutions:

1. Drop some covariates.
2. Generalized Inverse ($\hat{\boldsymbol{\beta}}$ not unique) 
3. Penalized regression methods, such as Ridge regression, which minimizes

$$\min_{\boldsymbol{\beta}} \| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \|^2 + \lambda \sum_{j=1}^p \beta_j^2, \quad \text{$$

where λ is a tuning parameter that can be decided by cross-validation. In other words, we penalize the size of the regression coefficients. The solution is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \lambda I_{p \times p})^{-1} \mathbf{X}'\mathbf{y}$ 

Hat Matrix From now on, we assume \mathbf{X} is full rank, unless otherwise specified.

The *hat matrix* is

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$



The hat matrix has the special properties of idempotency ($\mathbf{HH} = \mathbf{H}$) and symmetry. In addition, $\text{rank}(\mathbf{H}) = r = \text{rank}(\mathbf{X})$, and $\text{rank}(\mathbf{I} - \mathbf{H}) = n - r$. The prediction value can be written as

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y},$$

\mathbf{H} is called the hat matrix because $\mathbf{H}\mathbf{y} = \hat{\mathbf{y}}$. The residuals

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

Estimating σ^2

The MLE of σ^2 is $\frac{SSE}{n} = \frac{\hat{\epsilon}'\hat{\epsilon}}{n}$.

$E\left[\frac{\hat{\epsilon}'\hat{\epsilon}}{n}\right] = \sigma^2 \left[\frac{n-r}{n}\right]$, so the MLE of σ^2 is biased. 

Recall: Calculating Sample Variance



When we calculate a variance, first we calculate a mean, say \bar{y} , and then we find the distance of each point from the mean, $y_i - \bar{y}$, $i = 1, \dots, n$. (Note: by definition of the mean, $\sum_{i=1}^n (y_i - \bar{y}) = 0$; for this reason, raw deviations are not a useful variance measure.) The *sum of squares* $\sum_{i=1}^n (y_i - \bar{y})^2$ is a useful measure of variability because it increases as the data are more dispersed about the mean. However, this measure also depends on n , and therefore it is not as useful in comparing groups. For comparison purposes, we convert the *sum of squares* to a *variance* by dividing by $n - 1$, where n is the number of subjects in a group.

Why not divide the sum of squares by n ?

The reason we do not divide by n is that we do not have n *independent* pieces of information about the variance. First, we calculated a mean, and then we calculated deviations from the mean. If we calculate the first $n - 1$ deviations, then we know the last $\sum_{i=1}^n (y_i - \bar{y}) = 0$. The independent pieces of information contributing to a statistic are called the *degrees of freedom*.

By similar logic, $\hat{\sigma}^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{n-r} = \frac{SSE}{n-r}$ is an unbiased estimate of σ^2 .

$\hat{\sigma}^2$ is a *quadratic form* in \mathbf{y} :

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\hat{\epsilon}'\hat{\epsilon}}{n-r} = \frac{(\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})}{n-r} = \frac{\mathbf{y}' [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \mathbf{y}}{n-r} \\ &= \frac{\mathbf{y}' [\mathbf{I} - \mathbf{H}] \mathbf{y}}{n-r}\end{aligned}$$

Example: Ozone Exposure Assessment

One common problem in environmental epidemiology is determining personal exposures to environmental toxicants, such as ozone in the air. Adverse health effects associated with ozone exposure include increased incidence of cough, chest pain, and other respiratory symptoms. Although outdoor ozone concentrations are monitored by the Environmental Protection Agency (EPA), it is more difficult to determine indoor concentrations. Personal exposures, which vary based on the proportion of time spent outdoors, at home, in the workplace, and in other areas, are even more difficult to measure. Using outdoor ozone concentrations as a crude approximation of personal exposure can lead to substantial measurement error, which can in turn lead to biased parameter estimates.

We consider data from a study conducted in State College, Pennsylvania, in which children wore small ($2\text{ cm} \times 3\text{ cm}$) personal

ozone samplers. Investigators wish to model personal ozone exposures ($O_{PERSONAL}$) measured by the samplers as a function of outdoor ($O_{OUTDOOR}$) ozone concentrations (measured at a central State College site), home indoor ozone concentrations (O_{HOME}) for each child, and the proportion of time each child spent outdoors ($TIME_{OUTDOORS}$).

The data we consider include 64 measurements of personal ozone exposure (in parts per billion or ppb) along with the corresponding measurements of outdoor ozone concentrations, home indoor ozone concentrations, and the proportion of time spent outdoors.

This model is written in the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ as follows:

$$\mathbf{y}_{64 \times 1} = \mathbf{X}_{64 \times 4} \boldsymbol{\beta}_{4 \times 1} + \boldsymbol{\varepsilon}_{64 \times 1}$$

$$\begin{bmatrix} 26.29 \\ 3.30 \\ 29.28 \\ 28.55 \\ 38.28 \\ \vdots \end{bmatrix}_{64 \times 1} = \begin{bmatrix} 1 & 35.88 & 22.29 & 0.57 \\ 1 & 34.37 & 22.27 & 0.17 \\ 1 & 45.96 & 23.40 & 0.00 \\ 1 & 92.56 & 7.14 & 0.26 \\ 1 & 30.44 & 35.38 & 0.69 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}_{64 \times 4} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \vdots \end{bmatrix}_{64 \times 1}$$

-
- β_0 is called the intercept, which is the expected value of $O_{PERSONAL}$ when all other predictors $(O_{OUTDOOR}, O_{HOME}, TIME_{OUTDOORS})$ take the value zero.
 - β_1 is the slope for outdoor ozone. It is interpreted as the expected ppb increase in personal exposure for a one ppb increase in outdoor ozone concentration.
 - β_2 is the slope for home indoor ozone. It is interpreted as the expected ppb increase in personal exposure for a one ppb increase in home indoor ozone concentration.
 - β_3 is the slope for the proportion of time spent outdoors. It is interpreted as the expected ppb increase in personal exposure for an additional one percent of time spent outdoors.

The following R code may be used to obtain parameter estimates for the ozone data.

```
> ozone = read.table("ozone.txt", header = T, sep = " ") # space delimited
> n = nrow(ozone) # number of rows of ozone (obs)
> X = model.matrix(~ outdoor + home + time_out, data = ozone) # predictor matrix
> y = ozone$personal # response
> head(X)

  (Intercept) outdoor home time_out
1          1 35.87771 22.29      0.57
2          1 43.79189 13.97      0.90
3          1 49.81255 18.96      0.55 
4          1 34.37366 22.27      0.17
5          1 45.95496 23.40      0.00
6          1 64.76558 39.62      0.30

> bhat = solve(t(X) %*% X) %*% t(X) %*% y # from notes
> yhat=X%*%bhat; # predicted values
> ehat=y-yhat; # residuals
> p=ncol(X); # num columns of X
> df=n-p; # df
> sse = t(y) %*% y - t(bhat) %*% t(X) %*% y # SSE
> mse=sse/df; # MSE
```

The output from the R is given below.

```
> print(bhat)
```

```
          [,1]
(Intercept) 3.78348593
outdoor      0.09142005
home         0.59543659
time_out     13.64453832
```

```
> print(mse)
```

```
          [,1]
[1,] 169.1365
```

The same estimates may be obtained from SAS PROC REG.

```
proc reg data=ozone;  
model personal=outdoor home time_out;  
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: personal Personal Ozone Exposure (ppb)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5034.90667	1678.30222	9.92	<.0001
Error	60	10148	169.13652		
Corrected Total	63	15183			

Root MSE	13.00525	R-Square	0.3316
Dependent Mean	23.54578	Adj R-Sq	0.2982
Coeff Var	55.23389		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value
Intercept	Intercept	1	3.78349	4.34206	0.87
outdoor	Outdoor Ozone Concentration (ppb)	1	0.09142	0.09042	1.01
home	Home Indoor Ozone Concentration (ppb)	1	0.59544	0.16478	3.61
time_out	Proportion of Time Spent Outdoors	1	13.64454	7.70973	1.77

Parameter Estimates

Variable	Label	DF	Pr > t
Intercept	Intercept	1	0.3870
outdoor	Outdoor Ozone Concentration (ppb)	1	0.3160
home	Home Indoor Ozone Concentration (ppb)	1	0.0006
time_out	Proportion of Time Spent Outdoors	1	0.0818

Hypothesis Testing

The General Linear (Univariate) Hypothesis, GLH

For testing, we assume i.i.d.Gaussian errors. β is the matrix of primary parameters, and $\theta_{a \times 1} = C_{a \times p} \beta_{p \times 1}$ is a matrix of secondary parameters, defined by C , the *contrast matrix*. Each row of C defines a new scalar parameter in terms of the β 's, e.g., $\beta_1 - \beta_2$.



Let θ_0 be a matrix of known constants (the hypothesized values). Most often θ_0 is taken to be the zero matrix. The (univariate) general linear hypothesis is

$$H_0 : \theta_{a \times 1} = \theta_0$$

$$H_A : \theta_{a \times 1} \neq \theta_0.$$

Example: Choosing Contrast Matrix and Secondary Parameter Matrix

For the ozone data, consider the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with $O_{PERSONAL}$ as the response and $O_{OUTDOOR}$, O_{HOME} , and $TIME_{OUTDOORS}$ as predictors. Suppose we hypothesize that the outdoor and home exposures have the same effect on personal exposure levels. Thus we have $H_0 : \beta_1 - \beta_2 = 0$. This implies the secondary parameter $\theta = (\beta_1 - \beta_2)$ with corresponding contrast matrix $\mathbf{C} = \begin{bmatrix} 0 & 1 & -1 & 0 \end{bmatrix}$ and $\theta_0 = 0$.

Suppose our hypothesis is that all the slopes are the same. What are the values of θ , \mathbf{C} , and θ_0 ?

Estimability of a Parameter

Choosing appropriate \mathbf{C} and $\boldsymbol{\theta}_0$, coupled with fitting appropriate models, allows testing hypotheses about all *estimable* parameters.

(Searle, 1971, p. 180): “A (linear) function of the parameters is defined to be estimable if it is identically equal to some linear function of the expected value of the vector of observations, \mathbf{y} .”

Thus a scalar parameter, $\theta_i = \mathbf{C}_{1 \times p} \boldsymbol{\beta}_{p \times 1}$, is estimable

$$\Leftrightarrow \mathbf{C}_{1 \times p} \boldsymbol{\beta}_{p \times 1} = \mathbf{t}'_{1 \times n} E(\mathbf{y}_{n \times 1}),$$

for \mathbf{t} a vector of constants.

More generally, for a vector we need $\boldsymbol{\theta}_{\mathbf{a} \times 1} = \mathbf{T}_{\mathbf{a} \times \mathbf{n}} E(\mathbf{y}_{n \times 1})$

There always exist $r = \text{rank}(\mathbf{X})$ distinct and estimable parameters (which are not necessarily elements of $\boldsymbol{\beta}$ but may be linear combinations of elements).





If $\text{rank}(\mathbf{X}) = r = p$, then $\hat{\boldsymbol{\beta}}$ exists (uniquely), $\boldsymbol{\beta}$ is estimable, and any (nonzero) \mathbf{C} gives estimable $\boldsymbol{\theta}$. This is usually the case with continuous predictors unless some predictors are collinear.

If $\text{rank}(\mathbf{X}) = r < p$, $\boldsymbol{\beta}$ is not estimable (although as many as r elements may be), and for $\hat{\boldsymbol{\theta}} = \mathbf{C}\boldsymbol{\beta}$, we must check estimability.

To show set of parameters

$$\boldsymbol{\theta}_{a \times 1} = \mathbf{C}_{a \times p} \boldsymbol{\beta}_{p \times 1} = \mathbf{T}_{a \times n} E(\mathbf{y}_{n \times 1})$$



is estimable, it suffices to show that $\mathbf{C}_{a \times p} = \mathbf{T}_{a \times n} \mathbf{X}_{n \times p}$.



Estimable $\hat{\boldsymbol{\theta}}$ shares the optimality of $\hat{\boldsymbol{\beta}}$ (whatever r is): BLUE for least squares and MLE with Gaussian errors.



Show that $H_0 : \beta_1 = \beta_2$ is estimable for the ozone data.



Testability of a Hypothesis

Consider the likelihood ratio (LR) test. Let $\mathbf{M}_{a \times a} = \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'$. Define GLH testability as the (unique) existence of the LR test.

θ is testable \Leftrightarrow

- C is full rank a (no redundancies), and
- θ is estimable,

OR, equivalently,

- M is full rank a , and
- θ is estimable.

If X is full rank, then θ is testable \Leftrightarrow

C is full rank a OR M is full rank a (because any θ is estimable)

Show that $H_0 : \beta_1 = \beta_2$ is testable for the ozone data.

Computation of Test Statistic and p-Value

Define the sums of squares hypothesis as

$$SSH_{1 \times 1} = (\hat{\theta} - \theta_0)' M^{-1} (\hat{\theta} - \theta_0). \quad \text{[Talk]}$$

With HILE Gauss, the likelihood ratio statistic equals

$$\begin{aligned} F_{obs} &= \frac{SSH/a}{SSE/(n-r)} = \frac{(\hat{\theta} - \theta_0)' M^{-1} (\hat{\theta} - \theta_0)/a}{\hat{\sigma}^2} \quad \text{[Talk]} \\ &= \frac{MSH}{MSE} \end{aligned}$$

Under H_0 : $\theta = \theta_0$, SSH and SSE are scaled χ^2 random variables, with $SSH/\sigma^2 \sim \chi^2(a)$, independently of $SSE/\sigma^2 \sim \chi^2(n-r)$. It can be shown that if $z_1 \sim \chi^2_{d_1}$, $z_2 \sim \chi^2_{d_2}$, and $z_1 \perp z_2$, then $\frac{z_1/d_1}{z_2/d_2}$ follows an F_{d_1, d_2} distribution. Thus

$$F_{obs} = \frac{[SSH/\sigma^2]/a}{[SSE/\sigma^2]/(n-r)} = \frac{SSH/a}{SSE/(n-r)} \sim F(a, n-r).$$

The p-value equals the probability of observed or more extreme data arising under the null, that is,

$$\text{p-val} = \Pr\{F(a, n - r) \geq F_{obs}\} = 1 - \Pr\{F(a, n - r) < F_{obs}\}.$$

Reject H_0 if $F_{obs} > f_{crit} = F^{-1}(1 - \alpha, a, n - r)$.

Obtain f_{crit} in SAS as FINV(prob, df_1, df_2), the value of an F statistic with df_1 numerator and df_2 denominator degrees of freedom, such that $\Pr\{F \leq f_{crit}\} = \text{prob}$. In R, use qf(prob, df_1, df_2). (To get p-values in R, use 1 - pf(crit, df_1, df_2).

All linear model GLH tests correspond to comparing two models, the “full” model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, and a reduced model defined by constraints. This concept lies at the heart of any LR test and is critical in understanding any particular GLH test.

Example: Computing a GLH Test

For the ozone data, again consider the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with $O_{PERSONAL}$ as the response and $O_{OUTDOOR}$, O_{HOME} , and $TIME_{OUTDOORS}$ as predictors. Suppose we wish to test that all

slopes are equal. Then $H_0 : \boldsymbol{\theta} = \begin{bmatrix} \beta_1 - \beta_2 \\ \beta_2 - \beta_3 \end{bmatrix} = \mathbf{0}$, with corresponding contrast matrix $\mathbf{C} = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}$.

The additional code needed to fit the model, along with PROC REG code, is given below.

```
> C = matrix(c(0,1,-1,0,0,0,1,-1), nrow = 2, byrow = T) # contrast matrix
> print(C)

 [,1] [,2] [,3] [,4]
[1,]    0    1   -1    0
[2,]    0    0    1   -1

> M=C %*% solve(t(X)%*%X)%*%t(C)
> thetahat=C%*%bhat
> ssh=t(theta)%*%solve(M)%*%theta
> f_obs=(ssh/nrow(theta))/mse
> p=1-pf(f_obs,2,60)
```

The results from R are below.

```
> print(ssh)
```

```
[,1]
```

```
[1,] 1237.324
```

```
> print(f_obs)
```

```
[,1]
```

```
[1,] 3.657767
```

```
> print(p)
```

```
[,1]
```

```
[1,] 0.03170141
```

We reject the null hypothesis and conclude that not all slopes are identical. At least one slope is not equal to the others.

This output corresponds to PROC REG below.

```
proc reg data=ozone;  
model personal=outdoor home time_out;  
test outdoor-home=0, home-time_out=0;  
run;
```

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable personal

Source	DF	Mean		
		Square	F Value	Pr > F
Numerator	2	618.66202	3.66	0.0317
Denominator	60	169.13652		

Wald Tests

For a single coefficient β_j , we can test $H_0 : \beta_j = 0$ if β_j is estimable. SAS automatically reports *Wald* tests for parameters in a regression model. These are tests of the hypothesis $H_0 : \beta_j = 0$ for each j .

We know that in large samples (or in small samples if our errors are exactly normal), $\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$. The variance of $\hat{\beta}_j$ is the j^{th} diagonal element of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

Using properties of the standard normal distribution, we can base our test on the ratio


$$t = \frac{\hat{\beta}_j - 0}{\sqrt{\text{var}(\hat{\beta}_j)}}.$$

Usually, we do not know σ^2 exactly and obtain an estimate as $\hat{\sigma}^2 = \frac{SSE}{dfE}$. If we do know σ^2 exactly, then $t \sim N(0, 1)$. If we estimate σ^2 from the data, then $t \sim t_{dfE}$, a Student's t distribution with dfE degrees of freedom.

One and Two-Sided Tests

One-sided tests exist only for scalar hypotheses, not for vector hypotheses. Let $a = \# \text{ rows of } C$. If $a = 1$, then θ is a scalar (1×1) and $F_{obs}(1, n - r) = t^2(n - r)$, where $t(n - r)$ is t-statistic with d.f. of $n - r$.

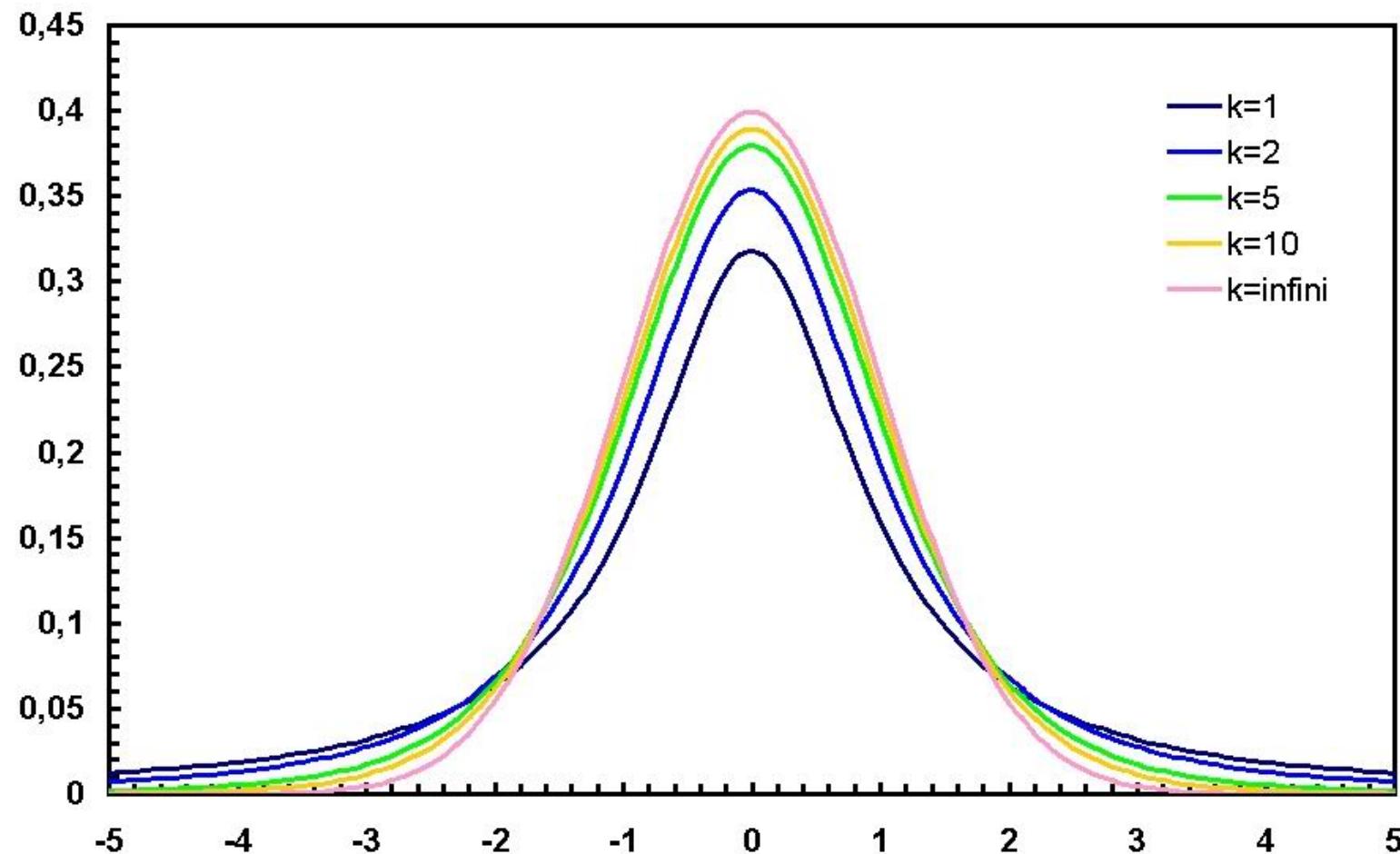
- Two-sided test: $H_0 : \theta = \theta_0$ vs. $H_A : \theta \neq \theta_0$
- One-sided tests:
$$\begin{cases} H_0: \theta = \theta_0 & \text{vs. } H_A: \theta > \theta_0 \\ H_0: \theta = \theta_0 & \text{vs. } H_A: \theta < \theta_0 \end{cases}$$

Conducting a test of size α by t-test:

- A two-sided t-test, $H_A: \theta \neq \theta_0$, uses the $\alpha/2$ and $(1 - \alpha/2)$ critical values.
- A one-sided t-test,, $H_A: \theta < \theta_0$, uses the α critical value.
- A one-sided t-test,, $H_A: \theta > \theta_0$, uses the $(1 - \alpha)$ critical value.

In all cases, one rejects H_0 if the test statistic is farther from zero than

the critical value.

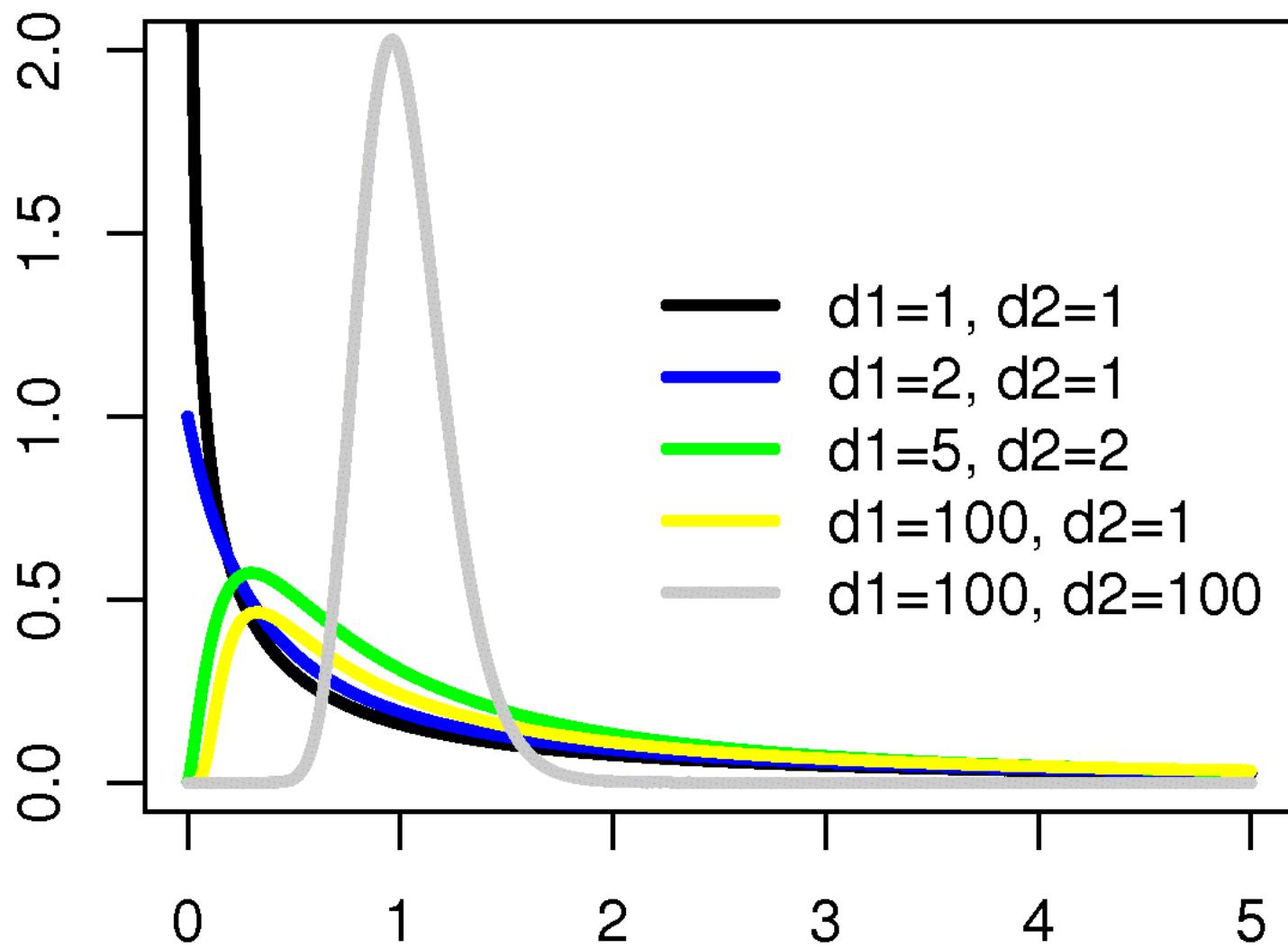


F-tests

- A two-sided F test, $H_A: \theta \neq \theta_0$, uses the α critical value.
- A one-sided F test, $H_A: \theta < \theta_0$, uses the 2α critical value and also requires appropriate sign of the difference.

Recall the definition of a CDF: $F_X(x) = \Pr\{X \leq x\}$.

$$F_F(f) = \Pr\{F \leq f\} = \int_0^f f_F(u)du$$



Next: Distributional Results

Reading Assignment:

- Muller and Fetterman Chapter 3: “Some Distributions for the GLM” (Required)

Lecture 6: Some Distributional Results for the GLM

Reading Assignment:

- Muller and Fetterman Chapter 3: “Some Distributions for the GLM”

In analysis with the GLM, we use three kinds of distributions: multivariate Gaussian, χ^2 , and F .

For now, assume all assumptions hold:

- HILE for estimation and
- Gaussian errors for testing.

A Full Rank Basis for Less Than Full Rank Models

If \mathbf{X} is $n \times p$, with $n \geq p$ and $\text{rank}(\mathbf{X}) = r \leq p$, then $\text{rank}(\mathbf{X}'\mathbf{X}) = r$.

If \mathbf{X} is less than full rank ($r < p$), then *collinearity* exists among columns of \mathbf{X} . If \mathbf{X} is less than full rank, then we say the model is also less than full rank. Also, $r = \text{rank}(\mathbf{X}) = \text{rank}(\mathbf{X}'\mathbf{X})$ is the # of estimable parameters.

For every less than full rank model, there exists a corresponding full rank model with r estimable parameters. That is, for less than full rank \mathbf{X} , there exists a $p \times r$ matrix \mathbf{V}_+ such that

$$\mathbf{X}_{n \times p} = \mathbf{X}_{*,(n \times r)} \mathbf{V}'_{+,(r \times p)} \quad \text{💡}$$

with $\text{rank}(\mathbf{X}_*) = \text{rank}(\mathbf{V}_+) = r < p$. \mathbf{X}_* provides a *full rank basis* for \mathbf{X} .

Suppose that we have the model

$$\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times 1} + \boldsymbol{\varepsilon}_{n \times 1},$$

where $\text{rank}(\mathbf{X}) = r < p$.

Then, defining $\mathbf{X}_{*,(n \times r)} = \mathbf{X}_{n \times p} \mathbf{V}_{+,(p \times r)}$ with corresponding parameter vector $\boldsymbol{\beta}_{*(r \times 1)}$, an equivalent full-rank model is given by

$$\mathbf{y}_{n \times 1} = \mathbf{X}_{*,(n \times r)} \boldsymbol{\beta}_{*(r \times 1)} + \boldsymbol{\varepsilon}_{n \times 1},$$

with $\widehat{\boldsymbol{\beta}}_* = (\mathbf{X}'_* \mathbf{X}_*)^{-1} \mathbf{X}'_* \mathbf{y}$. 

Many possible choices of the matrix \mathbf{V}_+ exist, such as the set of eigenvectors of $\mathbf{X}' \mathbf{X}$ corresponding to non-zero eigenvalues.

Every parameter estimable in the original (less than full rank) model is also estimable in the full rank model, and any estimable parameter is expressible as a linear combination of the β_* 's.

More About the Multivariate Gaussian Distribution

If Σ is not full rank (e.g., covariance matrix of residuals), then the multivariate Gaussian distribution is said to be *singular normal*. In this case, the density does not exist. (The multivariate Gaussian density exists only when Σ_z is non-singular. For example, if $y \sim N_1(\mu, 0)$, we have a discrete distribution with $P(y = \mu) = 1$, a point mass at μ .)

We define a singular multivariate Gaussian distribution for a vector \mathbf{z} in terms of a particular linear transformation \mathbf{Uz} that leads to a full rank covariance matrix $\mathbf{U}\Sigma\mathbf{U}'$ so that we can define a density for the transformed random vector \mathbf{Uz} (with redundancies eliminated).

Definition of Singular and Nonsingular Multivariate Gaussian

1. Full Rank Case

If $\text{rank}(\boldsymbol{\Sigma}_z) = n$ and the density of \mathbf{z} is

$$p(\mathbf{z}) = (2\pi)^{-n/2} |\boldsymbol{\Sigma}_z|^{-1/2} \exp[-(\mathbf{z} - \boldsymbol{\mu}_z)' \boldsymbol{\Sigma}_z^{-1} (\mathbf{z} - \boldsymbol{\mu}_z)/2],$$

then \mathbf{z} is distributed multivariate Gaussian, indicated

$$\mathbf{z}_{n \times 1} \sim \mathcal{N}_n(\boldsymbol{\mu}_{z,(n \times 1)}, \boldsymbol{\Sigma}_{z,(n \times n)}).$$

For example, with $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and HILE Gauss, $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ and $\mathbf{y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$.

2. Less than Full Rank Case

If $1 \leq \text{rank}(\boldsymbol{\Sigma}_z) = n_* < n$, then a fixed matrix $\mathbf{U}_{n_* \times n}$ exists such that $\text{rank}(\mathbf{U}\boldsymbol{\Sigma}_z\mathbf{U}') = n_*$ (full rank).

If the density of \mathbf{Uz} is

$$p(\mathbf{Uz}) = (2\pi)^{-\frac{n_*}{2}} |\mathbf{U}\boldsymbol{\Sigma}_z \mathbf{U}'|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{Uz} - \mathbf{U}\boldsymbol{\mu}_z)'(\mathbf{U}\boldsymbol{\Sigma}_z \mathbf{U}')^{-1}(\mathbf{Uz} - \mathbf{U}\boldsymbol{\mu}_z)\right],$$

then \mathbf{z} is distributed as a singular multivariate Gaussian, indicated by

$$\mathbf{z}_{n \times 1} \sim \mathcal{SN}_n(\boldsymbol{\mu}_{z,(n \times 1)}, \boldsymbol{\Sigma}_{z,(n \times n)}).$$

We must add side conditions to specify $\mathbf{U}_{n_* \times n}$ uniquely.

Example: Pick \mathbf{U} to contain the eigenvectors for non-zero eigenvalues of $\boldsymbol{\Sigma}_z$.

Sampling Distributions of Estimators

β Estimators $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

From properties of the multivariate Gaussian distribution,

$$\begin{aligned} E[\hat{\beta}] &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] E(\mathbf{y}) \\ &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \mathbf{X}\beta = \beta, \end{aligned}$$

and

$$\begin{aligned} \text{Cov}(\hat{\beta}) &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \text{Cov}(\mathbf{y}) [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \\ &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] (\sigma^2 \mathbf{I}_n) [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \\ &= \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}, \end{aligned}$$

with

$$\hat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1})$$

because $\hat{\beta}$ is a full rank linear transformation of the normal random vector \mathbf{y} .

 For \mathbf{X} less than full rank and appropriate choice of generalized inverse, we have

$$\hat{\boldsymbol{\beta}} \sim \mathcal{SN}_p \left[(\mathbf{X}'\mathbf{X})^{-}(\mathbf{X}'\mathbf{X})\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-} \right],$$

which is a singular multivariate Gaussian because $(\mathbf{X}'\mathbf{X})$ is less than full rank.

For a Gaussian distribution, we know that most ($> 95\%$) of its mass lies within two standard deviations of the mean. So if our null hypothesis is for no effect of the covariate corresponding to β_j and

$$|\hat{\beta}_j| > 0 + 1.96 * \sqrt{\sigma^2(\mathbf{X}'\mathbf{X})_{j,j}^{-1}},$$

where $(\mathbf{X}'\mathbf{X})_{j,j}^{-1}$ refers to the $(j, j)^{th}$ element of $(\mathbf{X}'\mathbf{X})^{-1}$, then we have evidence that the covariate of interest has an effect on the response.

Example: Covariance Matrix of $\hat{\beta}$

The following R and PROC REG code below may be used to obtain the estimated covariance matrix of $\hat{\beta}$ for the ozone data with $O_{PERSONAL}$ as the outcome and $O_{OUTDOOR}$, O_{HOME} , and $TIME_{OUTDOORS}$ as predictors.

```
> ozone = read.table("ozone.txt", header = T, sep = " ") # space delimited
> n = nrow(ozone) # number of rows of ozone (obs)
> X = model.matrix(~ outdoor + home + time_out, data = ozone) # predictor matrix
> y = ozone$personal # response
> head(X)

  (Intercept)  outdoor  home time_out
1          1 35.87771 22.29      0.57
2          1 43.79189 13.97      0.90
3          1 49.81255 18.96      0.55
4          1 34.37366 22.27      0.17
5          1 45.95496 23.40      0.00
6          1 64.76558 39.62      0.30

> bhat = solve(t(X) %*% X) %*% t(X) %*% y # from notes
> yhat=X%*%bhat; # predicted values
> ehat=y-yhat; # residuals
> p=ncol(X); # num columns of X
> df=n-p; # df
> sse = t(y) %*% y - t(bhat) %*% t(X) %*% y # SSE
> mse=sse/df; # MSE
> covbhat=solve(t(X) %*% X)*as.numeric(mse) # scalar multiplication
> print(covbhat)
```



	(Intercept)	outdoor	home	time_out
(Intercept)	18.8534457	-0.183111754	-0.18665969	-15.18741728
outdoor	-0.1831118	0.008175203	-0.00831528	-0.06888717
home	-0.1866597	-0.008315280	0.02715354	0.07755804
time_out	-15.1874173	-0.068887173	0.07755804	59.43995287



We may also obtain the estimated covariance matrix of the β 's using the COVB option in SAS PROC REG.

```
proc reg;
model personal=outdoor home time_out/covb;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: personal

Analysis of Variance

Source	DF	Sum of Squares		Mean Square		
				F Value	Pr > F	
Model	3	5034.90667	1678.30222	9.92	<.0001	
Error	60	10148	169.13652			
Corrected Total	63	15183				
Root MSE		13.00525	R-Square	0.3316		
Dependent Mean		23.54578	Adj R-Sq	0.2982		
Coeff Var		55.23389				

Parameter Estimates

Variable	DF	Parameter Estimate		Standard Error	
		t Value	Pr > t		
Intercept	1	3.78349	4.34206	0.87	0.3870
outdoor	1	0.09142	0.09042	1.01	0.3160
home	1	0.59544	0.16478	3.61	0.0006

time_out	1	13.64454	7.70973	1.77	0.0818
----------	---	----------	---------	------	--------

Model: MODEL1

Dependent Variable: personal Personal Ozone Exposure (ppb)

Covariance of Estimates

Variable	Label	Intercept	outdoor
Intercept	Intercept	18.853445739	-0.183111754
outdoor	Outdoor Ozone Concentration (ppb)	-0.183111754	0.0081752029
home	Home Indoor Ozone Concentration (ppb)	-0.186659692	-0.00831528
time_out	Proportion of Time Spent Outdoors	-15.18741728	-0.068887173

Covariance of Estimates

Variable	Label	home	time_out
Intercept	Intercept	-0.186659692	-15.18741728
outdoor	Outdoor Ozone Concentration (ppb)	-0.00831528	-0.068887173
home	Home Indoor Ozone Concentration (ppb)	0.0271535425	0.0775580389
time_out	Proportion of Time Spent Outdoors	0.0775580389	59.439952865

θ Estimator

Again using properties of the multivariate Gaussian distribution, for $\mathbf{C}_{a \times p}$ a matrix of constants and $\boldsymbol{\theta} = \mathbf{C}\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\theta}} \sim N_a(\boldsymbol{\theta}, \sigma^2 \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}').$$

We should be careful to verify that $\boldsymbol{\theta}$ is both estimable and testable.

Example: Estimating θ

Suppose we wish to test the hypothesis that the $O_{OUTDOOR}$, O_{HOME} , and $TIME_{OUTDOORS}$ coefficients are all equal in the ozone data. So

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix},$$

$$\boldsymbol{\theta} = \begin{bmatrix} \beta_1 - \beta_2 \\ \beta_2 - \beta_3 \end{bmatrix},$$

and

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}_2(\boldsymbol{\theta}, \sigma^2 \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}').$$

We may obtain the estimated covariance matrix of $\hat{\boldsymbol{\theta}}$ with the following additional R code.

```
> C=matrix(c(0, 1, -1, 0, 0, 0, 1, -1), nrow = 2, byrow = T);
> print(C)

 [,1] [,2] [,3] [,4]
[1,]    0    1   -1    0
[2,]    0    0    1   -1

> covhat=as.numeric(mse)*C%*%solve(t(X) %*% X)%*%t(C) # sigma^2 * M
> print(covhat)

 [,1]      [,2]
[1,] 0.05195931 0.1109764
[2,] 0.11097639 59.3119903
```

Predicted Values: Conditional Means and Future Observations

For the GLM

$$\begin{aligned} E(\mathbf{y}_{n \times 1}) &= E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= \mathbf{X}\boldsymbol{\beta} = \mu(\mathbf{y} \mid \mathbf{X}). \end{aligned}$$

- We write the estimator of the expected values as

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} = [\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \mathbf{y} \\ &= \mathbf{H}\mathbf{y}. \end{aligned}$$

Prediction is often an important goal of modeling. For example, an actuary might want to predict medical costs for health insurance patients covered by a given insurance plan given their characteristics (including BMI, smoking status, and age).

Caution: prediction outside the range of observed data can be extremely dangerous!

To find the distribution of $\hat{\mathbf{y}}$,



$$\begin{aligned} E(\hat{\mathbf{y}}) &= \mathbf{H}E(\mathbf{y}) \\ &= [\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}. \end{aligned}$$

In addition, for $\hat{\mathbf{y}}$ an estimated conditional mean response,

$$\begin{aligned} \text{cov}(\hat{\mathbf{y}}) &= \mathbf{H}(\sigma^2\mathbf{I})\mathbf{H}' \\ &= \sigma^2\mathbf{H}\mathbf{H}' \stackrel{\text{?}}{=} \sigma^2\mathbf{H} \\ &= \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \end{aligned}$$

as \mathbf{H} is symmetric and idempotent.

As a transformation of a **singular** multivariate Gaussian ($\mathbf{H}_{n \times n}$ has rank $p < n$ when \mathbf{X} is full rank), the distribution of $\hat{\mathbf{y}}$ is given by

$$\hat{\mathbf{y}} \sim SN_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{H}).$$

In contrast, consider predicting future observations at time f with covariate values \mathbf{X}_f . In this case, we predict an *individual outcome* rather than a mean outcome. The predicted future response $\hat{\mathbf{y}}_f$ involves variance due to

1. estimating β (the fitted line is not the exact true line)
2. observing ϵ_f in the future sample (even if the fitted line is the exact true line, there is still variability about it), and
3. changing the design (predictor values).

Now

$$\hat{\mathbf{y}}_f = \mathbf{X}_f \hat{\beta},$$

where \mathbf{X}_f is $n_f \times p$, while \mathbf{X} is $n \times p$. ($\mathbf{X}_f \neq \mathbf{X}$). These predictions will have additional error (as opposed to estimated conditional means) ϵ_f : the errors to be observed at future time f ($\epsilon_f \neq \epsilon$), and ϵ_f are independent of the errors at the current time.

Thus we have



$$\hat{\mathbf{y}}_f \sim \mathcal{SN}_{n_f} \{ \mathbf{X}_f \boldsymbol{\beta}, \sigma^2 [\mathbf{X}_f (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'_f + \mathbf{I}] \}$$



$\mathbf{X}_f = \mathbf{X}$ implies

$$\hat{\mathbf{y}}_f \sim \mathcal{SN}_n \{ \mathbf{X} \boldsymbol{\beta}, \sigma^2 [\mathbf{H} + \mathbf{I}] \},$$



so we see the additional uncertainty in predicting a particular future observation rather than estimating the mean.

It would be unambiguous to discuss estimated conditional means, $\hat{\boldsymbol{\mu}} | \mathbf{X}$, or predicted future observations, $\hat{\mathbf{y}}_f$; however, describing $\hat{\mathbf{y}} = \mathbf{X} \hat{\boldsymbol{\beta}}$ as predicted values has become standard, so that you must rely on the context to decide whether estimated conditional means or predicted future observations are being discussed.

Definitions and Properties of Residuals

The residuals $\hat{\varepsilon}$ are defined as

$$\begin{aligned}\hat{\varepsilon} &= (\mathbf{y} - \hat{\mathbf{y}}) \\ &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= (\mathbf{I} - \mathbf{H})\mathbf{y},\end{aligned}$$

with

$$\hat{\varepsilon} \sim \mathcal{SN}_n(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H})).$$

$\hat{\varepsilon}_i$ is not independent of $\hat{\varepsilon}_j$ ($i \neq j$) unless $p = 0$ or $n \rightarrow \infty$.

$\hat{\varepsilon}_i \sim \mathcal{N}[0, \sigma^2(1 - h_i)]$, where $h_i = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'$, the (i, i) element of $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

If the model spans an intercept, $\sum_{i=1}^n \hat{\varepsilon}_i = 0$.

Residual Variance

Let $r = \text{rank}(\mathbf{X})$. Recall that

$$\begin{aligned}\hat{\sigma}^2 &= \frac{SSE}{n - r} = \frac{\hat{\boldsymbol{\varepsilon}}' \hat{\boldsymbol{\varepsilon}}}{n - r} \\ &= \frac{\mathbf{y}'(\mathbf{I} - \mathbf{H})'(\mathbf{I} - \mathbf{H})\mathbf{y}}{n - r} = \frac{\mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}}{n - r}\end{aligned}$$

as $(\mathbf{I} - \mathbf{H})$ is symmetric and idempotent.

Now

$$\hat{\sigma}^2 \left(\frac{n - r}{\sigma^2} \right) \sim \chi^2(n - r).$$



Because $E(\chi^2(\nu)) = \nu$, $\hat{\sigma}^2$ is unbiased. In addition, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are statistically independent.

Standardized Residuals

Replacing σ^2 with $\hat{\sigma}^2$ leads some distributions to change from Gaussian to Student's T .

Because $\hat{\varepsilon}_i \sim \mathcal{N}[0, \sigma^2(1 - h_i)]$, it follows that

$$\frac{\hat{\varepsilon}_i}{\sqrt{\sigma^2(1 - h_i)}} \sim \mathcal{N}(0, 1).$$

Define the standardized residual as

$$r_i = \frac{\hat{\varepsilon}_i}{\sqrt{\hat{\sigma}^2(1 - h_i)}} \quad \text{💡}$$

The standardized residuals do not follow a T distribution; in fact,

$$\frac{r_i^2}{n - r} \sim \text{Beta}\left(\frac{1}{2}, \frac{n - r - 1}{2}\right),$$

and the r_i are bounded between $-\sqrt{n - r}$ and $\sqrt{n - r}$, where $r = \text{rank}(\mathbf{X})$.

Example: Calculating Standardized Residuals

To calculate the standardized residuals for the ozone example, we add the following R code to the previous code. In addition, we may obtain the standardized residuals from PROC REG by using the “R” option with the model statement.

```
> H=X%*%solve(t(X) %*% X) %*% t(X) # calculate Hat Matrix  
> h_i=diag(H) # get the diagonal  
> r_i=e_hat/(sqrt(as.numeric(mse)*(1-h_i)));  
> head(r_i,5)  
[ ,1]  
1 -0.1441837  
2 -1.1486537  
3 -0.4461650  
4 -1.4976573  
5  0.5791377
```

The first five residuals are printed above. You may also get these in R from your fitted linear model using the MASS package.

```
> out = lm(personal ~ outdoor + home + time_out, data = ozone)  
> # equivalent to out = lm(y ~ X - 1)
```

```
> summary(out)
```

Call:

```
lm(formula = personal ~ outdoor + home + time_out, data = ozone)
```

Residuals:

Min	1Q	Median	3Q	Max
-25.930	-7.855	-4.257	4.880	36.295

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.78349	4.34206	0.871	0.387032
outdoor	0.09142	0.09042	1.011	0.316031
home	0.59544	0.16478	3.613	0.000619 ***
time_out	13.64454	7.70973	1.770	0.081845 .

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 13.01 on 60 degrees of freedom

Multiple R-squared: 0.3316, Adjusted R-squared: 0.2982

F-statistic: 9.923 on 3 and 60 DF, p-value: 2.094e-05

```
> library(MASS) #load MASS package, or install using install.packages("MASS")
```

```
> r_i MASS = stdres(out) # stdres function from MASS
> head(r_i MASS,5)

      1          2          3          4          5
-0.1441837 -1.1486537 -0.4461650 -1.4976573  0.5791377
```

The following PROC REG code obtain the residuals.

```
proc reg;  
model personal=outdoor home time_out/r;  
run;
```

The first five residuals are below. Note that the standardized residuals are labeled 'student residual.'

The REG Procedure

Model: MODEL1
Dependent Variable: personal Personal Ozone Exposure (ppb)

Output Statistics

Obs	Dep Var personal	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual
1	26.2900	28.1131	3.0429	-1.8231	12.644	-0.144
4	3.3000	22.5059	2.1640	-19.2059	12.824	-1.498
5	29.2800	21.9179	2.7454	7.3621	12.712	0.579
9	28.5500	20.0446	5.9850	8.5054	11.546	0.737
13	38.2800	37.0471	5.1369	1.2329	11.948	0.103

Jackknifing

Jackknifing usually involves computing a statistic with one observation deleted from a sample, once for each observation. Let the subscript $(-i)$ indicate having deleted the i th observation. For example, $\mathbf{y}_{(-i)}$ and $\mathbf{X}_{(-i)}$ have $n - 1$ rows.

Compute

- $\widehat{\boldsymbol{\beta}}_{(-i)} = (\mathbf{X}'_{(-i)} \mathbf{X}_{(-i)})^{-1} \mathbf{X}'_{(-i)} \mathbf{y}_{(-i)}$
- $\widehat{\mathbf{y}}_{(-i)} = \mathbf{X} \widehat{\boldsymbol{\beta}}_{(-i)}$
- $\widehat{\boldsymbol{\varepsilon}}_{(-i)} = \mathbf{y}_{(-i)} - \widehat{\mathbf{y}}_{(-i)}$ and
- $\widehat{\sigma}_{(-i)}^2 = \widehat{\boldsymbol{\varepsilon}}'_{(-i)} \widehat{\boldsymbol{\varepsilon}}_{(-i)} / (n - r - 1),$

where $r = \text{rank}(\mathbf{X})$. 

If $\widehat{\boldsymbol{\beta}}_{(-i)}$ differs greatly from $\widehat{\boldsymbol{\beta}}$, then we see that observation i has a good deal of influence on the analysis.

Studentized Residuals

Standardizing the jackknifed residuals (by dividing by an estimate of the standard deviation) yields a set of residuals (the *Studentized residuals*) which follow a Student's T distribution:

$$r_{(-i)} = \frac{\widehat{\varepsilon}_i}{\sqrt{\widehat{\sigma}_{(-i)}^2(1 - h_i)}} \sim T(n - r - 1)$$

Chapter 7 in MF has extensive discussion of this most important tool for assumption evaluation.

Example: Calculating Studentized Residuals

The following code may be added to the previous code to obtain the studentized residuals in PROC IML. The code uses the following formula for the studentized residuals given in Muller and Fetterman:

$$r_{(-i)} = r_i \left(\frac{(n - r) - 1}{(n - r) - r_i^2} \right)^{\frac{1}{2}},$$

where $r = \text{rank}(\mathbf{X})$.

```
> library(Matrix) #install.packages("Matrix")
> r_i2=r_i^2; # square each element of r_i
> r = rankMatrix(X) # full rank
> r_mi=r_i*sqrt((n-r-1)/(n-r-r_i2))
> head(r_mi,5)

      [,1]
1 -0.1430019
2 -1.1517756
3 -0.4431671
4 -1.5136869
5  0.5759032
```



The following SAS PROC REG code may be used to obtain the studentized residuals.

```
proc reg;  
model personal=outdoor home time_out;  
output out=resid student=standresid rstudent=studresid;  
run;
```



```
proc print data=resid;  
var personal standresid studresid;  
run;
```

The first five observations are given below.

Obs	personal	standresid	studresid
1	26.29	-0.14418	-0.14300
4	3.30	-1.49766	-1.51369
5	29.28	0.57914	0.57590
9	28.55	0.73663	0.73379
13	38.28	0.10319	0.10233

Next: Multiple Regression

Reading Assignment:

- Muller and Fetterman, Chapter 4: “Multiple Regression”

Lecture 7: Multiple Regression: General Consideration

Reading Assignment:

- Muller and Fetterman, Chapter 4: “Multiple Regression”

Why use more than one covariate in a model?

- Why not fit separate models for every covariate?
- Omitting an important covariate, x_2 , can cause you to miss significant relationships between x_1 and y or even to make completely wrong conclusions!

Example: Math Ability

Suppose that you hypothesize that taller children are better at math than shorter children. You take a random sample of 32 children of various ages in Ephesus Elementary. These children take a math test and have their heights measured. When you fit a linear model using height to predict math test score, you find that height is highly significant.

Does this make sense? Recall the definition of a confounder: a *confounder* is a factor that is associated with the exposure and independently affects the outcome.

In this case, age is a potential confounder because it is associated with height and, independently of height, is related to math test scores.

When we fit the *multiple regression* model (sometimes called a *multivariable* regression model) with both age and height as predictors, we see that age is an important predictor of mathematical ability. In addition, after accounting for age, height is unimportant. Our conclusion is that after accounting for age, height is not related to mathematical ability in our sample of elementary school children.

For a general multiple regression model, we write

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{10} & \dots & x_{1,p-1} \\ \vdots & & \\ x_{n0} & \dots & x_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$


The corresponding scalar version of the model is

$$y_i = \sum_{j=0}^{p-1} x_{ij} \beta_j + \varepsilon_i$$

Assume we fit a model and test hypotheses with a continuous, interval scale response and with linear combinations of one or more continuous variables as predictors. All GLH tests can be understood in terms of comparisons of two models: a full model and a reduced model. These tests may be conducted by comparing the *sums of squares* from the two models.

Definitions of Basic Sums of Squares

For the model $\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$, we have

- $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$,
- $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, and
- $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}}$.

We are already familiar with the sum of squares for error, given by

$$\begin{aligned} SSE &= \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) \\ &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2. \end{aligned}$$

The *uncorrected total sum of squares*, USS , is given by

$$USS(\text{total}) = \mathbf{y}'\mathbf{y} = \sum_{i=1}^n y_i^2.$$

The *uncorrected model sum of squares* (or *uncorrected regression sum of squares*) is given by


$$USS(\text{model}) = \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{y}'\mathbf{H}\mathbf{y} = \sum_{i=1}^n \hat{y}_i^2.$$

The difference between the uncorrected total sum of squares and the uncorrected model sum of squares is the sum of squares for error:

$$USS(\text{total}) - USS(\text{model}) = \mathbf{y}'\mathbf{I}\mathbf{y} - \mathbf{y}'\mathbf{H}\mathbf{y} = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y} = SSE.$$

So we see that

$$\begin{aligned} USS(\text{total}) &= USS(\text{model}) + SSE \\ \mathbf{y}'\mathbf{y} &= \mathbf{y}'\mathbf{H}\mathbf{y} + \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}. \end{aligned}$$

Example: Calculating Uncorrected Sums of Squares

Uncorrected sums of squares may be calculated using the following R code.

```
> ozone = read.table("ozone.txt", header = T, sep = " ") # space delimited
> n = nrow(ozone) # number of rows of ozone (obs)
> X = model.matrix(~ outdoor + home + time_out, data = ozone) # predictor matrix
> y = ozone$personal # response
> head(X)

  (Intercept)  outdoor  home time_out
1          1 35.87771 22.29      0.57
2          1 43.79189 13.97      0.90
3          1 49.81255 18.96      0.55
4          1 34.37366 22.27      0.17
5          1 45.95496 23.40      0.00
6          1 64.76558 39.62      0.30

> bhat = solve(t(X) %*% X) %*% t(X) %*% y # from notes
> yhat=X%*%bhat; # predicted values
> ehat=y-yhat; # residuals
> p=ncol(X); # num columns of X
```

```
> df=n-p; # df
> sse = t(y) %*% y - t(bhat) %*% t(X) %*% y # SSE
> mse=sse/df; # MSE
> H=X%*%solve(t(X) %*% X) %*% t(X) # calculate Hat Matrix
> uss_t=t(y)%*%y
> uss_m=t(y)%*%H%*%y
> print(uss_t)

[,1]
[1,] 50664.94

> print(uss_m);

[,1]
[1,] 40516.75

> print(uss_m + sse);

[,1]
[1,] 50664.94
```

We may also get the same results in SAS using PROC GLM, a more general form of PROC REG, using the int option to calculate the uncorrected sums of squares, printing them in an ANOVA table.

```
proc glm data=ozone;
model personal=outdoor home time_out/int;
run;
```

The GLM Procedure

Dependent Variable: personal Personal Ozone Exposure (ppb)

Source	DF	Sum of		F Value	Pr > F
		Squares	Mean Square		
Model	4	40516.75081	10129.18770	59.89	<.0001
Error	60	10148.19129	169.13652		
Uncorrected Total	64	50664.94210			

Corrected sums of squares are defined in relationship to the intercept.

The Nature of the Intercept

It is often (but not always) appropriate to include a constant predictor equal to 1.0 so that $x_0 = \mathbf{J}$. Then

$$y_i = \beta_0 + x_{i1}\beta_1 + \dots + x_{i,p-1}\beta_{p-1} + \varepsilon_i.$$

In such models, we call β_0 the intercept. The name reflects the fact that if $x_{i0} = 1$ and all other predictors take the value zero, then


$$E(y_i) = E(\beta_0) + E\left(\sum_{j=1}^{p-1} x_{ij}\beta_j\right) + E(\varepsilon_i) = \beta_0,$$

and β_0 is the y intercept of the fitted line (the value of y at which the regression function intercepts the vertical axis). In addition, β_0 equals the response (y) value predicted if all x_{ij} , $j \in \{1, 2, \dots, p-1\}$, are 0.

For \bar{x}_j the average of the j^{th} column of \mathbf{X} , $\beta_0 = \mu_y - \sum_{j=1}^{p-1} \bar{x}_j \beta_j$ and $\hat{\beta}_0 = \bar{y} - \sum_{j=1}^{p-1} \bar{x}_j \hat{\beta}_j$.

Example: Ozone Data

For the ozone data, we have the model

$$E(personal_i) = \beta_0 + \beta_1 outdoor_i + \beta_2 home_i + \beta_3 timeout_i,$$

$i = 1, \dots, 64$. The parameter estimates from this model are given by $\hat{\beta} = (3.78, 0.09, 0.60, 13.64)'$. In addition, in our dataset, we observe outdoor ozone concentrations ranging from 11.60-104.10 ppb, home ozone concentrations ranging from 0.80-46.04 ppb, and percent time spent outdoors ranging from 0-90%. In this model, are we particularly interested in the value of the intercept?

Models That Span but May Not Include an Intercept

Any model that includes an intercept has \mathbf{J}_n as a column in \mathbf{X} .

Models with an intercept allow computing the corrected sums of squares, which exclude the portion of the sums of squares due to the intercept and are usually preferred over the uncorrected sums of squares.

When using dummy variables, it is often convenient to code the model such that no column of \mathbf{X} is $\mathbf{1}$ although $\mathbf{X}_{n \times p} \mathbf{t}_{p \times 1} = \mathbf{1}$, for $\mathbf{t}_{p \times 1}$ a vector of constants.

Such a model *spans* an intercept, even though the design matrix, \mathbf{X} , does not include an intercept.

For example, let



$$X_1 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad \beta_1 = \begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix}, \quad X_2 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \beta_2 = \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix}.$$

Notice that $X_2 = X_1 \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ and

$$\beta_1 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix} = \begin{bmatrix} \alpha_0 \\ \alpha_0 + \alpha_1 \end{bmatrix}$$

so that the vector $t = (1, 1)'$ finds the hidden intercept.

Corrected Sums of Squares



Corrected sums of squares are adjusted (corrected) for the intercept and measure the effect above and beyond it. In addition, the corrected sums of squares are invariant to location shifts of the response or predictors (e.g.,centering a covariate).

Uncorrected sums of squares are always well-defined. Corrected sums of squares and associated statistics entirely exclude the intercept but involve comparing models which all span an intercept. Thus corrected sums of squares are defined only if the model includes or spans an intercept.

The correction term or *sum of squares due to the intercept*,
 $n\bar{y}^2 = \frac{\mathbf{y}' \mathbf{J}_n \mathbf{J}_n' \mathbf{y}}{n} = SSI$, corrects for location.

The corrected total sum of squares is computed as

$$\begin{aligned} CSS(\text{total}) &= USS(\text{total}) - SSI \\ &= \mathbf{y}'\mathbf{y} - \frac{\mathbf{y}'\mathbf{J}_n\mathbf{J}'_n\mathbf{y}}{n} \\ &= \mathbf{y}' \left[\mathbf{I}_n - \frac{\mathbf{J}_n\mathbf{J}'_n}{n} \right] \mathbf{y} \\ &= \sum_{i=1}^n (y_i - \bar{y})^2. \end{aligned}$$

The corrected model sum of squares is also computed by subtracting the SSI :

$$\begin{aligned} CSS(\text{model}) &= USS(\text{model}) - SSI \\ &= \mathbf{y}' \mathbf{H} \mathbf{y} - \frac{\mathbf{y}' \mathbf{J}_n \mathbf{J}'_n \mathbf{y}}{n} \\ &= \mathbf{y}' \left[\mathbf{H} - \frac{\mathbf{J}_n \mathbf{J}'_n}{n} \right] \mathbf{y} \\ &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2. \end{aligned}$$

So we see that

$$\begin{aligned} CSS(\text{total}) &= CSS(\text{model}) + SSE \\ \mathbf{y}' \left[\mathbf{I} - \frac{1}{n} \mathbf{J}_n \mathbf{J}'_n \right] \mathbf{y} &= \mathbf{y}' \left[\mathbf{H} - \frac{1}{n} \mathbf{J}_n \mathbf{J}'_n \right] \mathbf{y} + \mathbf{y}' (\mathbf{I} - \mathbf{H}) \mathbf{y}. \end{aligned}$$

Example: Computing Corrected Sums of Squares

The additional R code for computing the corrected sum of squares in R is given below.

```
> one = matrix(1, n) # column vec of 1s
> I_n = diag(as.vector(one)) # diagonal matrix of 1s
> css_t=t(y)%*%(I_n-(one%*%t(one))/n)%*%y
> css_m=t(y)%*%(H-(one%*%t(one))/n)%*%y
> print(css_t)

[,1]
[1,] 15183.1

> print(css_m)

[,1]
[1,] 5034.907

> print(css_m + sse)

[,1]
[1,] 15183.1
```

We may also obtain this from PROC GLM

```
proc glm data=ozone;  
model personal=outdoor home time_out;  
run;
```

```
The GLM Procedure  
Dependent Variable: personal Personal Ozone Exposure (ppb)
```

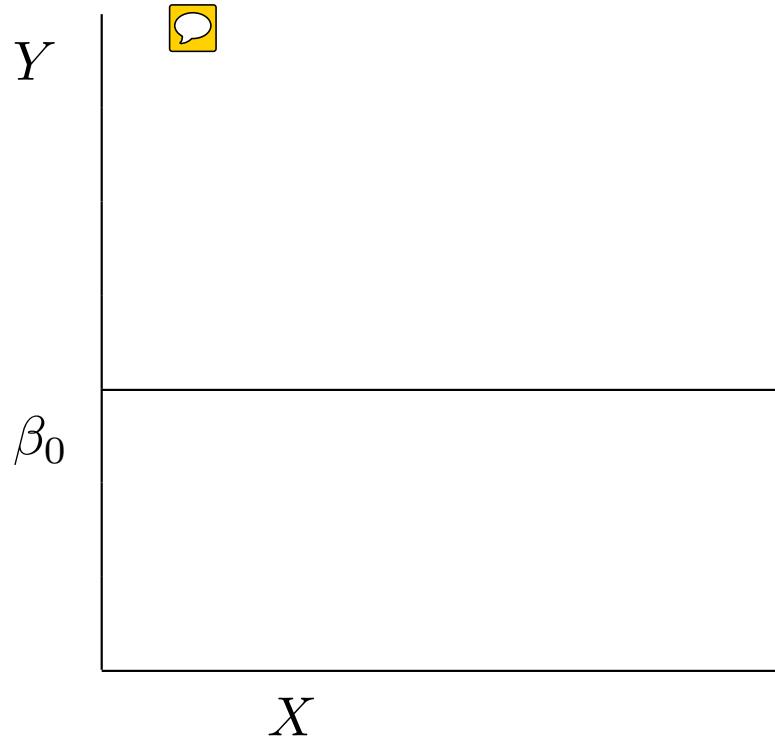
Source	DF	Sum of		F Value	Pr > F
		Squares	Mean Square		
Model	3	5034.90667	1678.30222	9.92	<.0001
Error	60	10148.19129	169.13652		
Corrected Total	63	15183.09796			

The Intercept Only Model

Consider the following model with only an intercept:

$$\mathbf{y}_{n \times 1} = \beta_0 + \boldsymbol{\varepsilon}_{n \times 1}$$
$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \beta_0 + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

It is easy to show that $\widehat{\beta}_0 = \bar{y} = \sum_{i=1}^n \frac{y_i}{n}$, i.e., that the estimate of β_0 is the average response. This model is sometimes called the *grand mean model*. The intercept only model predicts a constant ($\widehat{y}_i = \widehat{y}_j = \bar{y}$ for all i, j), and the predicted regression line has zero slope. For the ozone data, an intercept only model expects that personal exposure would be the same for all students. The estimate of the intercept is simply the mean personal exposure, so that we have $\widehat{\beta}_0 = 23.55$.



Intercept-only Model

Although this model is not very useful by itself, it is very useful in model selection and testing. We will use it to answer questions like the following: “Does adding home exposure to this basic model give us any additional useful information about personal exposure?”

For the intercept only model, we have 

$$\bullet \quad \hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{J}_n\hat{\beta}_0 = \mathbf{J}_n\bar{y} = \begin{bmatrix} \bar{y} \\ \vdots \\ \bar{y} \end{bmatrix}_{n \times 1},$$

$$\bullet \quad \hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}} = \begin{bmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix}_{n \times 1},$$

$$\bullet \quad SSE = \hat{\boldsymbol{\varepsilon}}' \hat{\boldsymbol{\varepsilon}} = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{y}' \left(\mathbf{I} - \frac{1}{n} \mathbf{J}_n \mathbf{J}_n' \right) \mathbf{y}, \quad \text{$$

$$\bullet \quad USS(\text{total}) = \mathbf{y}'\mathbf{y}$$

$$\bullet \quad USS(\text{model}) = \mathbf{y}'\mathbf{H}\mathbf{y} = \mathbf{y}' \left(\frac{1}{n} \mathbf{J}_n \mathbf{J}_n' \right) \mathbf{y},$$

-

$$\begin{aligned} CSS(\text{total}) &= USS(\text{total}) - SSI \\ &= \mathbf{y}' \left(\mathbf{I} - \frac{1}{n} \mathbf{J}_n \mathbf{J}'_n \right) \mathbf{y} = SSE, \text{ and} \end{aligned}$$

-

$$\begin{aligned} CSS(\text{model}) &= USS(\text{model}) - SSI \\ &= \mathbf{y}' \mathbf{H} \mathbf{y} - \frac{1}{n} \mathbf{y}' \mathbf{J}_n \mathbf{J}'_n \mathbf{y} \\ &= \mathbf{y}' \left(\frac{1}{n} \mathbf{J}_n \mathbf{J}'_n \right) \mathbf{y} - \frac{1}{n} \mathbf{y}' \mathbf{J}_n \mathbf{J}'_n \mathbf{y} \\ &= 0. \end{aligned}$$



Example: Computing Sums of Squares for the Intercept Only Model

```
proc glm data=ozone;  
model personal= /;  
run;
```

```
*****
```

The GLM Procedure

Dependent Variable: personal Personal Ozone Exposure (ppb)

Source	DF	Sum of		F Value	Pr > F
		Squares	Mean Square		
Model	1	35481.84414	35481.84414	147.23	<.0001
Error	63	15183.09796	241.00155		
Uncorrected Total	64	50664.94210			

From the SAS output, what are the values of the following?

- SSE : 
- $USS(\text{total})$: 
- $USS(\text{model})$: 
- SSI : 
- $CSS(\text{total})$: 
- $CSS(\text{model})$: 

What test is provided in the SAS output? 

The Null Model



An even simpler model exists. The “null” model assumes the intercept is zero and all slopes are zero so that we have no parameters in the model (except for σ^2) and thus 0 model degrees of freedom. Call

$$\mathbf{y} = \boldsymbol{\varepsilon}$$

the *null model*, with $p = 0$ and $\beta = \emptyset$.

Necessarily, $\hat{\mathbf{y}} = \mathbf{0}$ and $\hat{\boldsymbol{\varepsilon}} = \mathbf{y}$. We do not define corrected sums of squares for the null model because it does not span an intercept. For this model, we have

$$SSE = \hat{\boldsymbol{\varepsilon}}' \hat{\boldsymbol{\varepsilon}} = \mathbf{y}' \mathbf{y} = USS(\text{total})$$



and $USS(\text{model}) = USS(\text{total}) - SSE = 0$.

Like the intercept only model, this model is not very interesting by itself. However, it is also useful for purposes of model selection and testing.

Overall ANOVA Table for Multiple Regression

Typically, with all continuous predictors, \mathbf{X} has full rank unless an error has been made. By default, for now we will assume that \mathbf{X} less than full rank implies an error. Writing $\hat{\boldsymbol{\beta}}$ presumes full rank \mathbf{X} , while $\tilde{\boldsymbol{\beta}}$ allows either full rank ($\text{rank}(\mathbf{X}) = p$) or less than full rank \mathbf{X} .

After fitting a model, one wishes to summarize the analysis and decide whether the variables in \mathbf{X} are useful predictors (reduce the variability of $\mathbf{y} \mid \mathbf{X}$).

All regression tests considered are special cases of the GLH, which we have already studied.



Overall ANOVA Table for Model:

$$\text{Wingspan}_i = \beta_0 + \beta_1 \text{Height}_i + \beta_2 \text{Waist}_i + \epsilon_i$$

Source	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F_{obs}</i>	<i>p</i>
Intercept	1	83.13			
Regression (uncorrected)	3	83.69			
Regression (corrected)	2	0.56			
Error (residual)	27	0.37			
Total (uncorrected)	30	84.06			
Total (corrected)	29	0.94			

Source indicates the variables which contribute the information.

Degrees of freedom (df) gives the dimension of the source (number of contributing variables).

Sums of squares (*SS*) are defined in the preceding section.

Mean Square (MS) = SS/df are sums of squares adjusted for the

sample size.

$$F_{obs} = MS(\text{source})/MS(\text{error}).$$

Under HILE Gauss and H_0 : $\sigma_{\text{source}}^2 = \sigma_{\text{error}}^2$,

$F_{obs} \sim F(df_{\text{source}}, df_{\text{error}})$, and

$$p = \Pr\{F_{obs} \geq F(df_{\text{source}}, df_{\text{error}})\}.$$

Overall ANOVA Table Formulas for Full Model
(Assuming X ($n \times p$) Spans an Intercept)

Source	<i>SS</i> Form	Scalar Form	Quadratic Form
Intercept	$SSE_{\emptyset} - SSE_0$	$n\bar{y}^2$	$\mathbf{y}'(\frac{\mathbf{J}\mathbf{J}'}{n})\mathbf{y}$
Regression (uncorrected)	$SSE_{\emptyset} - SSE_{p-1}$	$\sum_{i=1}^n \hat{y}_i^2$	$\mathbf{y}'\mathbf{H}\mathbf{y}$
Regression (corrected)	$SSE_0 - SSE_{p-1}$	$\sum_{i=1}^n \hat{y}_i^2 - n\bar{y}^2$	$\mathbf{y}'(\mathbf{H} - \frac{\mathbf{J}\mathbf{J}'}{n})\mathbf{y}$
Error (residual)	SSE_{p-1}	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	$\mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}$
Total (uncorrected)	SSE_{\emptyset}	$\sum_{i=1}^n y_i^2$	$\mathbf{y}'\mathbf{y}$
Total (corrected)	SSE_0	$\sum_{i=1}^n y_i^2 - n\bar{y}^2$	$\mathbf{y}'(\mathbf{I} - \frac{\mathbf{J}\mathbf{J}'}{n})\mathbf{y}$

Corrected sums of squares and associated statistics are not defined if a model does not span an intercept. Corrected sums of squares and tests always include the intercept in models and exclude it from tests, which is generally our objective.

Usual (“Corrected”) Overall Test for Regression

Consider testing whether the predictors in \mathbf{X} have any value. This corresponds to the hypothesis that all slopes are zero, which corresponds to no predictive contribution (explanatory value) of the covariates, except possibly the intercept.

In this test we compare the full model,

$$y_i = \beta_0 + \sum_{j=1}^{p-1} x_{ij}\beta_j + \varepsilon_i,$$

to the reduced model,

$$y_i = \beta_0 + \varepsilon_i,$$

to test $H_0 : \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$.

$$\begin{aligned}
 F_{obs} &= \frac{MS(\text{hypothesis})}{MSE} = \frac{SSH/dfH}{SSE/dfE} \\
 &= \frac{[SSE(\text{reduced}) - SSE(\text{full})]/[dfE(\text{reduced}) - dfE(\text{full})]}{SSE(\text{full})/dfE(\text{full})} \\
 &= \frac{CSS(\text{Regression})/(p-1)}{SSE(\text{full})/(n-p)}.
 \end{aligned}$$

Reject the hypothesis if $F_{obs} \geq F_F^{-1}(1 - \alpha, p-1, n-p) = f_{crit}$.

The usual test of overall regression assumes model spans an intercept and excludes the intercept from the test.



Overall **Corrected** ANOVA Table Formulas for Full Model (Assuming \mathbf{X} ($n \times p$) Spans an Intercept)

Source	Scalar Form	Quadratic Form	df
Intercept	$n\bar{y}^2$	$\mathbf{y}'(\frac{\mathbf{J}\mathbf{J}'}{n})\mathbf{y}$	1
Regression (corrected)	$\sum_{i=1}^n \hat{y}_i^2 - n\bar{y}^2$	$\mathbf{y}'(\mathbf{H} - \frac{\mathbf{J}\mathbf{J}'}{n})\mathbf{y}$	$p - 1$
Error (residual)	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	$\mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}$	$n - p$
Total (corrected)	$\sum_{i=1}^n y_i^2 - n\bar{y}^2$	$\mathbf{y}'(\mathbf{I} - \frac{\mathbf{J}\mathbf{J}'}{n})\mathbf{y}$	$n - 1$

So $CSS(\text{total}) = CSS(\text{model}) + SSE$. To find $USS(\text{total})$ and $USS(\text{model})$, simply add SSI to the corresponding corrected values (and adjust the degrees of freedom accordingly).

Example: Conducting the “Corrected” Overall Test for Regression

```
proc glm data=ozone;  
model personal= outdoor home time_out;  
run;  
*****
```

The GLM Procedure

Dependent Variable: personal Personal Ozone Exposure (ppb)

Source	DF	Sum of			F Value	Pr > F
		Squares	Mean Square			
Model	3	5034.90667	1678.30222		9.92	<.0001
Error	60	10148.19129	169.13652			
Corrected Total	63	15183.09796				

“Uncorrected” Overall Test for Regression

Consider testing whether the variables in \mathbf{X} , including the intercept, have any value as predictors.

This corresponds to the hypothesis that all slopes and the intercept are zero, which corresponds to no contribution of any predictor.

In order to do this, compare the full model,

$$y_i = \beta_0 + \sum_{j=1}^{p-1} x_{ij}\beta_j + \varepsilon_i,$$

to the reduced model,

$$y_i = \varepsilon_i.$$

This yields $H_0: \beta_0 = \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$.

The test statistic for the uncorrected overall test is

$$\begin{aligned} F_{obs} &= \frac{[SSE(\text{reduced}) - SSE(\text{full})]/[dfE(\text{reduced}) - dfE(\text{full})]}{SSE(\text{full})/dfE(\text{full})} \\ &= \frac{USS(\text{Regression})/p}{SSE(\text{full})/(n - p)}. \end{aligned}$$

Reject the hypothesis if $F_{obs} \geq F_F^{-1}(1 - \alpha, p, n - p) = f_{crit}$.

This test is well defined whether or not the model spans an intercept, but it rarely has scientific value.

Overall Uncorrected ANOVA Table Formulas for Full Model
 (Assuming \mathbf{X} ($n \times p$) Spans an Intercept)

Source	Scalar Form	Quadratic Form	df
Regression			
(uncorrected)	$\sum_{i=1}^n \hat{y}_i^2$	$\mathbf{y}' \mathbf{H} \mathbf{y}$	p
Error (residual)	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	$\mathbf{y}' (\mathbf{I} - \mathbf{H}) \mathbf{y}$	$n - p$
Total (uncorrected)	$\sum_{i=1}^n y_i^2$	$\mathbf{y}' \mathbf{y}$	n

So $USS(\text{total}) = USS(\text{model}) + SSE$, and SSI is included in the regression sum of squares.

Strength of Association

Decomposing Response Variance

Statistical “significance” does not measure scientific importance.

We may decompose the variance into variance explained by the model and random error, and then we may ask: what fraction of Y variance does X predict?

Usual “Corrected” R^2

For a model spanning an intercept, define the proportion of variance in Y predictable from the X 's, adjusted for the intercept, as

$$R_c^2 = \frac{CSS(\text{Regression})}{CSS(\text{Regression}) + SSE(\text{full})}$$
$$= \frac{CSS(\text{Regression})}{CSS(\text{total})}$$

R_c^2 estimates ρ_c^2 , the population ratio of model to total variance, with $0 \leq \rho_c^2 \leq 1$ and $0 \leq R_c^2 \leq 1$.

Facts about R_c^2 :

- R_c^2 is the maximum likelihood estimate of ρ_c^2 under HILE Gauss.
- R_c^2 is biased in general: $E[R_c^2] \geq \rho_c^2$, with equality for $\rho_c^2 = 1$ or $n \rightarrow \infty$.
- R_c^2 is invariant with respect to full rank linear transformation of the response or predictors (including location and scale changes).
- The corrected test for overall regression,

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$$

holds if and only if

$$H_0 : \rho_c^2 = 0$$

is true.

“Uncorrected” R^2

For any model define the proportion of variance in Y predictable from the X ’s, including the intercept if spanned, as

$$R_u^2 = \frac{USS(\text{Regression})}{USS(\text{Regression}) + SSE(\text{full})} ,$$
$$= \frac{USS(\text{Regression})}{USS(\text{total})}$$

where $0 \leq R_u^2 \leq 1$ and $0 \leq \rho_u^2 \leq 1$.

Facts about R_u^2 :

- R_u^2 may vary due to any linear transformation of Y or X 's.
- The uncorrected test for overall regression,

$$H_0 : \beta_0 = \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0,$$

holds if and only if

$$H_0 : \rho_u^2 = 0$$

is true.

- R_u^2 is a biased estimate of ρ_u^2
- Uncorrected R^2 is defined for any GLM, while corrected R^2 is defined only for models that span an intercept.
- Neither corrected nor uncorrected R^2 is always best, although corrected R^2 should be the default choice.

Comparing Corrected and Uncorrected R^2

$$R_c^2 = \frac{CSS(\text{Model})}{CSS(\text{Total})}$$
$$R_u^2 = \frac{USS(\text{Model})}{USS(\text{Total})} = \frac{CSS(\text{Model}) + SSI}{CSS(\text{Total}) + SSI}$$

$$0 \leq R_c^2 \leq 1$$

$$0 \leq R_u^2 \leq 1$$

As always, $SSI = n\bar{y}^2$.

Note: R^2 always increases when additional predictors are added to the model, whether or not they are practically or statistically important.

Exercise: Computing Corrected and Uncorrected R^2

Compute corrected and uncorrected R^2 for the ozone data.

Adjusted R^2

One potential problem with using R^2 (either corrected or uncorrected) to assess model adequacy is that R^2 never decreases when you add additional predictors to the model. (Even if a predictor is meaningless, R^2 will never be smaller when it is added.) Although we will later learn how to conduct hypothesis tests about whether an increase in R^2 is statistically significant, some investigators prefer to use the *adjusted R^2* , defined as

$$R^2_{adj} = 1 - \frac{SSE/(n - r)}{CSS(\text{Total})/(n - 1)},$$

which is adjusted for the degrees of freedom. It will only increase on adding a variable to the model if the variable reduces the mean square for error. The fact that adjusted R^2 penalizes us for adding terms that are not useful makes it helpful in evaluating and comparing a set of candidate regression models.

Next: Testing

Reading Assignment:

- Muller and Fetterman, Chapter 5: “Testing Hypotheses in Multiple Regression” (Required)

Lecture 8: Testing Hypotheses in Multiple Regression

Reading Assignment:

- Muller and Fetterman, Chapter 5: “Testing Hypotheses in Multiple Regression”

After fitting a model, one seeks to draw inferences about parameters. Correlations and confidence intervals measure scientific importance, while tests and p-values assess statistical “significance”. The two concepts are not necessarily the same!



Review of GLH Concepts

- Assume HILE Gauss and let $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.
- We estimate the $\boldsymbol{\beta}$ as $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ or $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.
- Compute $\hat{\mathbf{y}} = \mathbf{X}\tilde{\boldsymbol{\beta}}$, $SSE = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{y}' [\mathbf{I} - \mathbf{H}] \mathbf{y}$, and $\hat{\sigma}^2 = MSE = \frac{SSE}{dfE}$, with $r = \text{rank}(\mathbf{X})$ and $dfE = n - r$.
- Define $\mathbf{C}_{a \times p}$, which implies $\boldsymbol{\theta} = \mathbf{C}\boldsymbol{\beta}$, and state the GLH as $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$. If \mathbf{X} is less than full rank, check estimability.
- Estimate secondary parameters: $\hat{\boldsymbol{\theta}} = \mathbf{C}\tilde{\boldsymbol{\beta}}$.
- Compute $\mathbf{M} = \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'$ and
$$SSH = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathbf{M}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$
- Compute $F_{obs} = \frac{SSH/dfH}{SSE/dfE}$ for $dfH = a$. reject H_0 if $F_{obs} > F_F^{-1}(1 - \alpha, a, n - r)$.

Although all the tests we consider fall into the framework of the General Linear Hypothesis (GLH), we will discuss several different subclasses of hypothesis tests in the GLH framework, including the following.

1. *Overall tests*, which measure the contribution of the entire set of predictors
2. *Addition of one variable tests*, which measure the contribution of one variable beyond others
3. *Intercept tests*, which measure the value of the intercept in predicting the response
4. *Addition of a group of variables tests*, which measure the contribution of a group of variables beyond that of others
5. GLH tests, which include the above tests and others not included in the above categories.

Calculating Test Statistic:

A GLM test statistic can be expressed in terms of SSE 's (and SSR 's) or $\{(\mathbf{X}'\mathbf{X})^{-1}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \mathbf{C}, \boldsymbol{\theta}_0\}$. Under HILE Gauss, the likelihood ratio test statistic follows an F distribution.

All tests compare two models: the full model and the reduced model (this is the basic idea of likelihood ratio tests, called the *likelihood ratio principle*).

We say that the reduced model is *nested* in the full model (so that its parameters are a subset of the parameters in the full model).

Sometimes, investigators wish to compare non-nested models.

(Example: An investigator wants to compare a model that uses weight as a continuous predictor to a model that uses indicator variables for overweight, underweight, or normal weight.) This is much more difficult and involves model-selection criteria other than F-tests. We

will only consider *nested* models for the present.

It is important to recognize that tests decompose a fixed amount of variance, σ_y^2 , so that sums of squares are not created or erased but are simply moved into different locations (from SSE to SSR or vice versa). An important consequence of this fact is that SSR increases $\Leftrightarrow SSE$ decreases.

Choosing an Error Term

In many cases, a series of nested models are compared. Tests are conducted for each model in the series, comparing that model to another.

An important question is “What error term should be used?”.

- It is safest to use the error term from the largest model in the entire series instead of the larger of any given pair because $SSE/(n - r)$ from the largest model “guarantees” (by assumption) an unbiased estimate of σ^2 .
- If terms beyond those in the smaller model of a pair are unimportant, then the SSE from the smaller model allows a more powerful test because the error df are $(n - r + d)$ rather than $(n - r)$, where d indicates the number predictors by which the models differ.

-
- If terms beyond those in the smaller model of a pair are important, but we use the SSE from the smaller, inadequate model, then we will inflate (bias upwards) the estimate of σ^2 . This may substantially reduce power.
 - A model that is too small (under-fitting) gives biased $\hat{\beta}$ and $\hat{\sigma}^2$ and possibly leads to a large power loss for any n .
 - A model that is too large gives unbiased $\hat{\beta}$ and $\hat{\sigma}^2$ and usually only a small power loss that goes to zero as n goes to ∞ .
 - Because $(n - r + d)/(n - r)$ should be near 1.0 with a sufficient amount of data, the choice of error term should not really matter.
 - As a general rule, we recommend using $\hat{\sigma}^2$ from the largest model in a series. This largest model is often called (somewhat loosely) the *full model*.

Comparing Two Models

If two models differ only by the addition or deletion of one or more variables, then

$$USS(\text{larger}) \geq USS(\text{smaller})$$

$$CSS(\text{larger}) \geq CSS(\text{smaller})$$

$$SSE(\text{larger}) \leq SSE(\text{smaller}),$$

because the larger model explains more of the variability in the data and thus has a larger SS model (and smaller SSE) than the smaller model. The SS Total of the two models should be identical.

Note that CSS may not be defined for all models. 

Test Class I: Overall Tests

Corrected Overall Test

The *corrected overall test* compares the full model,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{p-1} x_{ip-1} + \varepsilon_i,$$

to the intercept-only model,

$$y_i = \beta_0 + \varepsilon_i.$$

If the model does not span an intercept this test is not defined.

This test corresponds to

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$$

and

$$H_0 : \rho_c^2 = 0.$$

Let $SSE(\text{reduced}) = SSE(\beta_0)$ be the sum of squares for error for the intercept-only model, and let $SSE(\text{full}) = SSE(\beta_0, \beta_1, \dots, \beta_{p-1})$ be the SSE for the full model. (Generally, $SSE(\beta_i, \beta_j)$ is the SSE for a model including the two parameters β_i and β_j .) Then the sum of squares for the hypothesis is given by

$$\begin{aligned}SSH &= SSE(\beta_0) - SSE(\beta_0, \beta_1, \dots, \beta_{p-1}) \\&= SSE(\text{reduced model}) - SSE(\text{full model}) \\&= CSS(\text{full model}) - CSS(\text{reduced model}) \\&= CSS(\text{full model}) - 0.\end{aligned}$$



The F statistic is

$$\begin{aligned}
F_{obs} &= \frac{\frac{SSE(\text{reduced}) - SSE(\text{full})}{dfE(\text{reduced}) - dfE(\text{full})}}{SSE(\text{full})/dfE(\text{full})} \\
&= \frac{\frac{SSE(\beta_0) - SSE(\beta_0, \beta_1, \dots, \beta_{p-1})}{dfE(\beta_0) - dfE(\beta_0, \beta_1, \dots, \beta_{p-1})}}{SSE(\beta_0, \beta_1, \dots, \beta_{p-1})/dfE(\beta_0, \beta_1, \dots, \beta_{p-1})} \\
&= \frac{CSS(\beta_0, \dots, \beta_{p-1})/(p-1)}{SSE(\beta_0, \dots, \beta_{p-1})/(n-p)},
\end{aligned}$$

for full rank \mathbf{X} .

The hypothesis is rejected if

$$F_{obs} > F_F^{-1}(1 - \alpha, p - 1, n - p).$$

We may also obtain the same test via the GLH formulation with

$$F_{obs} = \frac{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathbf{M}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) / dfH}{\mathbf{y}' (\mathbf{I} - \mathbf{H}) \mathbf{y} / dfE},$$

where

- $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' M^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is the *SSH*,
- $dfH = a$ is the hypothesis degrees of freedom or the number of rows of $\boldsymbol{\theta}$ (unless we pick a redundant $\boldsymbol{\theta}$),
- $\mathbf{y}' (\mathbf{I} - \mathbf{H}) \mathbf{y}$ is the *SSE*, and
- $dfE = \text{rank}(\mathbf{I} - \mathbf{H})$ is the degrees of freedom for error.

To conduct the corrected overall test, we pick

$$\mathbf{C} = \begin{bmatrix} \mathbf{0}_{p-1} & | & \mathbf{I}_{p-1} \end{bmatrix} = \begin{bmatrix} 0 & 1 & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & 1 \end{bmatrix}_{(p-1) \times p}$$

and $\boldsymbol{\theta}_0 = \mathbf{0}$, which imply that $\boldsymbol{\theta} = [\beta_1, \dots, \beta_{p-1}]'$ and $a = p - 1$.

The following R and SAS code may be used to construct the corrected overall test for the ozone data, which is a test of

$$H_0 : \beta_{OUTDOOR} = \beta_{HOME} = \beta_{TIMEOUT} = 0.$$

```
> ozone = read.table("ozone.txt", header = T, sep = " ") # space delimited
> n = nrow(ozone) # number of rows of ozone (obs)
> X = model.matrix(~ outdoor + home + time_out, data = ozone) # predictor matrix
> y = ozone$personal # response
> head(X)

(Intercept)  outdoor  home time_out
1           1 35.87771 22.29      0.57
2           1 43.79189 13.97      0.90
3           1 49.81255 18.96      0.55
4           1 34.37366 22.27      0.17
5           1 45.95496 23.40      0.00
6           1 64.76558 39.62      0.30

> bhat = solve(t(X) %*% X) %*% t(X) %*% y # from notes
> yhat=X%*%bhat; # predicted values
> ehat=y-yhat; # residuals
> p=ncol(X); # num columns of X
> df=n-p; # df
```

```
> I_n = diag(rep(1, n))
> H=X%*%solve(t(X) %*% X) %*% t(X)
> sse = t(y) %*%(I_n - H)%*% y # SSE
> mse=sse/df; # MSE
> C=matrix(c(0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1), nrow = 3, byrow = T)
> print(C)

[,1] [,2] [,3] [,4]
[1,] 0 1 0 0
[2,] 0 0 1 0
[3,] 0 0 0 1

> M=C%*%solve(t(X)%*%X)%*%t(C)
> thetahat=C%*%bhat
> ssh=t(theta)%*%solve(M)%*%theta
> msh=ssh/nrow(theta)
> f_obs=msh/mse
> p=1-pf(f_obs,3,60)
> print(f_obs)

[,1]
[1,] 9.922767

> print(p)
```

```
[,1]
[1,] 2.094332e-05
```

We may also obtain the same values in SAS after running the following

```
proc glm data=ozone;
model personal= outdoor home time_out;
run;
```

The GLM Procedure

Dependent Variable: personal Personal Ozone Exposure (ppb)

Source	DF	Sum of Squares	Mean Square	F Value
Model	3	5034.90667	1678.30222	9.92
Error	60	10148.19129	169.13652	
Corrected Total	63	15183.09796		

Source	Pr > F
Model	<.0001
Error	
Corrected Total	

We reject the null hypothesis that outdoor concentration of ozone, home concentration of ozone, and proportion of time spent outdoors have no effect on personal ozone exposure. We conclude that at least one of our predictors (outdoor concentration of ozone, home concentration of ozone, or proportion of time spent outdoors) is related to personal ozone exposure.

Heuristics

For data that are exactly normal, this F test is a *likelihood ratio test*. You will learn more about the likelihood ratio test in BIOS 661, so our discussion will be brief.

Our general approach is to fit two models: one full model containing all the parameters of interest, and one smaller model with only an intercept. Then, we compare their maximized likelihoods (or log-likelihoods).

For a fixed value of σ^2 , the maximized log-likelihood of the intercept-only model is

$$\max \log L(\beta_0) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2} \frac{SSE(\beta_0)}{\sigma^2},$$

and the maximized log-likelihood of the larger model is

$$\max \log L(\beta_0, \dots, \beta_{p-1}) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2} \frac{SSE(\beta_0, \dots, \beta_{p-1})}{\sigma^2}.$$

To compare the log-likelihoods, we calculate minus twice their difference (denoted $-2 \log \lambda$), which is given by

$$\frac{SSE(\beta_0) - SSE(\beta_0, \dots, \beta_{p-1})}{\sigma^2}.$$

We see that this criterion is used to judge whether the additional predictors associated with $(\beta_1, \dots, \beta_{p-1})$ result in a significant reduction in the SSE. This likelihood ratio criterion, $-2 \log \lambda$, has an exact F distribution when our data are exactly normal. In BIOS 661, you will learn more about the asymptotic distribution of this criterion, useful when the data are *not* exactly normal.

Uncorrected Overall Test (rarely used)

The *uncorrected overall test* is a test of whether all parameters (including an intercept, if there is one) are equal to zero. This test is defined whether or not the model includes an intercept. It compares the null model,

$$y_i = \varepsilon_i,$$

with the full model,

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{p-1} x_{ip-1} + \varepsilon_i$$

and tests both

$$H_0 : \beta_0 = \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$$

and

$$H_0 : \rho_u^2 = 0.$$

The test statistic and p-value are location dependent.

The F statistic is

$$\begin{aligned} F_{obs} &= \frac{\frac{SSE(\text{reduced}) - SSE(\text{full})}{dfE(\text{reduced}) - dfE(\text{full})}}{SSE(\text{full})/dfE(\text{full})} \\ &= \frac{\frac{SSE(\text{null}) - SSE(\beta_0, \beta_1, \dots, \beta_{p-1})}{dfE(\text{null}) - dfE(\beta_0, \beta_1, \dots, \beta_{p-1})}}{SSE(\beta_0, \beta_1, \dots, \beta_{p-1})/dfE(\beta_0, \beta_1, \dots, \beta_{p-1})} \\ &= \frac{USS(\beta_0, \dots, \beta_{p-1})/p}{SSE(\beta_0, \dots, \beta_{p-1})/(n-p)}, \end{aligned}$$

for full rank \mathbf{X} , the hypothesis is rejected if

$$F_{obs} > F_F^{-1}(1 - \alpha, p, n - p).$$

We may also obtain the same test via the GLH formulation with

$$F_{obs} = \frac{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathbf{M}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) / dfH}{\mathbf{y}' (\mathbf{I} - \mathbf{H}) \mathbf{y} / dfE},$$

where $dfH = p$ and $dfE = n - p$ when \mathbf{X} is full rank.

To conduct the uncorrected overall test, we pick

$$\mathbf{C} = [\mathbf{I}_p] = \begin{bmatrix} 1 & & & 0 \\ \vdots & \ddots & & \\ 0 & & & 1 \end{bmatrix}_{p \times p}$$

and $\boldsymbol{\theta}_0 = \mathbf{0}$. Thus $\boldsymbol{\theta} = [\beta_0, \beta_1, \dots, \beta_{p-1}]'$ and $a = p$.

Model Pools

Added-Last Pool

With p parameters, 2^p distinct models can be defined. We usually only consider subsets formed by some rule. For added-last tests, model j has β_j deleted and contains $p - 1$ parameters.

An *added-last test* compares the full model to a reduced model that is obtained by deleting the j th variable from the full model. In this subset, there exist p reduced models (corresponding to deleting each of the p parameters). Each reduced model has $p - 1$ parameters, and the full model has p parameters. The model pool for added-last testing is provided below.

Added-Last Model Pool

Model

-0	$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{ip-1} + \varepsilon_i$
-1	$y_i = \beta_0 + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{ip-1} + \varepsilon_i$
:	
-j	$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{j-1} x_{ij-1} + \cdots + \beta_{p-1} x_{ip-1} + \varepsilon_i$
:	
full	$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{j-1} x_{ij-1} + \beta_j x_{ij} + \cdots + \beta_{p-1} x_{ip-1} + \varepsilon_i$

Added-In-Order Pool

An *added-in-order test* compares parameters added in sequence. In this subset, there exist $p + 1$ models (the null model and p models corresponding to adding each of the p parameters). The $(p + 1)$ th model is the full model and has p parameters, the p th model has one fewer parameter, the $(p - 1)$ th model has two fewer parameters, etc.

The model pool for added-in-order testing is provided below.

Added-in-Order Model Pool 

Model

$$\emptyset \quad y_i = \varepsilon_i$$

$$0 \quad y_i = \beta_0 + \varepsilon_i$$

$$1 \quad y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$$

:

$$j-1 \quad y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{j-1} x_{ij-1} + \varepsilon_i$$

$$j \quad y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{j-1} x_{ij-1} + \beta_j x_{ij} + \varepsilon_i$$

:

$$p-1 \quad y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{j-1} x_{ij-1} + \beta_j x_{ij} + \dots + \beta_{q-1} x_{iq-1} + \varepsilon_i$$

Model $p-1$ is the full model.

Test Class 2: Addition of One Variable

Added-Last Test

The *added-last test* seeks to assess the usefulness of one predictor, above and beyond all others. An added-last test for variable j compares the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{j-1} x_{ij-1} + \beta_{j+1} x_{ij+1} + \dots + \beta_{p-1} x_{ip-1} + \varepsilon_i$$

to the full model and corresponds to testing the hypothesis

$H_0 : \beta_j = 0$ when β_j is the final variable added to the model.

In this test class, there are $(p - 1)$ reduced models (eliminating each of the $(p - 1)$ non-intercept parameters) with $(p - 1)$ parameters each (including the intercept). We compare each reduced model to the full model to produce an added-last-test for each predictor.

Added-last tests do not depend on the order of variables in the model. Added-last tests are printed in SAS as the Type III SS tests.

Denote the sum of squares for error for the reduced model (without the j th variable) by

$$SSE(\beta_0, \beta_1, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_{p-1}).$$

Then the F statistic is

$$\begin{aligned} F_{obs} &= \frac{\frac{SSE(\text{reduced}) - SSE(\text{full})}{dfE(\text{reduced}) - dfE(\text{full})}}{SSE(\text{full})/dfE(\text{full})} \\ &= \frac{\frac{SSE(\beta_0, \beta_1, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_{p-1}) - SSE(\beta_0, \dots, \beta_{p-1})}{dfE(\beta_0, \beta_1, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_{p-1}) - dfE(\beta_0, \dots, \beta_{p-1})}}{SSE(\beta_0, \dots, \beta_{p-1})/dfE(\beta_0, \dots, \beta_{p-1})} \\ &= \frac{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathbf{M}^{-1} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) / dfH}{\mathbf{y}' (\mathbf{I} - \mathbf{H}) \mathbf{y} / dfE}, \end{aligned}$$

where

$$\mathbf{C} = \begin{bmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 \end{bmatrix}_{1 \times p}$$

has a 1 in the position corresponding to the variable we wish to test, and 0 elsewhere. If we let $\theta_0 = 0$, then we have $\theta = \beta_j$ and $a = 1$.

$T_{obs} = F_{obs}^{\frac{1}{2}}$ follows a Student's T distribution with $(n - p)$ df if \mathbf{X} is full rank. Many regression programs report T_{obs} as a test of the j th regression coefficient equaling zero, so the usual T test for a regression coefficient provides an added last test.

We reject the hypothesis if $F_{obs} \geq F_F^{-1}(1 - \alpha, 1, n - p)$.

SAS code and output for an added-last test of the variable $O_{OUTDOOR}$ is provided.

```
> C=matrix(c(0, 1, 0, 0), nrow = 1, byrow = T)
> print(C)

 [,1] [,2] [,3] [,4]
[1,]    0    1    0    0

> M=C%*%solve(t(X)%*%X)%*%t(C)
> thetahat=C%*%bhat
> ssh=t(thetahat)*solve(M)%*%thetahat
> msh=ssh/nrow(thetahat)
> f_obs=msh/mse
> p=1-pf(f_obs,1,60)
> print(f_obs)

 [,1]
[1,] 1.022314

> print(p)

 [,1]
[1,] 0.3160309

proc glm data=ozone;
model personal= outdoor home time_out;
run;
```

The GLM Procedure

Source	DF	Type I SS	Mean Square	F Value
outdoor	1	2419.043105	2419.043105	14.30
home	1	2086.106656	2086.106656	12.33
time_out	1	529.756909	529.756909	3.13

Source	Pr > F
outdoor	0.0004
home	0.0009
time_out	0.0818

Source	DF	Type III SS	Mean Square	F Value
outdoor	1	172.910655	172.910655	1.02
home	1	2208.421320	2208.421320	13.06
time_out	1	529.756909	529.756909	3.13

Source	Pr > F
outdoor	0.3160
home	0.0006

	time_out	0.0818		
Standard				
Parameter	Estimate	Error	t Value	Pr > t
Intercept	3.78348593	4.34205547	0.87	0.3870
outdoor	0.09142005	0.09041683	1.01	0.3160
home	0.59543659	0.16478332	3.61	0.0006
time_out	13.64453832	7.70973105	1.77	0.0818

Again, note that the t value squared is equal to the F statistic for the added-last test (Type III SS). We fail to reject the null hypothesis and conclude that outdoor ozone concentration does not make a significant contribution to personal exposure above and beyond the intercept, home ozone, and the proportion of time spent outdoors.

Caution: Added-Last Tests

Suppose you want to estimate a person's weight given the length of their legs. (Legs account for 30-35% of human body weight.) First, you fit the model

$$E(\text{weight}) = \beta_0 + \beta_1 \text{ right leg length} .$$

The results of this model fit are provided below.

```
proc glm;
model weight=rleg;
run;
*****
          The GLM Procedure
```

Dependent Variable: weight

Source	DF	Sum of		F Value
		Squares	Mean Square	
Model	1	3627.670184	3627.670184	125.75
Error	98	2827.099916	28.847958	
Corrected Total	99	6454.770100		

Source	Pr > F
--------	--------

	Model	<.0001
R-Square	Coeff Var	Root MSE weight Mean
0.562016	6.969661	5.371030 77.06300
Source	DF	Type I SS Mean Square F Value
rleg	1	3627.670184 3627.670184 125.75
	Source	Pr > F
	rleg	<.0001
Source	DF	Type III SS Mean Square F Value
rleg	1	3627.670184 3627.670184 125.75
	Source	Pr > F
	rleg	<.0001
		Standard
Parameter	Estimate	Error t Value Pr > t
Intercept	-3.731074382	7.22481246 -0.52 0.6067
rleg	1.008948511	0.08997309 11.21 <.0001

The added-last test, given by the “Type III” results, leads us to conclude that right leg length is significantly related to weight.

Now consider the results from fitting the model

$$E(\text{weight}) = \beta_0 + \beta_1 \text{ right leg length} + \beta_2 \text{ left leg length}.$$

```
proc glm;
model weight=rleg lleg;
run;
*****  
The GLM Procedure
```

Dependent Variable: weight

Source	DF	Sum of		F Value
		Squares	Mean Square	
Model	2	3698.009886	1849.004943	65.06
Error	97	2756.760214	28.420208	
Corrected Total	99	6454.770100		

Source	Pr > F
Model	<.0001

R-Square	Coeff Var	Root MSE	weight Mean
0.572911	6.917795	5.331061	77.06300

Source	DF	Type I SS	Mean Square	F Value
--------	----	-----------	-------------	---------

rleg	1	3627.670184	3627.670184	127.64
lleg	1	70.339702	70.339702	2.47

Source	Pr > F
rleg	<.0001
lleg	0.1189

Source	DF	Type III SS	Mean Square	F Value
rleg	1	88.17311509	88.17311509	3.10
lleg	1	70.33970223	70.33970223	2.47

Source	Pr > F
rleg	0.0813
lleg	0.1189

Parameter	Estimate	Standard		
		Error	t Value	Pr > t
Intercept	-1.828644701	7.27229423	-0.25	0.8020
rleg	9.433109580	5.35550484	1.76	0.0813
lleg	-8.447148369	5.36937194	-1.57	0.1189

The added-last tests of right leg length ($p=0.08$) and left leg length (0.12) indicate that neither significantly predicts weight! This occurs because after you adjust for the effect of the length of one leg, the other provides no additional useful information. Such situations arise often in biostatistics, especially when predictors may be highly correlated.

With an added-in-order test (“Type I” in SAS), which we will discuss next, instead of evaluating each variable in the presence of all others, we will evaluate variables in a fixed order. Using that type of test for the leg data, you find that the leg added first (here, the right leg) does significantly predict weight, while the addition of the second leg to the model does not sharpen our prediction over a model already containing one leg. Due to the high correlation of left and right leg lengths (here, greater than 0.99), this result is not surprising.

Added-in-Order Test

The *added-in-order test* seeks to assess the contribution of predictor j above and beyond all of the preceding $j - 1$ predictors (without the $j + 1, j + 2$, etc.predictors in the model).

Added-in-order SS are mutually exclusive and together exhaustive pieces of the model SS (i.e.,if you add the SS for each predictor, the resulting sum is the model SS). Results do depend on order of inclusion except when all predictors are uncorrelated. If all predictors are uncorrelated, then added-last SS coincide with added-in-order.

Added-in-order tests are available from SAS type I SS. The sum of squares for the hypothesis for the added-in-order test for the variable j is given by

$$\begin{aligned}SSH &= SSE(\beta_0, \dots, \beta_{j-1}) - SSE(\beta_0, \dots, \beta_{j-1}, \beta_j) \\&= SSE(\text{smaller}) - SSE(\text{larger}).\end{aligned}$$

The recommended F statistic is given by

$$\begin{aligned} F_{obs} &= \frac{\frac{SSE(\text{smaller}) - SSE(\text{larger})}{dfE(\text{smaller}) - dfE(\text{larger})}}{SSE(\text{full})/dfE(\text{full})} \\ &= \frac{\frac{SSE(\beta_0, \beta_1, \dots, \beta_{j-1}) - SSE(\beta_0, \dots, \beta_j)}{dfE(\beta_0, \beta_1, \dots, \beta_{j-1}) - dfE(\beta_0, \dots, \beta_j)}}{SSE(\beta_0, \dots, \beta_{p-1})/(n-p)} \\ &= \frac{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' M^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) / dfH}{\mathbf{y}' (\mathbf{I} - \mathbf{H}) \mathbf{y} / dfE}, \end{aligned}$$

where

$$\mathbf{C} = \begin{bmatrix} 0 & \dots & 0 & 1 \end{bmatrix}_{1 \times (j+1)}$$

and multiplies the $\beta_{(j+1) \times 1}$ vector corresponding to the larger model.

If we let $\theta_0 = 0$, then we have $\theta = \beta_j$ and $a = 1$.

(Although it is possible to use the MSE from the larger of the two models being compared as the denominator of the F test, the MSE from the full model is preferred.)

Test Class 3: Tests of the Intercept

General Features

While tests of the intercept are generally not recommended, they may be obtained by treating the intercept variable (a column of 1's) just like any other variable in Class 2 (Addition of One Variable Tests).

Muller and Fetterman discuss examples of situations in which tests of the intercept are valid.

Test Class 4: Addition of a Group of Variables

The tests in this class are conducted just like tests in Class 2, with the exception that now we are testing more than one variable (and thus have more than one degree of freedom for the hypothesis).

Group Added-Last Tests

We assume predictors fall into two groups $\{X_1, \dots, X_{g_1}\}$ and $\{X_{g_1+1}, \dots, X_{p-1}\}$, with g_1 variables in the first group and $g_2 = [(p - 1) - g_1]$ variables in the second group. Note that $p = 1 + g_1 + g_2$.

Group Added-Last Pool

Model

$$1 \quad y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{g_1} x_{ig_1} + \varepsilon_i$$

$$2 \quad y_i = \beta_0 + \beta_{g_1+1} x_{ig_1+1} + \dots + \beta_{p-1} x_{ip-1} + \varepsilon_i$$

$$3 \quad y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{g_1} x_{ig_1} + \beta_{g_1+1} x_{ig_1+1} + \dots + \beta_{p-1} x_{ip-1} + \varepsilon_i$$

The added-last test of the first group compares model 2 and model 3, with corresponding hypothesis

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_{g_1} = 0.$$

The added last test of the second group compares model 1 and model 3, with corresponding hypothesis

$$H_0 : \beta_{g_1+1} = \beta_{g_1+2} = \dots = \beta_{p-1} = 0.$$

Always compare the appropriate reduced model to the full model. The methods generalize to any number of groups.

These tests are often called *lack-of-fit* tests and measure the adequacy of a smaller model compared to a larger model.

The F statistic for an added-last test of the first group is given by

$$\begin{aligned} F_{obs} &= \frac{\frac{SSE(\text{reduced}) - SSE(\text{full})}{dfE(\text{reduced}) - dfE(\text{full})}}{SSE(\text{full})/dfE(\text{full})} \\ &= \frac{\frac{SSE(\beta_{g_1+1}, \beta_{g_1+2}, \dots, \beta_{p-1}) - SSE(\beta_0, \dots, \beta_{p-1})}{dfE(\beta_{g_1+1}, \beta_{g_1+2}, \dots, \beta_{p-1}) - dfE(\beta_0, \dots, \beta_{p-1})}}{SSE(\beta_0, \dots, \beta_{p-1})/(n-p)} \\ &= \frac{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathbf{M}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) / dfH}{\mathbf{y}' (\mathbf{I} - \mathbf{H}) \mathbf{y} / dfE}, \end{aligned}$$

where

$$\mathbf{C} = \begin{bmatrix} \mathbf{0}_{\mathbf{g}_1 \times 1} & \mathbf{I}_{g_1} & \mathbf{0}_{\mathbf{g}_1 \times \mathbf{g}_2} \end{bmatrix}$$

and $\boldsymbol{\theta}_0 = \mathbf{0}$, which implies $\boldsymbol{\theta} = (\beta_1, \dots, \beta_{g_1})$ and $a = g_1$.

Exercise: Computing a Group Added Last Test

Given the following SAS code, compute the group added-last test of no effect of outdoor ozone concentration and home ozone concentration for the ozone data.

```
proc glm data=ozone;
model personal=outdoor home time_out;
run;
proc glm data=ozone;
model personal=time_out;
run;
```

```
*****
```

The GLM Procedure

Dependent Variable: personal Personal Ozone Exposure (ppb)

Source	DF	Sum of		
		Squares	Mean Square	F Value
Model	3	5034.90667	1678.30222	9.92
Error	60	10148.19129	169.13652	
Corrected Total	63	15183.09796		

	Source		Pr > F
	Model		<.0001
R-Square	Coeff Var	Root MSE	personal Mean
0.331613	55.23389	13.00525	23.54578

The GLM Procedure

Dependent Variable: personal Personal Ozone Exposure (ppb)

Source	DF	Sum of		F Value
		Squares	Mean Square	
Model	1	574.49202	574.49202	2.44
Error	62	14608.60594	235.62268	
Corrected Total	63	15183.09796		

	Source		Pr > F
	Model		0.1235
R-Square	Coeff Var	Root MSE	personal Mean
0.037838	65.19217	15.35001	23.54578

Group Added-In-Order Tests

The *group added-in-order test* seeks to assess the usefulness of a group of predictors above and beyond that of preceding predictors.

Group Added-in-Order Pool

Model

0	$y_i = \beta_0$	$+ \varepsilon_i$
1	$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{g_1} x_{ig_1}$	$+ \varepsilon_i$
2	$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{g_1} x_{ig_1} + \beta_{g_1+1} x_{ig_1+1} + \dots + \beta_{p-1} x_{ip-1}$	$+ \varepsilon_i$

An added-in-order test of the first group compares models 0 and 1, with corresponding hypothesis

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_{g_1} = 0.$$

Listing order strongly affects the tests except when all predictors are uncorrelated (then added-last tests coincide with added-in-order tests).

Added-in-order sums of squares are mutually exclusive and together exhaustive, while added-last sums of squares typically does not overlap.

The F statistic for an added-in-order test of the first group is given by

$$\begin{aligned} F_{obs} &= \frac{\frac{SSE(\text{smaller}) - SSE(\text{larger})}{dfE(\text{smaller}) - dfE(\text{larger})}}{SSE(\text{full})/dfE(\text{full})} \\ &= \frac{\frac{SSE(\beta_0) - SSE(\beta_0, \dots, \beta_{g_1})}{dfE(\beta_0) - dfE(\beta_0, \dots, \beta_{g_1})}}{SSE(\beta_0, \dots, \beta_{p-1})/(n-p)} \\ &= \frac{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathbf{M}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) / dfH}{\mathbf{y}' (\mathbf{I} - \mathbf{H}) \mathbf{y} / dfE}, \end{aligned}$$

where

$$\mathbf{C} = \begin{bmatrix} \mathbf{0}_{g_1 \times 1} & \mathbf{I}_{g_1} \end{bmatrix}$$

and $\boldsymbol{\theta}_0 = \mathbf{0}$, which implies $\boldsymbol{\theta} = (\beta_1, \dots, \beta_{g_1})$ and $a = g_1$.

As with a single variable test, the in-order test for the last group always coincides with the corresponding added-last test.

Exercise: Computing a Group Added-In-Order Test

Given the additional SAS code and output below, compute the group added-in-order test of no effect of outdoor ozone concentration and home ozone concentration for the ozone data.

```
proc glm data=ozone;
model personal=/;
run;
proc glm data=ozone;
model personal=outdoor home;
run;
```

```
*****
```

The GLM Procedure

Dependent Variable: personal Personal Ozone Exposure (ppb)

Source	DF	Sum of		
		Squares	Mean Square	F Value
Model	1	35481.84414	35481.84414	147.23
Error	63	15183.09796	241.00155	
Uncorrected Total	64	50664.94210		

Source	Pr > F
Model	<.0001
Error	
Uncorrected Total	

The GLM Procedure

Dependent Variable: personal Personal Ozone Exposure (ppb)

Source	DF	Sum of		F Value
		Squares	Mean Square	
Model	2	4505.14976	2252.57488	12.87
Error	61	10677.94820	175.04833	
Corrected Total	63	15183.09796		

Source	Pr > F
Model	<.0001

R-Square	Coeff Var	Root MSE	personal Mean
0.296721	56.19089	13.23058	23.54578

Source	DF	Type I SS	Mean Square	F Value
outdoor	1	2419.043105	2419.043105	13.82
home	1	2086.106656	2086.106656	11.92

Source	Pr > F
--------	--------

	outdoor		0.0004
	home		0.0010
Source	DF	Type III SS	Mean Square F Value
outdoor	1	240.247837	240.247837 1.37
home	1	2086.106656	2086.106656 11.92
	Source		Pr > F
	outdoor		0.2459
	home		0.0010

Test Class 5: Other GLH Tests

As we have seen so far, the GLH test is a powerful tool to test many hypotheses.

Example: Laboratory Instrument Validation

A hospital laboratory has obtained a new machine to measure CO bound to hemoglobin. Investigators wish to know if the new machine is producing the same readings as the old one.

If so, the predicted regression line should be a line with slope one through the origin.

Ideally,

$$\begin{aligned}\mathbf{new}_i &= \mathbf{old}_i + \varepsilon_i \\ &= 0 + 1 \cdot \mathbf{old}_i + \varepsilon_i \\ &= \beta_0 + \beta_1 \mathbf{old}_i + \varepsilon_i \\ \mathbf{new} &= \begin{bmatrix} \mathbf{1} & \mathbf{old} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \varepsilon\end{aligned}$$

The equivalence of the new and old machines implies $\beta_0 = 0$ and $\beta_1 = 1$. So our null hypothesis is

$$H_0: \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Setting $\mathbf{C} = \mathbf{I}_2$ yields $\boldsymbol{\theta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$.

Choosing $\boldsymbol{\theta}_0 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ defines a GLH which compares two models,

$$1 \quad \text{new}_i = \text{old}_i + \varepsilon_i$$

$$2 \quad \text{new}_i = \beta_0 + \beta_1 \cdot \text{old}_i + \varepsilon_i ,$$

with test statistic

$$F_{obs} = \frac{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathbf{M}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) / dfH}{\mathbf{y}' (\mathbf{I} - \mathbf{H}) \mathbf{y} / dfE},$$

where $dfH = a = 2$.

We note that PROC GLM will only test a GLH with $\boldsymbol{\theta}_0 = \mathbf{0}$. In order to conduct this test, we must either use R or PROC REG with the TEST statement.

Multiple Testing

In multiple regression, we may wish to conduct $p - 1$ tests (one for each predictor in the model). This multiple testing often inflates α above the desired nominal level, which is generally $\alpha = 0.05$.

The Bonferroni inequality says that

$$\begin{aligned}\Pr \{A_1 \cup A_2\} &= \Pr \{A_1\} + \Pr \{A_2\} - \Pr \{A_1 \cap A_2\} \\ \Pr \{A_1 \cup A_2\} &\leq \Pr \{A_1\} + \Pr \{A_2\}.\end{aligned}$$

If $\Pr \{A_1\} = \Pr \{A_2\} = \alpha_k$, then $\Pr \{A_1 \cup A_2\} \leq 2\alpha_k$.

More generally

$$\Pr \{A_1 \cup A_2 \cup A_3 \dots \cup A_K\} \leq K\alpha_k.$$

To ensure an upper bound of α overall, we use $\alpha_k = \alpha/K$.

For independent events (this rarely applies to multiple comparisons!)

$$\Pr \{A_1 \cup A_2 \cup A_3 \dots \cup A_K\} = 1 - (1 - \alpha_k)^K.$$

The Bonferroni correction is accurate for small K , is almost always accurate for independent events, but is least accurate for highly correlated events. The poor accuracy results in less powerful tests, but the Bonferroni correction does guarantee control of the type I error rate.

Exercise: Bonferroni Correction

Suppose we will test 10 predictors and want an overall $\alpha = 0.05$. At which level should we conduct each test?

Interaction

X_1 and X_2 interact (in predicting Y) if one must know the level of X_1 in order to describe relationship of Y to X_2 (and hence must know the level of X_2 in order to describe relationship of Y to X_1).

For example, one cancer treatment may be beneficial to patients with a certain genotype of a gene but not to other patients.

Interaction Model Pool

Model

1	$y_i = \beta_0 + \beta_3 x_{i1} x_{i2} + \varepsilon_i$	avoid
2	$y_i = \beta_0 + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i$	avoid
3	$y_i = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i1} x_{i2} + \varepsilon_i$	avoid
4	$y_i = \beta_0 + \varepsilon_i$	
5	$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$	
6	$y_i = \beta_0 + \beta_2 x_{i2} + \varepsilon_i$	
7	$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$	
8	$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i$	

Analysis plan:

First, test the interaction, $H_0: \beta_3 = 0$, with $\mathbf{C} = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}$ and $\theta_0 = 0$, to compare model 8 to model 7. If significant, quit.

The interaction test asks if $y_i = \alpha_0 + \alpha_1 x_{i1} + \varepsilon_i$ is the same for all x_{i2} .

For $x_{i2} \neq x'_{i2}$, the null hypothesis corresponds to parallel lines with distinct intercepts (for $\beta_3 = 0$).

If β_3 is significant, both variables are important, even if β_1 and/or β_2 is not significant. If β_3 is significant, retain both variables, even if β_1 and/or β_2 is not significant.

Always include *main effects* (here, x_1 and x_2) in the model when an interaction term between them is also in the model.

Interpreting Interaction Terms

With an interaction term in the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i,$$

it is difficult to interpret β_1 and β_2 in isolation. However, we may express the effect of a one-unit increase in x_1 with x_2 held constant as

$$\begin{aligned} E(y_i | x_{i1} + 1, x_{i2}) - E(y_i | x_{i1}, x_{i2}) &= \\ &\quad \beta_0 + \beta_1(x_{i1} + 1) + \beta_2 x_{i2} \\ &\quad + \beta_3(x_{i1} + 1)x_{i2} \\ &\quad - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}) \\ &= \beta_1 + \beta_3 x_{i2}. \end{aligned}$$

Similarly, we can find the effect of a one-unit increase in x_2 with x_1 held constant.

Next: Correlation

Reading Assignment:

- Muller and Fetterman, Chapter 6: “Correlations” (Required)

Lecture 9: Correlations

Reading Assignment:

- Muller and Fetterman, Chapter 6: “Correlations”

For two random variables X and Y , recall that the correlation ρ is defined as

$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}},$$

where $-1 \leq \rho \leq 1$. We estimate the population correlation, ρ , using *Pearson's coefficient of correlation*, given by

$$R = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left(\sum_{i=1}^n (X_i - \bar{X})^2\right) \left(\sum_{i=1}^n (Y_i - \bar{Y})^2\right)}}.$$

Consider the simple linear model $\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \boldsymbol{\varepsilon}$.

- The *squared correlation coefficient*, R^2 , measures the strength of the linear relationship between the variables \mathbf{y} and \mathbf{x} . $R^2 \rightarrow 1$ indicates a stronger linear relationship, and $R^2 \rightarrow 0$ indicates a weaker linear relationship.
- The *sample correlation coefficient* R is a general measure of the *linear* relationship between two random variables.
 - If $R \approx 0$, there is little evidence of a linear association between two variables. (This does not mean there is no association between the two variables.)
 - As $R \rightarrow 1$, the linear association is more positive (i.e., subjects with high values of x will tend to have high values of y as well).
 - As $R \rightarrow -1$, the linear association is more negative (i.e., subjects with high values of x will likely have low values of y).

We can interpret R as the expected change in the response (in units of standard deviation) associated with a change of one standard deviation in the predictor of interest.

To see this, standardize the response by subtracting its mean and dividing by its standard deviation, standardize the predictor x in the same way, and then regress the standardized response on the standardized predictor, omitting an intercept from the model. The resulting estimate of the regression coefficient is R . We verify this using the ozone data below.

```
proc corr;  
var personal outdoor;  
run;  
  
proc standard data=ozone mean=0 std=1 out=ozone2;  
var personal outdoor;  
run;  
  
proc glm data=ozone2;  
model personal=outdoor/noint;  
run;  
*****
```

Pearson Correlation Coefficients, N = 64
Prob > |r| under H0: Rho=0

	personal	outdoor
personal	1.00000	0.39916
		0.0011
outdoor	0.39916	1.00000
		0.0011

The GLM Procedure

Dependent Variable: personal

Source	DF	Sum of Squares	Mean Square	F Value
Model	1	10.03745850	10.03745850	11.94
Error	63	52.96254150	0.84067526	
Uncorrected Total	64	63.00000000		

Source	Pr > F
Model	0.0010
Error	
Uncorrected Total	

R-Square	Coeff Var	Root MSE	personal Mean
0.159325	-2.8725E17	0.916883	-0.000000

NOTE: No intercept term is used: R-square is not corrected for the mean.

Source	DF	Type I SS	Mean Square	F Value
outdoor	1	10.03745850	10.03745850	11.94

Source	Pr > F
outdoor	0.0010

Source	DF	Type III SS	Mean Square	F Value
outdoor	1	10.03745850	10.03745850	11.94

Source	Pr > F
outdoor	0.0010

Parameter	Estimate	Standard		
		Error	t Value	Pr > t
outdoor	0.3991550301	0.11551646	3.46	0.0010

Example: Correlation Matrix for Ozone Data

The correlation matrix for the ozone data is provided below.

The CORR Procedure

4 Variables: personal outdoor home time_out

Pearson Correlation Coefficients, N = 64

	personal	outdoor	home	time_out
personal	1.00000	0.39916	0.53000	0.19452
outdoor	0.39916	1.00000	0.55582	0.07818
home	0.53000	0.55582	1.00000	-0.00714
time_out	0.19452	0.07818	-0.00714	1.00000

Pearson Correlation Coefficients, N = 64

How do we interpret these values?

These correlations may not give us exactly the information we want. Suppose that based on these correlations, we decide to perform a regression of personal exposure on home exposure. Before considering whether outdoor ozone or time spent outdoors should be added next, it would be helpful to know how associated each is with the response *after* we have controlled for home ozone.

The simple correlations do not tell us about

1. the relationship between $O_{PERSONAL}$ and $(O_{OUTDOOR}, O_{HOME}, \text{ and } TIME_{OUT})$ as a group (multiple correlation coefficient),
2. the relationship between $O_{PERSONAL}$ and O_{HOME} after controlling for $TIME_{OUT}$ (partial correlation coefficient), or
3. the relationship between $O_{PERSONAL}$ and the combined effects of $O_{OUTDOOR}$ and O_{HOME} after controlling for $TIME_{OUT}$ (multiple partial correlation coefficient).

Interpreting ρ^2

Corrected ρ^2

Consider decomposing the variance of Y , namely σ_y^2 . Because $\sigma_y^2 \geq 0$, it is sensible to ask what fraction (proportion) of the variance is explained by using a linear model to predict Y .

For a model that spans the intercept, one can define the *corrected correlation coefficient*

$$\rho_c^2 = \frac{\sigma^2(\beta_0) - \sigma^2(\beta_0 + \beta_1 X_1 + \beta_2 X_2 \dots + \beta_{p-1} X_{p-1})}{\sigma^2(\beta_0)}$$

$$\hat{\rho}_c^2 = R_c^2 = \frac{CSS(\text{Regression})}{CSS(\text{Regression}) + SSE}$$

$$= \frac{CSS(\text{Regression})}{CSS(\text{total})}$$

Under HILE Gauss, $\hat{\rho}_c^2$ is the MLE. Both $0 \leq \rho_c^2 \leq 1$ and $0 \leq R_c^2 \leq 1$.

R_c^2 equals the squared univariate correlation between Y and \hat{Y} and is provided for the ozone data below.

Example: R_c^2 for ozone data

Verify that the value labeled “R-Square” on the SAS output is the corrected R_c^2 . How do we interpret this value?

```
proc glm data=ozone;
model personal= outdoor home time_out;
run;
*****
```

Source	DF	Sum of		
		Squares	Mean Square	F Value
Model	3	5034.90667	1678.30222	9.92
Error	60	10148.19129	169.13652	
Corrected Total	63	15183.09796		

Source	Pr > F
Model	<.0001

R-Square	Coeff Var	Root MSE	personal Mean
0.331613	55.23389	13.00525	23.54578

“Uncorrected” ρ^2

The value of uncorrected correlations lies in evaluating models without an intercept and in evaluating the value of an intercept. We follow the default of considering only corrected correlations. Note that the discussion of partial correlations generalizes to uncorrected correlations with, as always, special attention paid to the presence or absence of the intercept.

Generalizing the Application of ρ^2

All GLH tests correspond to comparing two nested models, and all GLH tests correspond to comparing two correlations. We will show that the standard F test (with the SSE estimated from the larger model and not the full model in this case) can be written in terms of the R^2 from the models being compared.

$$\begin{aligned}
 F_{obs} &= \frac{\frac{[SSE(\text{smaller}) - SSE(\text{larger})]}{[dfE(\text{smaller}) - dfE(\text{larger})]}}{\frac{SSE(\text{larger})}{dfE(\text{larger})}} \\
 &= \frac{\frac{[CSS(\text{larger}) - CSS(\text{smaller})]}{[dfE(\text{smaller}) - dfE(\text{larger})]}}{\frac{SSE(\text{larger})}{dfE(\text{larger})}} = \frac{\frac{[R^2(\text{larger}) - R^2(\text{smaller})]}{a}}{\frac{[1 - R^2(\text{larger})]}{dfE(\text{larger})}}.
 \end{aligned}$$

We assume the smaller model contains $\{X_1, \dots, X_{g_1}\}$ and the larger model differs by the addition of $\{X_{g_1+1}, \dots, X_{p-1}\}$.

Let $\Delta R^2 = R^2(\text{larger}) - R^2(\text{smaller})$, with

$$0 \leq \Delta R^2 \leq 1.$$

If all variables are scaled to have unit variance, ΔR^2 equals the covariance (with Y) of a group of predictors, adjusted for predictors already in the model. Examining that relationship leads to the study of partial correlations.

Correlation Formulae

Let $\mathbf{v}_j = \{v_{ij}\}$ indicate an $n \times 1$ vector of i.i.d. observations.

- Compute the mean as $\bar{v}_j = \mathbf{J}'_n \mathbf{v}_j / n$.
- Compute the sample variance as

$$s_j^2 = \sum_{i=1}^n (v_{ij} - \bar{v}_j)^2 / n = \mathbf{v}'_j \mathbf{v}_j / n - \bar{v}_j^2,$$

with $s_j^2 \left(\frac{n}{(n-1)} \right)$ an unbiased estimator of the population variance.

- Compute the sample covariance for two vectors j and j' as

$$c_{jj'} = \sum_{i=1}^n (v_{ij} - \bar{v}_j)(v_{ij'} - \bar{v}_{j'}) / n = \mathbf{v}'_j \mathbf{v}_{j'} / n - \bar{v}_j \bar{v}_{j'},$$

with $c_{jj'} \left(\frac{n}{(n-1)} \right)$ an unbiased estimator of the population covariance.

-
- Compute the sample correlation coefficient

$$r(v_{ij}, v_{ij'}) = \frac{c_{jj'}}{s_j s_{j'}} .$$

Partial Correlation

Partial correlations involve a predictor, X , a response, Y , and nuisance variables, Z .

Partial correlations describe the strength of the linear relationship between two variables, Y and X , after controlling for the effects of other variables Z . When multiple X and Z variables are involved, we call the partial correlation a *multiple partial correlation*. The *order* of a partial correlation depends on the number of variables for which we control: *first-order* partials control for $Z_{(n \times 1)}$, *second-order* partials control for $Z_{(n \times 2)}$, and p^{th} -*order* partials control for $Z_{(n \times p)}$.

Example: Why do we want to consider a partial correlation? Consider computing the correlation between hair length and height in the general population.

Consider the following table of partial correlations for the ozone data.

<i>Order</i>	<i>Controlling Variables</i>	<i>Form of Correlation</i>	<i>Correlation Estimate</i>
0		$r_{PERS, HOME}$	0.53
0		$r_{PERS, OUT}$	0.40
0		$r_{PERS, TIME}$	0.19
1	HOME	$r_{PERS, OUT HOME}$	0.15
1	HOME	$r_{PERS, TIME HOME}$	0.23
1	OUT	$r_{PERS, HOME OUT}$	0.40
1	OUT	$r_{PERS, TIME OUT}$	0.18
1	TIME	$r_{PERS, HOME TIME}$	0.54
1	TIME	$r_{PERS, OUT TIME}$	0.39
2	HOME,OUT	$r_{PERS, TIME HOME, OUT}$	0.22
2	HOME,TIME	$r_{PERS, OUT HOME, TIME}$	0.13
2	OUT,TIME	$r_{PERS, HOME OUT, TIME}$	0.42

-
- Which variable has the greatest linear association with personal exposure?
 - After that variable, what is the next most important variable?
 - How do we tell whether the linear association is "enough" to warrant inclusion in a statistical model? When do we stop adding variables?

At some point, we need to decide whether a particular partial correlation coefficient is significantly different from zero. Previously, we used F tests to determine whether adding a variable to a regression model was worthwhile, given that other variables were in the model. This type of test is also called a *partial F test*. This test is exactly equivalent to a test of significance for the corresponding partial correlation coefficient.

That is, a test of whether the population partial correlation coefficient $\rho_{PERS, TIME|HOME}$ is equal to 0 is exactly equivalent to a test of $H_0 : \beta_{TIME} = 0$ given that home exposure is already in the model.

Insight of Partial Correlation

Consider fitting two models, $\mathbf{y} = \mathbf{z}\beta + \boldsymbol{\varepsilon}$ and $\mathbf{x} = \mathbf{z}\gamma + \boldsymbol{\varepsilon}$, to create $\widehat{\beta}$ and $\widehat{\gamma}$ as well as predicted values $\widehat{\mathbf{y}} = \mathbf{z}\widehat{\beta}$ and $\widehat{\mathbf{x}} = \mathbf{z}\widehat{\gamma}$.

Let $\widehat{\varepsilon}_{y_i} = y_i - \widehat{y}_i$ and $\widehat{\varepsilon}_{x_i} = x_i - \widehat{x}_i$.

Describe $r(\widehat{\varepsilon}_{y_i}, \widehat{\varepsilon}_{x_i})$, the correlation between these residuals, as the sample value of the (full) *partial correlation* between Y and X , controlling both for \mathbf{Z} . We use the notation $r_{YX|Z}$ for this partial correlation.

The partial correlation describes the simple linear correlation between Y and X after removing the linear effects of \mathbf{Z} .

Adjusting for Confounders

A very important application of partial F tests concerns controlling for confounders. Suppose that we have one main study variable of interest, \mathbf{x} , and q control variables (or confounders) $\mathbf{z}_1, \dots, \mathbf{z}_q$. We can evaluate the effect of \mathbf{x} on the outcome of interest, controlling for the potential confounders, \mathbf{z} , using the model

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{z}_1 + \dots + \beta_q \mathbf{z}_q + \beta_{q+1} \mathbf{x} + \boldsymbol{\epsilon}.$$

We then test the hypothesis $H_0 : \beta_{q+1} = 0$, which is equivalent to testing whether the population partial correlation between \mathbf{y} and \mathbf{x} , adjusting for the confounders \mathbf{Z} , is zero.

When we are interested in several variables in \mathbf{X} , we must determine which of these are important, and we may need to rank order them by their relative importance. We will discuss this strategy further when we talk about model selection.

Semi-Partial Correlation

We consider semi-partial correlations when we know that only one of X or Y is affected by the nuisance variables Z . For example, in a randomized clinical trial, the treatment variable X should be unaffected by nuisance factors Z (such as age or tumor type), although these nuisance factors may have an effect on the outcome.

Describe $r(\hat{\varepsilon}_{y_i}, x_i) = r_{X(Y|Z)}$ as the semi-partial correlation between Y and X , controlling only Y for Z .

Also define $r(y_i, \hat{\varepsilon}_{x_i}) = r_{Y(X|Z)}$ as the semi-partial correlation between Y and X , controlling X for Z .

Example: Semi-Partial Correlations in Ozone Data

Suppose we believe that while time spent outdoors may be related to personal ozone exposure, time spent outdoors and home ozone concentration are unlikely to be related ($r=-0.007$). In this case, when computing the correlation between personal ozone exposure and home ozone concentrations, we may wish to correct personal ozone exposure (but not home ozone concentrations) for the effect of time spent outdoors.

First, we will compute the partial correlation between personal ozone exposure and home ozone concentration, controlling both variables for the effect of time spent outdoors.

```
proc corr data=ozone nosimple /* noprobs */;  
var personal home;  
partial time_out;  
run;  
*****
```

Pearson Partial Correlation Coefficients, N = 64

Prob > |r| under H0: Partial Rho=0

	personal	home
personal	1.00000	0.54175
Personal Ozone Exposure (ppb)		<.0001
home	0.54175	1.00000
Home Indoor Ozone Concentration (ppb)	<.0001	

So the partial correlation between personal ozone exposure and home ozone concentration, controlling both variables for the effect of time spent outdoors is 0.54.

Next, we control only personal ozone exposure (and not home ozone concentration) for the effect of time spent outdoors.

```
proc corr data=semipart nosimple noprob;
var home e_p;
run;
*****
Pearson Correlation Coefficients, N = 64
```

	home	e_p
home	1.00000	0.54173
Home Indoor Ozone Concentration (ppb)		
e_p	0.54173	1.00000

The partial and semi-partial correlation coefficients are virtually identical. The simple correlation between home ozone concentration and personal ozone exposure is also similar at 0.53, suggesting that the linear influence of time spent outdoors on personal ozone exposure and/or on home ozone concentrations has little effect on the correlation between personal ozone exposure and home ozone exposure.

Choosing the Proper Correlation Coefficient

Considering the proper correlation coefficient to describe the relationship among variables Y and X along with nuisance (or confounding) variables Z depends on the nature of their relationship. The table below can be used to determine which correlation is appropriate.

<i>Nuisance Relationship</i>	<i>Preferred Correlation</i>
Neither X nor Y affected by Z	r_{xy}
Both X and Y affected by Z	$r_{yx z}$
Only X affected by Z	$r_{y(x z)}$
Only Y affected by Z	$r_{x(y z)}$

Computing Partial Correlations

Consider the following correlation matrix for three variables X , Y , and Z .

$$\boldsymbol{R} = \begin{bmatrix} X & Y & Z \\ 1 & r_{xy} & r_{xz} \\ r_{yx} & 1 & r_{yz} \\ r_{zx} & r_{zy} & 1 \end{bmatrix} \quad \begin{matrix} X \\ Y \\ Z \end{matrix}$$

Note that $r_{yx} = r_{xy}$.

A **full partial correlation** may be computed as:

$$r_{yx|z} = \frac{r_{yx} - r_{yz}r_{xz}}{\sqrt{(1 - r_{yz}^2)(1 - r_{xz}^2)}}.$$

In the same setting, semi-partials may be computed as:

$$r_{y(x|z)} = \frac{r_{yx} - r_{yz}r_{xz}}{\sqrt{(1)(1 - r_{xz}^2)}}$$

$$r_{x(y|z)} = \frac{r_{yx} - r_{yz}r_{xz}}{\sqrt{(1 - r_{yz}^2)(1)}}.$$

The (1) term represents the variance of a standardized variable.

Summary: Partial Correlation Coefficient

- The partial correlation $r_{YX|Z_1, \dots, Z_q}$ measures the strength of the linear relationship between X and Y while controlling for \mathbf{Z} .
- The square of the partial correlation $r_{YX|Z_1, \dots, Z_q}$ measures the proportion of the sum of squares for error in a model containing only \mathbf{Z} that is accounted for by the addition of X to a regression model already containing \mathbf{Z} .
- The partial F statistic is used to test $H_0 : \rho_{YX|Z_1, \dots, Z_q} = 0$.
- The partial correlation $r_{YX|Z}$ can be defined as the correlation of the residuals of the straight-line regressions of Y on Z and of X on Z .
- Multiple partial correlations are tested using group-wise partial F tests.

Next: Assumption Diagnostics

Reading Assignment:

- Muller and Fetterman, Chapter 7: “GLM Assumption Diagnostics”
- Weisberg, Chapter 9: “Regression Diagnostics”

Lecture 10: Assumption Diagnostics

Reading Assignment:

- Muller and Fetterman, Chapter 7: “GLM Assumption Diagnostics”
(Required)

Wait, what are those assumptions again?

Homogeneity, Independence, Linearity, Existence, and Gaussian distribution.

Checking model assumptions focuses on analysis of residuals and outliers. Ethically, one must avoid allowing exploratory analysis (data snooping) to bias confirmatory analysis (inflate α). It is, however, acceptable to examine outliers or to transform variables to get the best test of the Gaussian distribution of residuals.

The First Step: Get to Know Your Data

Checking the Basics

- Determine the sampling unit (child, family, mouse, basketball team)
- Investigate the data collection procedure (counting number of vegetable servings consumed daily, measuring tumor size, measuring body mass index (BMI))
- Obtain units of measurement for all variables (number, cubic centimeters, kilograms per square meter)
- Obtain plausible range of values (0-10 servings, 0-1000 cubic centimeters, 17-40 kg per square meter)
- Obtain typical values for all variables (2 servings, 25 cubic cm, 25 kg per square meter)

Examining Individual Values

- Print the first 50 or so observations (PROC PRINT
DATA=OZONE (OBS=50);)
- Calculate the minimum and maximum values for all numeric
variables (PROC MEANS MIN MAX;)
- Generate frequency tables for character variables (PROC FREQ)

Summarizing the Data

- Report descriptive statistics (mean, median, quartiles, largest and
smallest few observations - PROC UNIVARIATE)
- Generate appropriate graphical displays (histograms, stem and leaf
plots, etc.using PROC UNIVARIATE)
- Look for patterns and unusual values

Check correlations, covariances, plot X_j versus Y , X_j versus $X_{j'}$

Example: Basic Diagnostics for the Ozone Data

The following SAS code is used to provide basic diagnostics.

```
proc print data=ozone (OBS=50);
run;

proc means min max data=ozone;
run;

proc univariate plot normal data=ozone;
var personal;
run;

proc capability data=ozone;
qqplot personal outdoor home time_out;
histogram time_out;
run;

proc corr noprob data=ozone;
var personal outdoor home time_out;
run;
```

The art in reporting these diagnostics lies in doing enough to provide necessary information with sufficiently few pages to be read (if you are

bored with your report, certainly everyone else is!). See Example 7.1 in the textbook for a nice illustration.

Selected diagnostics from the ozone data are presented below.

1. Basics: the sampling units in the ozone data are young children.
Personal ozone exposures were measured using a portable monitor attached to the child's clothing, outdoor ozone exposures were measured at EPA monitoring stations, indoor (home) exposures were measured using monitors in each child's home, and the proportion of time spent outdoors was reported by a parent.
Ozone exposures are measured in parts per billion. The plausible range of time spent outdoors would be between 0 and 1.
2. Individual values: in the list of data, one subject spent 90% of time outdoors, leading the biostatistician to make a note to ask the investigator if this variable is "proportion of play time spent outdoors" because 90% of one day would be 21.6 hours.
3. Data Summary: personal ozone exposures range from 0.46 to

67.18 ppb, with one outlier on the boxplot (67.18 ppb, which is not disturbing in a sample of 64 observations) and a mean and median in the range of 22-24 ppb. This variable does not look perfectly Gaussian, which is not disturbing as the Gaussian assumption applies not to Y unconditionally but to Y conditional on the covariates. (We should describe the covariates in a similar manner, noting that we need not assume they follow Gaussian distributions.)

4. Further details: personal ozone exposure appears to be most highly correlated with home exposure, and we do not see evidence of collinearity of covariates from the correlations.

Violations of Independence

The independence assumption applies to all statistical tests we have discussed thus far. The consequences of violating the independence assumption range from finding spurious relationships to missing significant ones.

Common violations of the independence assumption include

- repeated measures
 - weight loss of individuals is monitored over time
 - CD-4 counts of AIDS patients are recorded over the course of the disease
- nested data
 - schools are randomly assigned different physical education programs with student fitness as the outcome

-
- family data
 - researchers at NIEHS are studying sisters and daughters of women with breast cancer
 - pregnant rats are exposed to toxins and litters are studied for malformations

In general, independence must be evaluated by careful consideration of the study design. Warning signs include “too many” data points (often lab scientists run multiple assays on tissues from one specimen, for example) or “strange” results.

Residual Analysis

Recall that the raw residuals are estimates of errors for the model
 $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

The HILE Gauss assumptions refer to errors ($\boldsymbol{\varepsilon}$), which correspond to $Y|X$ and not Y unconditionally.

Thus to evaluate the assumptions we study the residuals $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}}$,
with $\hat{\varepsilon}_i = y_i - \hat{y}_i$.

Example: Residuals

Suppose we plan to sample undergraduate males at UNC and measure their lung function.

What might the density (histogram) of their lung function look like?

(What fraction smoke?) Y may be bimodal (if so, certainly not normal!). However, errors, if we include the amount smoked in our model, may be Gaussian.

Properties of Residuals

Recall that the raw residuals, which are differences between the data points and fitted values, follow a singular normal distribution

$$\hat{\boldsymbol{\varepsilon}} \sim \mathcal{SN}_n\{\mathbf{0}, \sigma^2[\mathbf{I} - \mathbf{H}]\}$$

and that we can estimate σ^2 using the residuals as follows:

$$\hat{\sigma}^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2 / (n - r).$$

The covariance among residuals, $\sigma^2[\mathbf{I} - \mathbf{H}]$, implies non-zero correlation and non-independence among residuals, though such correlations are modest in many settings.

If the model spans an intercept, then $\bar{\hat{\varepsilon}} = \sum_{i=1}^n \hat{\varepsilon}_i / n = 0$. In this case, it is clear that residuals are *not* independent!

Standardized Residuals



Because $\text{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_i)$, the residuals may have different variances because the variance of the residuals depends on the pattern of covariate values.

It would be easier to use the residuals to evaluate model assumptions if they all had the same variance. For this reason, we construct *standardized residuals*

$$r_i = \frac{\hat{\varepsilon}_i}{\sqrt{\hat{\sigma}^2(1 - h_i)}}.$$

The standardized residuals (also called internally studentized residuals) will follow a Gaussian distribution with mean 0 and variance 1 in large samples. Because of this property, one rule of thumb is to examine any observations with standardized residuals > 2 in absolute value in further detail (the justification is that 95% of the mass of the Gaussian distribution lies between -1.96 and 1.96, so that studentized residuals with absolute values greater than 2 represent observations in the

extreme tails of the distribution). Standardizing residuals thus helps us to assess their magnitude relative to the precision of the estimated regression analysis. Although these residuals are standardized (and thus have mean zero and variance one), they do depend on a variance estimate that may be affected by outliers. Because of this, we may miss those outliers in a residual analysis.

The jack-knifed or studentized residuals provide a solution to this problem based on more robust variance estimates. Recall that the studentized residuals (our preference) are defined as

$$r_{(-i)} = \frac{\widehat{\varepsilon}_i}{\sqrt{\widehat{\sigma}_{(-i)}^2(1 - h_i)}} \sim T(n - r - 1),$$

where the jack-knifed estimate of σ^2 is more robust to outliers than the usual estimate. These residuals are also called **externally studentized residuals**.

Note that a sample set of standardized or studentized residuals have approximate mean zero and approximate variance one. As $(n - r)$ becomes large, $T \rightarrow$ Gaussian. Studentized residuals can still be thought of in “standard deviation” units so that the usual cutoff of 2 is approximately valid.

Evaluating Assumptions with (or Without) Residuals



- Homogeneity: violations seen in the pattern of residuals.
- Independence: assessed through logic of sampling scheme.
- Linearity: examine pattern of residuals.
- Existence: (finite sample...).
- Gaussian distribution: distributional assessment involves box plot of residuals, histogram of residuals, and test of Gaussian distribution of residuals. (The discrepancy between T and Gaussian random variables somewhat inflates the probability of rejecting the null...why?)

Sometimes, Gaussian errors, homogeneity, and linearity come and go as a package.

Evaluating Heterogeneity and Linearity



Plotting $\{r_{(-i)}\}$ vs. $\{\hat{y}_i\}$ in an *R/P plot* provides the most useful diagnostic because predicted values capture all the information in predictors that is available as a linear combination of them. The R/P plot allows us to assess both heterogeneity and linearity. The R/P plot should resemble a rectangular cloud with no obvious trends or pattern. When examining an R/P plot, look for the following indicators of problems:

- any trend, such as a tendency for negative residuals for small predicted values and positive residuals for large predicted values, which may indicate non-linearity
- non-constant spread of the residuals, such as a tendency for tightly clustered residuals for small predicted values and widely dispersed residuals for large predicted values, which may indicate heteroscedasticity.

One issue in multiple regression is the detection of *multivariate* outliers. The R/P plot replaces a vector of predictors with one univariate predicted value. This is just one way to move from the multivariate predictor space to a univariate space.

A *probability plot* or *Q-Q plot* graphs the ordered residuals versus the expected order statistics of the standard normal distribution. To create a Q-Q plot, we order the residuals from smallest to largest, letting $r_{(1)}$ be the smallest, $r_{(i)}$ the i^{th} smallest, and $r_{(n)}$ be the largest. As you may recall from BIOS 550/660, $r_{(i)}$ is called an *order statistic*.

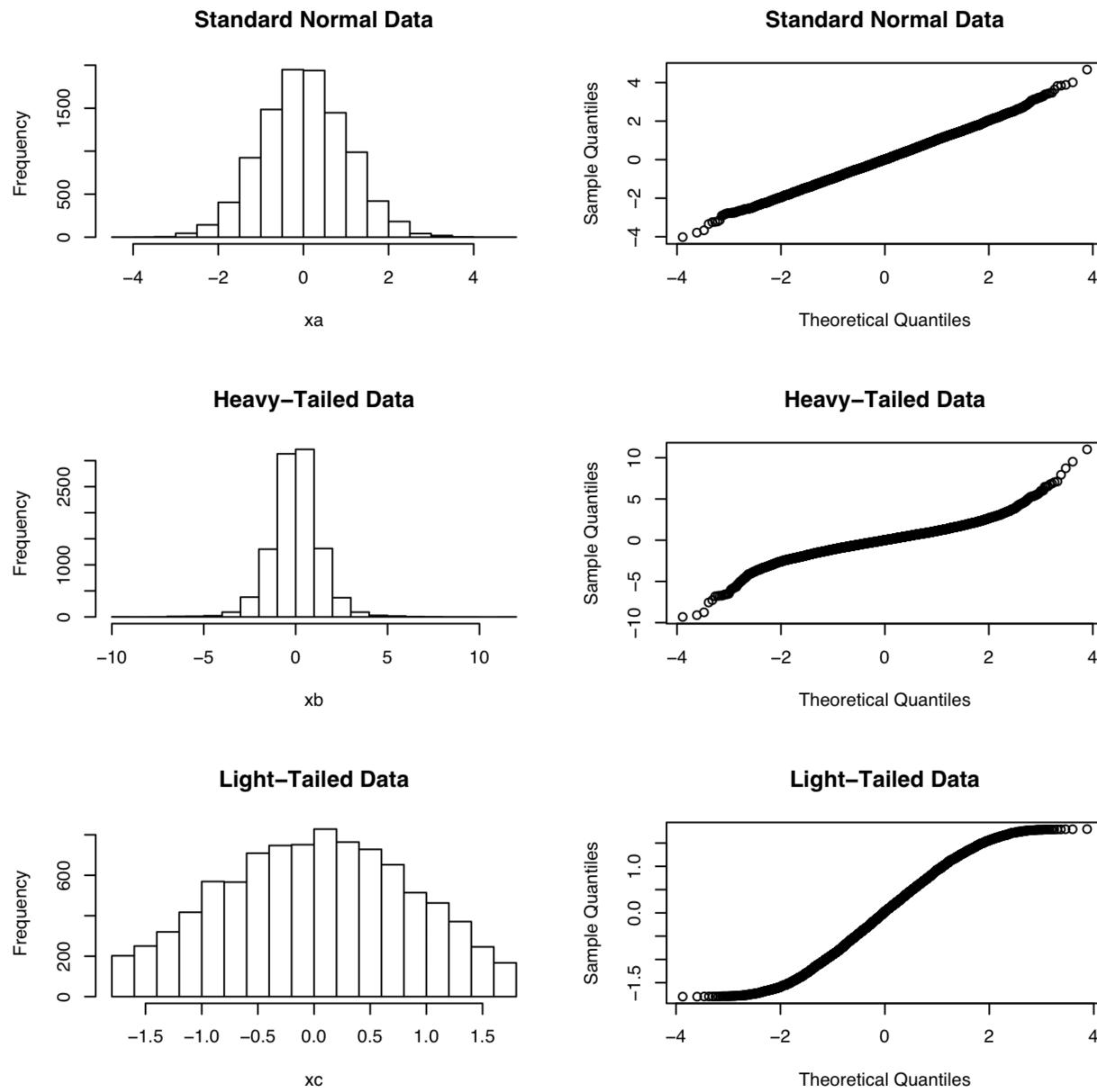
Next, imagine taking a sample of size n from a standard normal distribution and ordering it to obtain the order statistics $z_{(i)}$, $i = 1, \dots, n$. Using results from probability theory, we can obtain the expected values of these order statistics. (For example, we expect the median to be 0, the 5^{th} percentile to be -1.96, and the 95^{th} percentile to be 1.96.)

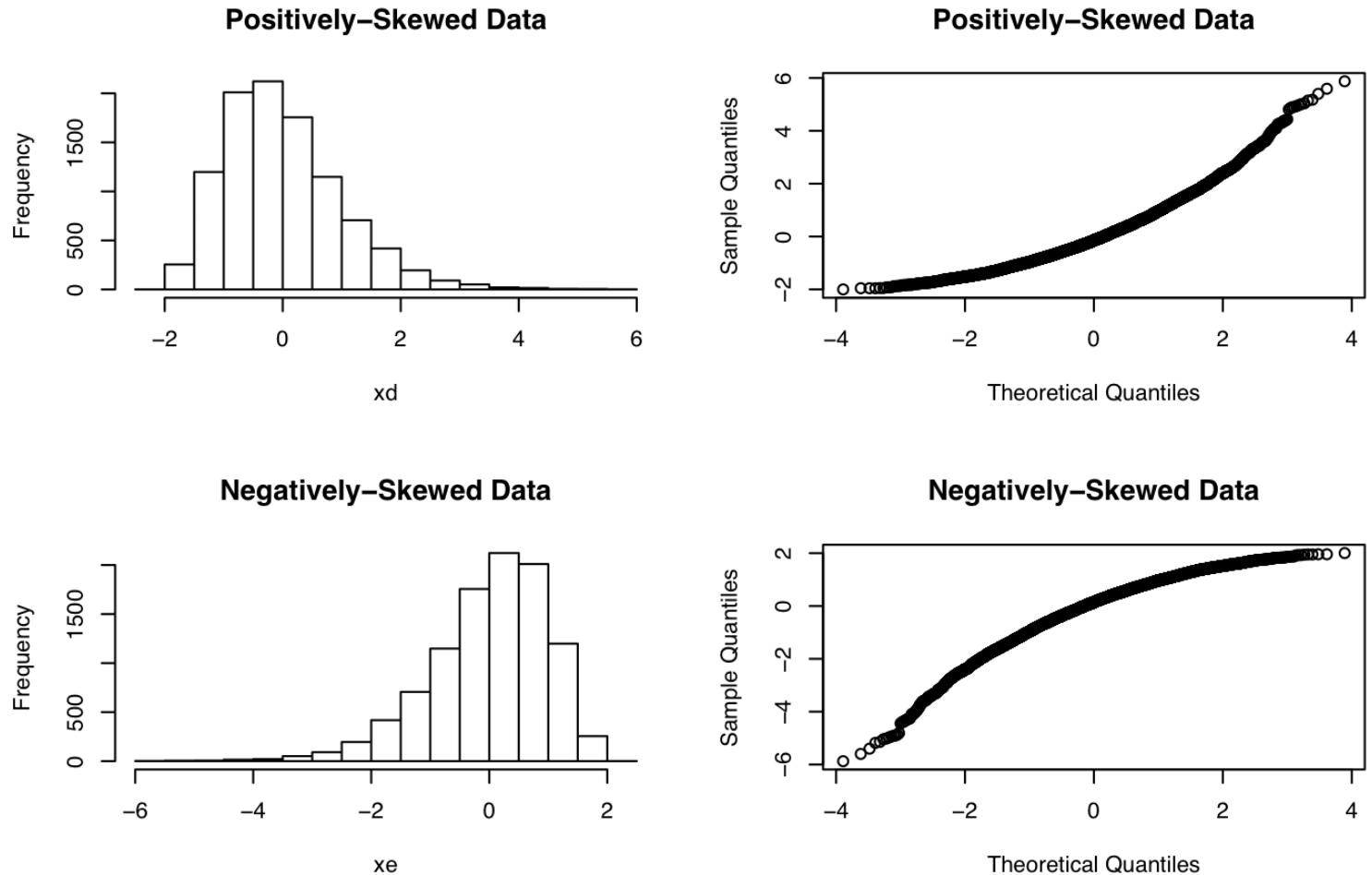
If our observations come from a Gaussian distribution, we expect our observed order statistics to be pretty close to the expected order statistics from the standard normal distribution. Specifically, if we plot the observed order statistics against the expected ones, we expect to

see a straight line through the origin with unit slope. When this is not the case, the observed order statistics deviate from what we would expect from a standard normal distribution, and the Gaussian errors assumption may be violated.

For example, if the residuals are from a distribution with heavy tails, we expect to see values consistently above a straight line through the origin with unit slope in the upper tail and consistently below the line in the lower tail. If the residuals are from a distribution with light tails, then we expect to see values consistently below the line in the upper tail and above the line in the lower tail.

A curve that is concave upwards indicates positively-skewed data, and a curve that is concave downwards indicates negatively-skewed data.





The following code creates an R/P plot as well as a Q-Q plot of the studentized residuals.

```
proc reg data=ozone;
model personal=outdoor home time_out;
output out=out rstudent=studresid predicted=predicted h=leverage;
run;

proc plot data=out;
plot studresid*predicted/vref=0;
run;

proc capability data=out;
qqplot studresid;
run;
```

Now we consider a variety of example R/P plots in order to see what types of plots are acceptable and what types of plots are cause for concern.

Figure 1: Acceptable residuals: simulated $x_i \sim \mathcal{N}(0, 1)$ independent of $\varepsilon_i \sim \mathcal{N}(0, 1)$ with true model $y_i = x_i + \varepsilon_i$ and fit valid model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

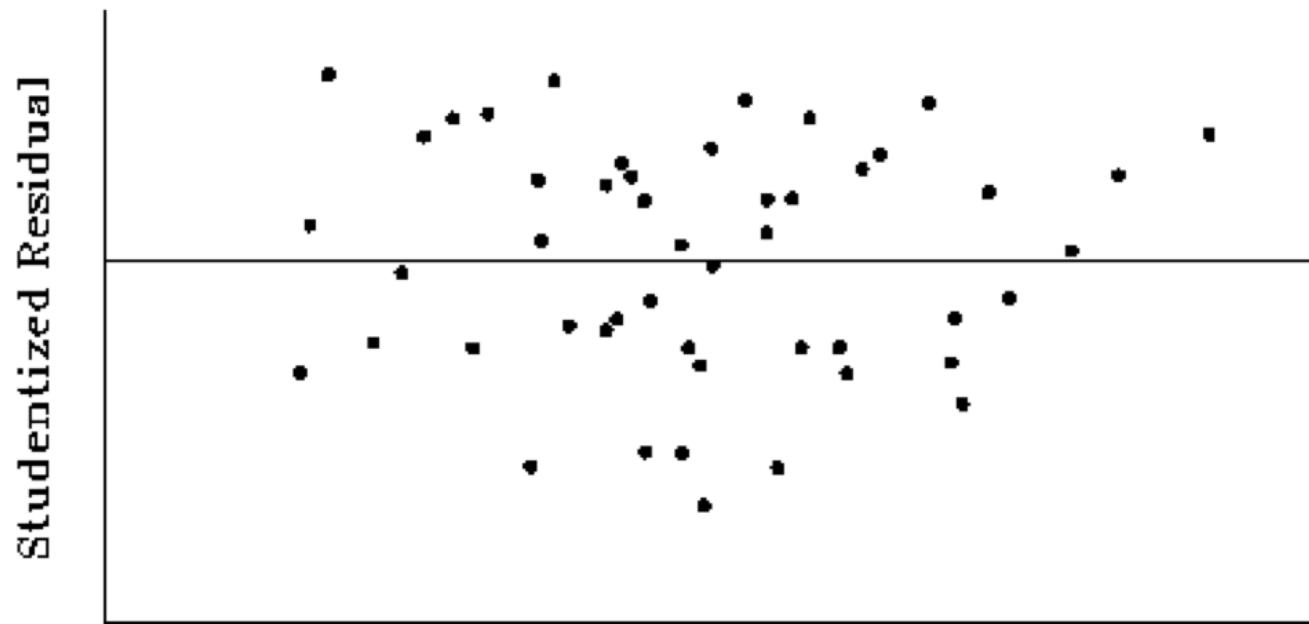


Figure 2: Unacceptable residuals (violation of homogeneity): simulated $x_i \sim U(0, 1)$ independent of $\varepsilon_i \sim \mathcal{N}(0, 1)$ with true model $y_i = x_i + x_i \varepsilon_i$ and fit invalid model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

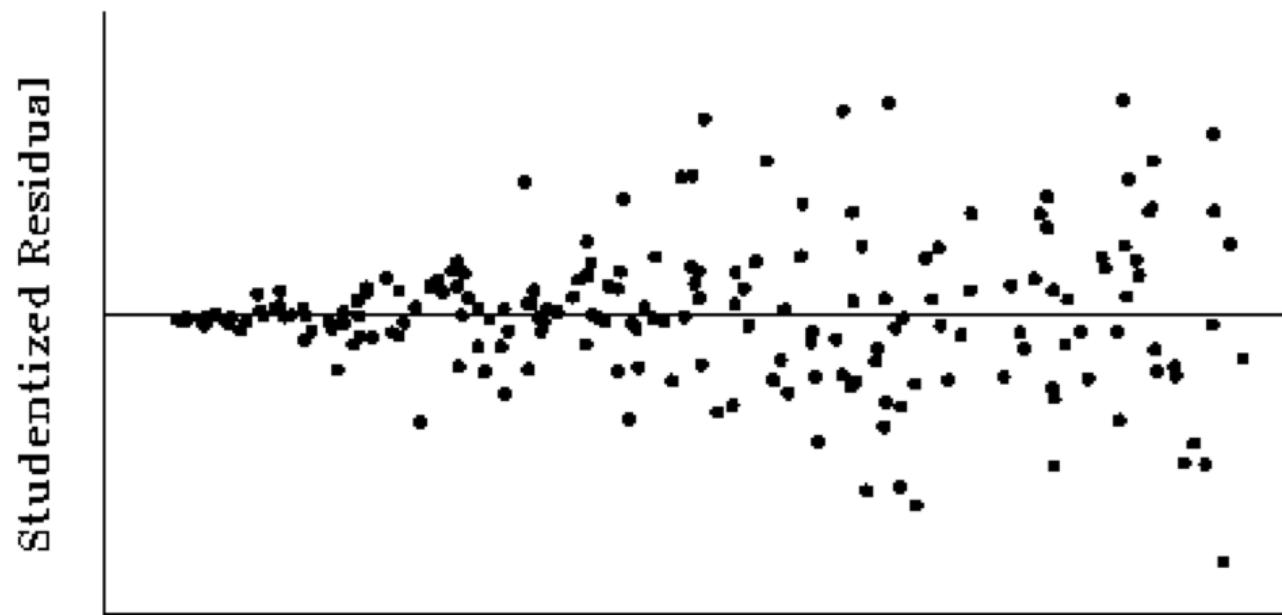


Figure 3: Unacceptable residuals (violation of homogeneity): simulated $x_i \sim U(0, 1)$ independent of $\varepsilon_i \sim \mathcal{N}(0, 1)$ with true model $y_i = x_i + (1 - x_i)\varepsilon_i$ and fit invalid model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

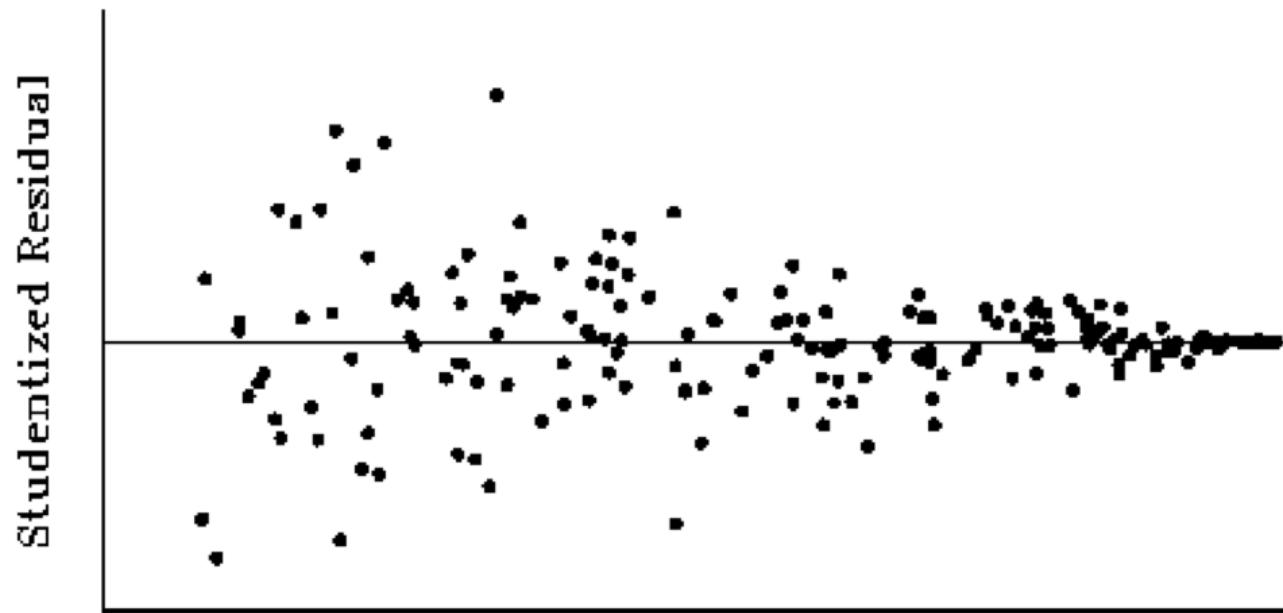


Figure 4: Unacceptable residuals (violation of homogeneity): simulated $x_i \sim U(0, 1)$ independent of $\varepsilon_i \sim \mathcal{N}(0, 1)$ with true model $y_i = x_i + 2 | 0.5 - x_i | \varepsilon_i$ and fit invalid model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

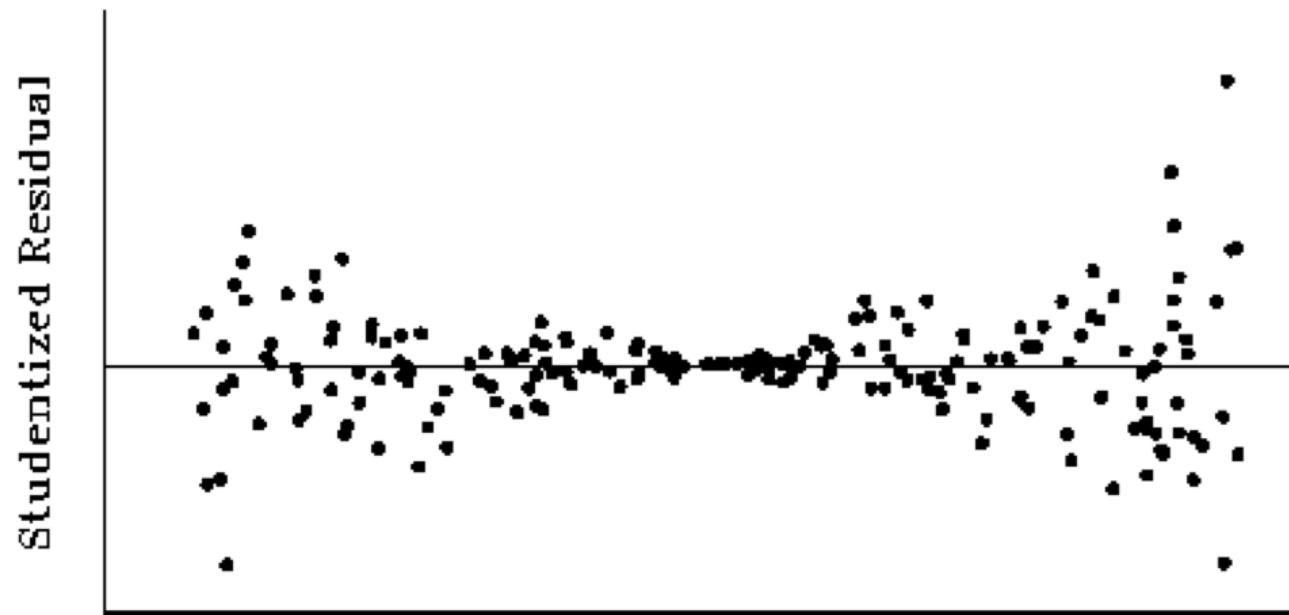


Figure 5: Unacceptable residuals (violation of homogeneity): simulated $x_i \sim U(0, 1)$ independent of $\varepsilon_i \sim \mathcal{N}(0, 1)$ with true model $y_i = x_i + (1 - |0.5 - x_i|)\varepsilon_i$ and fit invalid model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

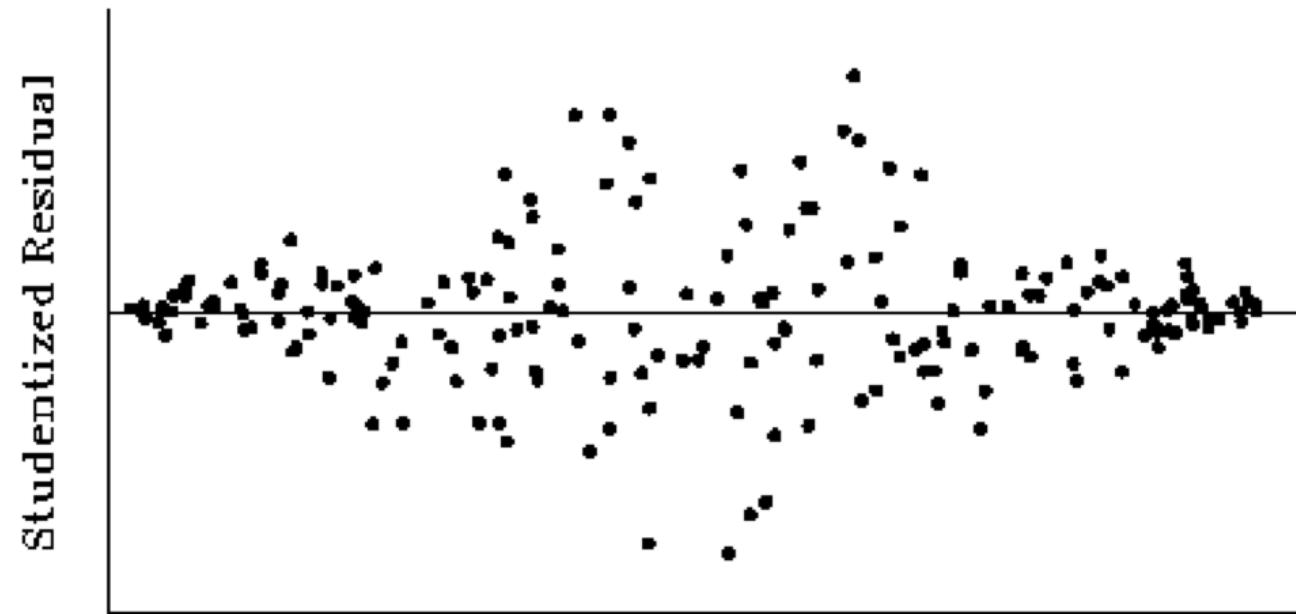
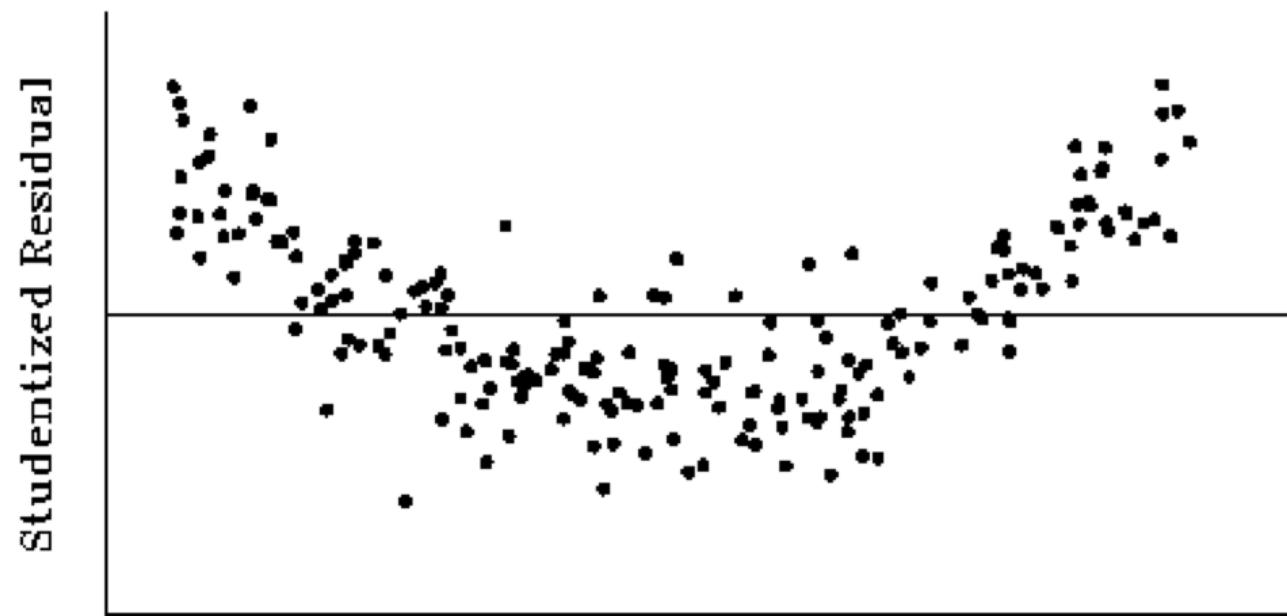


Figure 6: Unacceptable residuals (violation of linearity): simulated $x_i \sim U(0, 1)$ independent of $\varepsilon_i \sim \mathcal{N}(0, 1)$ with true model $y_i = x_i^2 + 0.05\varepsilon_i$ and fit invalid model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.



Some analysts plot the response versus each individual predictor, a tactic that ignores the impact of other predictors. A better idea is to create the *partial plot* of $\{y\}$ versus $\{x_j | \{x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_p\}\}$, the residuals from a regression with x_j as response and the other covariates as the predictors. The univariate correlation in the plot is the semi-partial correlation



$$r(y, x_j | \{x_1, x_2, \dots, x_{j-1}, x_{j+1}, \dots, x_p\}).$$

These plots are most useful for identifying culprits in known problems and not for detecting any new problems.

Evaluating Gaussian Distribution

SAS PROC UNIVARIATE provides the following information:

- a Q-Q (“normal”) plot (use PROC CAPABILITY to get a nicer one),
- a box and whisker plot,
- a stem and leaf plot (a very basic histogram),
- moments (mean, variance, skewness, kurtosis),
- the largest and smallest 5 values with ID’s, and
- quartiles.

Example: Ozone Data Residuals

The following code is used to evaluate the residuals for the ozone data from the model

$$O_{PERSONAL} = \beta_0 + \beta_1 O_{OUTDOOR} + \beta_2 O_{HOME} + \beta_3 TIME_{OUT} + \epsilon.$$

```
proc reg data=ozone;
```

```
model personal=outdoor home time_out;
output out=out rstudent=studresid predicted=predicted;
run;

proc univariate plot normal data=out;
id subject;
var studresid;

proc capability data=out;
qqplot studresid;
run;

proc capability data=out;
histogram studresid/kernel;
run;
```

```
*****
```

```
The UNIVARIATE Procedure
```

```
Variable: studresid (Studentized Residual without Current Obs)
```

```
Moments
```

N	64	Sum Weights	64
Mean	0.00618379	Sum Observations	0.39576231
Std Deviation	1.0298702	Variance	1.06063263
Skewness	0.90708362	Kurtosis	0.893602

Uncorrected SS	66.8223027	Corrected SS	66.8198554
Coeff Variation	16654.3631	Std Error Mean	0.12873377

Basic Statistical Measures

	Location	Variability	
Mean	0.00618	Std Deviation	1.02987
Median	-0.33120	Variance	1.06063
Mode	.	Range	5.12982
		Interquartile Range	1.02182

Tests for Location: Mu0=0

Test	-Statistic-	-----p Value-----
Student's t	t 0.048035	Pr > t 0.9618
Sign	M -4	Pr >= M 0.3817
Signed Rank	S -133	Pr >= S 0.3779

Tests for Normality

Test	--Statistic---	-----p Value-----
Shapiro-Wilk	W 0.935601	Pr < W 0.0024
Kolmogorov-Smirnov	D 0.148292	Pr > D <0.0100

Cramer-von Mises	W-Sq	0.279391	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.565583	Pr > A-Sq	<0.0050

We see that all the normality tests agree that the residuals do not appear to be normal. (P-value shopping is not advised – it is best to pick one test before doing the analysis.) The differences among the tests are described below. For all tests, the null hypothesis is H_0 : the data are normally distributed.

- Shapiro-Wilks: roughly, a measure of the straightness of a Q-Q plot (appropriate for small sample sizes)
- Kolmogorov-Smirnov: largest discrepancy between the empirical cdf and the estimated hypothesized one (based on observed mean and variance)
- Cramer-von Mises: considers squared difference between empirical and estimated cdf
- Anderson-Darling: considers a weighted squared difference between empirical and estimated cdf

We can look at the quantiles, boxplot, and histogram to check whether residuals are skewed.

Quantiles (Definition 5)

Quantile	Estimate
100% Max	3.054968
99%	3.054968
95%	2.155071
90%	1.483757
75% Q3	0.388099
50% Median	-0.331204
25% Q1	-0.633717
10%	-0.907937
5%	-1.151776
1%	-2.074857
0% Min	-2.074857

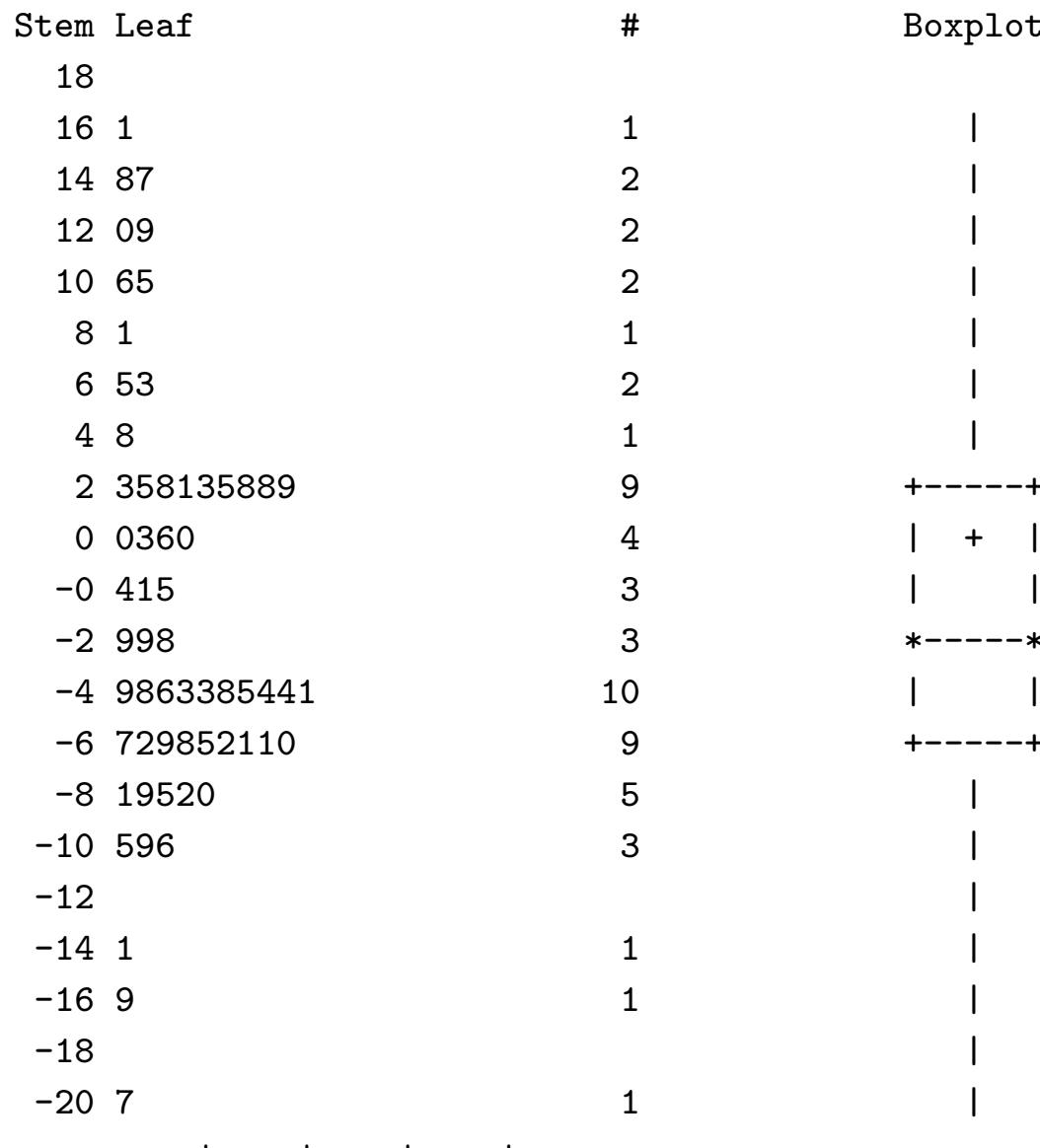
We see some evidence of skewness in the quantiles.

In addition, we can look at the most extreme residuals to gauge whether any outliers deserve further attention.

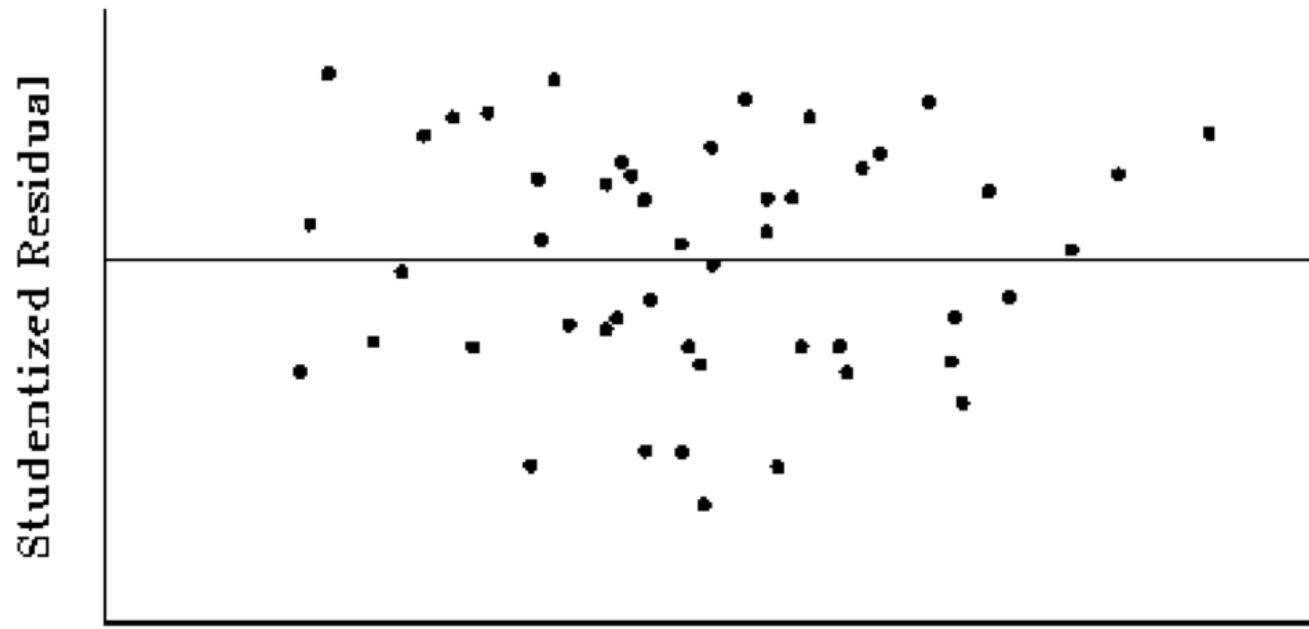
Extreme Observations

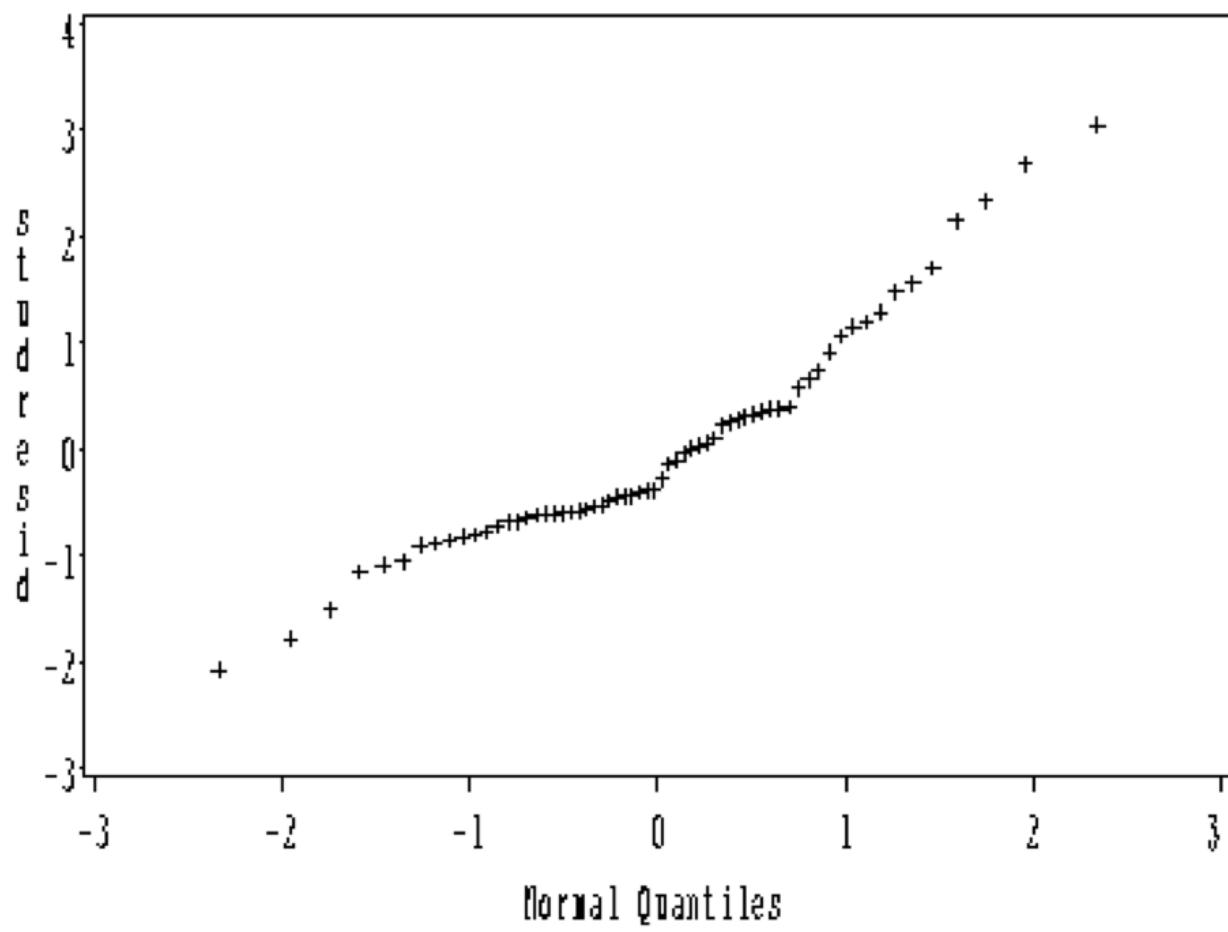
-----Lowest-----			-----Highest-----		
Value	subject	Obs	Value	subject	Obs
-2.07486	25	25	1.70690	49	49
-1.78561	36	36	2.15507	64	64
-1.51369	4	4	2.33176	17	17
-1.15178	2	2	2.67945	39	39
-1.08553	20	20	3.05497	26	26

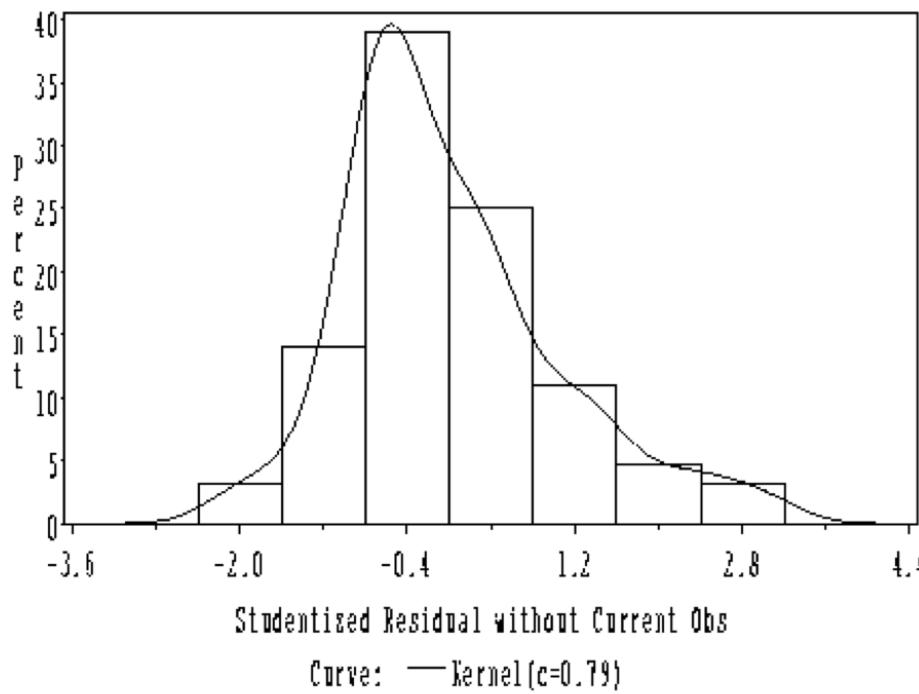
There do seem to be some outliers on the right.



Multiply Stem.Leaf by 10**-1







From this output, we see that the data appear to be skewed to the right a little bit, and there is some peakedness in the distribution. Again, we see several outliers in the right tail of the distribution. A *kernel density estimate* is superimposed on the histogram in PROC CAPABILITY in order to facilitate interpretation.

Evaluating Extreme Residuals



One residual must be the most extreme, but too many extreme residuals (or too extreme residuals) are evidence of poor model fit.

Testing residuals (the null hypothesis is $H_0: E[r_{(-i)}] = 0$) leads to a

multiple comparisons issue: we do n tests (one for each residual), so

- we should use a Bonferroni correction (use α/n instead of α and the $\alpha/(2n)$ critical value for T).



Any positive correlation among residuals makes the correction conservative.

If $(n - r) \rightarrow \infty$ then $\sigma^2(\mathbf{I} - \mathbf{H}) \rightarrow \sigma^2 \mathbf{I}$, implying that correlations among residuals converge to zero for large samples.

Example: Extreme Residuals in Ozone Data

Based on previous analyses, do any residuals appear to be extreme? If so, which ones? What do you conclude based on a test of the hypothesis that any extreme residuals come from a population with mean 0?

Outliers

An *outlier* is a value (of a predictor or a response) much larger in absolute value than next nearest value. On a box plot, outliers are often marked using 0 or *. 

Least squares estimation is rather sensitive to outliers.

An outlier may be an anomaly or merely a chance event (e.g., a child reports 27 vegetable servings per day, or height is reported as 60 feet rather than 60 inches).

Automatically discarding extreme observations is WRONG! Exceptions: verifiable instrument malfunction, recording error.

Consider whether an outlier is plausible, implausible, or impossible.

- A woman weighs 250 pounds, 400 pounds, or 1000 pounds...
- Body temperature of 36.8 (centigrade or Fahrenheit? average

person, heart surgery patient, or crocodile?)

Extreme residuals may indicate an anomaly, while extreme values of Y or X_j need not be important.

Leverage

A *leverage* value depends only on \mathbf{X} and measures how extreme the i th observation is in terms of the predictor space. An observation with high leverage has the potential to have great influence on the model fit. 

The hat matrix, $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, has i th diagonal element h_i , the *leverage* for the i th observation.

$$h_i = \mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_i, \text{with } \mathbf{X}_i \text{ } 1 \times p, \text{ the } i\text{th row of } \mathbf{X}$$

For $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, with $s_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / n$,


$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{ns_x^2} = \left[1 + \frac{(x_i - \bar{x})^2}{s_x^2} \right] / n.$$
  

With an intercept and all $p - 1$ predictors mean zero and uncorrelated,

$$h_i = \frac{1}{n} + \sum_{j=1}^{p-1} \frac{x_{ij}^2}{ns_{x_j}^2} = \left[1 + \sum_{j=1}^{p-1} \frac{x_{ij}^2}{s_{x_j}^2} \right] / n.$$

 We can prove $\sum_{i=1}^n h_i = r$, the rank of \mathbf{X} , so full rank \mathbf{X} implies $\bar{h} = \sum_{i=1}^n h_i / n = p/n$.

A handy rule of thumb is to examine further any subjects with $h_i > \frac{2p}{n}$, which is leverage greater than twice the average value.

With an intercept and $(p - 1)$ multivariate Gaussian predictors (NOT part of HILE Gauss!) with each row i.i.d., we see that a function of

the leverage values follow an F distribution,

$$F_i = \frac{(h_i - 1/n)/(p-1)}{(1-h_i)/(n-p)} = F(p-1, n-p).$$



Multiple testing leads to using a Bonferroni correction and using α/n .

This F statistic tests the hypothesis that the observation in question comes from a multivariate Gaussian distribution with the same mean as the other observations.

If the model does not span an intercept, then

$$F_i = \frac{h_i/p}{(1-h_i)/(n-p)} \sim F(p, n-p).$$

Example: Leverage Values for Ozone Data

The following code is used to print and test leverage values.

```
proc reg data=ozone;
model personal=outdoor home time_out;
output out=out rstudent=studresid predicted=predicted h=leverage;
run;
proc sort data=out;
by descending leverage;
run;
data out_1;
set out (obs=10);
p=4; n=64;
F=((leverage-(1/n)/(p-1))/((1-leverage)/(n-p)));
pvalue=1-probf(F,p-1,n-p);
if pvalue <=0.05/n then BONF="*";
else BONF=" "; *Bonferroni correction;
label Bonf="Signif at 0.05/n?";
run;
proc print data=out_1 uniform label noobs;
var subject personal outdoor home time_out leverage F pvalue Bonf;
run;
```

Personal

Home Indoor

Proportion

subject	Ozone Exposure (ppb)	Outdoor Ozone Concentration (ppb)	Ozone Concentration (ppb)	of Time Spent Outdoors
9	28.55	92.563	7.14	0.26
13	38.28	30.435	35.38	0.69
2	14.63	43.792	13.97	0.90
12	37.58	104.100	45.10	0.42
35	11.81	20.649	5.71	0.78
7	31.97	88.863	44.12	0.05
42	61.19	54.318	46.04	0.45
17	53.18	70.003	11.13	0.65
8	32.45	81.916	45.74	0.33
30	13.94	71.878	9.51	0.38

Leverage	F	pvalue	Signif at 0.05/n?
0.21178	15.7246	.000000117	*
0.15601	10.7210	.000009731	*
0.15214	10.3980	.000013240	*
0.14355	9.6918	.000026235	*
0.13310	8.8519	.000060328	*
0.12067	7.8783	.000162823	*

0.11869	7.7259	.000190717	*
0.11679	7.5798	.000222079	*
0.09578	6.0096	.001193472	
0.09544	5.9852	.001225971	

How do you interpret the results?

Mahalanobis Distance

Assume the model spans an intercept and let \mathbf{X}^* indicate the $p - 1$ predictors other than the intercept (removing the intercept column from the $\mathbf{X} = [\mathbf{J}_n \quad \mathbf{X}^*]$ matrix).

Define \mathbf{C}^* as the $(p - 1) \times (p - 1)$ sample covariance matrix for the predictors in \mathbf{X}^* :

$$\mathbf{C}^* = \frac{\mathbf{X}^{*\prime} \left(\mathbf{I}_n - \frac{\mathbf{J}_n \mathbf{J}'_n}{n} \right) \mathbf{X}^*}{n}.$$

(Note that the sample variances of the predictors lie on the diagonal of \mathbf{C}^* .)

Compute the unbiased estimate of population covariances as

$$\widehat{\boldsymbol{\Sigma}}^* = \left(\frac{n}{n-1} \right) \mathbf{C}^*.$$

The *Malahanobis distance*



$$m_i = (\mathbf{X}^{*'}_i - \bar{\mathbf{x}}^*)' \widehat{\boldsymbol{\Sigma}}^{-1} (\mathbf{X}^{*'}_i - \bar{\mathbf{x}}^*)$$

is the deviation of one observation's predictors from the center of the predictor space.

If all predictors are uncorrelated (in the sample at hand), then

$$m_i = \sum_{j=1}^{p-1} \frac{(x_{ij}^* - \bar{x}_j^*)^2}{\widehat{\sigma}_{x_j^*}^2} .$$

The Essential Equivalence of Leverage and Mahalanobis Distance

Both leverage and Mahalanobis distance measure the “extremeness” of an observation in \mathbf{X} space.

For any model that spans an intercept,

$$h_i = n^{-1} + (n - 1)^{-1}m_i.$$

Thus only leverage or Mahalanobis distance need be examined (leverage is more commonly used).



Influence: Cook's Distance



Cook proposed an influence measure based on the extent to which parameter estimates would change if we had deleted the i^{th} observation from the sample. *Cook's distance* assesses the impact of deleting one observation from the sample.

Why is this a sensible diagnostic?

Cook's statistic measures the standardized shift in predicted values and the shift in $\hat{\beta}$ due to deleting the i th observation:

$$\begin{aligned} D_i &= \frac{(\hat{\mathbf{y}}_{(-i)} - \hat{\mathbf{y}})'(\hat{\mathbf{y}}_{(-i)} - \hat{\mathbf{y}})}{p\hat{\sigma}^2} \\ &= \frac{(\hat{\boldsymbol{\beta}}_{(-i)} - \hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\hat{\boldsymbol{\beta}}_{(-i)} - \hat{\boldsymbol{\beta}})}{p\hat{\sigma}^2} \\ &= r_i^2 \cdot \frac{h_i}{p(1 - h_i)} \end{aligned}$$

where r_i is standardized residual. As a quadratic form, $D_i \geq 0$. Distributional results are not straightforward, and there is no perfect rule of thumb for evaluating any particular D_i . One proposed rule of thumb is that values of D_i close to 1 are indicative of excessive influence.

DFBETAS and DFFITS

Two other common measures of influence are DFBETAS and DFFITS.

The *DFFITS* statistic is a scaled measure of the change in the predicted value of the i^{th} observation if that observation were omitted from the analysis. That is,

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(-i)}}{\sqrt{\hat{\sigma}_{(-i)}^2} h_i}.$$

A general rule of thumb is to examine further any observations with $DFFITS_i > 2\sqrt{\frac{p}{n}}$.

The *DFBETAS* statistic is a normalized measure of the effect of observations on the estimated regression coefficients. There are multiple DFBETAS statistics for each subject (one for each regression

coefficient). They are computed as


$$DFBETAS_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j(-i)}}{\sqrt{\hat{\sigma}_{(-i)}^2 (\mathbf{X}'\mathbf{X})_j^{-1}}}.$$

One typically examines observations with $DFBETAS_{ij} > \frac{2}{\sqrt{n}}$.

Concluding Comments

“One should be cautioned that deleting the most deviant observations will in all cases slightly improve, and sometimes substantially improve, the fit of the model. One must be careful not to data snoop simply in order to polish the fit of the model by discarding troublesome data points.” (KKM, 1988, p201)

KKM: Kleinbaum, Kupper, Muller, and Nizam, *Applied Regression Analysis and Other Multivariable Methods*

One *must* report any deletions, and observations should be deleted only in the most extreme circumstances.

Leverage evaluates design and extremeness in predictor space, while residuals evaluate model adequacy. Cook's D_i values evaluate influence (both of above).

It may be argued that a measure of adequacy of a sample is that no

observation is influential, even though it may be high in leverage and have an extreme residual.

Use $\{h_i\}$, $\{m_i\}$, $\{r_{(-i)}^2\}$, $\{\hat{\beta}_{(-i)}\}$, and $\{\hat{\sigma}_{(-i)}^2\}$ to define “bad data,” not y and X .

Next: Computation Diagnostics

Reading Assignment:

- Muller and Fetterman, Chapter 8: “GLM Computation Diagnostics”

Lecture 11: Computation Diagnostics

Reading Assignment:

- Muller and Fetterman, Chapter 8: “GLM Computation Diagnostics” (Required)

Naive acceptance of computer output may lead to major errors, even though all assumptions prove valid.

The culprit is finite precision arithmetic, which is typically 7-15 decimal digits of accuracy for a single computation. Regression computations accumulate sums and cross-products so that numerical inaccuracy accumulates.

In the ideal situation, you are interested in a specific number of predictors and know the model beforehand. In many observational studies, investigators have a large number of measurements of predictors or exposures and want to evaluate them all in a model (e.g., positive anxiety (promotion at work, engagement), negative anxiety (fired, illness of family member), impact of positive anxiety, impact of negative anxiety, and potential interactions of these with two stress hormones!).

Including a lot of variables in a model often results in multicollinearity, a high degree of correlation among several predictors. This happens when too many variables have been put into the model, and a number of these variables measure similar phenomena. Collinearity does not

-  cause a violation of the HILE Gauss assumptions. However, it
 - 1. tends to inflate the variances of predicted values, and
 - 2. tends to inflate the variances of parameter estimates (bad!).

Single Variable Problems and Solutions

Disparity in location and scale (mean and variance) can create substantial inaccuracy. Examine summary statistics (mean, variance, minimum, and maximum) to find potential location or scale problems. The easiest remedy is to scale predictors to have range $\in \{-10, +10\}$, a common scientific practice.

Examples of location or scale problems:

- Consider children's birth year, ranging from 1980 to 1990. This creates a location problem (and *collinearity* with the intercept), but replacing birth year with age avoids the location problem.
- Consider children's spirometry performance measured by forced vital capacity (FVC, the maximum amount of air that can be forcibly expired). Recording values in mL may give a range of $\{1000, 4000\}$ mL, leading to very large elements in $\mathbf{X}'\mathbf{X}$. Using liters (dividing FVC by 1000) eliminates the scaling problem.

Reducing location and scale disparities never hurts and may substantially improve accuracy. In extreme cases, creating means of zero (centering) and sums of squares or variances of one (normalizing) may be necessary, though using convenient center points and scales often suffices. For example, use human body temperatures as $T - 37$ (C).

Collinearity

Predictors in a model are *collinear* whenever the columns of \mathbf{X} contain some amount of redundancy.

Mathematically, collinearity corresponds to linear dependence among columns of \mathbf{X} . Collinearity exists along a continuum (from nonexistent to moderate to severe to disastrous). Often, collinearity involves more than two variables.

Classic Example: Socioeconomic Status

Outcome: anything

Predictors: gender, race, age, education, socioeconomic status

Collinearity diagnostics identify variables containing little or no information above and beyond that in the other predictors.

Recall that the # of linearly independent columns in \mathbf{X} equals $r = \text{rank}(\mathbf{X})$, and a full rank model has $\text{rank}(\mathbf{X}) = r = p$. 

Some ANOVA (Chap. 12) designs use purposefully less than full rank \mathbf{X} , but for now the focus centers on unintentionally less than full rank designs (i.e., problems).

Uncontrolled collinearity creates computational difficulties and confuses testing and interpretation.

Loosely, consider

- Rank: the number of distinct dimensions of information.
- Eigenanalysis: creating sets of regression coefficients (eigenvectors) needed to produce new variables which are uncorrelated, with successively maximum variances (eigenvalues). 

Consider the columns of \mathbf{X} as a collection of vectors:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_0 & \mathbf{X}_1 & \dots & \mathbf{X}_{p-1} \end{bmatrix}.$$

Here $\mathbf{X}_0 = \mathbf{J}_n$ whenever the model contains an intercept.

Recall that the columns of a matrix \mathbf{X} are linearly dependent if there exists a vector $\boldsymbol{\delta} \neq 0$ such that $\mathbf{X}\boldsymbol{\delta} = \mathbf{0}$. So if $\exists \{\delta_0, \delta_1, \dots, \delta_{p-1}\}$, not all zero, such that $\sum_{j=0}^{p-1} \delta_j \mathbf{X}_j = \mathbf{0}$, then the columns of \mathbf{X} are linearly dependent. Otherwise, the columns of \mathbf{X} are linearly independent. If the columns of \mathbf{X} are linearly independent, then \mathbf{X} is full rank.

Also recall that an inner product matrix, $\mathbf{X}'\mathbf{X}$, has non-negative eigenvalues.

The # of non-zero (strictly positive) eigenvalues of $\mathbf{X}'\mathbf{X}$ equals r , and if $r < p$ (\mathbf{X} less than full rank), then \mathbf{X} contains an exact collinearity.

Only predictor properties determine collinearity. 

Exact collinearity rarely occurs, but too much collinearity causes a loss

of numerical, statistical, and scientific accuracy.

Matrices Providing Information about Collinearity

Difficulties with $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ arise from computing cross-products. Four versions of the inner product matrix $\mathbf{X}'\mathbf{X}$ provide information about collinearity and accuracy.

SSCP Matrix

The $p \times p$ *sums-of-squares and cross-products (SSCP) matrix* is given by

$$\mathbf{X}'\mathbf{X} = \left\{ \sum_{i=1}^n x_{ij}x_{ij'} \right\}.$$

Dividing by n yields $\mathbf{X}'\mathbf{X}/n$, the average cross-products matrix.

Scaled SSCP Matrix



Consider the $p \times p$ matrix

$$\text{D}_s = \begin{bmatrix} \sum x_{i0}^2 & & & 0 \\ & \sum x_{i1}^2 & & \\ & & \ddots & \\ 0 & & & \sum x_{i,p-1}^2 \end{bmatrix},$$

which contains the diagonal elements of the SSCP matrix.

Define the $p \times p$ scaled SSCP matrix of average scaled (but not centered) cross-products as

$$(\mathbf{X}'\mathbf{X})_s = \mathbf{D}_s^{-0.5} (\mathbf{X}'\mathbf{X}) \mathbf{D}_s^{-0.5}$$

This matrix has 1's on the diagonal. Examining the eigenvalues and eigenvectors of the scaled SSCP matrix is a good way to detect collinearity with the intercept.

Covariance Matrix

Partition the covariate matrix \mathbf{X} into $\begin{bmatrix} \mathbf{J}_n & \mathbf{X}^* \end{bmatrix}$, where \mathbf{X}^* is the $n \times (p - 1)$ matrix excluding the intercept.

The $(p - 1) \times (p - 1)$ covariance matrix adjusts for predictor means.

Recall that the column means of \mathbf{X}^* are defined by

$$\overline{\mathbf{X}^*} = \mathbf{X}^{*\prime} \mathbf{J}_n / n = \left\{ \sum_{i=1}^n x_{ij}^* / n \right\} .$$

We define the covariance matrix \mathbf{C} by

$$\begin{aligned}
 \mathbf{C} = \{c_{jj'}\} &= \left\{ \sum_{i=1}^n (x_{ij}^* - \bar{x}_j^*)(x_{ij'}^* - \bar{x}_{j'}^*) \right\} / n \\
 &= \mathbf{X}^{*\prime} (\mathbf{I}_n - \mathbf{J}_n \mathbf{J}_n' / n) \mathbf{X}^* / n \\
 &= \mathbf{X}^{*\prime} \mathbf{X}^* / n - \overline{\mathbf{X}^*} \overline{\mathbf{X}^*}' \\
 &= \mathbf{X}_c^{*\prime} \mathbf{X}_c^* / n,
 \end{aligned}$$

where $\mathbf{X}_c^* = \{(x_{ij}^* - \bar{x}_j^*)\}$ contains the $p - 1$ columns of centered data. Note that the diagonal of \mathbf{C} contains sample variances of the predictors.

The covariance matrix contains average centered cross-products and excludes the intercept. It is also an inner product matrix as shown above.



If \mathbf{X} is full rank, then \mathbf{C} has rank $p - 1$.

Correlation Matrix

Extract the diagonal elements of the covariance matrix \mathbf{C} :

$$\mathbf{D}_c = \begin{bmatrix} c_{11} & & & 0 \\ & c_{22} & & \\ & & \ddots & \\ 0 & & & c_{p-1,p-1} \end{bmatrix}.$$

Define

$$\begin{aligned} \mathbf{R} &= \mathbf{D}_c^{-0.5} \mathbf{C} \mathbf{D}_c^{-0.5} \\ &= \mathbf{D}_c^{-0.5} \mathbf{X}_{c'}^* \mathbf{X}_c^* \mathbf{D}_c^{-0.5} / n \\ &= \{r_{jj'}\} = \left\{ \frac{c_{jj'}}{\sqrt{c_j c_{j'}}} \right\} = \left\{ \frac{c_{jj'}}{s_j s_{j'}} \right\}. \end{aligned} \quad \square$$

Note that $c_{jj} = s_j^2$, the sample variance.

\mathbf{R} contains average centered and scaled cross-products and is an inner

product matrix.

Also, $\mathbf{X}^* {}_c \mathbf{D}_c^{-0.5} = \mathbf{Z} = \{(x_{ij}^* - \bar{x}_j^*)/s_j\}$ equals centered and scaled data.

Examining the eigenvalues and eigenvectors of the correlation matrix is a good way to detect collinearity among predictors other than the intercept.

Eigenanalysis!

An eigenanalysis is the best way to detect and describe collinearity.

Recall that the eigenvalues of the square matrix A are the roots the characteristic equation:

$$|A - \lambda I| = 0.$$

Equivalently, write $AX = \lambda X$.

Also recall that the rank of a (square) matrix equals the # of non-zero eigenvalues, so a matrix is full rank if and only if the matrix has no zero eigenvalues.

For the special case of a symmetric A , we define the spectral decomposition:

$$A = V \Lambda V'$$



where $\Lambda = \text{diag}(\boldsymbol{\lambda})$ represents the diagonal matrix of eigenvalues, ordered so that $\lambda_1 \geq \lambda_2 \geq \dots \lambda_p$. The k th column of \mathbf{V} has an eigenvector corresponding to k th eigenvalue, with the eigenvectors and eigenvalues in the same order.

\mathbf{V} is a full rank orthogonal matrix, regardless of the rank of \mathbf{A} .

Since we usually scale eigenvectors to unit length, $\mathbf{V}\mathbf{V}' = \mathbf{V}'\mathbf{V} = \mathbf{I}$, $\mathbf{V}^{-1} = \mathbf{V}'$ and $(\mathbf{V}')^{-1} = \mathbf{V}$.

We will consider eigenanalysis of inner-product matrices like $\mathbf{X}'\mathbf{X}$, \mathbf{C} , and \mathbf{R} , which are all symmetric and positive semi-definite (so that all eigenvalues are ≥ 0). The rank of these matrices equals the number of nonzero eigenvalues, and the number of zero eigenvalues equals the number of linear dependences. Remember that the rank of \mathbf{X} is the same as the rank of $\mathbf{X}'\mathbf{X}$.

The eigenvalues of $\mathbf{X}'\mathbf{X}$, \mathbf{C} , and \mathbf{R} are variances of linear

combinations of X_j 's, and eigenvectors provide regression coefficients to create new variables (variates) having the eigenvalues as variances. The relative size of eigenvalues, especially largest to smallest, indicates the amount of collinearity, while the eigenvectors allow discovery of which variables overlap.

Troubles often reduce to a scaling problem or a variable with s^2 near zero.

Example: Collinearity

Consider records from one night at a hospital emergency room.

If 39 of 40 patients are male, then the indicator of gender (0 for female, 1 for male) is highly collinear with the intercept.

Consider the role of eigenvalues in determining the stability of $(\mathbf{X}'\mathbf{X})^{-1}$. Using the spectral decomposition,

$$\mathbf{X}'\mathbf{X} = \mathbf{V}\text{Diag}(\boldsymbol{\lambda})\mathbf{V}'$$

If \mathbf{X} has full rank, then $(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{V}\text{Diag}(\boldsymbol{\lambda}^{-1})\mathbf{V}'$. If any eigenvalues are 0, we cannot compute the usual inverse. A zero eigenvalue implies some linear combination of the predictors has a zero variance and provides no additional information.



Principal Component Analysis



Principal component analysis describes variables using eigenanalysis.

Let $\mathbf{V} = [\mathbf{v}_1 \dots \mathbf{v}_p]$ indicate the $p \times p$ matrix of eigenvectors of $\mathbf{X}'\mathbf{X}/n$, with each column of \mathbf{V} a $p \times 1$ eigenvector.

Define $\mathbf{X}_* = \mathbf{X}\mathbf{V}$ as the $n \times p$ matrix of principal component scores. Each column represents a new variable (a variate).

The j th element of k th eigenvector (\mathbf{v}_k) provides the coefficient for the j th column (variable) in \mathbf{X} to compute the k th principal component score.

Using eigenvectors of $\mathbf{C} = \mathbf{X}'_c\mathbf{X}_c/n$ allows computing principal component scores for centered data, while using eigenvectors of $\mathbf{R} = \mathbf{D}_c^{-0.5}\mathbf{X}'_c\mathbf{X}_c\mathbf{D}_c^{-0.5}/n$ allows computing principal component scores for centered and scaled data.

Except in special cases, the eigenvalues, eigenvectors, and principal component scores of \mathbf{C} , \mathbf{R} , and the SSCP matrices have no simple correspondences between them.

Example: Eigenanalysis of Correlation Matrix for Ozone Data

```
proc princomp data=ozone;  
var outdoor home time_out;  
run;
```

```
*****
```

The PRINCOMP Procedure

Observations	64
Variables	3

Simple Statistics

	outdoor	home	time_out
Mean	44.94541180	19.83328125	0.2817187500
StD	21.90644120	11.98360958	0.2135754184

Correlation Matrix

outdoor	Outdoor Ozone Concentration (ppb)
home	Home Indoor Ozone Concentration (ppb)
time_out	Proportion of Time Spent Outdoors

Correlation Matrix

	outdoor	home	time_out
outdoor	1.0000	0.5558	0.0782
home	0.5558	1.0000	-.0071
time_out	0.0782	-.0071	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue 	Difference	Proportion	Cumulative
1	1.56035176	0.55838231	0.5201	0.5201
2	1.00196944	0.56429064	0.3340	0.8541
3	0.43767880		0.1459	1.0000



Eigenvectors

outdoor	Outdoor Ozone Concentration (ppb)
home	Home Indoor Ozone Concentration (ppb)
time_out	Proportion of Time Spent Outdoors

Eigenvectors



	Prin1	Prin2	Prin3
outdoor	0.707674	0.012227	-.706434
home	0.700809	-.139231	0.699629
time_out	0.089803	0.990184	0.107099

None of the eigenvalues are too close to zero, indicating that there is not much sign of collinearity among the predictors.

Now, we also examine collinearity with the intercept by conducting an eigenanalysis of the scaled SSCP matrix.

```
data ozone; set ozone; int=1; run;

proc princomp data=ozone noint;
var int outdoor home time_out;
run; 
*****
```

The PRINCOMP Procedure

Observations	64
Variables	4

Simple Statistics

	int	outdoor	home	time_out
Mean	1.000000000	44.94541180	19.83328125	0.2817187500
UStd	1.000000000	49.92478235	23.12405860	0.3525155138

Uncorrected Correlation Matrix

	int	outdoor	home	time_out
int	1.0000	0.9003	0.8577	0.7992
outdoor	0.9003	1.0000	0.8966	0.7399
home	0.8577	0.8966	1.0000	0.6832

time_out	0.7992	0.7399	0.6832	1.0000
----------	--------	--------	--------	--------

Eigenvalues of the Uncorrected Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	3.44397079	3.09272276	0.8610	0.8610
2	0.35124803	0.23327663	0.0878	0.9488
3	0.11797140	0.03116161	0.0295	0.9783
4	0.08680979		0.0217	1.0000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4
int	0.517493	-.013105	-.672666	-.528724
outdoor	0.515072	-.282728	-.212037	0.780901
home	0.500584	-.471532	0.649407	-.324566
time_out	0.465099	0.835195	0.284308	0.072810



Condition Number and Condition Index

The *condition index* for the k th eigenvalue equals $\sqrt{\lambda_1/\lambda_k}$. The maximum condition index, called the *condition number*, $\max_k \sqrt{\lambda_1/\lambda_k}$, involves the first and last eigenvalue.

(Some authors exchange the names of condition number and condition index, so be careful!)

Note that $1 \leq \text{CI}_k < \infty$, where large values of the condition index indicate greater collinearity.

Example: Eigenanalysis of Average SSCP, Scaled SSCP, C, and R
Using PROC PRINCOMP, we conduct an eigenanalysis for the ozone data. Note the VARDEF option, which is used to specify that n and not $n - 1$ is used in calculations for average SSCP and covariance matrices.



```
/* SSCP */
proc princomp data=ozone noint cov vardef=n;
var int outdoor home time_out;
run;
/* Scaled SSCP */
proc princomp data=ozone noint;
var int outdoor home time_out;
```

```
run;
/* Covariance Matrix */
proc princomp data=ozone noint cov vardef=n;
var outdoor home time_out;
run;
/* Correlation Matrix */
proc princomp data=ozone;
var outdoor home time_out;
run;
```

The condition indices (calculated by hand) are provided in a table.

Matrix Type	Eigenvalue	Condition Index
Average SSCP	2939.11	1.00
	88.99	5.75
	0.19	23.90
	0.04	267.81
Scaled SSCP	3.44	1.00
	0.35	3.13
	0.12	5.40
	0.09	6.30

Matrix Type	Eigenvalue	Condition Index
Covariance	2938.29	1.00
	88.99	5.75
	0.06	229.10
Correlation	1.56	1.00
	1.00	1.25
	0.44	1.89

We next consider interpretation of these values.

Interpretation of Condition Number

One rule of thumb is that $\text{CN} > 30$ implies moderate to severe collinearity. This corresponds to a ratio of variances of roughly 1000, so single computations may involve numbers varying by 3 decimal places. Loss of precision occurs as such errors accumulate.

Note that λ_k very near zero indicates a redundancy in the predictors, and the eigenvector can identify the culprits. The relative sizes of $\mathbf{v}_k = \{v_{jk}\}_{p \times 1}$ give the relative importance of the columns of \mathbf{X} in determining the undesirable k th variate. The k th variate has zero variance, and hence no predictive value. We interpret the $\{v_{jk}\}$ as regression coefficients.

Consider the SSCP and scaled SSCP matrices. A λ_k near zero indicates either collinearity with the intercept or among other predictors. Consider carefully the value of the eigenvector (with a bad condition number) weighting the intercept. A relatively large value

indicates a collinearity with the intercept.



Many simple mistakes may create collinearity with the intercept:

- using birth year rather than age,
- analyzing FVC in mL, rather than FVC in L,
- including variables with near zero variance (39/40 males),
- including redundant codings (including male and female codes), and
- some combination of the above.

C and R diagnose collinearity in predictors other than the intercept. C involves the relative scales of the variables, and R does not. Since few tests vary due to a change in scale or location, we prefer R , assuming no loss of precision.

R_j^2 , Tolerance, and VIF

Correlation and Collinearity

The simplest type of collinearity for variables other than the intercept is a correlation of 1.0 between two predictors.

Example: Exact Collinearity

Let X_1 represent temperature in F, and let X_2 represent temperature in C. Then $x_{i1} = 32 + 1.8x_{i2}$.

Fitting the model $x_{i1} = \beta_0 + \beta_1 x_{i2} + \varepsilon_i$ yields $\hat{\sigma}^2 = 0$ and $r^2(x_1, x_2) = 1$.

Suppose we use X_j as the response in a regression with the $p - 2$ predictors (excluding the intercept),

$$\{X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_{p-1}\}.$$

Then we compute the *squared multiple correlation*

$$R_j^2 = R^2(X_j, \{X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_{p-1}\}).$$

Clearly $R_1^2 = R_2^2 = 1$.

Large values (close to 1) of R_j^2 imply worse collinearity by indicating the extent of redundancy of a predictor with the remaining ones, and values close to zero indicate little or no collinearity.

Define *tolerance* as $1 - R_j^2$. Tolerance close to 1 is good, while tolerance close to 0 is bad.

Define the *variance inflation factor* as

$$\text{VIF}_j = \frac{1}{1 - R_j^2} = \frac{1}{\text{tolerance}}.$$

The name comes from the fact that $\text{var}(\hat{\beta}_j)$ is proportional to VIF_j .

The variance inflation factor shows how multicollinearity has increased

the instability of the coefficient estimates. The variance of the parameter estimate is larger (by the VIF) than it would be if there were no multicollinearity.

A VIF close to 1 is good, and $\text{VIF} \rightarrow \infty$ is bad.

Summary:

Good	Diagnostic	Bad
0	$\leq R_j^2$	≤ 1
1	$\geq 1 - R_j^2$	≥ 0
1	$\leq \frac{1}{1-R_j^2}$	$\leq \infty$

Eliminating a redundant variable does not reduce prediction!

As a rule of thumb,

$R_j^2 > .90$ merits some attention, and

$R_j^2 > .98$ suggests the need for removal of a useless variable.

Example: Diagnosing Collinearity in Ozone Data

Suppose that we fit the model $y_i = \beta_0 + \beta_1 OUTDOOR_i + \beta_2 HOME_i + \beta_3 HOME_i^2 + \beta_4 HOME_i^3 + \beta_5 TIMEOUT_i + \varepsilon_i$ to the ozone data so that we can investigate whether a polynomial in home provides a better fit to the data. If so, we obtain the following regression model output.

```
data ozone; set ozone;
home2=home*home;
home3=home*home2;
run;

proc reg;
model personal=outdoor home home2 home3 time_out/tol vif;
run;
*****
```

```
The REG Procedure
Model: MODEL1
Dependent Variable: personal
```

Analysis of Variance

Source	DF	Sum of Squares		Mean Square		
				F Value	Pr > F	
Model	5	5109.91477	1021.98295	5.88	0.0002	
Error	58	10073	173.67557			
Corrected Total	63	15183				

Root MSE	13.17860	R-Square	0.3366
Dependent Mean	23.54578	Adj R-Sq	0.2794
Coeff Var	55.97012		

Parameter Estimates

Variable	DF	Parameter	Standard	t Value	Pr > t
		Estimate	Error		
Intercept	1	2.34851	8.45710	0.28	0.7822
outdoor	1	0.10435	0.09388	1.11	0.2709
home	1	0.56187	1.46423	0.38	0.7026
home2	1	0.01058	0.07103	0.15	0.8821
home3	1	-0.00024100	0.00098237	-0.25	0.8071
time_out	1	13.70723	7.95522	1.72	0.0902



Parameter Estimates			
Variable	DF	Tolerance	Variance Inflation
Intercept	1	.	0
outdoor	1	0.65176	1.53431
home	1	0.00895	111.68575
home2	1	0.00160	624.92088
home3	1	0.00415	241.25289
time_out	1	0.95497	1.04715

Is there evidence of collinearity?

To investigate further, we conduct eigenanalyses of the scaled SSCP and correlation matrices.

```
/* Scaled SSCP matrix */  
proc princomp data=ozone noint;  
var int outdoor home home2 home3 time_out;  
run;  
  
/* correlation matrix */  
proc princomp data=ozone;
```

```
var outdoor home home2 home3 time_out;
run;
```

The PRINCOMP Procedure

Observations	64
Variables	6

Simple Statistics

	int	outdoor	home
Mean	1.000000000	44.94541180	19.83328125
UStd	1.000000000	49.92478235	23.12405860

Simple Statistics

	home2	home3	time_out
Mean	534.7220859	17251.08446	0.2817187500
UStd	788.6788104	31240.89600	0.3525155138

Uncorrected Correlation Matrix

	int	outdoor	home	home2	home3	time_out
int	1.0000	0.9003	0.8577	0.6780	0.5522	0.7992
outdoor	0.9003	1.0000	0.8966	0.7937	0.7039	0.7399
home	0.8577	0.8966	1.0000	0.9459	0.8610	0.6832
home2	0.6780	0.7937	0.9459	1.0000	0.9771	0.5338
home3	0.5522	0.7039	0.8610	0.9771	1.0000	0.4355
time_out	0.7992	0.7399	0.6832	0.5338	0.4355	1.0000

Eigenvalues of the Uncorrected Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.81211368	3.97703127	0.8020	0.8020
2	0.83508241	0.60164583	0.1392	0.9412
3	0.23343658	0.14484901	0.0389	0.9801
4	0.08858757	0.05833043	0.0148	0.9949
5	0.03025714	0.02973452	0.0050	0.9999
6	0.00052262		0.0001	1.0000

The PRINCOMP Procedure

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6
int	0.406570	0.394870	-.446481	-.481845	0.489442	-.087778
outdoor	0.428643	0.164398	-.405351	0.785773	-.086223	-.007426
home	0.447257	-.107067	-.121952	-.352657	-.660918	0.460919
home2	0.421044	-.411815	0.117787	-.107096	-.154855	-.777050
home3	0.387170	-.549914	0.264047	0.103256	0.539792	0.419447
time_out	0.351780	0.577573	0.733402	0.062342	-.029063	-.007115

The PRINCOMP Procedure

Observations 64
Variables 5



Simple Statistics

	outdoor	home	home2
Mean	44.94541180	19.83328125	534.7220859
StD	21.90644120	11.98360958	584.3126423

Simple Statistics

home3 time_out

Mean	17251.08446	0.2817187500
StD	26251.89177	0.2135754184

Correlation Matrix

	outdoor	home	home2	home3	time_out
outdoor	1.0000	0.5558	0.5730	0.5697	0.0782
home	0.5558	1.0000	0.9642	0.9037	-.0071
home2	0.5730	0.9642	1.0000	0.9836	-.0181
home3	0.5697	0.9037	0.9836	1.0000	-.0116
time_out	0.0782	-.0071	-.0181	-.0116	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	3.31662552	2.30161843	0.6633	0.6633
2	1.01500708	0.44509651	0.2030	0.8663
3	0.56991057	0.47248772	0.1140	0.9803
4	0.09742285	0.09638887	0.0195	0.9998

5	0.00103398	0.0002	1.0000
---	------------	--------	--------

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5
outdoor	0.389979	0.167238	0.905409	0.012712	-.004641
home	0.524979	-.038764	-.227787	0.749446	0.330663
home2	0.539900	-.050271	-.226048	-.094914	-.803663
home3	0.529906	-.042101	-.208732	-.655080	0.494700
time_out	0.004661	0.982970	-.183526	-.005519	-.006083

Detecting Numerical Inaccuracy

Which Matrix Should Be Examined?

Consider the ordered list of matrices: R , C , SSCP, and X . The order corresponds to the following rank order of:

- most to least processed,
- least to most information about scaling problems,
- least to most information about location problems, and
- easiest to hardest in which to see collinearity (beyond intercept).

A general strategy for investigating collinearity is provided below.

1. Look at descriptive statistics (mean, variance) for all predictors. If necessary, change location or scale and remove variables with zero variance.
2. Examine eigenvalues and eigenvectors of scaled SSCP matrix to check for collinearity with the intercept and eliminate problems (by changing location or scale or deleting variables) if necessary.
3. Examine eigenvalues and eigenvectors of **R** to discover collinearity among other predictors.

Location Problems

Consider a full rank GLM that includes an intercept.

In theory, slopes are location invariant. In addition, a full rank GLM which spans an intercept (and a GLH test which does not) has an F statistic and p-value invariant to location. In practice, finite precision arithmetic may create violations.

Any variation in F and p-value due to location implies a problem (for a model spanning an intercept and a GLH test which does not).

Only the scaled SSCP, SSCP, and \mathbf{X} directly detect location problems. Eigenanalysis of the scaled SSCP matrix, means, histograms, and extreme values for y and \mathbf{X} help detect location problems.

Extreme ratios of data values and SSCP or scaled SSCP eigenvalues (or elements) warn of location-induced inaccuracy.

Scaling Problems

Slope estimates may vary due to changes in scale.

In theory, a full rank GLM which spans an intercept (and a GLH test which does not involve the intercept) has F statistic and p-value invariant to change in scale of \mathbf{y} and/or \mathbf{X} .

Any variation in F statistic and p-value due to change of scale indicates a problem (if model spans an intercept and test does not).

\mathbf{C} , scaled SSCP, SSCP and \mathbf{X} yield information about scale problems.

Extreme ratios of values of \mathbf{C} (especially the variances) and histograms and extreme values of \mathbf{y} and \mathbf{X} can detect scale problems.

Collinearity Problems

In theory a GLH test that does not span an intercept in a full rank GLM which spans an intercept has an F statistic and p-value invariant to full rank linear transformation of \mathbf{y} and/or the columns of \mathbf{X}

Any variation in F statistic and p-value due to a full rank linear transformation diagnoses a problem (if the model spans an intercept and the test does not).

\mathbf{R} demonstrates invariance to location and scale changes.

\mathbf{R} , its eigenvalues, and R_j^2 's detect collinearity problems.

The eigenvector for the smallest eigenvalue of \mathbf{R} identifies the culprits.

Recommendations

1. Conduct data validation as part of data entry and file creation.
2. Treat non-independence of observations, if necessary. (For example, use multivariate methods, such as repeated measures ANOVA or random effects models.)
3. Minimize location and scale problems by roughly aligning variable ranges of variables. Centering and scaling may be necessary.
4. Minimize collinearity.
 - Eliminate predictors with near zero variance.
 - Eliminate redundant variables defined by unimportance or R_j^2 .
 - Avoid unnecessary collinearity: center polynomials or use orthogonal polynomials (Chapter 9), and use cell-mean or effect coding (Chapter 12).
5. Treat “linearity” (model specification error). Find useful predictors and transformations (Chapter 10).

6. Treat non-normality and heterogeneity. Consider transformations.

See Chapter 11 for a general treatment of exploratory regression.

We will address all of these issues in the coming weeks.

To summarize, first, clean the data. Then scale, center, and code it sensibly. If necessary, transform and delete variables. Do not be satisfied with the model unless all diagnostics create no substantial cause for concern. Chapter 11 centers on using this approach while creating statistically defensible estimates and hypothesis tests.

Next: Polynomial Regression

Reading Assignment:

- Muller and Fetterman, Chapter 9: “Polynomial Regression”
(Required)

Lecture 12: Polynomials and Splines

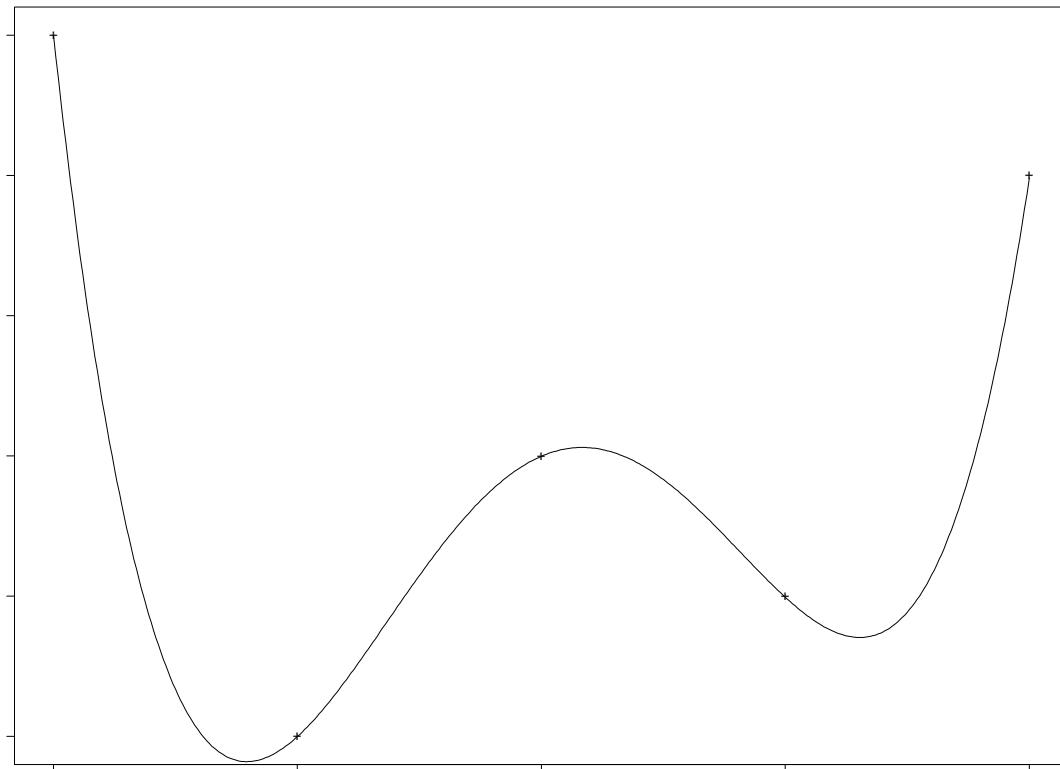
Reading Assignment:

- Muller and Fetterman, Chapter 9: “Polynomial Regression”

Motivation

Response-predictor relationships are often not linear. In most cases, we want to use tests and models with maximum power to detect association between a predictor and a response. The tests and models that we have discussed so far assume a linear relationship between $E(y)$ and x , and power for detecting an association between y and x will be reduced when this is not true. (A linear model may have good power if the trend is nonlinear but still monotone, but power will likely be terrible for U-shaped or umbrella-shaped relationships.)

A simple way to capture non-linear relationships between an exposure and response of interest is to add polynomial terms to a linear model. With d distinct (x, y) pairs, a polynomial of order at most $d - 1$ will pass exactly through all the points. Consider the following example, in which $\mathbf{x} = (1, 2, 3, 4, 5)'$ and $\mathbf{y} = (5, 0, 2, 1, 4)'$.



Why wouldn't we always want to use this type of model, which perfectly fits the data?

Polynomial Models

	Order	Regression Model	Shape
0th	Zero	$y_i = \beta_0 + \varepsilon_i$	Horizontal Line
1st	Linear	$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$	Line
2nd	Quadratic	$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$	Parabola
3rd	Cubic	$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$	Ogive etc.

The number of points of inflection is one fewer than the number of monotone pieces of the curve, which is the order (degree) of the polynomial. The maximum order that can be fitted is one less than the number of distinct X values, but one should be able to defend the use of any model with higher than a 3rd order polynomial due to difficulties in interpretation.

We will use the GLM for estimation of parameters, but how?

Example: Romanesque Cathedrals in England

Consider the following data collected by Stephen J. Gould on the dimensions of medieval English cathedrals. (The Romanesque period in English architecture roughly dates to 1066-1180.)

Cathedral	Nave Height (ft)	Total Length (ft)
Durham	75	502
Canterbury	80	522
Gloucester	68	425
Hereford	64	344
Norwich	83	407
Peterborough	80	451
St. Albans	70	551
Winchester	76	530
Ely	74	547

The nave height is the height of the center of the church, as seen (for Canterbury Cathedral) on the following page.



UNC Biostatistics 663, Spring 2019 Lecture 12

We can fit a polynomial model, say a third order polynomial, to the cathedral data, treating length as the dependent variable and height as the independent variable. The \mathbf{X} matrix is given by

$$\mathbf{X} = \begin{bmatrix} 1 & 75 & 75^2 & 75^3 \\ 1 & 80 & 80^2 & 80^3 \\ 1 & 68 & 68^2 & 68^3 \\ 1 & 64 & 64^2 & 64^3 \\ 1 & 83 & 83^2 & 83^3 \\ 1 & 80 & 80^2 & 80^3 \\ 1 & 70 & 70^2 & 70^3 \\ 1 & 76 & 76^2 & 76^3 \\ 1 & 74 & 74^2 & 74^3 \end{bmatrix}.$$

Variable added-last tests for each term are not recommended for

model selection. (Why?)

To test association (i.e., is x related to y ?) when linearity is not assumed, we will conduct a *groupwise* test of all terms involving the predictor.



We fit a cubic model to the cathedral data below.

```
data romanesque; set romanesque;
heightsq=height*height; heightcu=heightsq*height;
run;
proc reg data=romanesque;
model length=height heightsq heightcu/tol vif ss1;
output out=out pred=pred;
run;
*****
The REG Procedure
Model: MODEL1
Dependent Variable: length
```

Analysis of Variance

Source	DF	Sum of		Mean		
		Squares	Square	F Value	Pr > F	
Model	3	34106	11369	7.41	0.0275	
Error	5	7676.04705	1535.20941			
Corrected Total	8	41782				

Root MSE	39.18175	R-Square	0.8163
Dependent Mean	475.44444	Adj R-Sq	0.7061

Coeff Var 8.24108

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-14512	30517	-0.48	0.6544
height	1	478.38312	1253.74118	0.38	0.7185
heightsq	1	-4.69654	17.10590	-0.27	0.7946
heightcu	1	0.01324	0.07752	0.17	0.8711

Parameter Estimates

Variable	DF	Type I SS	Tolerance	Variance Inflation
Intercept	1	2034427	.	0
height	1	3035.20888	0.00000317	315582
heightsq	1	31026	7.844759E-7	1274736
heightcu	1	44.79835	0.00000310	323080

We consider three basic inferential questions.

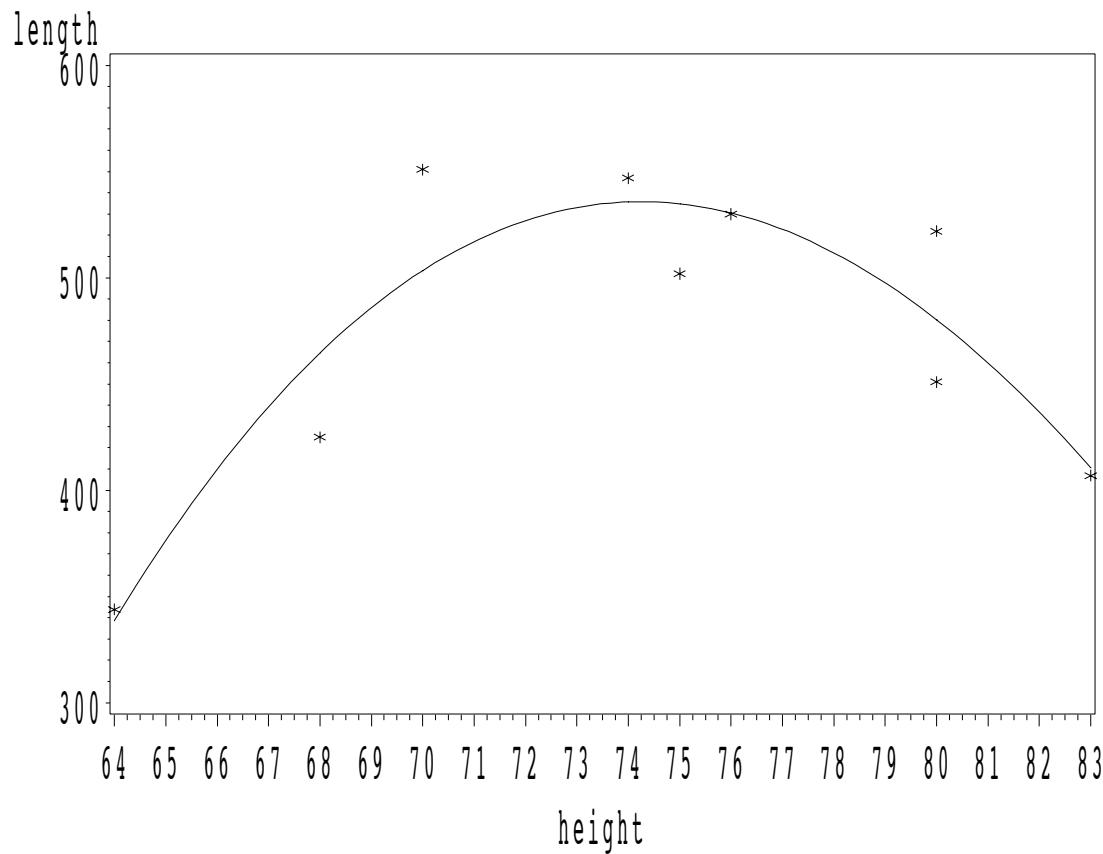
1. Are height and length related?
2. Is a cubic model significantly better than the quadratic model? If not, is a quadratic model significantly better than a linear model?
3. Should we have considered even higher-order polynomial terms?

How do you interpret the tolerance and VIF values?

Tolerance is a measure of collinearity, which equals to $1-R^2$ of predicting this covariate with all the other covariates. $VIF = 1/\text{tolerance}$.

The following code may be used to obtain a plot of observed and predicted length values for the cubic model.

```
symbol1 color=black i=none v=star;
symbol2 color=black v=point line=1 i=RC;
/* RL for linear, RQ for quadratic */
proc gplot data=out;
plot length*height=1 pred*height=2/overlay;
run;
```



What do you think about the model fit based on the observed and predicted nave lengths? What do you think about inferences outside the range of the data?

Limitations of Natural Polynomials

Fitting higher-order natural polynomial models can be dangerous! Because independent variables in a polynomial model are functions of the same basic variable x , they are correlated, and computational difficulties due to collinearity may result. In fact, collinearity is almost certainly present if the order of the polynomial is high. Techniques like centering and the use of orthogonal polynomials will help, though there is no remedy for difficulty of interpretation for high-order polynomials in most settings.

With a polynomial model, extrapolation beyond the range of the data is even more dangerous than usual.

Orthogonal Polynomials

Orthogonal polynomials provide one good solution to numerical problems in a polynomial regression by using an alternate coding scheme. Natural polynomials yield the model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i,$$

while orthogonal polynomials yield the model

$$y_i = \gamma_0 + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \gamma_3 z_{i3} + \delta_i.$$

The orthogonal polynomial scores ($\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$) are the solution to a system of equations under the constraints that the new predictors ($\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$)

1. contain the same information as the original variables ($\mathbf{x}, \mathbf{x}^2, \mathbf{x}^3$),
2. are linear combinations of original data,
3. are mean zero (except for the intercept),

-
- 4. are mutually orthogonal (uncorrelated), and
 - 5. z_1 captures all information in x beyond location of response (intercept), z_2 captures all information in x^2 , beyond linear term in x and the intercept, etc.



Finding the values of z_1, z_2, z_3 corresponds to solving a system of simultaneous equations for each subject:



$$z_{i1} = a_{01} + a_{11}x_i$$

$$z_{i2} = a_{02} + a_{12}x_i + a_{22}x_i^2$$

⋮

$$z_{ip-1} = a_{0p-1} + \sum_{j=1}^{p-1} a_{jp-1}(x_i)^j.$$

Clearly, the new variables are linear combinations of the original ones.
We may write

$$\begin{aligned}
 x_i &= b_{01} + b_{11}z_{i1} \\
 x_i^2 &= b_{02} + b_{12}z_{i1} + b_{22}z_{i2} \\
 &\vdots \\
 x^{p-1} &= b_{0p-1} + \sum_{j=1}^{p-1} b_{jp-1}z_{ij}.
 \end{aligned}$$

Thus we may (without losing any information) write either

$$y_i = \beta_0 + \beta_1 x_i + \dots + \beta_{p-1} x_i^{p-1} + \varepsilon_i$$

or

$$y_i = \alpha_0 + \alpha_1 z_{i1} + \dots + \alpha_{p-1} z_{ip-1} + \varepsilon_i.$$

The parameters β and α will not be numerically identical, but the overall F tests from these models are exactly the same. In addition, testing $H_0 : \alpha_k = 0$ for the orthogonal polynomial model is equivalent to an added-in-order test of $H_0 : \beta_k = 0$ in the natural polynomial

model.

If the predictor values are equally spaced and there are an equal number of observations at each of d values, finding the orthogonal predictors reduces to finding $d - 1$ vectors, each of length d . The $d - 1$ vectors are known as the orthogonal polynomial coefficients. The table in Appendix B.10 of Muller and Fetterman may be used to obtain coefficients if predictors are equally spaced and each predictor value has the same number of replicates.

Other Ways to Improve Fit of Natural Polynomial Models

- Scaling: For example, if a weight variable is measured in grams with a range of 1000-2000, use kg with a range of 1-2. (Presence of the intercept leads to preferring a range of 1-10 to avoid collinearity.)
- Centering: Often dramatically reduces collinearity (always eliminates any with intercept) and is strongly recommended as habitual technique.
- Pseudo-Centering: Some “nice” number often provides most of the numerical advantage of centering, while simplifying interpretation. So if a predictor is weight or blood pressure, we might pseudo-center it at a “healthy” value of the predictor (average in a healthy population say).



Model Selection in Polynomial Regression

Muller and Fetterman prefer *backwards selection*, starting at the largest desirable polynomial model (they recommend starting at a cubic) and evaluating smaller ones with added-in-order tests. They recommend stopping at the highest order polynomial that is significant and including all lower order terms. Model diagnostics are very important in polynomial models as in all other regression models.

More Flexible Models

Often, y does not behave linearly in all the predictors. The simplest way to describe a nonlinear effect of x is to include polynomial terms in the model. However, nonlinear effects may not follow polynomial forms. In such cases, transformations might be able to induce linearity, but often the transformation is not known or does not exist.

Linear Splines

Spline functions are piecewise polynomials used to fit curves. Within intervals of x , they are polynomials, and they are connected across the different intervals. The most simple type of spline function is a *linear spline function*, which is a piecewise linear function. Linear spline functions may also be fit in the framework of the GLM.

Divide the x axis into intervals with endpoints k_1, k_2, \dots, k_K , called *knots*. The linear spline function is given by

$$\begin{aligned} E(\mathbf{y}) &= \beta_0 + \beta_1 \mathbf{x} + \beta_2 (\mathbf{x} - k_1 \mathbf{J}_n)_+ + \beta_3 (\mathbf{x} - k_2 \mathbf{J}_n)_+ \\ &\quad + \dots + \beta_{K+1} (\mathbf{x} - k_K \mathbf{J}_n)_+, \end{aligned}$$



where

$$(u)_+ = \begin{cases} u & u > 0 \\ 0 & u \leq 0. \end{cases}$$

The number of knots varies depending on the amount of data available.

We can rewrite the linear spline function as follows:

$$E(y_i) = \begin{cases} \beta_0 + \beta_1 x_i & x_i \leq k_1 \\ \beta_0 + \beta_1 x_i + \beta_2(x_i - k_1) & k_1 < x_i \leq k_2 \\ \beta_0 + \beta_1 x_i + \beta_2(x_i - k_1) + \beta_3(x_i - k_2) & k_2 < x_i \leq k_3 \\ \vdots \\ \beta_0 + \beta_1 x_i + \beta_2(x_i - k_1) + \dots + \beta_{K+1}(x_i - k_K) & k_K < x_i. \end{cases}$$

We can fit this model in SAS by creating new variables
 $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_K)$ where

$$\begin{aligned} \mathbf{x}_1 &= \mathbf{x} \\ \mathbf{x}_2 &= (\mathbf{x} - k_1)_+ \\ &\vdots && \vdots \\ \mathbf{x}_K &= (\mathbf{x} - k_K)_+. \end{aligned}$$

Then, we can fit the model



$$E(\mathbf{y}) = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \dots + \beta_{K+1} x_K$$

in any standard software package.

To test linearity in x , we simply test the hypothesis

$$H_0 : \beta_2 = \beta_3 = \dots = \beta_{K+1} = 0.$$



To construct a linear spline with knots at $k_1 = 70$ and $k_2 = 77$ for the cathedral data, we use the following \mathbf{X} :

$$\mathbf{X} = \begin{bmatrix} 1 & 75 & 5 & 0 \\ 1 & 80 & 10 & 3 \\ 1 & 68 & 0 & 0 \\ 1 & 64 & 0 & 0 \\ 1 & 83 & 13 & 6 \\ 1 & 80 & 10 & 3 \\ 1 & 70 & 0 & 0 \\ 1 & 76 & 6 & 0 \\ 1 & 74 & 4 & 0 \end{bmatrix}.$$

Then we can fit the model using the following SAS code.

```
data romanesque; set romanesque;
ht_a=0;
ht_b=0;
if height>70 then ht_a=height-70;
if height>77 then ht_b=height-77;
run;

proc reg;
model length=height ht_a ht_b;
run;
*****
```

The REG Procedure
Model: MODEL1
Dependent Variable: length

Analysis of Variance

Source	DF	Sum of Squares		Mean Square		
				F Value	Pr > F	
Model	3	35677	11892	9.74	0.0157	
Error	5	6105.05660	1221.01132			
Corrected Total	8	41782				

Root MSE	34.94297	R-Square	0.8539
Dependent Mean	475.44444	Adj R-Sq	0.7662
Coeff Var	7.34954		

Parameter Estimates

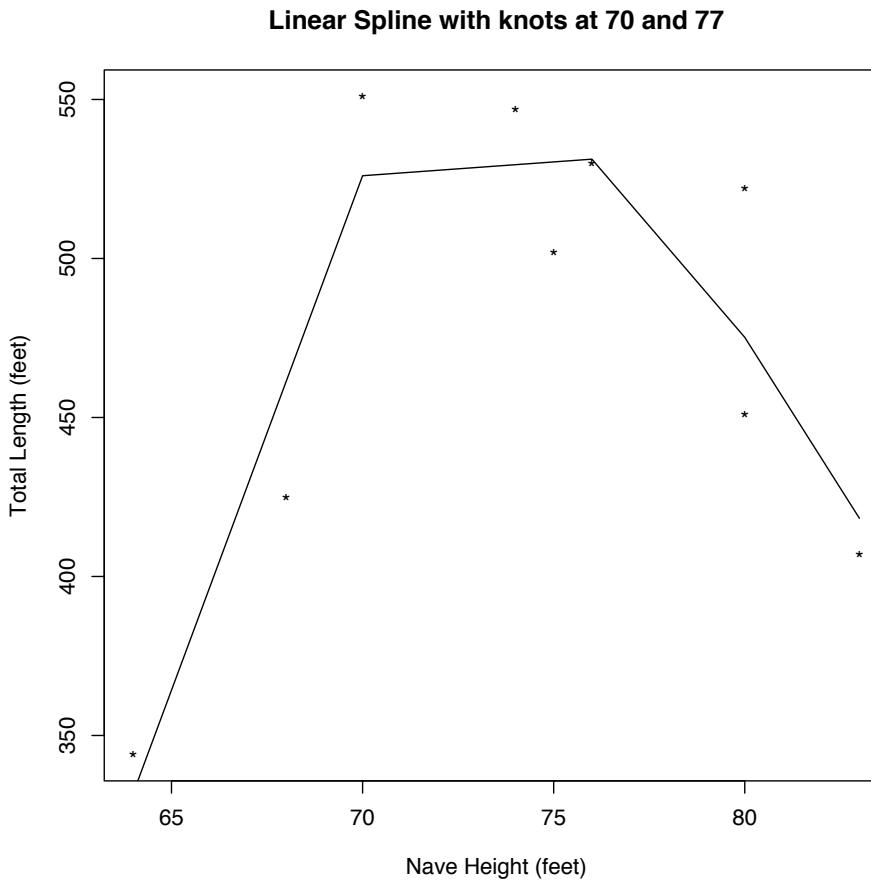
Variable	DF	Parameter		Standard	
		Estimate	Error	t Value	Pr > t
Intercept	1	-1738.74528	534.64321	-3.25	0.0226
height	1	32.35377	7.92328	4.08	0.0095

ht_a	1	-31.48585	12.78943	-2.46	0.0571
ht_b	1	-19.83333	12.62576	-1.57	0.1770

Parameter Estimates

Variable	DF	Type I SS
Intercept	1	2034427
height	1	3035.20888
ht_a	1	29629
ht_b	1	3012.97872

The plot of fitted values vs. height is below.



Test linearity using this model. How do you interpret the parameter estimates?

Cubic Spline Functions

Although linear splines are simple and are good approximations for some relationships, they are not smooth, and they will not fit curved functions very well. To do so, we must use piecewise polynomials of higher orders. Cubic polynomials have nice properties and do a good job of fitting sharp curves. To make cubic splines smooth at the knots, we force the first and second derivatives of the function to agree at the knots.

Spline Functions in SAS

PROC GAM (generalized additive model) in SAS allows us to fit spline functions for covariates. These splines are slightly different from cubic spline functions but are based on the same ideas that we have discussed.

Categorization of Predictors

In epidemiology, categorizing continuous predictors is almost a default practice. The thought is that creating categories of exposure (or dichotomizing exposure into “exposed” and “unexposed” groups) avoids assumptions of linearity. Although categorization does make interpretation simple, it can make unnatural assumptions about the exposure effect and may also lead to power losses.

For example, suppose that we decide to divide churches into three groups: short (≤ 70 feet), medium (> 70 and ≤ 77 feet), and tall (> 77 feet). We can then define two variables to indicate whether churches are medium or tall:

$$medium_i = \begin{cases} 1 & 70 < height \leq 77 \\ 0 & \text{otherwise} \end{cases}$$

$$tall_i = \begin{cases} 1 & height > 77 \\ 0 & \text{otherwise.} \end{cases}$$

We can then fit the model

$$E(length) = \beta_0 + \beta_1 medium_i + \beta_2 tall_i + \varepsilon_i,$$

$i = 1, \dots, n$. In this model, we have

$$E(length|\text{short nave}) = \beta_0$$

$$E(length|\text{medium nave}) = \beta_0 + \beta_1$$

$$E(length|\text{tall nave}) = \beta_0 + \beta_2.$$

What does this model assume about the effect of increasing nave height among churches with short naves?

Below are SAS code and output from fitting this model.

```
data romanesque; set romanesque;  
medium=1; tall=0;  
if height<70.1 then medium=0;  
if height>77 then medium=0;  
if height>77 then tall=1;  
run;
```

```
proc reg;   
model length=medium tall/ss1;  
run;  
*****
```

```
The REG Procedure  
Model: MODEL1  
Dependent Variable: length
```

Analysis of Variance

Source	DF	Sum of Squares		Mean Square		
		F Value	Pr > F			
Model	2	12254	6126.77778	1.24	0.3530	
Error	6	29529	4921.44444			
Corrected Total	8	41782				

Root MSE	70.15301	R-Square	0.2933
Dependent Mean	475.44444	Adj R-Sq	0.0577
Coeff Var	14.75525		

Parameter Estimates

Variable	DF	Parameter	Standard	t Value	Pr > t
		Estimate	Error		
Intercept	1	440.00000	40.50286	10.86	<.0001
medium	1	86.33333	57.27969	1.51	0.1825
tall	1	20.00000	57.27969	0.35	0.7389

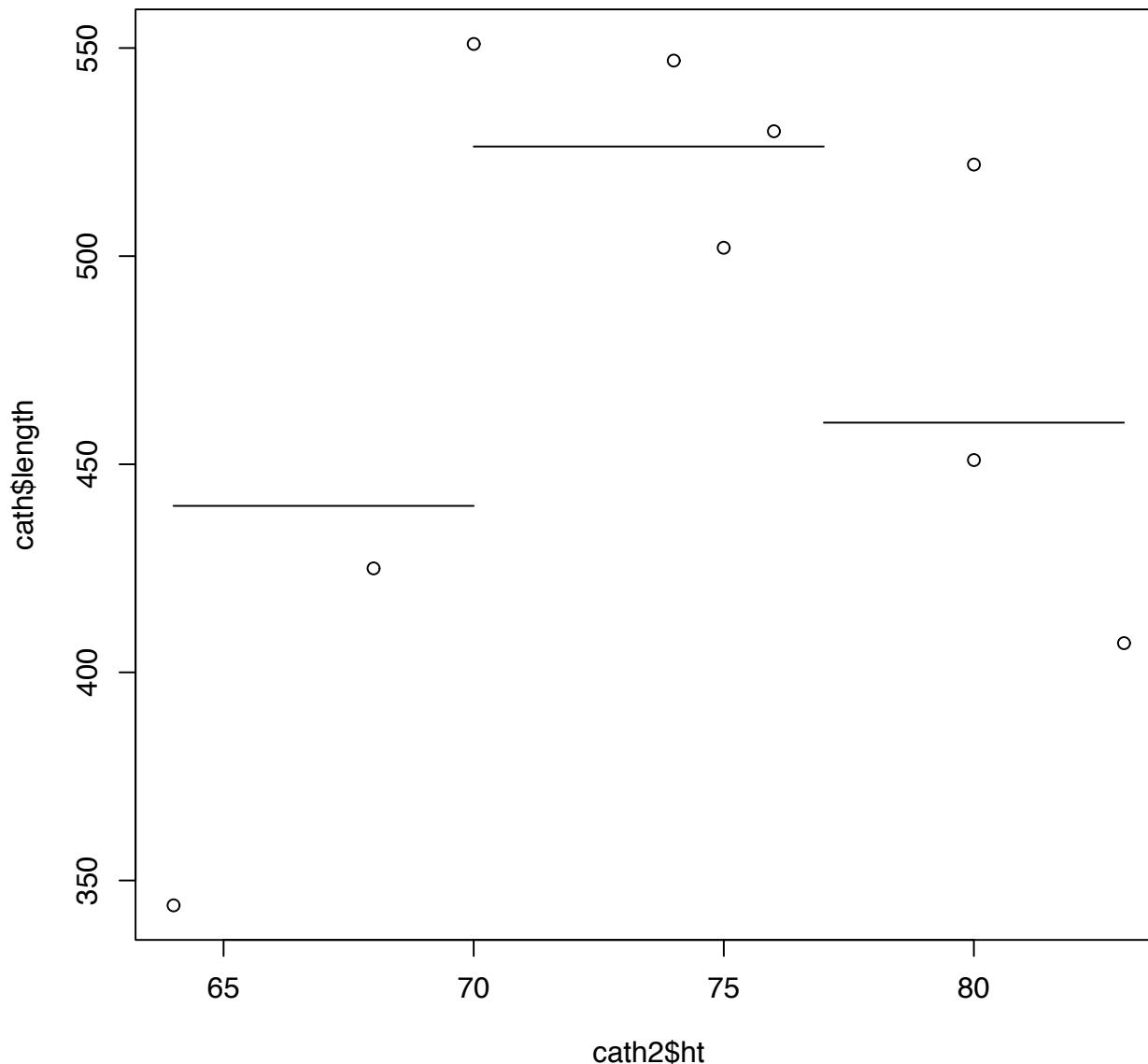
Parameter Estimates

Variable	DF	Type I SS
Intercept	1	2034427
medium	1	11654
tall	1	600.00000

A plot of the model fit is provided.



Fit of model $E(y) = b_0 + b_1 * \text{medium} + b_2 * \text{tall}$



Test association in this model. In addition, give proper interpretations of all parameter estimates, and provide estimated conditional means for cathedrals with nave heights of 69.9 and 70.1 feet, respectively.

Nonparametric Regression

Nonparametric smoothers are tools that help determine the shape of the relationship between variables. These tools work best when interest is in one continuous predictor and one continuous response at a time.

Moving Average

The most simple nonparametric smoother is a *moving average*.

Suppose our data are $\mathbf{x} = (1, 2, 3, 5, 8)'$ and

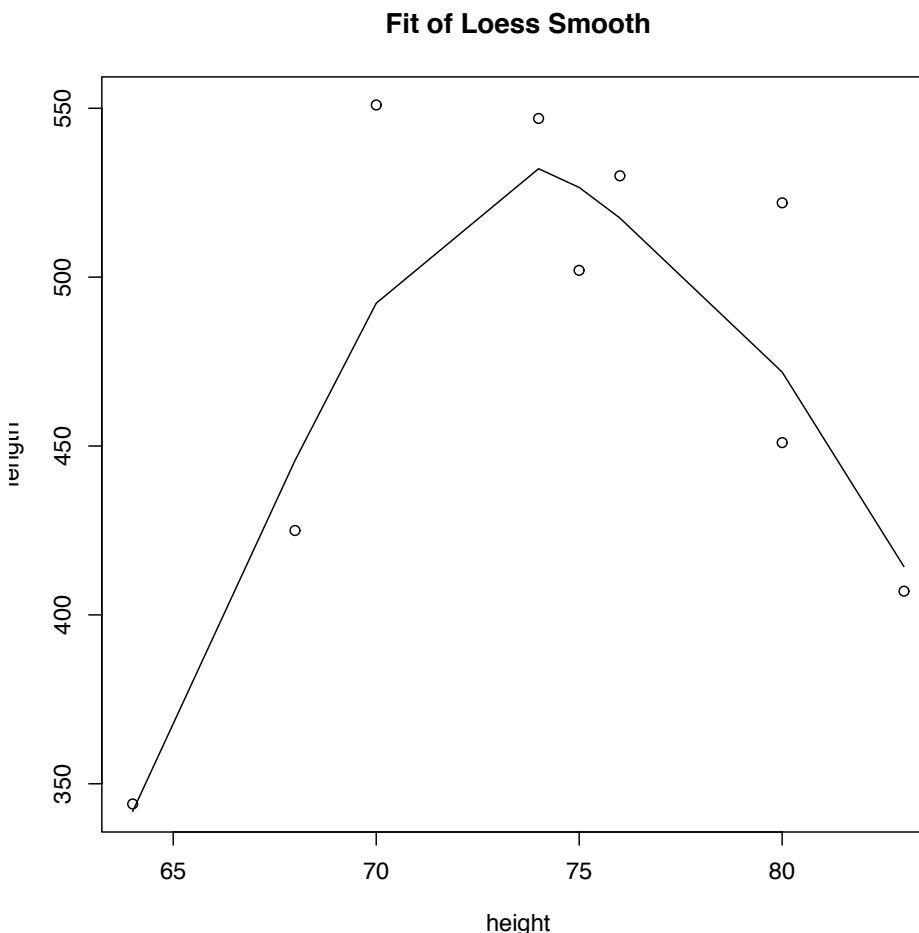
$\mathbf{y} = (2.1, 3.8, 5.7, 11.1, 17.2)'$. To smooth their relationship, we could estimate $E(Y|X=2)$ by $\frac{2.1+3.8+5.7}{3}$. This is problematic at the upper and lower limits X , and estimates are sensitive to the interval width, which here is 3 data points. Moving average smoothers are also called moving flat line smoothers.

Loess

A moving least squares linear regression smoother is superior to the moving average approach. The smoother *loess* is the most popular smoother of this type. To get smoothed values of $E(Y)$ at $X = x$, we take all data with X values within a suitable interval (often $\frac{2}{3}$ of data points) about x , and then we fit a linear regression using only these points. The predicted value from this regression at $X = x$ will be used as our estimate of $E(Y|X = x)$. Then, we move to the next observed value of X and repeat the process.

Actually, *weighted least squares* estimates are typically used so that points closest to $X = x$ are given the most weight, and points further away are given less weight. For this reason, *loess* is called a locally weighted least squares method. Loess calculates a smoothed value for $E(Y)$ at each observed value of X , and estimates for other X 's are obtained by interpolation.

Next is a plot of the model fit using nonparametric regression with a loess smoother for nave height. Like splines, loess smoothers may be fit in PROC GAM.



Summary

Relationships between a response and predictor are often not linear. Nonlinear effects can be accommodated in the GLM framework by using

- polynomial terms (orthogonal polynomials are more stable than natural polynomials),
- transformations (often not known in advance),
- simple splines, or
- categorization of predictors (power loss; interpretation difficulties).

Smoothing splines and loess smoothers are useful in determining shapes of relationships and may be the best choice for presenting analysis results if relationships are highly nonlinear or when thresholds are of interest. However, such smoothers are often not the most powerful approach when relationships are monotone or can be well-represented by lower-order polynomials.

Next: Transformations

Reading Assignment:

- Muller and Fetterman, Chapter 10: “Transformations”
- Weisberg, Chapter 8: “Transformations”

Lecture 13: Transformations

Reading Assignment:

- Muller and Fetterman, Chapter 10: “Transformations”
- Weisberg, Chapter 8: “Transformations”

Transformation of the response and/or predictor variables may correct violations of homogeneity, linearity, and Gaussian distribution of errors (often these three assumptions stand or fall as a group).

Transformation may also simplify a model by linearizing a relationship.

Using transformations often boils down to trial and error, and we use diagnostics on the residuals to gauge the value of a transformation.

Always bear in mind that nonlinear models are part of the complete picture and may be the best alternative.

For regression, we consider three groups of transformations: linear, monotone, or non-monotone. This grouping is used to indicate invariance of statistical properties. For a T test, means and variances change under linear transformations, but the test statistic and p-value do not. Test statistics and p-values for many non-parametric methods do not change with monotone transformations of the data.

A transformation has statistical benefit or cost only if it changes probabilities (and hence inferences).

Transformations may help in model fitting and may provide scientific insight. If height^2 works better than height , then perhaps the true relationship involves surface area.

Variance Stabilizing Transformations of the Response

Variance stabilizing transformations are useful in some cases for treating heteroscedasticity.

When observations are amounts or measurements (think of a ratio scale variable), the standard deviation is often proportional to the mean. As an example, think of counting the money in a coke machine versus counting all the money in a Wachovia office.

Standard variance stabilizing transformations include the following.

Data	Distribution	Transformation
Count	Poisson	$\sqrt{y_i}$
Amount	Gamma	$\log y_i$
Proportion	Binomial/ n	$\sin^{-1}(\sqrt{y_i})$



Despite these transformations, there are much better modeling strategies for count data and proportions. *Generalized Linear Models*  are more appropriate for counts and proportions/percentages since they accommodate data from distributions other than the normal distribution.

We will devote most of our energy to discussion of *linearizing* transformations.

log transformation

The natural logarithm is useful when one expects the effect to be proportional to the response. For example, consider a model with one predictor, x in which the response is expected to increase 100ρ percent for each 1-unit increase in x . In addition, suppose that the error, δ , is multiplicative. Then we can write the model

$$y = \gamma(1 + \rho)^x \delta.$$

As you can see, when $x = 0$, $y = \gamma\delta$, and when $x = 1$, $y = \gamma\delta + \gamma\rho\delta$, which is a $100\rho\%$ increase!

Taking logs on both sides, we have

$$\begin{aligned}\log(y) &= \log(\gamma) + x \log(1 + \rho) + \log(\delta) \\ &= \alpha + \beta x + \varepsilon,\end{aligned}$$

where $\alpha = \log(\gamma)$, $\beta = \log(1 + \rho)$, and $\varepsilon = \log(\delta)$. Thus taking logs transforms the complex multiplicative model into a simple linear model.

Solving for ρ in terms of β , we have $\rho = \exp(\beta) - 1$. So a one-unit increase in x corresponds to a $100(\exp(\beta) - 1)\%$ increase in y . For small β , say $|\beta| < 0.10$, you can interpret $\hat{\beta}$ as a $100\hat{\beta}$ percent effect on the response.

Power Transformation of the Response

Box-Cox Transformations

The square root transformation involves taking $y^{\frac{1}{2}}$. We may also wish to consider other powers of y in models of the form $y_i^\pi \equiv \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i$.  

Although we can fit this model for various values of π , it is difficult to decide which transformation to use since the models are not directly comparable (the SSE's are not in the same units). To solve this problem, Box and Cox (1964) introduced a family of transformations of the response variable:

$$Y_i(\pi) = \begin{cases} \frac{Y_i^\pi - 1}{\pi \cdot Y^{*(\pi-1)}} & \pi \neq 0 \\ Y^* \cdot \ln(Y_i) & \pi = 0. \end{cases}$$


where $Y^* = \left(\prod_{i=1}^N Y_i \right)^{1/N}$, the geometric mean of the response, and the natural logarithm for $\pi = 0$ reflects the limit as $\pi \rightarrow 0$.

This basically corresponds to the transform y^π for $\pi \neq 0$ and $\log(y)$ for $\pi = 0$, except that the Box-Cox transformation puts the SSE's on the same scale so that they can be compared.

Box and Cox recommended choosing π to minimize the SSE.

Below is a simple example to illustrate the Box-Cox transformation.

Data were generated from the model $y = e^{x+\epsilon}$ where $\epsilon \sim N(0, 1)$.

The transformed data can be fit with a linear model $\log(y) = x + \epsilon$.

SAS code is below:

```
data x;
  do x = 1 to 8 by 0.025;
    y = exp(x + normal(7));
    output;
  end;
run;

proc transreg data=x details;
  title2 Defaults;
  model boxcox(y) = identity(x);
run;
```

The TRANSREG Procedure

Transformation Information for BoxCox(y)

Lambda	R-Square	Log Like
-3.00	0.03	-4601.01
-2.75	0.04	-4266.08
-2.50	0.04	-3934.11
-2.25	0.05	-3605.75
-2.00	0.06	-3281.88
-1.75	0.07	-2963.74
-1.50	0.10	-2653.14
-1.25	0.14	-2352.72
-1.00	0.21	-2066.32
-0.75	0.34	-1799.25
-0.50	0.52	-1558.55
-0.25	0.71	-1360.28
0.00	0.79	-1275.31 <- best lambda
0.25	0.70	-1382.62
0.50	0.51	-1589.03
0.75	0.34	-1834.53
1.00	0.22	-2105.88
1.25	0.15	-2397.35
1.50	0.11	-2704.64
1.75	0.08	-3024.24
2.00	0.06	-3353.38
2.25	0.05	-3689.91
2.50	0.04	-4032.18

2.75	0.03	-4378.97
3.00	0.03	-4729.37

Regression assumption diagnostics must always be examined as well.

Note: all of these transformations are not well-defined if y is not strictly positive. In such a case, you may encounter problems for $\pi \leq 0$.

Ladder of Power Transformations

Alternatively, the “ladder of power transformations” below can be used to guide the choice of transformations.

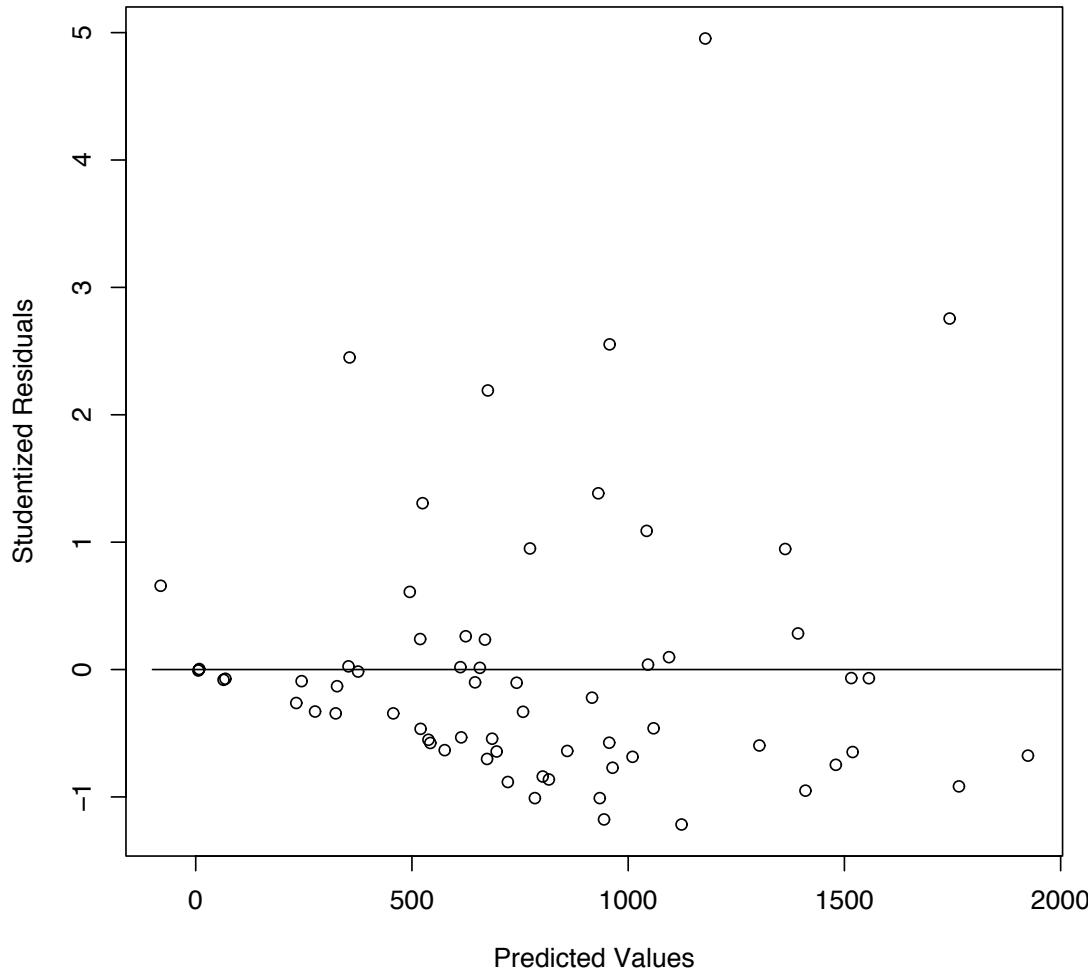
Half-Steps on the Ladder of Power Transformations

π	Transform	Description
	:	
-2	y^{-2}	
-3/2	$y^{-3/2}$	
-1	y^{-1}	reciprocal
-1/2	$y^{-1/2}$	$= 1/\sqrt{y}$
"0"	$\lim_{\pi \rightarrow 0} y^\pi$	$= \ln y$
1/2	$y^{1/2}$	square root
1	y^1	identity
3/2	$y^{3/2}$	
2	y^2	square
	:	

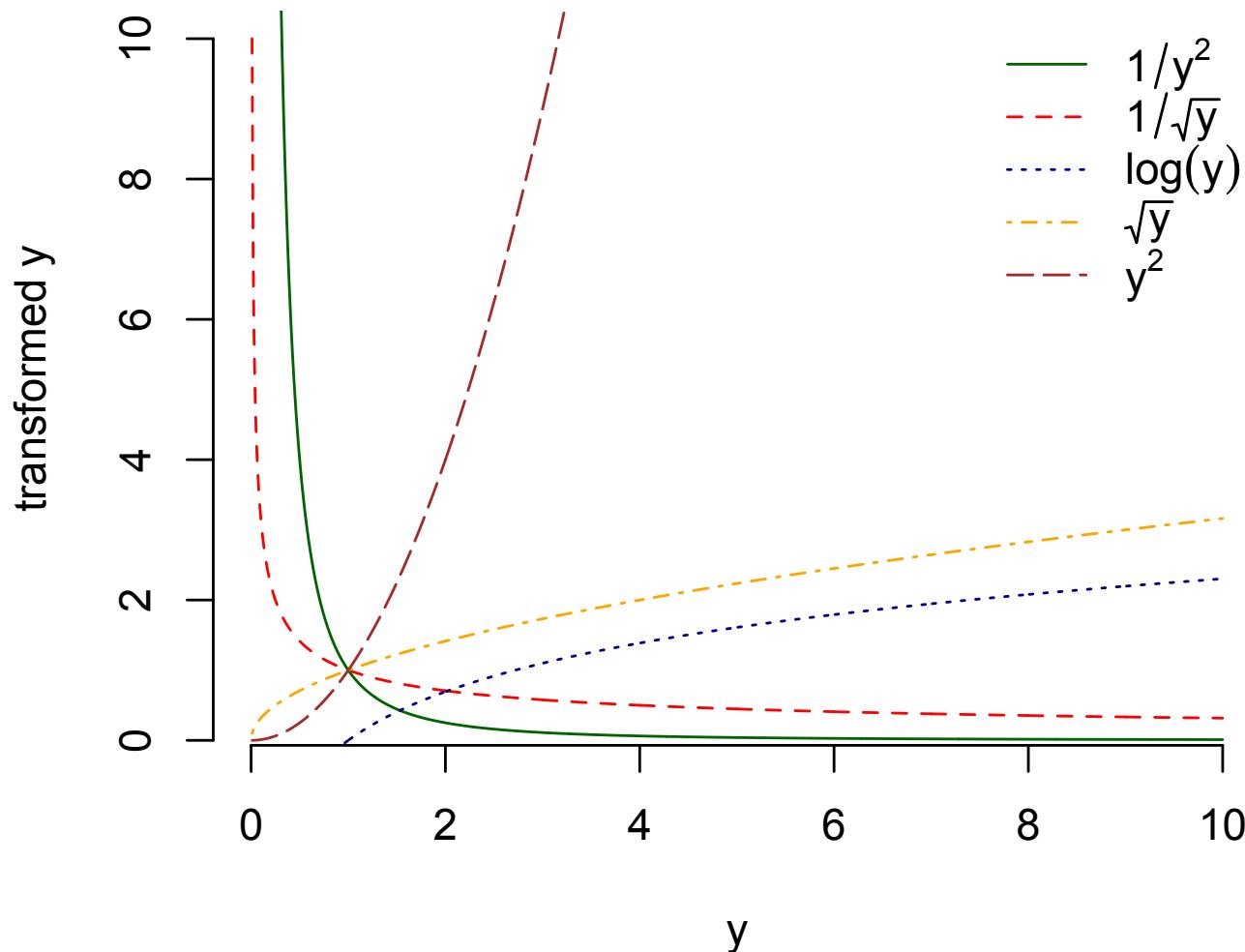
A crude but effective selection strategy is outlined below.

1. Fit the proposed model and check studentized residuals, including the R/P plot, the frequency histogram for skewness, and a test of Gaussian distribution. Multi-modality suggests an important predictor (or predictors) is missing from the model. If the residuals look good, keep the current model.
2. Based on the R/P plot and frequency histogram, select a grid of π values. Positive skewness in the histogram or a fan shape opening to the right on the R/P plot implies moving up the ladder from the value 1 (choosing values of $\pi < 1$), while negative skewness or a fan shape opening to the left on the R/P plot implies moving down ($\pi > 1$). Ratio scale variables often require $\pi < 1$.
3. Check studentized residuals for each value of π and examine other appropriate diagnostics, including tests of the Gaussian distribution of residuals.
4. Note that if you pick $\pi < 0$, the ladder direction reverses due to

the reciprocal transformation in addition to the power transformation. So if $\pi < 0$ for the current model, an R/P plot with a fan opening to the right implies moving down the ladder, and a fan opening to the left implies moving up the ladder. For example, consider the following R/P plot for $\pi = 2$ in the ozone data.



In this case, the fan opens to the right, and we move up the ladder.



5. To select among models with residuals that look ok, choose the

model with the smallest SSE.

6. Failure to obtain an acceptable model implies a more serious problem such as inadequate predictors or an entirely inappropriate model.

Example: Power Transformations for Ozone Data

The SAS code below can be used to program the Box-Cox transformation for the ozone data for selected values of π in the range $[-2, 2]$.

```
data ozonex;  
merge gmean ozone;   
y_2=((personal**2)-1)/(2*(geomeany**(2-1)));  
y_1_5=((personal**1.5)-1)/(1.5*(geomeany**(1.5-1)));  
y_1=((personal**1)-1)/(1*(geomeany**(1-1)));  
y_5=((personal**0.5)-1)/(0.5*(geomeany**(0.5-1)));  
y_0=geomeany*log(personal);  
y_m5=((personal**(-0.5))-1)/(-0.5*(geomeany**(-0.5-1)));  
y_m1=((personal**(-1))-1)/(-1*(geomeany**(-1-1)));  
y_m1_5=((personal**(-1.5))-1)/(-1.5*(geomeany**(-1.5-1)));  
y_m2=((personal**(-2))-1)/(-2*(geomeany**(-2-1)));  
run;
```

```
proc glm data=ozonex;
model y_2=outdoor home time_out;
run;
```

```
proc glm data=ozonex;
model y_1_5=outdoor home time_out;
run;
```

```
proc glm data=ozonex;
model y_1=outdoor home time_out;
run;
```

```
proc glm data=ozonex;
model y_5=outdoor home time_out;
run;
```

```
proc glm data=ozonex;
model y_0=outdoor home time_out;
run;
```

```
proc glm data=ozonex;
```

```
model y_m5=outdoor home time_out;
run;

proc glm data=ozonex;
model y_m1=outdoor home time_out;
run;

proc glm data=ozonex;
model y_m1_5=outdoor home time_out;
run;

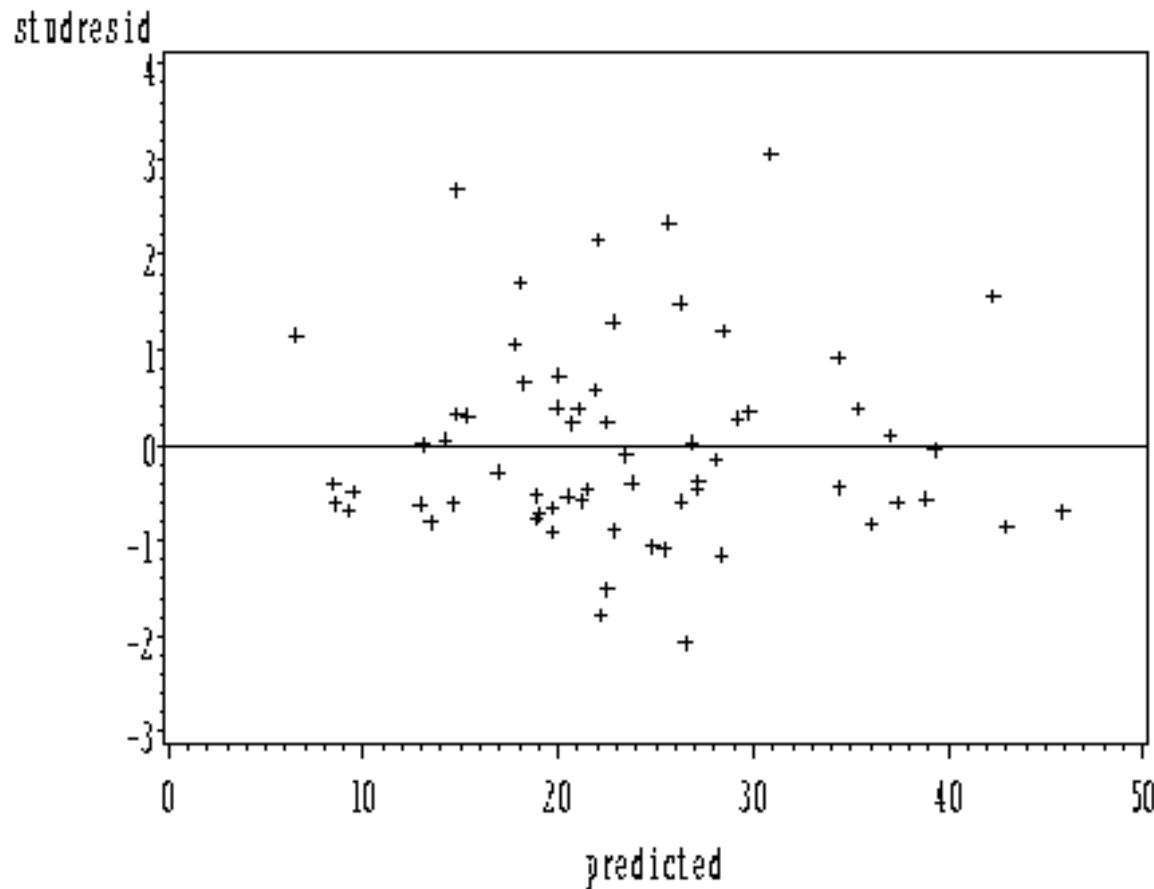
proc glm data=ozonex;
model y_m2=outdoor home time_out;
run;
```

In the interest of time, we just look at the SSE for each model in the table below.

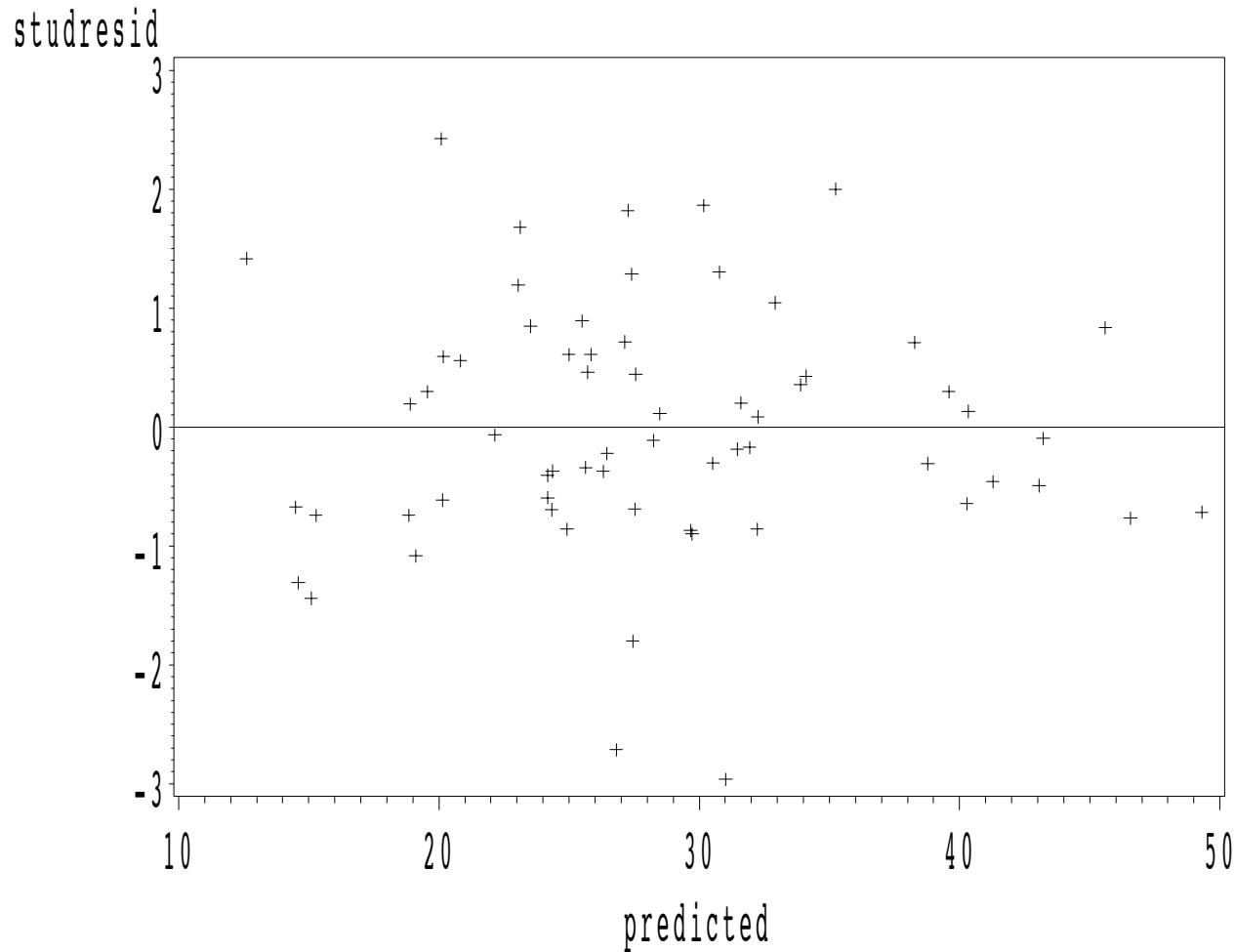
SSE for Ozone Data

π	SSE
2.0	37788
1.5	17671
1.0	10148
0.5	8578
0.0	15126
-0.5	69083
-1.0	637807
-1.5	8767407
-2.0	152261093

Let's examine the residuals for the untransformed data and for the square root transformed data. For the untransformed data, we have



and for the square root transformation, we have



Which transformation is best?

Alternatively, we can try the following SAS code:

```
proc transreg data=ozone ss2 details;
title2 Defaults;
model boxcox(personal) =
    identity(outdoor home time_out);
run;
```

Transformation Information
for BoxCox(personal)

Lambda	R-Square	Log Like
-3.00	0.05	-667.601
-2.75	0.05	-617.486
-2.50	0.05	-568.075
-2.25	0.05	-519.492
-2.00	0.05	-471.896
-1.75	0.05	-425.492
-1.50	0.05	-380.551
-1.25	0.06	-337.442
-1.00	0.07	-296.686
-0.75	0.09	-259.033

-0.50	0.14	-225.559
-0.25	0.19	-197.696
0.00	0.26	-176.954
0.25	0.31	-164.160
0.50	0.34	-158.805 < 
0.75	0.34	-159.364
1.00	0.33	-164.183
1.25	0.31	-171.987
1.50	0.29	-181.931
1.75	0.27	-193.471
2.00	0.25	-206.253
2.25	0.23	-220.036
2.50	0.21	-234.650
2.75	0.19	-249.966
3.00	0.17	-265.885

Comments

- Regression assumption diagnostics should guide model choice.
- Remember that Gaussian distribution, homogeneity, and linearity often stand or fall together.
- Aesthetics often dictate using parallel transformations, as for a baseline covariate (if we choose to transform personal ozone exposures, then we may also wish to transform home and outdoor exposures using the same transformation).
- Choosing some transformations (say $y^{\frac{3}{2}}$) will make models difficult to interpret.
- Ratio scale variables often need transformation.
- The nature of any zeros affects the type of transformation. Log and reciprocal transformations often work well for blood, urine, or water assays of natural compounds with non-zero background levels. Values below a detection threshold that are recorded as

zero are instead informatively censored and not missing or equal to zero. Non-natural compounds, such as some pharmaceuticals in blood, have true zeros.

- Some transformations may not work well for some data (we do not wish to take the square root of a negative number or the log of zero).

Avoid back transformation because it may lead to potential bias. For example, a scientist was studying a population with a chronic disease. A depression scale was used as the response in a GLM. Using the strategy described above, the statistician chose \sqrt{Y} . The scientist complimented the statistician on the analysis. However, in the draft manuscript, the scientists included means, standard deviations, and back transformed predicted values (in the original metric). The statistician convinced the scientist to stick with transformed data.
Why?

Nonlinear functions usually imply 

 $E[f(X)] \neq f(E[X]).$

For example, from Jensen's inequality we know that

$$E[Y^2] > (E[Y])^2.$$

Transforming Predictor Variables

Box and Tidwell suggest the following procedure to check whether a predictor should be transformed. To test whether a transform x^λ should be used in the model,

1. add the covariate $a_i = x_i \log(x_i)$ to a model already containing x_i
2. let $\hat{\gamma}$ be the estimated coefficient of a_i and test to see if it is significantly different from zero (using a t-test say)
 - (a) if not significant, no transform is needed
 - (b) if significant, a preliminary guess at the proper transform is given by $\hat{\lambda} = \frac{\hat{\gamma}}{\hat{\beta}} + 1$, where $\hat{\beta}$ is the estimated coefficient of x in the model containing both x and a .

Why does this technique work?

Consider a true model,

$$y = \alpha + \beta x^\lambda + \varepsilon.$$

We will use Taylor's theorem to expand x^λ around $\lambda = 1$. Recall from calculus that Taylor's theorem says

$$f(b) = f(a) + f'(a)(b - a) + \frac{f''(a)}{2!}(b - a)^2 + \dots$$

In addition, recall that $\frac{\partial}{\partial x} c^x = c^x \log(c)$ for $c > 0$. Using these facts,

$$\begin{aligned} f(\lambda) &= x^\lambda, & \frac{\partial}{\partial \lambda} f(\lambda) &= x^\lambda \log(x) \\ f(\lambda = 1) &= x, & f'(\lambda = 1) &= x \log(x) \end{aligned}$$

Using a first-order Taylor series expansion with $b = \lambda$, $a = 1$,

$$x^\lambda \approx x + (\lambda - 1)x \log(x).$$

Substituting this into the above model, we have

$$\begin{aligned}y &= \alpha + \beta(x + \lambda x \log(x) - x \log(x)) + \varepsilon \\&= \alpha + \beta x + \beta(\lambda - 1)x \log(x) + \varepsilon \\&= \alpha + \beta x + \gamma x \log(x) + \varepsilon,\end{aligned}$$

where $\gamma = \beta(\lambda - 1)$. If γ is not significant, then either $\beta = 0$ (no effect of the covariate x) or $\lambda = 1$ (no need for a transformation). If γ is significant, then we solve to get $\lambda = \frac{\gamma}{\beta} + 1$.

Next: Model Selection

Reading Assignment:

- Muller and Fetterman, Chapter 11: “Selecting the Best Model”
- Weisberg, Chapter 10: “Model Selection”

Lecture 14: Selecting the Best Model

Reading Assignment:

- Muller and Fetterman, Chapter 12: “Selecting the Best Model”
(Required)

Selecting the best model, while perhaps one of the most common data analysis tasks, is an **exploratory** activity that is usually accompanied by type I error rate inflation (every time we compare two models, we compromise the type I error rate of the model eventually chosen). A **confirmatory** analysis requires specifying the variables, model, and tests without knowledge of the data at hand. (This is typically the strategy used for FDA drug approval, a situation in which a great deal is known about the action of the drug through clinical trials and studies of related drugs so that such a model may be specified in advance.)

Often, investigators have many potential predictors and wish to find the best subset, and in such situations, it is important to bear in mind that p-values from hypothesis tests must be interpreted with caution in an exploratory analysis. (Alternatively, one may wish to use a correction such as the Bonferroni correction for all hypothesis tests conducted.)

Model Selection Strategy Overview

Assuming that our goal is to use purely exploratory analysis to find the best-fitting model for the data at hand, our model selection strategy consists of four steps.

1. Specify the maximum model under consideration.
2. Specify a criterion for model selection.
3. Specify a strategy for applying the criterion.
4. Conduct the analysis.

In addition, if we wish to have confidence in our model for data outside the current sample, we should evaluate the reliability of the model chosen. In a purely exploratory analysis, one often simply interprets significance of coefficients in the final model with caution, refraining from using $p \leq 0.05$ after conducting a series of tests on the data in order to choose that final model. (The practice of doing this anyway is widespread.) However, other methods may be used to assess

model reliability more formally, and we will discuss two of these, split-sample analysis and cross-validation more generally, in detail.

Specifying the Maximum Model

We will use *maximum model* or *full model* to refer to the model with the greatest number of predictors of any model. In model selection, we will consider more parsimonious models than this model. In some cases, models under consideration will be *nested* in the full model, which means that they can be created simply by deleting variables from the full model. In other cases, smaller models may not be nested in the maximum model. A common goal is to find a more parsimonious model that still describes the data (almost) as well as the full model.

When selecting the full model, one must be sure not to select a model that is too large. As a bare minimum, we need $n - p > 0$, where p is the number of columns in \mathbf{X} . In order to have fairly stable estimates, we would want $n > 5p$ or $n > 10p$. 

Consider a model with 5 covariates, x_1, x_2, x_3, x_4, x_5 , and an intercept. Several possible “full models” are listed below.

$$E(\mathbf{y}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \quad (1)$$

$$\begin{aligned} E(\mathbf{y}) = & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_2^2 \\ & + \beta_7 x_3^2 \end{aligned} \quad (2)$$

$$E(\mathbf{y}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_1 x_2 \quad (3)$$

$$\begin{aligned} E(\mathbf{y}) = & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_1 x_2 \\ & + \beta_7 x_1 x_3 + \beta_8 x_1 x_4 + \beta_9 x_1 x_5 + \beta_{10} x_1 x_2^2 + \beta_{11} x_1 x_3^2 \end{aligned} \quad (4)$$

$$E(\mathbf{y}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 \log(x_5) \quad (5)$$

Suppose that we select Model 2 as our best model. Model 1 would be

-  a candidate model that is *nested* in Model 2, while Model 5 is a *non-nested* candidate model.

Specifying a Model Selection Criterion

Next, we need to choose a criterion to help us define which model is “best”. Some criteria are used to evaluate models with respect to a null or intercept-only model, while other criteria are used to evaluate models with respect to the full model or any larger model (this latter class of criteria is restricted to models *nested* in the full model).

Comparing a Candidate Model to the Null or Intercept Model

Adjusted R^2

One intuitive model selection criteria is R^2 . Recall that

$$R^2 = \frac{SSE(\beta_0) - SSE(\beta_0, \dots, \beta_{p-1})}{SSE(\beta_0)}$$

and provides us with an estimate of the percent of variability explained by the model under consideration. One problem with R^2 is that it never decreases when additional variables are added to a model, so that it will always favor selection of the largest model. An *adjusted R^2* is given by

$$R_A^2 = 1 - \frac{n-1}{n-p}(1 - R^2),$$

where p is the number of columns of \mathbf{X} . This adjusts the usual R^2 for the number of covariates, adding a penalty for models with “too many” covariates. Typically, the model with the largest R_A^2 is said to be the best model.

F Tests

The test of corrected overall regression, given by

$$F_p = \frac{[SSE(\beta_0) - SSE(\beta_0, \dots, \beta_p)]/(p-1)}{MSE(\beta_0, \dots, \beta_p)},$$



may be used to test whether the model under consideration offers significant improvement over the intercept-only model. For models without an intercept, the test of uncorrected overall regression may be used to test whether the model under consideration offers significant improvement over the null model.

Single Model Criteria: AIC and SBC/BIC

The Akaike Information Criterion (AIC) and the Schwarz Criterion or Bayesian Information Criterion (SBC in SAS and BIC many other places) are general metrics. In linear regression,

$$AIC = n \log \left(\frac{SSE}{n} \right) + 2p$$

$$SBC = n \log \left(\frac{SSE}{n} \right) + \log(n)p. \quad \square$$

For both criteria, **smaller is better.** Increasing the SSE increases the AIC and SBC, while increasing the number of predictors also increases each measure. AIC tends to favor models that are too large, while the SBC places a greater penalty on larger models. These criteria allow comparison of nested and non-nested models. NOTE: In older versions of SAS, some procedures used “bigger is better” definitions of AIC and SBC, so take care if you are using SAS versions before 8.1.

Comparing a Candidate Model to the Full Model

Mallows C_p

Mallows C_p compares a candidate model to the full model and is given by

$$C_p = \frac{SSE(\beta_0, \dots, \beta_{p-1})}{MSE(\text{full})} - n + 2p.$$

Based on the rationale that a model not omitting useful variables should have $SSE(\beta_0, \dots, \beta_{p-1}) \approx (n - p)\sigma^2$, we seek a model with $C_p \approx p$. 

F Tests

A groupwise test of a candidate model versus the full model, given by

$$F_p = \frac{[SSE(\beta_0, \dots, \beta_{p-1}) - SSE(\text{full})]/[df(\text{candidate}) - df(\text{full})]}{MSE(\text{full})},$$

may be used to test whether the model under consideration is sufficiently close to the reduced model, provided that the model under consideration is nested in the full model. (Note that this test may be used to test the adequacy of any model nested in a larger model.)

Specifying the Selection Strategy

All Possible Regressions

Suppose our full model contains p predictors. An *all possible regressions* strategy involves fitting all possible models. In this case, each of the p predictors (including the intercept) could be in or out of the model, so that we would consider 2^p models in our selection process (including a null model). This strategy may get out of hand very quickly, since there are $2^5 = 32$ possible models with 5 columns in \mathbf{X} , $2^8 = 256$ possible models with 8 columns in \mathbf{X} , and $2^{10} = 1024$ possible models with 10 columns in \mathbf{X} . Although SAS will print the “top” models from all possible ones with using some criteria (R^2 for example), the time required for an all possible regressions strategy for other criteria (such as F tests against a maximum model) make this strategy infeasible when p is large.

Backward Elimination

Backward elimination begins with the full model and deletes variables with little value. The procedure is described below.

1. Specify the full model and set $p = p_{full}$, the number of columns of the \mathbf{X} matrix in the full model.
2. Fit all $p - 1$ variable models defined by deleting a single variable from the base model. (Typically, we do not consider eliminating the intercept.)
3. For each model, compute an added-last test for the candidate variable.
4. Find the minimum F statistic out of the set of $p - 1$ models (maximum p-value).
 - (a) If the corresponding variable is “significant” at a specified level  (typically ranging from 0.05 to 0.20), stop and select the model with p predictors as the best model.

-
- (b) If the corresponding variable is not significant, delete the variable in question and set $p = p - 1$.
5. Repeat until a model is chosen.
- Forward Selection** 
- Forward selection* begins with the intercept-only (or null) model and adds variables with predictive value. The procedure is described below.
1. Fit the intercept-only (or null) model as the base model so that $p = 1$.
 2. Fit all $p + 1$ variable models defined by adding a single variable to the base model.
 3. For each model, compute an added-in-order test for the candidate variable.
 4. Find the maximum F statistic out of the set of models (minimum p-value).
 - (a) If the corresponding variable is “significant” at a specified level

(typically ranging from 0.05 to 0.20), add the variable in question and set $p = p + 1$.

- (b) If the corresponding variable is not significant, stop and choose the model with p predictors.

5. Repeat until a model is chosen.



Stepwise Selection

Strictly speaking, *stepwise selection* refers to a selection procedure with both forward and backward steps so that addition and deletion of variables may be considered. However, this term is sometimes loosely used to refer to forward selection or backward elimination as well.

SAS implementation: The stepwise method is a modification of the forward-selection technique and differs in that variables already in the model do not necessarily stay there. As in the forward-selection method, variables are added one by one to the model, and the statistic for a variable to be added must be significant at the SLENTRY= level. After a variable is added, however, the stepwise method looks at all



the variables already included in the model and deletes any variable that does not produce an statistic significant at the SLSTAY= level. Only after this check is made and the necessary deletions are accomplished can another variable be added to the model.

Note: It is important to note that these three strategies do not consider all possible models and might actually miss a "best" model. In addition, it is entirely possible that they may select three different models as "best" models. Famous quote from George Box: all models are wrong, but some are useful.



```
data a; do i = 1 to 500;
x1 = 10 + 5*rannor(0); * Normal(10, 25);
x2 = exp(3*rannor(0)); * lognormal;
x3 = 5+10*ranuni(0); * uniform;
x4 = 100 + 50*rannor(0); * Normal(100, 2500);
x5 = x1 + 3*rannor(0); * normal bimodal;
x6 = 2*x2 + ranexp(0);
* lognormal and exponential mixture;
x7 = 0.5*exp(4*rannor(0)); * lognormal;
x8 = 10 +8*ranuni(0); * uniform;
x9 = x2 + x8 + 2*rannor(0);
* lognormal, uniform and normal mix
x10 = 200 +90*rannor(0); * normal(200, 8100);
*x10 = 5*x2 + rannor(0);
y = 3*x2 - 4*x8 + 5*x9 + 3*rannor(0);
* true model with no intercept term;
output; end;
```



```
/*run all possible models*/
proc reg data=a outest=est;
model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 /
selection=adjrsq sse aic ;
output out=out p=p r=r; run; quit;
```

```
/*run all possible models without intercept*/
proc reg data=a outest=est0;
model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 /
noint selection=adjrsq sse aic ;
output out=out0 p=p r=r; run; quit;
```

```
/*forward selection*/ 
proc reg data=a outest=est1;
model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 /
slentry=0.15 selection=forward
ss2 sse aic;
```

```
output out=out1 p=p r=r; run; quit;

/*backward selection*/
proc reg data=a outest=est2;
model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 /
slstay=0.15 selection=backward
ss2 sse aic;
output out=out1 p=p r=r; run; quit;

/*stepwise selection*/ 
proc reg data=a outest=est3;
model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 /
slstay=0.15 slentry=0.15 selection=stepwise
ss2 sse aic;
output out=out3 p=p r=r; run; quit;

/*stepwise group selection*/
```

```
proc reg data=a outest=est4;
model y={x1 x2} x3 x4 x5 x6 x7 x8 x9 x10 /
selection=stepwise slstay=0.15 slentry=0.15
groupnames=x1 x2 x3 x4 x5 x6
x7 x8 x9 x10;
```

Comments

1. In any selection strategy, one must take steps to ensure that nonsensical models do not emerge as “winners”. For example, if x , x^2 , and x^3 are in the full model, then we do not wish to allow a backward selection strategy to delete x while letting x^2 and x^3 remain in the model. In such cases, we will define variable “groups” to be tested as units. In this case, a logical strategy would be to treat (x, x^2, x^3) as a group. In backward elimination, we would first test the group. If it was significant, we might then move to test x^3 , and if significant, we would retain all three variables in the model.
2. The p-values obtained using a series of F tests should not be viewed strictly but merely as guides due to inflated type I error rates due to multiple testing.
3. Ideally, the best error term is the one from the full model because it should have the least bias. In forward selection in particular, the

estimate of the error term may be biased upwards, leading the process to select a final model that is too small.

4. A group-wise, backward selection strategy based on F tests is superior in a wide variety of settings.



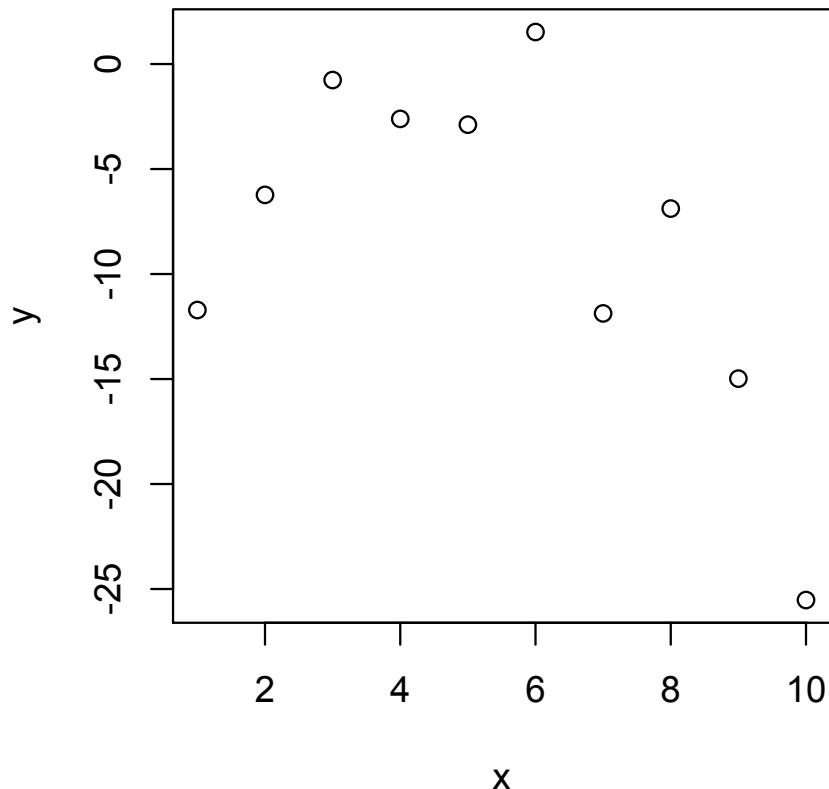
Fixed Tests Model Selection Strategy

If one wishes to avoid “data dredging” and to better preserve the type I error rate, a fixed tests implementation should be considered. This strategy is outlined below.

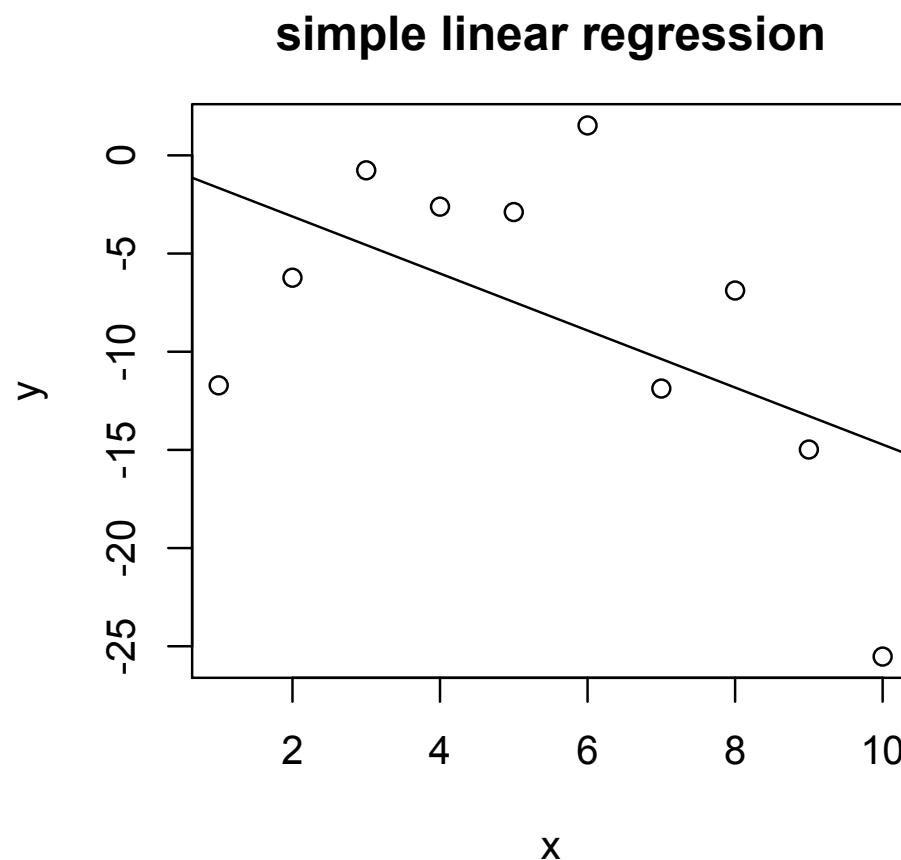
1. Very thoughtfully specify the maximum model.
2. Select groups and use a groupwise strategy, selecting a sequence for testing models and adjusting α for tests.
3. Assess diagnostics.
 - (a) Fit the full model and check for collinearity.
 - (b) Parsimoniously change or reduce the full model to eliminate collinearity.
 - (c) Conduct assumption diagnostics.
4. Conduct the *planned* tests.
5. Conduct assumption diagnostics on the resulting reduced model.

Assessing Reliability

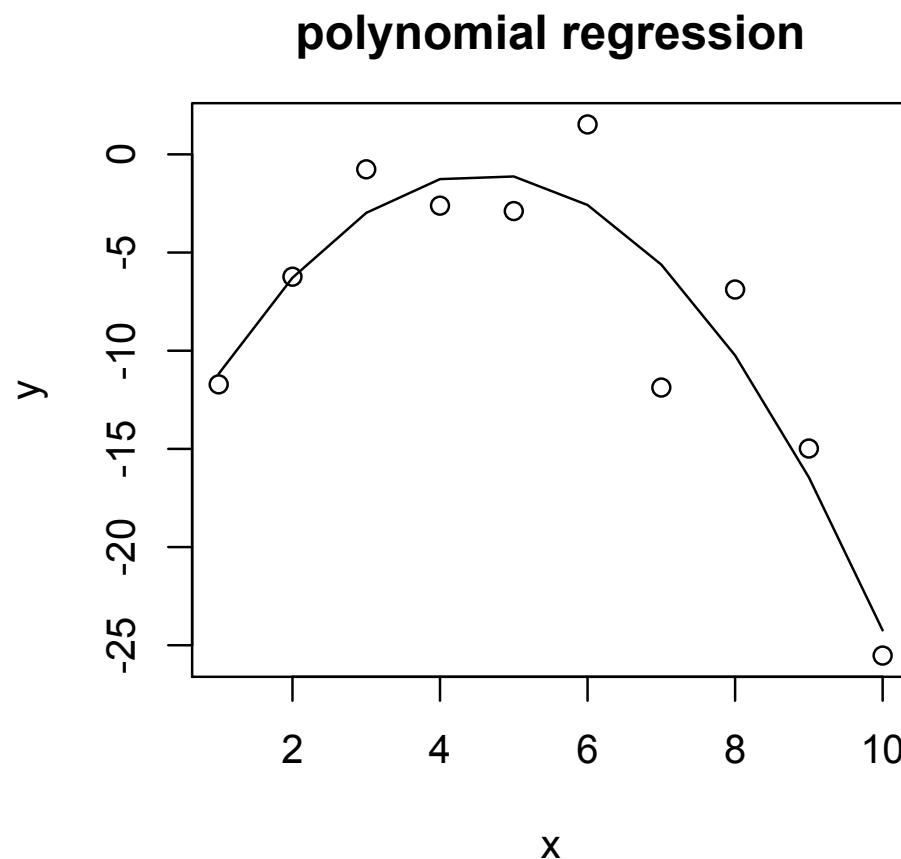
Why is assessing reliability an important step in model fitting?
Consider the following data.



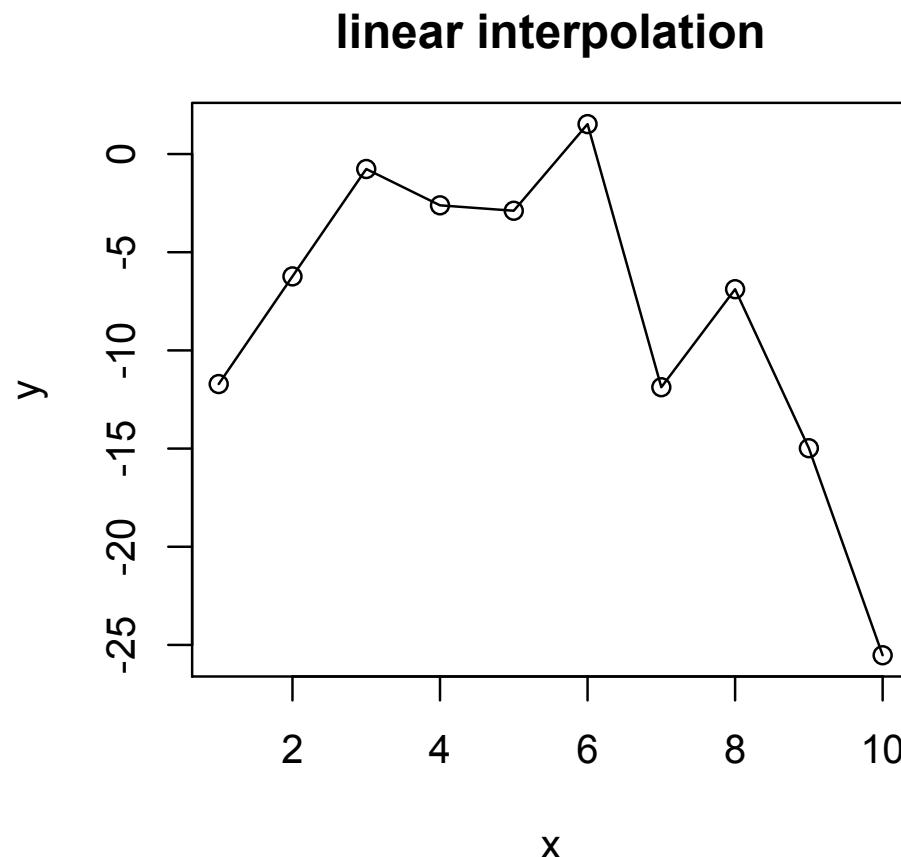
One possible model for the data is a simple linear regression, shown below.



Another approach is a regression model with linear and quadratic terms in x .



A third approach is a linear nonparametric approach (“connect the dots”).



Why not choose the model with the best fit to the data?

How well will the model predict future data from the same distribution?

Split-Sample Analysis

The simplest model validation method is one-time *data-splitting*. In this method, the data are split into *training* (model development) and *test* (model validation) samples by a random process. This splitting must occur before any analysis begins. Then, regression is performed on the training set, with future performance evaluated in the test set.

The split-sample procedure is outlined below.

1. Before looking at the data, split the sample using a random process (this process may be stratified by subgroups defined by values of \mathbf{X} such as smoking status or race). The split fraction  could be 50-50(so that half the data fall into the training sample and half fall into the test sample). For large data sets, often a fraction as small as 10% could be placed in the training sample while still providing a sufficient sample size for exploratory data analysis. If the sample size is too small to allow a reasonable fraction in the training sample, the value of a split-sample

approach will be greatly diminished.

2. Conduct any and all desired exploratory analyses on the training data, including diagnostics.
3. Based on the exploratory analyses, select a “best” model based on selected criteria. The parameter estimates for this model will be called $\hat{\beta}_1$.
4. Compute the predicted values \hat{y} for this model and call them \hat{y}_1 .
5. Compute the squared multiple correlation, $R_1^2 = r^2(y_1, \hat{y}_1)$ for this model.
6. Compute the cross-validation correlation.
 - (a) Compute the predicted values in the test sample as $\hat{y}_2 = \mathbf{X}_2 \hat{\beta}_1$. (That is, use the $\hat{\beta}$ from the training data with the covariates from the test data to compute predicted values.)
 - (b) Compute the squared cross-validation correlation as $R_{*2}^2 = r^2(y_2, \hat{y}_2)$.

-
- (c) Compute the estimated shrinkage on cross-validation as $R_1^2 - R_{*2}^2$. Smaller shrinkages are better. (One rule of thumb is that shrinkage < 0.05 indicates results are generalizable and that shrinkage > 0.10 is cause for concern.) Often, one computes the proportion relative shrinkage by dividing by R_1^2 .
 - (d) The MSE may also be used to evaluate future performance.



-
1. Conduct regression diagnostics on the pooled data.
 - (a) If shrinkage is “small enough,” pool the data to provide best available estimates of β .
 - (b) If shrinkage is “poor,” pool data to conduct a second-round *exploratory* analysis.
 2. Report results honestly.

One big advantage of data-splitting is that hypothesis tests are confirmed in the test sample. In addition, this method is simple to implement. However, it has several disadvantages, including the following.

- Data-splitting greatly reduces the sample size for both model development and model testing.
- Different splits may lead to different results, and with small sample sizes, our test set may just be lucky or unlucky.
- Data-splitting does not validate the final model but rather a model

developed on only a subset of the data. The training and test sets are recombined for fitting the final model, which is not validated.

Cross-Validation

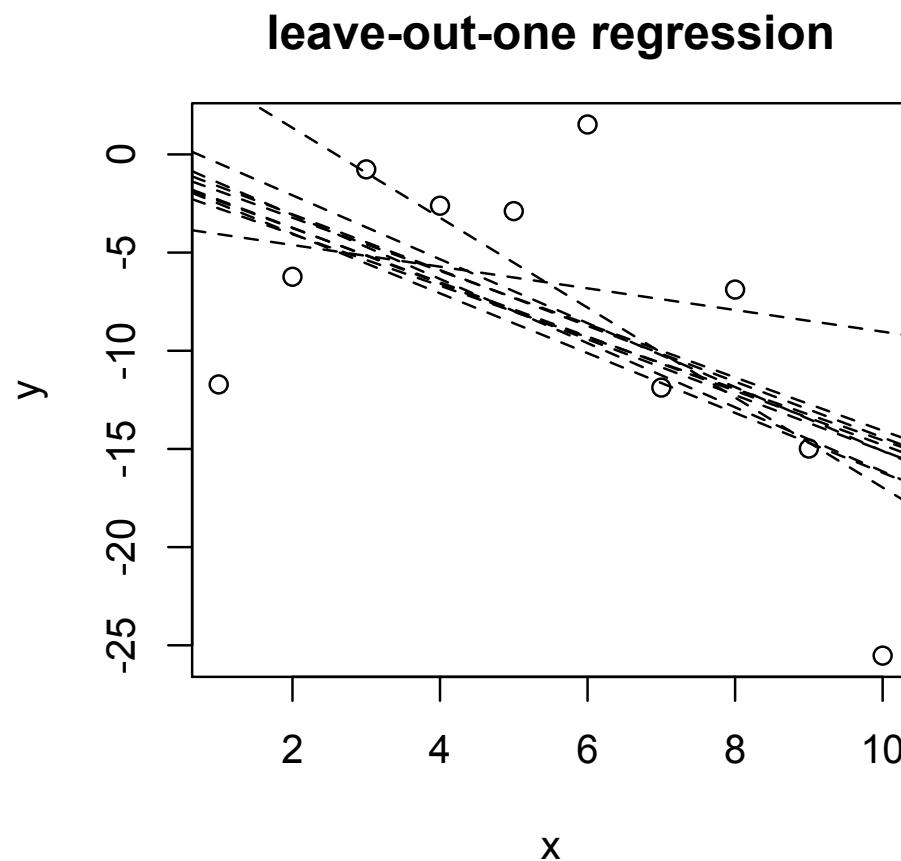
Cross-validation is a generalization of data-splitting that overcomes some of its disadvantages.

Leave-out-one cross-validation is similar to jackknifing. In leave-out-one cross-validation, one observation is omitted from the analysis, and the response for that one subject is predicted using a model obtained from the remaining $n - 1$ subjects. Note the error of that prediction. When you've done this n times (once for each subject), report the mean squared error over all the predictions. This method does not “waste” data, but is computationally expensive.





Consider leave-out-one cross-validation for the previous data. Omitting the j^{th} observation, $j = 1, \dots, 10$, we have the lines below for the linear models.



The errors are calculated as the distance from the fitted line omitting each subject to each subject's observed outcome. Summing these squared distances, we obtain SSE . The estimated error from the linear model can be compared to that of the quadratic and "connect-the-dots" models to see which model is preferred.

K-fold cross validation randomly breaks the dataset into k partitions. For each partition, train on the points not in the partition, and find the test-set errors for the partition of interest. Repeat for all k partitions, and then report the mean squared error over all the predictions. This type of cross-validation (often done for $k = 10$) is a compromise between the leave-out-one and the split sample approaches.

Final Comments

Model selection and validation tend to be personal topics with different investigators preferring a variety of different methods. The major consensus is that exploratory analyses must be treated with caution, and test results must be presented honestly (and not as “confirmatory” when extensive model exploration has been carried out).

Frank Harrell's comments:

Here are some of the problems with stepwise variable selection.

- It yields R-squared values that are badly biased to be high.
- The F and chi-squared tests quoted next to each variable on the printout do not have the claimed distribution.
- The method yields confidence intervals for effects and predicted values that are falsely narrow.
- It yields p-values that do not have the proper meaning, and the proper correction for them is a difficult problem.

-
- It gives biased regression coefficients that need shrinkage (the coefficients for remaining variables are too large).
 - It has severe problems in the presence of collinearity.
 - It is based on methods (e.g., F tests for nested models) that were intended to be used to test prespecified hypotheses.
 - Increasing the sample size doesn't help very much.

For Post-Selection Statistical Inference, you may check Robert Tibshirani's talk.

Coding Schemes for Regression

Reading Assignment:

- Muller and Fetterman, Chapter 12: “Coding Schemes for Regression” (Required)

Goals

1. Understand various coding schemes for ANOVA and relationships

between ANOVA and multiple regression models.

2. Understand basic strategy and techniques of multiple comparisons.
3. Recognize the impact of various amounts of missing data.

The study of ANOVA is motivated by desire to model and test hypotheses about two or more group means.

Regression or ANOVA?

In theory and computation, ANOVA is a special case of regression analysis with dummy variables as predictors.

An *indicator* or *dummy* variable is used to represent group membership. Suppose we are interested in the effect of two weight loss regimens, diet and exercise, over a period of two months, compared to subjects on neither regimen (control subjects). Then we can create three dummy variables to denote regimen membership as follows.



$$x_1 = \begin{cases} 1 & \text{diet} \\ 0 & \text{exercise or neither} \end{cases}$$
$$x_2 = \begin{cases} 1 & \text{exercise} \\ 0 & \text{diet or neither} \end{cases}$$
$$x_3 = \begin{cases} 1 & \text{neither} \\ 0 & \text{diet or exercise} \end{cases}$$

For analysis, we will use (x_1, x_2, x_3) (or some combination of them) in our **X** matrix to represent the diet regimens.

The name ANOVA reflects the expression of tests in terms of variances and the nature of computational strategies predating computers.

An ANOVA *coding scheme* defines a set of rules for representing all of the information in one or more categorical variables as one or more interval variables. We will discuss the following five coding schemes:

1. reference cell,
2. cell mean,
3. classical ANOVA,
4. effect, and
5. polynomial.

Each coding scheme merits consideration as well as a description of its relationship to other schemes. This chapter only considers estimation, and all consideration of testing will be left to subsequent chapters.

Classical ANOVA Coding

Classical ANOVA coding uses a less than full rank \mathbf{X} matrix ($\text{rank}(\mathbf{X}_{n \times (p+1)}) = p$) intentionally. This \mathbf{X} matrix is equal to the cell mean coding \mathbf{X} matrix with the addition of a column of 1's for the



intercept. The model is given by



$$\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times (p+1)} \boldsymbol{\beta}_{(p+1) \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & & & & \vdots \\ 1 & 1 & 0 & \cdots & \cdots & \cdots & \vdots \\ 1 & 0 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & & & \vdots \\ 1 & 0 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where \mathbf{X} has rank p (less than full rank).

Often, we represent $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$ as $\boldsymbol{\beta} = (\mu, \alpha_1, \dots, \alpha_p)$.

Expected Values for Classical ANOVA Coding

Group	$E[y_i]$	Mean
1	$1 \cdot \beta_0 + 1 \cdot \beta_1 + \cdots + 0 \cdot \beta_{p-1} = \beta_0 + \beta_1$	μ_1
2	$1 \cdot \beta_0 + 0 \cdot \beta_1 + 1 \cdot \beta_2 + \cdots + 0 \cdot \beta_{p-1} = \beta_0 + \beta_2$	μ_2
\vdots	\vdots	\vdots
p	$1 \cdot \beta_0 + 0 \cdot \beta_1 + \cdots + 1 \cdot \beta_p = \beta_0 + \beta_p$	μ_p

Group 1 is the reference group, and $\beta_0 + \beta_1 = \mu_1$ is the mean for group 1. The difference in mean response between group 2 and group 1 is $(\mu_2 - \mu_1) = ((\beta_0 + \beta_2) - (\beta_0 + \beta_1)) = (\beta_2 - \beta_1)$, and the difference in mean response between group 2 and group p is $(\mu_2 - \mu_p) = ((\beta_0 + \beta_2) - (\beta_0 + \beta_p)) = (\beta_2 - \beta_p)$.

Because \mathbf{X} is not full rank, our estimates of $\boldsymbol{\beta}$ are not unique because $\boldsymbol{\beta}$ is not estimable. However, we may obtain a unique solution by imposing an additional constraint. A common constraint is to require $\sum_{j=1}^p \beta_j = 0$. This ensures that β_0 is the grand mean because

$$\begin{aligned}\frac{(\beta_0 + \beta_1) + (\beta_0 + \beta_2) + \cdots + (\beta_0 + \beta_p)}{p} &= \frac{p\beta_0 + \sum_{j=1}^p \beta_j}{p} \\ &= \beta_0.\end{aligned}$$

Many authors describe $\beta_0 = \mu$ as the grand mean.

Classical ANOVA coding leads to a less than full rank model that may be numerically unstable. Because all parameters are not estimable, we must use the theory for the less than full rank model. However, it is easier just to use a full rank coding scheme.

SAS uses classical ANOVA coding but imposes the constraint that $\beta_j = 0$ for one j (a *reference cell* constraint).

Note that parameter estimates under different constraints may have different meanings. For example, $\hat{\beta}_0$ under the first constraint is the grand mean, while $\hat{\beta}_0$ under the second constraint is the mean for the reference group.

Reference Cell Coding

One Group

Suppose that a geneticist wishes to address the impact of a potentially toxic environmental exposure on mice. The intercept-only model assumes that all responses differ randomly from a common mean response, called the *grand mean*. Thus we have the model

$$\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times 1} \boldsymbol{\beta}_{1 \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$$
$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} \beta_0 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

This ANOVA model is often called the *grand mean model*. In this model, \mathbf{X} is full rank, and $E[y_i] = \beta_0$. Often, we replace β_0 with μ so that $E[y_i] = \mu$.

In this model,

$$\begin{aligned}\widehat{\beta}_0 &= (\mathbf{J}'_n \mathbf{J}_n)^{-1} \mathbf{J}'_n \mathbf{y} \\ &= n^{-1} \sum_{i=1}^n y_i \\ &= \bar{y}.\end{aligned}$$

Two Groups

Now suppose that the mice are from different mouse strains so that n_1 of the mice are black six mice, and $n - n_1 = n_2$ of the mice are Swiss

albino mice. Then we have the model

$$\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times 2} \boldsymbol{\beta}_{2 \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$$
$$\begin{bmatrix} y_1 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where

$$\mathbf{X} = \begin{bmatrix} \mathbf{J}_{n_1 \times 1} & \mathbf{0}_{\mathbf{n}_1 \times 1} \\ \mathbf{J}_{n_2 \times 1} & \mathbf{J}_{n_2 \times 1} \end{bmatrix}_{n \times 2}.$$

A dummy (indicator) variable indicates the species for each



observation; that is,

$$x_1 = \begin{cases} 0, & \text{black six} \\ 1, & \text{Swiss albino} \end{cases} .$$

Predictor values assigned by the scientist are often called treatment levels.

Three or More Groups

For p mouse species, we have the model

$$\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$$
$$\begin{bmatrix} y_1 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & & & & \vdots \\ 1 & 0 & \cdots & \cdots & \cdots & \vdots \\ 1 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & & & \vdots \\ 1 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where \mathbf{X} is full rank p .

Expected Values for Reference Cell Coding

Group	$E[y_i]$	Group Mean
1	$1 \cdot \beta_0 + 0 \cdot \beta_1 + \cdots + 0 \cdot \beta_{p-1} = \beta_0$	μ_1
2	$1 \cdot \beta_0 + 1 \cdot \beta_1 + \cdots + 0 \cdot \beta_{p-1} = \beta_0 + \beta_1$	μ_2
:	:	:
p	$1 \cdot \beta_0 + 0 \cdot \beta_1 + \cdots + 1 \cdot \beta_{p-1} = \beta_0 + \beta_{p-1}$	μ_p

Group 1 is the reference group, and $\beta_0 = \mu_1$ is the mean for group 1. The difference in mean response between group 2 and group 1 is $(\mu_2 - \mu_1) = ((\beta_0 + \beta_1) - \beta_0) = \beta_1$, and the difference in mean response between group 2 and group p is $(\mu_2 - \mu_p) = ((\beta_0 + \beta_1) - (\beta_0 + \beta_{p-1})) = (\beta_1 - \beta_{p-1})$.

Cell Mean Coding

With *cell mean coding*, all of the β 's equal group means. Cell mean coding is the most natural coding scheme and the easiest to understand (and the easiest to explain to investigators!). For p groups, we create p indicator variables and do not include an intercept in the

model, which is given by

$$\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$$
$$\begin{bmatrix} y_1 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & & & \vdots \\ 1 & 0 & \cdots & \cdots & \cdots & \vdots \\ 0 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & & & \vdots \\ 0 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \vdots \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where \mathbf{X} is full rank p .

Expected Values for Cell Mean Coding

Group	$E[y_i]$	Group Mean
1	$1 \cdot \beta_0 + 0 \cdot \beta_1 + \cdots + 0 \cdot \beta_{p-1} = \beta_0$	μ_1
2	$0 \cdot \beta_0 + 1 \cdot \beta_1 + \cdots + 0 \cdot \beta_{p-1} = \beta_1$	μ_2
:	:	:
p	$0 \cdot \beta_0 + 0 \cdot \beta_1 + \cdots + 1 \cdot \beta_{p-1} = \beta_{p-1}$	μ_p

The vector $\boldsymbol{\beta} = (\beta_0, \dots, \beta_{p-1})$ is often written $\boldsymbol{\beta} = (\mu_1, \dots, \mu_p)$ or $\boldsymbol{\beta} = (\alpha_1, \dots, \alpha_p)$.

Group 1 is the reference group, and $\beta_0 = \mu_1$ is the mean for group 1. The difference in mean response between group 2 and group 1 is $(\mu_2 - \mu_1) = (\beta_1 - \beta_0)$, and the difference in mean response between group 2 and group p is $(\mu_2 - \mu_p) = (\beta_1 - \beta_{p-1})$. Does this model span an intercept?

Effect Coding

Effect coding provides a useful coding scheme. This full rank scheme has design matrix \mathbf{X} with a column of 1's and $p - 1$ other columns, like the reference cell coding scheme. However, the *effect coding scheme* reassigned the value of 0 to -1 for dummy variable covariates for

subjects in the reference group. The model is given by

$$\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$$
$$\begin{bmatrix} y_1 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 & -1 & \cdots & -1 \\ \vdots & \vdots & & & & \vdots \\ 1 & -1 & -1 & -1 & \cdots & -1 \\ 1 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & & & \vdots \\ 1 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where \mathbf{X} is full rank p .

The effect coding indicator variable may be created in two steps.

1. Create a dummy variable indicating group membership, with the value 1 if the observation belongs to the group and 0 otherwise.
2. If the observation represents the reference group, then reassign the value of 0 to -1 .

Expected Values for Effect Coding

Group	$E[y_i]$	Mean
1	$1 \cdot \beta_0 - 1 \cdot \beta_1 - \cdots - 1 \cdot \beta_{p-1} = \beta_0 - \sum_{j=1}^{p-1} \beta_j$	μ_1
2	$1 \cdot \beta_0 + 1 \cdot \beta_1 + \cdots + 0 \cdot \beta_{p-1} = \beta_0 + \beta_1$	μ_2
:	:	:
p	$1 \cdot \beta_0 + 0 \cdot \beta_1 + \cdots + 1 \cdot \beta_{p-1} = \beta_0 + \beta_{p-1}$	μ_p

Group 1 is the reference group, and $\beta_0 - \sum_{j=1}^{p-1} \beta_j = \mu_1$ is the mean for group 1. The difference in mean response between group 2 and group 1 is $(\mu_2 - \mu_1) = ((\beta_0 + \beta_1) - (\beta_0 - \sum_{j=1}^{p-1} \beta_j)) = 2\beta_1 + \sum_{j=2}^{p-1} \beta_j$, and the difference in mean response between group 2 and group p is $(\mu_2 - \mu_p) = ((\beta_0 + \beta_1) - (\beta_0 + \beta_{p-1})) = (\beta_1 - \beta_{p-1})$.

Note that the mean of all group means is

$$\frac{(\beta_0 - \sum_{j=1}^{p-1} \beta_j) + (\beta_0 + \beta_1) + \cdots + (\beta_0 + \beta_p)}{p} = \frac{p\beta_0}{p} = \beta_0.$$

This parameter is the mean of the particular cells in the current design. This is different from the grand mean parameter in classical ANOVA coding (when there is no restriction), which represents the hypothetical mean of all observations in the population.

Relationships Among Coding Schemes

Any full rank X may be expressed as a full rank linear transform of any other (with both based on the same categorical predictors). Any parameter estimable or testable in one coding scheme is also estimable or testable in any other.

With p groups, only p parameters are estimable.

Parameters for any coding scheme may be expressed as linear functions of cell means.

Next: One-way ANOVA

Reading Assignment:

- Muller and Fetterman, Chapter 13: “One-Way ANOVA”

Lecture 16: One-Way ANOVA

Reading Assignment:

- Muller and Fetterman, Chapter 13: “One-Way ANOVA”

We use analysis of variance (ANOVA) to answer questions like the following.

- Do two or more groups differ in mean response?
- Does the new drug reduce symptoms when compared to existing drugs?
- Does group membership differentially predict response?

Now we shift from considering continuous predictors to considering categorical predictors. ANOVA models are a type of GLM used with

categorical predictors.

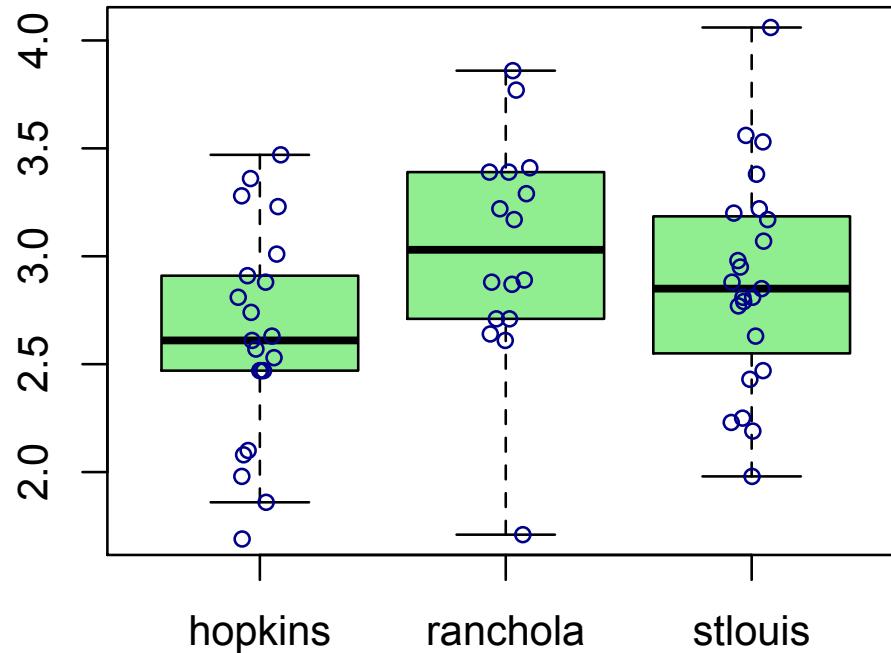
Specification of the Model

We want to determine whether two or more groups differ in location. The Gaussian errors assumption reduces this task to testing equality of means, generalizing Student's T test to three or more groups.

Example: Testing Equality of Cell Means

Pagano and Gauvreau (1993) present lung function data for heart patients at three medical centers: Johns Hopkins University, Rancho Los Amigos, and St. Louis University. For each center, investigators measured lung function as FEV_1 , the forced expiratory volume in 1 second in liters.

Below is a plot of the FEV_1 for each subject by center.



The FEV_1 of patients across centers does differ, but there is also a lot of variation within center. While Rancho Los Amigos does have the highest lung function on average, some patients at Johns Hopkins have higher FEV_1 than some of the Rancho Los Amigos patients, for example.

We need a statistical test to determine whether this is a statistically significant difference. In ANOVA (as well as regression more generally), we ask questions about *means* of groups by analyzing their *variances*...how does this work?

GLM Assumptions in ANOVA

Assumption	ANOVA Interpretation
H	within cell / between cell 
I	check sampling scheme 
L	automatically OK, given design
E	no problem in practice
Gauss	within cell 

Violation of the independence assumption may inflate α greatly. We must verify independence with careful thought about the sampling scheme, e.g., one individual may be sampled twice. To evaluate the assumption of Gaussian distribution of residuals, we may look at the Gaussian errors assumption within each group separately if the sample size in each group is large enough (≥ 30 say). Otherwise, we may look at studentized residuals pooled over all the groups.

The Primacy of Cell Means

Cell mean coding will provide our default choice, reflecting our primary interest in comparing means across groups. Recall that for cell mean coding with G groups, we create G indicator variables and do not include an intercept in the model, which is given by

$$\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times G} \boldsymbol{\beta}_{G \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & & & \vdots \\ 1 & 0 & \cdots & \cdots & \cdots & \vdots \\ 0 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & & & \vdots \\ 0 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \vdots \\ \beta_{G-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where \mathbf{X} is full rank G . Matrix \mathbf{X} can be summarized with the essence matrix, the matrix created by deleting any duplicate rows from the design matrix. The essence matrix helps us to determine rank and the

relationships between parameter definitions. What is the essence matrix of \mathbf{X} ? (answer: $\text{Es}(\mathbf{X}) = \mathbf{I}$).

Often, double indices are used to denote y 's by group as follows.



$$\begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{Gn_G} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & & & \vdots \\ 1 & 0 & \cdots & \cdots & \cdots & \vdots \\ 0 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & & & \vdots \\ 0 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \vdots \\ \beta_{G-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Parameters in other codings are linear transformations of cell means.

Consider **cell mean coding** for G groups, with $n = \sum_{g=1}^G n_g$. Then

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n_1 & & & & 0 \\ & n_2 & & & \\ 0 & & \ddots & & \\ & & & & n_G \end{bmatrix}_{G \times G}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} \sum_{j=1}^{n_1} y_{1j} \\ \sum_{n=1}^{n_2} y_{2j} \\ \vdots \\ \sum_{n=1}^{n_G} y_{Gj} \end{bmatrix}_{G \times 1}$$

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \begin{bmatrix} \bar{y}_{1\cdot} \\ \bar{y}_{2\cdot} \\ \vdots \\ \bar{y}_{G\cdot} \end{bmatrix}_{G \times 1}$$

Note that in *cell mean* coding, our estimates are the sample means for each group! In other coding schemes, estimates are functions of cell means.

$$\hat{\boldsymbol{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \begin{bmatrix} \mathbf{J}\bar{y}_{1\cdot} \\ \mathbf{J}\bar{y}_{2\cdot} \\ \vdots \\ \mathbf{J}\bar{y}_{G\cdot} \end{bmatrix}_{n \times 1}$$

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \begin{bmatrix} y_{11} - \bar{y}_{1\cdot} \\ \vdots \\ y_{1n_1} - \bar{y}_{1\cdot} \\ y_{21} - \bar{y}_{2\cdot} \\ \vdots \\ y_{Gn_G} - \bar{y}_{G\cdot} \end{bmatrix}_{n \times 1}$$

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}' \hat{\boldsymbol{\varepsilon}}}{(n - G)} = \frac{SSE}{(n - G)} = \frac{\sum_{i=1}^G \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2}{n - G}$$

(Usual) Overall Test

We wish to test the hypothesis that the means are the same for each group. For three groups, we have

$$H_0: \beta_0 = \beta_1 = \beta_2 \quad (\text{cell mean})$$

\Leftrightarrow

$$H_0: \beta_1 = \beta_2 = 0 \quad (\text{reference cell})$$

\Leftrightarrow

$$H_0: \beta_1 = \beta_2 = 0 \quad (\text{effect})$$

The overall test involves equality of G means, which implies $G - 1$ constraints so that $\boldsymbol{\theta}$ is $(G - 1) \times 1$ and $a = G - 1$.

We are testing $H_0: \boldsymbol{\theta} = \boldsymbol{\theta}_0$.

We state \mathbf{C} (unique to coding scheme), use $\boldsymbol{\theta}_0 = \mathbf{0}$, and then compute $SSH = (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathbf{M}^{-1} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ and $f_{obs} = (SSH/a)/\widehat{\sigma}^2$.

Reference Cell Coding

Recall that for reference cell coding with three groups, we have the model

$$\mathbf{y} = \mathbf{X} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \boldsymbol{\varepsilon},$$

where $\text{Es}(\mathbf{X}) = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$. To test the hypothesis of equal cell means, we have $\mathbf{C} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ implying that $\boldsymbol{\theta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$. We use $\boldsymbol{\theta}_0 = \mathbf{0}$.

Effect Coding

Recall that for effect coding with three groups, we have the model

$$\mathbf{y} = \mathbf{X} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \boldsymbol{\varepsilon},$$

where $\text{Es}(\mathbf{X}) = \begin{bmatrix} 1 & -1 & -1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$. To test the hypothesis of equal cell

means, we have $\mathbf{C} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ implying that $\boldsymbol{\theta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$. We use $\theta_0 = \mathbf{0}$.

If any $G - 1$ pairwise differences are zero, then all pairwise differences are zero, so the traditional overall test necessarily has $G - 1 = a$ degrees of freedom. All the above coding schemes give the same

p-value, F , and R^2 because they are full rank transformations of the rows; e.g., $\mathbf{C}_1 = \mathbf{T}\mathbf{C}_2$.

Traditional Source Table

Traditional ANOVA Source Table

Source	df	SS	Mean Square	F_{obs}	p
Between	$G - 1$		SSH/a		
Within	$n - G$		$\hat{\sigma}^2$		
(Error)					
Total	$n - 1$				
(Corrected)					

Estimating and Testing Cell Means in One-Way ANOVA

Using the essence matrix as the contrast matrix defines the cell means. For cell mean coding, the parameter estimates are the cell means. For other coding schemes, we use the parameter estimates provided to calculate the cell means.

Once you determine the cell means for each coding scheme, it is straightforward to test contrasts in the cell means.

Which Means are Different?

If we reject the null hypothesis that all means are equal, naturally we next want to know which means are in fact different.

There are $\binom{G}{2} = \frac{G(G-1)}{2}$ pairwise comparisons, and infinitely many others (if three or more means are involved).

An overall hypothesis that all means are equal tests all linear combinations of means and not just pairwise comparisons.

For the overall test with cell mean coding,

$$\mathbf{C} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}$$

creates a matrix of secondary parameters which are mean differences. Each row of \mathbf{C} defines a contrast matrix, a GLH, and an F test. Any

\mathbf{C} with a single row creates an F test with one numerator df and hence corresponds to a T .

$$\mathbf{C} = \begin{bmatrix} 1 & -1 & 0 \end{bmatrix} \quad H_0: \beta_0 = \beta_1$$

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & -1 \end{bmatrix} \quad H_0: \beta_0 = \beta_2$$

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & -1 \end{bmatrix} \quad H_0: \beta_1 = \beta_2$$

Contrasts

Describe a linear combination of cell means, $\sum_{i=1}^G c_i \mu_i$, as a **contrast** if $\sum_{i=1}^G c_i = 0$. Pairwise contrasts (pairwise comparisons) use coefficients such as $\mathbf{C} = [1 \ -1 \ 0]$ or $\mathbf{C} = [-1 \ 0 \ 1]$.

Complex contrasts, such as $\mathbf{C} =$

$$\left[-1 \quad \frac{1}{2} \quad \frac{1}{2} \right],$$

involve 3 or more means.

The null and alternative hypotheses in this case are

$$H_0 : \sum_{i=1}^G c_i \mu_i = 0$$

$$H_A : \sum_{i=1}^G c_i \mu_i \neq 0.$$

A planned (*a priori*) contrast requires choosing the number of contrasts and specifying the associated C matrices before seeing any data. Unplanned (*a posteriori*) contrasts include all others not specified in advance.

- ❑ For **balanced** data, a set of contrasts is **orthogonal** if CC' is diagonal (rows orthogonal).

Example: FEV_1 Data

The following SAS code creates the design matrix for **cell mean coding** and conducts the overall test of equality of group means for the lung function data.

```
data fev2;
set fev;
if center="hopkins" then h_ind=1; else h_ind=0;
if center="ranchola" then r_ind=1; else r_ind=0;
if center="stlouis" then s_ind=1; else s_ind=0;
run;

proc glm data=fev2;
model fev=h_ind r_ind s_ind/noint;
contrast Usual Overall Test h_ind 1 r_ind -1 s_ind  0,
                    h_ind 1 r_ind 0  s_ind -1;
run;
```

The results of the test are provided below.

The GLM Procedure

Dependent Variable: fev

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	482.5697439	160.8565813	633.19	<.0001
Error	57	14.4802561	0.2540396		
Uncorrected Total	60	497.0500000			
Source	DF	Type I SS	Mean Square	F Value	Pr > F
h_ind	1	144.8344048	144.8344048	570.13	<.0001
r_ind	1	147.1369000	147.1369000	579.19	<.0001
s_ind	1	190.5984391	190.5984391	750.27	<.0001
Source	DF	Type III SS	Mean Square	F Value	Pr > F
h_ind	1	144.8344048	144.8344048	570.13	<.0001
r_ind	1	147.1369000	147.1369000	579.19	<.0001
s_ind	1	190.5984391	190.5984391	750.27	<.0001
Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F

Usual Overall Test	2		1.58283723	0.79141861	3.12	0.0520
--------------------	---	---	------------	------------	------	--------

Parameter	Estimate	Standard		
		Error	t Value	Pr > t
h_ind	2.626190476	0.10998692	23.88	<.0001
r_ind	 3.032500000	0.12600585	24.07	<.0001
s_ind	2.878695652	0.10509614	27.39	<.0001

We see that there is a marginally significant difference among the cell means ($F = 3.12$ with 2 numerator and 57 denominator degrees of freedom yields $p = 0.05$). At least one center has average FEV that is different from the others.

What is the interpretation of the overall F test and the nine tests for the individual groups?

Step-Down Testing

Describe a test logically subsumed by another test as a **step-down test**. For example, having conducted an (overall) test for equality among a set of means, one may wish to step down to test equality of two particular means. In the population, the truth of the null hypothesis of the parent implies the truth of the null hypothesis of the child, and the truth of the alternative hypothesis for the child implies the truth of the alternative hypothesis for the parent.

Any linear combination of the rows of C represents a step-down test.

Properties of Cell Mean Estimates

If \mathbf{C} and \mathbf{X} are both orthogonal, then the contrast estimates $\hat{\theta}_k$ are uncorrelated and statistically independent (under HILE Gauss). This occurs when we estimate cell means with cell mean coding.

There are at most $G - 1$ mutually orthogonal contrasts (of mean differences) but infinitely many sets of contrasts are possible. There can be G mutually orthogonal contrasts if the set spans the grand mean. Recall that

$$\begin{aligned}\hat{\boldsymbol{\theta}} &\sim \mathcal{N}_a[\boldsymbol{\theta}, \sigma^2 \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{C}'] \\ &\sim \mathcal{N}_a[\boldsymbol{\theta}, \sigma^2 \mathbf{M}] \end{aligned}$$

Each row of \mathbf{C} defines a θ_k , with $\hat{\theta}_k \sim \mathcal{N}(\theta_k, \sigma^2 m_{kk})$ and

$$\frac{\hat{\theta}_k - \theta_0}{\sqrt{\sigma^2 m_{kk}}} \sim T(n - r),$$

where r is the rank of \mathbf{X} .

For cell mean coding the $\hat{\theta}_k$'s are independent for orthogonal C .

$\hat{\theta}_k$ tests are never exactly independent (for finite n) because they all use $\hat{\sigma}^2$.

We invert the test to create confidence intervals:

$$\hat{\theta}_k \pm (\hat{\sigma} \cdot \sqrt{m_{kk}}) \cdot t_{crit}$$

with $t_{crit} = F_T^{-1}(1 - \alpha/2; n - r)$

Conducting Multiple Comparisons

Choosing An Error Rate

The type I error rate for a collection of tests may be higher than α .

K independent tests with a nominal type I error rate of α lead to a probability of one or more false positives of $1 - (1 - \alpha)^K$. Using $\alpha = .05$ and $K = 10$ implies $1 - (1 - \alpha)^K \approx 0.40 >> 0.05$.

- For dependent tests, the Bonferroni inequality provides an upper bound of $K\alpha$, which equals 0.50 in this example. For the Bonferroni approach, test each comparison at $\alpha_k = \alpha/K = 0.05/10 = 0.005$.

Overview of Methods for Multiple Comparisons

Only a few of many techniques for controlling the error rate will be highlighted here.

The Bonferroni approach is completely general for planned (*a priori*) comparisons.

The development of alternative methods arose from a desire for

1. increased statistical power,
2. the ability to control unplanned comparisons, or
3. increased robustness to violation of GLM assumptions.

For a small number of planned contrasts, use the Bonferroni correction.

With an overall test size of .01 and four contrasts, $\alpha_k = \alpha/K$ implies a nominal size for each test of $.01/4 = .0025$.

Some authors recommend not using any correction in the special case of a set of planned and orthogonal contrasts because a significant overall test implies a trend of some sort, and the step-down tests merely need to identify the order of the trend. This approach reduces the chances of having a significant overall test with no particular trend significant.

Dunnett's (1955, 1964) test was designed to test each of $G - 1$ treatments against a control. Considering any other comparison with Dunnett's correction inflates test size.

Tukey's HSD (honest significant difference) method and generalizations of it (Tukey, 1953; Kramer, 1956) allow testing all pairwise comparisons.

Scheffé's (1953, 1959) test controls test size with any contrast, including complex contrasts (non-pairwise), whether planned or not. This test also protects against testing infinitely possible contrasts. It

has the desirable property of *coherence*. That is, the failure of the overall test ensures no significant step-down tests (e.g., mean comparison tests) exist, while a significant overall test ensures at least one significant step-down test.

For a small number of comparisons, a Bonferroni correction suffices. For a larger number or for unplanned comparisons, Scheffé's test is preferred.

Example Formulas

In general, $\hat{\boldsymbol{\theta}} \sim \mathcal{N}_a(\boldsymbol{\theta}, \sigma^2 \mathbf{M})$, with $\mathbf{M} = \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'$.

For a cell mean coding with G groups, $(\mathbf{X}'\mathbf{X}) = \text{Dg}(n_1, n_2, \dots, n_G)$, which implies $(\mathbf{X}'\mathbf{X})^{-1} = \text{Dg}(1/n_1, 1/n_2, \dots, 1/n_G)$. For $\mathbf{C}_k = \text{row}_k(\mathbf{C})$ and *cell mean coding*,

$$m_{kk} = \mathbf{C}_k(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'_k = \sum_{g=1}^G \frac{c_{kg}^2}{n_g}.$$

A Bonferroni correction for planned comparisons may be implemented with any linear models program merely by proper specification of the test size.

For other corrections, most of the work involves selecting the proper critical value based on determining the set of comparisons.

Scheffé's method for unplanned comparisons provides the most general and sturdy technique. Its downside is a lack of power.

Example: Testing Combinations of Means for FEV Data

We use the following SAS code to test all possible differences among cell means for the FEV data, using Scheffé's correction.

```
proc glm data=fev2;  
class center;  
model fev=center/noint;  
lsmeans center/ pdiff adjust=scheffe;  
run;
```

The LSMEANS statement produces the least squares estimates of CLASS variable means, where CLASS is used to indicate to SAS which variables are categorical. You may user the ORDER= option in the proc glm statement to alter how SAS chooses the reference level in proc glm (several options). In one-way ANOVA, these are just the usual means. In higher-way ANOVA models, the least squares means are simply the means of group means (thus if the design is not *balanced*, LSMEANS are not simple means).

This code yields the following additional output.

Least Squares Means
Adjustment for Multiple Comparisons: Scheffe

center	fev	LSMEAN	
		LSMEAN	Number
hopkins	2.62619048		1
ranchola	3.03250000		2
stlouis	2.87869565		3

Least Squares Means for effect center
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: fev

i/j	1	2	3
1		0.0603	0.2605
2	0.0603		0.6466
3	0.2605	0.6466	



We see some evidence that the mean FEV_1 at Rancho Los Amigos (3.03 liters) is higher than the mean FEV_1 at Johns Hopkins (2.63 liters), but the other means do not appear to differ significantly.

Trend Tests

Usually, orthogonal sets of tests are unappealing because $G - 1$ tests do not cover all interesting mean comparisons.

For example, if the categorical predictor can be interpreted as an interval scale variable (e.g., low, medium, or high dose), then the $G - 1$ orthogonal trend tests are appealing.

Assume we are interested in one species of tree, grown at 200, 400, or 600m, with corresponding hypothesis $H_0: \mu_{200} = \mu_{400} = \mu_{600}$.

Having rejected H_0 , what is trend across elevation?

First, we can plot the observed treatment means as a function of the factor levels. This plot may provide a general idea of any pattern that is present.

G levels imply $G - 1$ trends can be tested. With equal spacing and equal cell sizes, we may use the orthogonal polynomial coefficients in

Table B-10. Without these conditions, tests are more complicated.

Using the contrast matrices described here with cell mean coding allows computing exactly the same tests as available in polynomial coding, which equal tests in polynomial regression.

For *cell mean* coding, consider

$$C_O = \begin{bmatrix} -1 & 0 & 1 \\ 1 & -2 & 1 \end{bmatrix} \begin{array}{l} \text{linear} \\ \text{quadratic} \end{array}$$



This spans the matrix commonly used for the overall test, namely

$$C = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}$$

in that each equals a full rank transformation of the rows of the other ($C_O = TC$, with T a full rank matrix of dimension $a \times a$).

The GLH p-value and test statistic do not change with such a transformation.

Plotting the elements of a row of C_O against the column # displays the shape indicated by the labeling.

Separate one df tests provide a set of $G - 1$ orthogonal trend tests.

It seems reasonable to ignore any Bonferroni or other sort of correction in this setting if the trend tests represent the only planned comparisons.

Homogeneity

There are only G distinct groups, so comparing variances for G cells is useful for assessing homogeneity of variances. The Hartley's test involves computing the ratio of the largest group variance, $\max(s_j^2)$ to the smallest group variance, $\min(s_j^2)$. The resulting ratio, F_{\max} , is then compared to a critical value from a table of the sampling distribution of F_{\max} . If the computed ratio is less than the critical value, the groups are assumed to have similar or equal variances.

Bartlett's test is used to test the null hypothesis, H_0 that all k population variances are equal against the alternative that at least two are different. It is a modification of the normal-theory likelihood ratio test. Bartlett's test is sensitive to departures from normality. That is, if the samples come from non-normal distributions, then Bartlett's test may simply be testing for non-normality.

Newer tests (the Brown and Forsythe test or O'Brien test)(HOVTEST=BF or HOVTEST=OBRIEN in PROC GLM) are much more robust to the underlying distribution. They transform the original values of the dependent variable to derive a dispersion variable and then to perform analysis of variance on this variable.

The significance level for the test of homogeneity of variance is the p-value for the ANOVA F-test on the dispersion variable. All of the homogeneity of variance tests available in PROC GLM except Bartlett's use this approach.

If one of these tests rejects the assumption of homogeneity of variance, you should use Welch's ANOVA (which is a very messy formula. I certainly would never expect anyone to memorize this; but $G = 2$, this test is equivalent to Welch's t-test: $\frac{\bar{y}_1 - \bar{y}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$) instead of the usual ANOVA to test for differences between group means.



```
proc glm data=fev2;
class center;
model fev=center/noint;
means center/hovtest=bf welch;
run;
```

The GLM Procedure

Brown and Forsythes Test for Homogeneity of fev Variance

ANOVA of Absolute Deviations from Group Medians

Source	DF	Sum of		F Value	Pr > F
		Squares	Mean Square		
center	2	0.00818	0.00409	0.04	0.9586
Error	57	5.5105	0.0967		

Welchs ANOVA for fev

Source	DF	F Value	Pr > F
center	2.0000	3.02	0.0614
Error	35.5447		

How do we do this in R?

```
> library(car)
> FEV = read.table("FEV2.dat", header = T)
> head(FEV)

  center   fev
1 hopkins 3.23
2 hopkins 3.47
3 hopkins 1.86
4 hopkins 2.47
5 hopkins 3.01
6 hopkins 1.69

> table(FEV$center)

hopkins ranchola stlouis
      21       16       23

> summary(FEV$fev)

  Min. 1st Qu. Median     Mean 3rd Qu.    Max.
1.690  2.515  2.830  2.831  3.220  4.060

> # cell means predictor matrix
> X_cell = model.matrix(~ factor(center)-1, data = FEV)
```

```
> head(X_cell, 3)

  factor(center)hopkins factor(center)ranchola factor(center)stlouis
1                 1                   0                   0
2                 1                   0                   0
3                 1                   0                   0

> tail(X_cell, 3)

  factor(center)hopkins factor(center)ranchola factor(center)stlouis
58                0                   0                   1
59                0                   0                   1
60                0                   0                   1

> # factor() tells R center is categorical (factor variable),
> # and -1 suppresses the intercept
```

We can specify this cell means coding directly in lm()

```
> fit = lm(fev ~ factor(center)-1, data = FEV)
> summary(fit) # similar to SAS
```

Call:

```
lm(formula = fev ~ factor(center) - 1, data = FEV)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.32250	-0.32250	-0.02244	0.32630	1.18130

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
factor(center)hopkins	2.6262	0.1100	23.88	<2e-16 ***
factor(center)ranchola	3.0325	0.1260	24.07	<2e-16 ***
factor(center)stlouis	2.8787	0.1051	27.39	<2e-16 ***

Signif. codes:	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1			

Residual standard error: 0.504 on 57 degrees of freedom

Multiple R-squared: 0.9709, Adjusted R-squared: 0.9693

F-statistic: 633.2 on 3 and 57 DF, p-value: < 2.2e-16

Now lets check for HOV

```
> leveneTest(y=FEV$fev, group=FEV$center, center=median)
```

```
Levenes Test for Homogeneity of Variance (center = median)
```

	Df	F value	Pr(>F)
group	2	0.0423	0.9586
	57		

```
> # if reject the null, Welchs ANOVA
```

```
> oneway.test(fev ~ center,  
+               data=FEV,  
+               var.equal=FALSE)
```

```
One-way analysis of means (not assuming equal variances)
```

```
data: fev and center
```

```
F = 3.0211, num df = 2.000, denom df = 35.545, p-value = 0.06141
```

What if we wanted to perform the overall test, where we test for the equality of means? Specify contrast matrix for cell means coding

```
> # overall test
> C = matrix(c(1, -1, 0, 1, 0, -1), nrow = 2, byrow = T)
> print(C)

 [,1] [,2] [,3]
[1,]    1   -1    0
[2,]    1    0   -1

> linearHypothesis(fit, C)

Linear hypothesis test
```

Hypothesis:

```
factor(center)hopkins - factor(center)ranchola = 0
factor(center)hopkins - factor(center)stlouis = 0
```

Model 1: restricted model

Model 2: fev ~ factor(center) - 1

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	59	16.063			

```
2      57 14.480  2     1.5828 3.1153  0.052 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Now perform pairwise comparison of means with Scheffe p-value correction

```
> library(DescTools)
> ScheffeTest(aov(fev ~ factor(center), data = FEV))

Posthoc multiple comparisons of means : Scheffe Test
95% family-wise confidence level

$factor(center)
            diff      lwr.ci      upr.ci    pval
ranchola-hopkins 0.4063095 -0.01408874 0.8267078 0.0603 .
stlouis-hopkins  0.2525052 -0.12986364 0.6348740 0.2605
stlouis-ranchola -0.1538043 -0.56622280 0.2586141 0.6466

---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

What if we wanted to use reference cell coding?

```
> fit = lm(fev ~ factor(center), data = FEV)
> summary(fit)
```

Call:

```
lm(formula = fev ~ factor(center), data = FEV)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.32250	-0.32250	-0.02244	0.32630	1.18130

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.6262	0.1100	23.877	<2e-16 ***
factor(center)ranchola	0.4063	0.1673	2.429	0.0183 *
factor(center)stlouis	0.2525	0.1521	1.660	0.1024

Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’
	0.1 ‘ ’	1		

Residual standard error: 0.504 on 57 degrees of freedom

Multiple R-squared: 0.09854, Adjusted R-squared: 0.06691

F-statistic: 3.115 on 2 and 57 DF, p-value: 0.052

```
> # overall test for regression  
> anova(fit)
```

Analysis of Variance Table

Response: fev

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(center)	2	1.5828	0.79142	3.1153	0.052 .
Residuals	57	14.4803	0.25404		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Note that the overall corrected test for regression here is the same as the overall test in cell means encoding (utilizing the proper contrasts). The overall corrected test of regression in cell means coding is testing what exactly?

If you want to choose a different reference level, you can specify a different reference level in center using relevel()

```
> FEV$center = relevel(FEV$center, ref = "ranchola")
> fit = lm(fev ~ factor(center), data = FEV)
> summary(fit)
```

Call:

```
lm(formula = fev ~ factor(center), data = FEV)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.32250	-0.32250	-0.02244	0.32630	1.18130

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.0325	0.1260	24.066	<2e-16 ***
factor(center)hopkins	-0.4063	0.1673	-2.429	0.0183 *
factor(center)stlouis	-0.1538	0.1641	-0.937	0.3525

Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’
	0.1 ‘ ’			1

Residual standard error: 0.504 on 57 degrees of freedom

Multiple R-squared: 0.09854, Adjusted R-squared: 0.06691
F-statistic: 3.115 on 2 and 57 DF, p-value: 0.052

Next: Two-way ANOVA

- Muller and Fetterman, Chapter 14: “Complete, Two-Way Factorial ANOVA”

Lecture 16: Two-Way ANOVA

Reading Assignment:

- Muller and Fetterman, Chapter 14: “Complete, Two-Way Factorial ANOVA” (Required)

In two-way analysis of variance (ANOVA), we wish to evaluate the importance of all combinations of two categorical variables in predicting a Gaussian response.

First, we define several terms commonly used to describe studies well-suited to ANOVA:

Cell: group of observational units, all receiving same treatment.

Balanced: all cells (present) have the same number of observational units.

Complete: all cells have at least one observation.

Balanced, complete designs are the simplest and lead to unambiguous tests. We use this situation as our prototype. Other designs require careful treatment and thoughtful consideration, though they can be handled in the general linear model framework.

Model Concepts

Factorial Design

A **factorial** design includes all combinations of the levels of two or more treatments.

If factor A has a levels and factor B has b levels, then a two-way factorial contains $a \cdot b$ combinations (cells).

For example, let factor A represent drug dose ($a = 2$) and let factor B represent drug formulations ($b = 3$) for a new cholesterol-reducing compound.

Two-Way Factorial Design

N Replicates

	$DRUG_1$	$DRUG_2$	$DRUG_3$
$DOSE_1$			
$DOSE_2$			

Treatments in a factorial are completely crossed (with each other). This means that each drug is tested at each dose, and *vice versa*.

One observation per cell ($a \cdot b = 6$) gives one replicate. For a balanced design, let N equal the # of observations in each cell so that $n = Nab$. So for the balanced two-way factorial design in the previous table, each dose-drug combination is tested in N patients for a total of $6N$ patients in the study.

We generally like to see 5-10 replicates for any analysis (this allows estimation of interaction terms), though further assumptions may be made in order to make inferences for smaller studies.

Cell and Marginal Means

Define a row marginal mean for balanced data as $\mu_{i\cdot} = \sum_{j=1}^b \mu_{ij}/b$, which is a row average in the table, where μ_{ij} is the mean in the cell in row i and column j of the table.

Define a column marginal mean for balanced data as $\mu_{\cdot j} = \sum_{i=1}^a \mu_{ij}/a$, which is a column average across the rows.

Note that the first subscript indicates the row (first factor) and the second subscript indicates the column (second factor).

The marginal means of factor A, $\{\mu_{1\cdot}, \mu_{2\cdot}\}$, have been averaged across factor B, and the marginal means of factor B, $\{\mu_{\cdot 1}, \mu_{\cdot 2}, \mu_{\cdot 3}\}$, have been averaged across factor A.

Consider the following marginal means for the cholesterol drug example.



Two-Way Factorial Design

	$DRUG_1$	$DRUG_2$	$DRUG_3$	
$DOSE_1$	μ_{11}	μ_{12}	μ_{13}	$\mu_{1\cdot}$
$DOSE_2$	μ_{21}	μ_{22}	μ_{23}	$\mu_{2\cdot}$
	$\mu_{\cdot 1}$	$\mu_{\cdot 2}$	$\mu_{\cdot 3}$	$\mu_{\cdot \cdot}$

1. μ_{12} is the population mean cholesterol for the N subjects taking dose 1 of drug 2,
2. $\mu_{1\cdot}$ is the population mean cholesterol for the $3N$ patients taking dose 1, averaged over all three drugs,
3. $\mu_{\cdot 2}$ is the population mean cholesterol for the $2N$ patients taking drug 2, averaged over the two doses, and
4. $\mu_{\cdot \cdot}$ is the population mean cholesterol over all $n = 6N$ subjects in the study.

Primary Hypotheses

In a model without *interaction*, main effects describe marginal means.
To test equality of marginal means, we have the hypotheses:

$$H_0 : \mu_{1\cdot} = \mu_{2\cdot} = \dots = \mu_{a\cdot}$$

and

$$H_0 : \mu_{\cdot 1} = \mu_{\cdot 2} = \dots = \mu_{\cdot b},$$

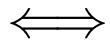
which correspond to testing if the factor A means are the same and if the factor B means are the same, respectively. These are called the tests of the *main effects* of factors A and B.

In the absence of interaction, we test main effects using contrasts like those in one-way ANOVA.



The **interaction** describes the difference among differences (is the dose effect the same for all three drugs?). If the interaction is significant, then the effect of factor B depends on the level of factor A or *vice versa*. Consider the hypothesis that the dose effect of drug A is the same as the dose effect of drug B:

$$H_0 : \mu_{11} - \mu_{21} = \mu_{12} - \mu_{22}$$



$$H_0 : (\mu_{11} - \mu_{21}) - (\mu_{12} - \mu_{22}) = 0.$$

If the null hypothesis is true, then the change in cholesterol going from dose 1 to dose 2 on drug A is the same as the change in cholesterol going from dose 1 to dose 2 on drug B. If the null hypothesis is false, then the dose effect is larger for one drug than for the other.

Coding and Computation

How *do* we code an interaction model? If we use classical ANOVA coding (like SAS), we have the model


$$E(y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij},$$

$i = 1, 2, j = 1, 2, 3, k = 1, \dots, N$. In this case, we have $2 \times 3 = 6$ dose-drug combinations but $1 + 2 + 3 + 6 = 12$ parameters! This model is grossly overparameterized, since we can fit only 6 cell means (thus 6 primary parameters are not estimable).

We could use cell means and fit the model

$$E(y_{ijk}) = \gamma_{ij},$$

in which each γ represents a cell mean for a given dose-drug combination. A more popular approach is to use reference cell coding, which sets one level of each drug and dose to be the reference. For the above example, if we set the highest level of each variable to be the

reference, we have the model

 $E(y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij},$

$i = 1, j = 1, 2, k = 1, \dots, N$. With only these six parameters in the model, everything is now estimable.

Reference Cell (Regression) Coding

Consider reference cell coding for our cholesterol drug example.

Suppose that we have two replicates and that our reference dose is dose 2 and our reference drug is drug 3. There is one column of 1's for the reference cell mean, $a - 1$ columns for the main effect of factor A, $b - 1$ columns for the main effect of factor B, and $(a - 1)(b - 1)$ columns for the interaction.

$$\begin{bmatrix}
 y_{111} \\
 y_{112} \\
 y_{121} \\
 y_{122} \\
 y_{131} \\
 y_{132} \\
 y_{211} \\
 y_{212} \\
 y_{221} \\
 y_{222} \\
 y_{231} \\
 y_{232}
 \end{bmatrix} =
 \begin{bmatrix}
 1 & \vdots & 1 & \vdots & 1 & 0 & \vdots & 1 & 0 \\
 1 & \vdots & 1 & \vdots & 1 & 0 & \vdots & 1 & 0 \\
 1 & \vdots & 1 & \vdots & 0 & 1 & \vdots & 0 & 1 \\
 1 & \vdots & 1 & \vdots & 0 & 1 & \vdots & 0 & 1 \\
 1 & \vdots & 1 & \vdots & 0 & 0 & \vdots & 0 & 0 \\
 1 & \vdots & 1 & \vdots & 0 & 0 & \vdots & 0 & 0 \\
 1 & \vdots & 0 & \vdots & 1 & 0 & \vdots & 0 & 0 \\
 1 & \vdots & 0 & \vdots & 1 & 0 & \vdots & 0 & 0 \\
 1 & \vdots & 0 & \vdots & 0 & 1 & \vdots & 0 & 0 \\
 1 & \vdots & 0 & \vdots & 0 & 1 & \vdots & 0 & 0 \\
 1 & \vdots & 0 & \vdots & 0 & 0 & \vdots & 0 & 0 \\
 1 & \vdots & 0 & \vdots & 0 & 0 & \vdots & 0 & 0
 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \end{bmatrix} + \boldsymbol{\varepsilon}$$

The interpretations of the parameters may be obtained by considering $E(y_{ij})$ for subjects on various doses (indexed by i) of various drugs (indexed by j). This means, for example, that we no longer have one parameter in our model that describes the dose effect – instead, we have 3 parameters that describe the dose effect; one for each drug.

In this example, μ represents the average cholesterol for subjects at dose 2 of drug 3. α_1 is the difference in cholesterol for subjects on dose 1 of drug 3 compared to subjects on dose 2 of drug 3. β_1 is the difference in cholesterol for subjects on dose 2 of drug 1 compared to subjects on dose 2 of drug 3. β_2 is the difference in cholesterol for subjects on dose 2 of drug 2 compared to subjects on dose 2 of drug 3.

The difference in dose effects for drug 1 versus drug 3 is given by γ_{11} , and the difference in dose effects for drug 2 versus drug 3 is given by γ_{12} .

Generating Cell Means

$\mathbf{C} = \text{Es}(\mathbf{X})$ gives the cell means, $\mathbf{C}\boldsymbol{\beta} = \text{Es}(\mathbf{X})\boldsymbol{\beta} = \boldsymbol{\theta} = \{\mu_{jk}\}$, for complete designs, whether balanced or not.

For the cholesterol example, the cell means are provided below.

$$\mathbf{C} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \end{bmatrix},$$



so that we have the following cell means.



Dose	Drug	Mean
1	1	$\mu + \alpha_1 + \beta_1 + \gamma_{11}$
1	2	$\mu + \alpha_1 + \beta_2 + \gamma_{12}$
1	3	$\mu + \alpha_1$
2	1	$\mu + \beta_1$
2	2	$\mu + \beta_2$
2	3	μ

Computing Estimates and Tests

Balanced, Complete Designs with Equal Cell Size

We can use simple formulae for hand computation in balanced designs, leading to simple interpretations.

Source	df	SS
A	$a - 1$	$bN \sum_{i=1}^a (\bar{y}_{i..} - \bar{y}...)^2$
B	$b - 1$	$aN \sum_{j=1}^b (\bar{y}_{.j.} - \bar{y}...)^2$
AB	$(a - 1)(b - 1)$	$N \sum_{i=1}^a \sum_{j=1}^b (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}...)^2$
Error	$ab(N - 1)$	$\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^N (y_{ijk} - \bar{y}_{ij.})^2$
Total	$abN - 1$	$\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^N (y_{ijk} - \bar{y}...)^2$

The SS formulae presented here are correct only with complete, balanced data, unless explicitly stated otherwise.

Unbalanced, Complete Data

With unbalanced, complete data, we revert to the general linear model framework and fit $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. By specifying the appropriate contrast matrices, we are able to carry out the appropriate tests.

One must fit multiple models in order to generate added-in-order tests.

Contrast Matrices for Marginal Means

Mean Estimation with Reference Cell Coding

Start with the matrix which computes cell means that correspond to the interaction in our cholesterol example: $C_{A \times B} = Es(\mathbf{X})$

$$C_{A \times B} \beta = Es(\mathbf{X}) \beta = \{\mu_{jk}\}$$

$$\theta_{A \times B} = \{\mu_{jk}\} = \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{23} \end{bmatrix}.$$

To estimate the marginal means of factor A (dose) for our example, we use

$$C_{A \text{ ref}} = \begin{bmatrix} 1 & : & 1 & : & \frac{1}{3} & \frac{1}{3} & : & \frac{1}{3} & \frac{1}{3} \\ 1 & : & 0 & : & \frac{1}{3} & \frac{1}{3} & : & 0 & 0 \end{bmatrix}$$

$$\text{to obtain } \theta_A = \begin{bmatrix} \mu_{1 \cdot} \\ \mu_{2 \cdot} \end{bmatrix}.$$

This becomes clear if we think about averaging the expected values by dose.



Similarly, for marginal means of factor B (drug), we have

$$C_{B \text{ ref}} = \begin{bmatrix} 1 & : & \frac{1}{2} & : & 1 & 0 & : & \frac{1}{2} & 0 \\ 1 & : & \frac{1}{2} & : & 0 & 1 & : & 0 & \frac{1}{2} \end{bmatrix}, \text{ leading to}$$

$$\theta_B = \begin{bmatrix} \mu_{.1} \\ \mu_{.2} \\ \mu_{.3} \end{bmatrix}.$$

The grand mean can be thought of as the average of all cell means or as the average of either set of marginal means:

$$C_{\text{grand mean ref}} = \left[1 \quad \frac{1}{2} \quad \frac{1}{3} \quad \frac{1}{3} \quad \frac{1}{6} \quad \frac{1}{6} \right].$$



Testing with Reference Cell Coding

One must take care when testing the main effects of factor A or factor B in the presence of their interaction. Specifically, testing that the α 's are equal to zero (or that the β 's are equal to zero) is not testing marginal means. In fact, this would be a test of simple main effects *at the reference level* of the other factor.

For factor A, we really wish to test the hypothesis

$$H_0 : \mu_{1\cdot} = \mu_{2\cdot}$$

Thus we wish to test the hypothesis

$$H_0 : \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_{1\cdot} \\ \mu_{2\cdot} \end{bmatrix} = 0.$$

Thus we may define the appropriate contrast matrix \mathbf{C} for factor A as

$$\begin{aligned}\mathbf{C} &= \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & : & 1 & : & \frac{1}{3} & \frac{1}{3} & : & \frac{1}{3} & \frac{1}{3} \\ 1 & : & 0 & : & \frac{1}{3} & \frac{1}{3} & : & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & : & 1 & : & 0 & 0 & : & \frac{1}{3} & \frac{1}{3} \end{bmatrix}.\end{aligned}$$

Similarly, for testing marginal means of factor B, we have the contrast matrix

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & : & \frac{1}{2} & : & 1 & 0 & : & \frac{1}{2} & 0 \\ 1 & : & \frac{1}{2} & : & 0 & 1 & : & 0 & \frac{1}{2} \\ 1 & : & \frac{1}{2} & : & 0 & 0 & : & 0 & 0 \end{bmatrix}$$
$$= \begin{bmatrix} 0 & : & 0 & : & 1 & 0 & : & \frac{1}{2} & 0 \\ 0 & : & 0 & : & 0 & 1 & : & 0 & \frac{1}{2} \end{bmatrix}.$$

The test of interaction is simply a test that the interaction terms γ are equal to zero.

Choosing and Interpreting Tests

Model Comparisons

All GLH tests correspond to comparing two models. Note that marginal means tests in reference cell coding do not correspond to column deletions.

Testing Strategies

All testing strategies are special cases of those in Chapters 5 and 11. The recommended strategy is a backwards group-wise approach. The first group contains the main effect of A variables, the second group contains the main effect of B variables, and the third group contains the interaction variables. Start by testing the interaction. If significant, stop, and fit the full model. If not significant, drop the interaction. If the interaction is dropped, proceed with testing of the main effects.

Step-Down Tests

Simple Main Effect Tests

A simple main effect equals a difference in cell means within a column ($\mu_{jk} - \mu_{j'k}$) or within a row ($\mu_{jk} - \mu_{jk'}$). These tests are of interest if the interaction terms are significant. The simple main effect (SME) test evaluates one effect while holding the other constant. There exist b SME tests for factor A and a SME tests for factor B. In our example, we may wish to use a simple main effect test to determine, for example, if there is a dose effect for drug A.

Second Level Step-Down Tests

If a SME with two or more numerator degrees of freedom is significant, additional tests are required to determine where the difference lies.

Missing Data

The pattern of cell sizes may have profound effects on computation and interpretation of test statistics in ANOVA. The complete, balanced case involves no missing data. The complete, unbalanced case involves at least some missing data, meaning that additional observations would be needed to make the design balanced. Careful specification of \mathbf{C} for marginal means tests “automatically” provides valid solutions with complete unbalanced data. The model corresponds to simply deleting some rows in \mathbf{X} and \mathbf{y} when missing data is simply due to some lack of balance. In this case, we have a model like

$$\mathbf{y} = \begin{bmatrix} \mathbf{J}_{n_1} \\ & \mathbf{J}_{n_2} \\ & & \mathbf{J}_{n_3} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} + \boldsymbol{\varepsilon}$$

with $n = n_1 + n_2 + n_3$. For this model $\hat{\boldsymbol{\beta}} = [\hat{\mu}_1 \quad \hat{\mu}_2 \quad \hat{\mu}_3]'$.

Strongly unbalanced data create additional difficulties.

With unbalanced data avoid automatic coding schemes provided by computer programs (such as the CLASS statement in PROC GLM).

More About Missing Data

The above comments apply when data is unbalanced *by design*.

Suppose we originally planned to treat 50 patients on dose 1 and 40 patients on dose 2. Alternatively, suppose that originally 50 patients were taking each dose, but 10 patients taking dose 2 dropped out of the study due to toxic side effects at that dose. In the second case, a *complete case analysis*, or an analysis that deletes some rows in \mathbf{X} and \mathbf{y} , may lead to biased parameter estimates.

The validity of any analysis with incomplete data depends on the reasons for which data (response or covariates) are missing.

A *complete case analysis* is valid only when data are **missing**

- ▀ **completely at random (MCAR)**, which means that the completely observed subjects are a completely random sample of all the subjects in the study. Complete case analysis is the default analysis in SAS PROC REG and SAS PROC GLM when data (responses or covariates) are missing, regardless of its validity for the data in question.



Other types of missingness include *missing at random (MAR)*, which means that missingness does not depend on the values of the variables that are missing but may depend on observed quantities. Examples of MAR mechanisms

- A subject may be removed from a trial if his/her condition is not controlled sufficiently well (according to pre-defined criteria on the

response).

- Two measurements of the same variable are made at the same time. If they differ by more than a given amount a third is taken. This third measurement is missing for those that do not differ by the given amount.

Nonignorable missingness, which means that the missingness may depend on the values of the variables that are missing.

In the latter two cases, the *complete case analysis* is not valid in general and may lead to considerable bias and faulty results. If a *complete case analysis* is carried out, deleted observations must be reported, and the investigator must seriously consider the validity of the analysis.

Even in the “ideal” MCAR situation, replacing missing data with simple imputed values (such as the average for all subjects) biases the

estimate of error variance downward and therefore inflates the type I error rate. Despite this relatively well-known fact, many investigators substitute means for missing values by default. “Making up” data in this manner can have extremely severe consequences and should be avoided.

Recently, SAS has implemented PROC MI and PROC MIANALYZE, which carry out multiple imputation, a more sophisticated method for handling missing observations. However, these procedures are still in the development stages and are not yet available for a wide class of problems (currently, PROC MI requires all variables to follow Gaussian distributions for general missing data patterns).

Example: Automobile Pollution Filter Noise

Associated Octel Company developed an automobile silencing filter, the Octel filter, that Texaco, Inc. stated (before the Air and Water Pollution Subcommittee of the Senate Public Works Committee) was just as effective as the standard silencing filter. The president of Texaco at the time presented data that we now analyze to back up his claim that the new filter (TYPE1=1 for standard, TYPE1=0 for Octel) was no more noisy than the standard filter. The dependent variable, NOISE, is noise level in decibels, with factor A as the vehicle size (small, medium, or large, with large as the reference level) and factor B as filter type (standard or Octel, with Octel as the reference level). Our goal is to evaluate the noise levels of the three car sizes at the two filter types.

First, we define the following variables for the reference cell coding scheme.



- SIZE1, which takes the value 1 for small cars and 0 otherwise
- SIZE2, which takes the value 1 for medium cars and 0 otherwise
- S1TYPE1, which takes the value 1 for small cars with the standard filter and zero otherwise
- S2TYPE1, which takes the value 1 for medium cars with the standard filter and zero otherwise

Next, let's take a look at the mean noise levels for each car size and filter type.

Size	Silencer/Filter Type	Average Noise Level
Small	Standard	825.83
Small	Octel	822.50
Medium	Standard	845.83
Medium	Octel	821.67
Large	Standard	775.00
Large	Octel	770.00

What is your impression based on these cell means?

Next, we fit the model

$$\begin{aligned}NOISE = \beta_0 + \beta_1 SIZE1 + \beta_2 SIZE2 + \beta_3 TYPE1 + \\ \beta_4 S1TYPE1 + \beta_5 S2TYPE1 + \varepsilon.\end{aligned}$$

The SAS code is below.

```
proc glm;
model noise=size1 size2 type1 s1type1 s2type1/solution;
estimate Grand Mean
    intercept 6 size1 2 size2 2 type1 3 s1type1 1 s2type1 1/divisor=6;
estimate Marg Mean: Small
    intercept 2 size1 2 size2 0 type1 1 s1type1 1 s2type1 0/divisor=2;
estimate Marg Mean: Medium
    intercept 2 size1 0 size2 2 type1 1 s1type1 0 s2type1 1/divisor=2;
estimate Marg Mean: Large
    intercept 2 size1 0 size2 0 type1 1 s1type1 0 s2type1 0/divisor=2;
estimate Marg Mean: Standard
    intercept 3 size1 1 size2 1 type1 3 s1type1 1 s2type1 1/divisor=3;
estimate Marg Mean: Octel
    intercept 3 size1 1 size2 1 type1 0 s1type1 0 s2type1 0/divisor=3;

contrast Interaction Silencer by Size
    intercept 0 size1 0 size2 0 type1 0 s1type1 1 s2type1 0,
    intercept 0 size1 0 size2 0 type1 0 s1type1 0 s2type1 1;
```

```
contrast Main Effect Vehicle Size
    intercept 0 size1 2 size2 0 type1 0 s1type1 1 s2type1 0,
    intercept 0 size1 0 size2 2 type1 0 s1type1 0 s2type1 1;

contrast Main Effect Silencer
    intercept 0 size1 0 size2 0 type1 3 s1type1 1 s2type1 1;

run;
```

```
/* ALTERNATIVELY, you can use the code below */  
/* DANGER DANGER -- this code uses the class statement */  
proc glm;  
class size type;  
model noise=size type size*type;  
run;
```

The output is provided below.

The GLM Procedure

Dependent Variable: noise

Source	DF	Sum of		
		Squares	Mean Square	F Value
Model	5	27911.80556	5582.36111	85.34
Error	30	1962.50000	65.41667	
Corrected Total	35	29874.30556		

Source	Pr > F
Model	<.0001
Error	

Corrected Total

R-Square	Coeff Var	Root MSE	noise	Mean
0.934308	0.998354	8.088057	810.1389	
Source	DF	Type I SS	Mean Square	F Value
size1	1	3542.01389	3542.01389	54.15
size2	1	22509.37500	22509.37500	344.09
type1	1	1056.25000	1056.25000	16.15
s1type1	1	253.12500	253.12500	3.87
s2type1	1	551.04167	551.04167	8.42
Source		Pr > F		
	size1	<.0001		
	size2	<.0001		
	type1	0.0004		
	s1type1	0.0585		
	s2type1	0.0069		
Source	DF	Type III SS	Mean Square	F Value
size1	1	8268.750000	8268.750000	126.40
size2	1	8008.333333	8008.333333	122.42
type1	1	75.000000	75.000000	1.15

Source		Pr > F		
	size1	<.0001		
	size2	<.0001		
	type1	0.2928		
Source	DF	Type III SS	Mean Square	F Value
s1type1	1	4.166667	4.166667	0.06
s2type1	1	551.041667	551.041667	8.42
Source		Pr > F		
	s1type1	0.8025		
	s2type1	0.0069		

Contrast	DF	Contrast SS		
Interaction Silencer by Size	2	804.16667		
Main Effect Vehicle Size	2	26051.38889		
Main Effect Silencer	1	1056.25000		
Contrast	Mean Square	F Value		
Interaction Silencer by Size	402.08333	6.15		
Main Effect Vehicle Size	13025.69444	199.12		
Main Effect Silencer	1056.25000	16.15		
Contrast	Pr > F			
Interaction Silencer by Size	0.0058			
Main Effect Vehicle Size	<.0001			
Main Effect Silencer	0.0004			
Parameter	Estimate	Error	t Value	Pr > t
Grand Mean	810.138889	1.34800951	600.99	<.0001
Marg Mean: Small	824.166667	2.33482095	352.99	<.0001
Marg Mean: Medium	833.750000	2.33482095	357.09	<.0001
Marg Mean: Large	772.500000	2.33482095	330.86	<.0001
Marg Mean: Standard	815.555556	1.90637333	427.80	<.0001

Marg Mean: Octel 804.722222 1.90637333 422.12 <.0001

Parameter	Estimate	Standard		
		Error	t Value	Pr > t
Intercept	770.0000000	3.30193546	233.20	<.0001
size1	52.5000000	4.66964191	11.24	<.0001
size2	51.6666667	4.66964191	11.06	<.0001
type1	5.0000000	4.66964191	1.07	0.2928
s1type1	-1.6666667	6.60387092	-0.25	0.8025
s2type1	19.1666667	6.60387092	2.90	0.0069

 Using the SAS output, we can construct the ANOVA table as follows.

Source	df	SS	MS	F	p
Size					
Silencer					
Size*Silencer					
Error					
Total					

What further tests, if any, should we consider?

Because the interaction is significant, a logical next step is to conduct simple main effect tests of the filter effect within each car size. The SAS code below may be used to conduct these SME tests.

```
proc glm;
model noise=size1 size2 type1 s1type1 s2type1/solution;
contrast SME Silencer at Size Small
    intercept 0 size1 0 size2 0 type1 1 s1type1 1 s2type1 0;
contrast SME Silencer at Size Medium
    intercept 0 size1 0 size2 0 type1 1 s1type1 0 s2type1 1;
contrast SME Silencer at Size Large
    intercept 0 size1 0 size2 0 type1 1 s1type1 0 s2type1 0;
run;
```

The additional output is provided below.

Contrast	DF	Contrast SS
SME Silencer at Size Small	1	33.33333
SME Silencer at Size Medium	1	1752.08333
SME Silencer at Size Large	1	75.00000

Contrast	Mean Square	F Value
----------	-------------	---------

SME Silencer at Size Small	33.33333	0.51
SME Silencer at Size Medium	1752.08333	26.78
SME Silencer at Size Large	75.00000	1.15

Contrast Pr > F

SME Silencer at Size Small	0.4808
SME Silencer at Size Medium	<.0001
SME Silencer at Size Large	0.2928

What do you conclude?

Alternatively (less logically?), you could have tested the SME of car size for each filter type. The SAS code and additional output are provided below.

```
proc glm;
model noise=size1 size2 type1 s1type1 s2type1/solution;
contrast SME Size at Standard Silencer
    intercept 0 size1 1 size2 -1 type1 0 s1type1 1 s2type1 -1,
    intercept 0 size1 1 size2 0 type1 0 s1type1 1 s2type1 0;
contrast SME Size at Octel Silencer
    intercept 0 size1 1 size2 -1 type1 0 s1type1 0 s2type1 0,
    intercept 0 size1 1 size2 0 type1 0 s1type1 0 s2type1 0;
run;
*****
```

Contrast	DF	Contrast SS
SME Size at Standard Silencer	2	16002.77778
SME Size at Octel Silencer	2	10852.77778

Contrast	Mean Square	F Value
----------	-------------	---------

SME Size at Standard Silencer	8001.38889	122.31
SME Size at Octel Silencer	5426.38889	82.95

Contrast Pr > F

SME Size at Standard Silencer	<.0001
SME Size at Octel Silencer	<.0001

What do you conclude?

Because the SME tests for size were significant for both filter types, we can now carry out step-down tests to determine specifically which car type means differ for each filter type.

 proc glm;
model noise=size1 size2 type1 s1type1 s2type1/solution;
contrast Step-Down of SME at Std: Small vs. Med
 intercept 0 size1 1 size2 -1 type1 0 s1type1 1 s2type1 -1;
contrast Step-Down of SME at Std: Small vs. Large
 intercept 0 size1 1 size2 0 type1 0 s1type1 0 s2type1 0;
contrast Step-Down of SME at Std: Med vs. Large
 intercept 0 size1 0 size2 1 type1 0 s1type1 0 s2type1 1;
contrast Step-Down of SME at Octel: Small vs. Med
 intercept 0 size1 1 size2 -1 type1 0 s1type1 0 s2type1 0;
contrast Step-Down of SME at Octel: Small vs. Large
 intercept 0 size1 1 size2 0 type1 0 s1type1 0 s2type1 0;
contrast Step-Down of SME at Octel: Med vs. Large
 intercept 0 size1 0 size2 1 type1 0 s1type1 0 s2type1 0;
run;

Contrast	DF	Contrast SS
Step-Down of SME at Std: Small vs. Med	1	1200.00000
Step-Down of SME at Std: Small vs. Large	1	8268.75000
Step-Down of SME at Std: Med vs. Large	1	15052.08333
Contrast	Mean Square	F Value
Step-Down of SME at Std: Small vs. Med	1200.00000	18.34
Step-Down of SME at Std: Small vs. Large	8268.75000	126.40
Step-Down of SME at Std: Med vs. Large	15052.08333	230.10
Contrast	Pr > F	
Step-Down of SME at Std: Small vs. Med	0.0002	
Step-Down of SME at Std: Small vs. Large	<.0001	
Step-Down of SME at Std: Med vs. Large	<.0001	

Contrast	DF	Contrast SS
Step-Down of SME at Octel: Small vs. Med	1	2.08333
Step-Down of SME at Octel: Small vs. Large	1	8268.75000
Step-Down of SME at Octel: Med vs. Large	1	8008.33333
Contrast	Mean Square	F Value
Step-Down of SME at Octel: Small vs. Med	2.08333	0.03
Step-Down of SME at Octel: Small vs. Large	8268.75000	126.40
Step-Down of SME at Octel: Med vs. Large	8008.33333	122.42
Contrast	Pr > F	
Step-Down of SME at Octel: Small vs. Med	0.8596	
Step-Down of SME at Octel: Small vs. Large	<.0001	
Step-Down of SME at Octel: Med vs. Large	<.0001	

What do you conclude?

For this analysis, we considered engine size as a *factor* instead of as a continuous or ordinal variable. Suppose instead we use one variable,

$$\text{Q } SIZE = \begin{cases} 1 & \text{small} \\ 2 & \text{medium} \\ 3 & \text{large} \end{cases}$$

to describe engine size. What are the implications for our analysis?

Next: Logistic Regression

Reading Assignment:

- Weisberg Chapter 12 and KKMN, Chapter 23: "Logistic Regression Analysis"

Lecture 17: Logistic Regression

Reading Assignment:

- Weisberg Chapter 12



Often, the response of interest in a scientific study is a binary variable, such as DISEASED/NOT DISEASED or DEAD/ALIVE. In this case, the linear regression model no longer holds because the errors will not follow a Gaussian distribution.

When studying linear regression, our models were of the form

$$E[\mathbf{y}] = \beta_0 + \beta_1 \mathbf{x}_1 + \dots + \beta_{p-1} \mathbf{x}_{p-1}.$$

The response \mathbf{y} was continuous, not discrete, and we wanted to predict the mean response and explain the variability among the observed outcomes.

When y is dichotomous, we observe only two possible values: “success” or “failure.” Often, we use a 0/1 coding scheme so that “success” means $y = 1$ and “failure” means $y = 0$. We can talk about the mean of y , which is the percentage of times that y takes the value 1, or the percentage of successes. We often denote this percentage as $p = E(y) = \Pr(y = 1) = \Pr(\text{success})$.

In this setting, we wish to estimate this probability p as well as the effect of various explanatory variables or covariates on this success probability. In order to do so, we use the *logistic regression* model.

Logistic Regression Model: Heuristics

Consider a study of cold incidence among French skiers (Pauling, PNAS, 1971), some of whom were given vitamin C.

	COLD	NO COLD	Total
VIT C	17	122	139
NO VIT C	31	109	140
Total	48	231	279



The **odds ratio** is a widely-used epidemiologic measure of association.



It compares two or more groups in predicting the outcome variable.

The **odds** are defined as the ratio of probabilities that an event (developing a cold) will occur divided by the probability that the same event will not occur. So if the probability of a cold is 0.25, then the odds of a cold are $\frac{0.25}{1.00-0.25} = \frac{1}{3}$. An odds of $\frac{1}{3}$ means that the

probability of a cold is one-third of the probability of no cold (in betting you often hear that the odds are “3 to 1” that the event will not occur). An *odds ratio* is just the ratio of two odds. When the *odds ratio* is one, then the two groups are equally likely to develop a cold.

In this sample, the probability of a cold for skiers taking vitamin C is $p_1 = \frac{17}{139} = 0.12$, and the corresponding probability for skiers not taking vitamin C is $p_2 = \frac{31}{140} = 0.22$. The *odds* of a cold for a skier taking vitamin C are $\frac{p_1}{1-p_1} = \frac{0.12}{1.00-0.12} = 0.14$, and for a skier not taking vitamin C, the odds are $\frac{p_2}{1-p_2} = \frac{0.22}{1.00-0.22} = 0.28$. We relate these odds using an *odds ratio*, defined as

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{0.14}{0.28} = 0.49$$

(after fixing some rounding error). This means that those skiers taking vitamin C have half the odds of a cold as skiers not taking vitamin C, or that vitamin C is protective against getting colds.

Aside: Why do we bother with the odds?

- ❑ A **prospective** study watches for outcomes, such as the development of a disease, during the study period and relates this to other factors such as suspected risk or protection factor(s). The study usually involves taking a cohort of subjects and watching them over a long period. The outcome of interest should be common; otherwise, the number of outcomes observed will be too small to be statistically meaningful. All efforts should be made to avoid sources of bias such as the loss of individuals to follow up during the study. Prospective studies usually have fewer potential sources of bias and confounding than retrospective studies.
- ❑ A **retrospective** study looks backwards and examines exposures to suspected risk or protection factors in relation to an outcome that is established at the start of the study. Many valuable case-control studies, such as Lane and Claypon's 1926 investigation of risk factors for breast cancer, were retrospective investigations. If the outcome of interest is uncommon, however, the size of prospective investigation

required to estimate relative risk is often too large to be feasible. In retrospective studies the odds ratio provides an estimate of relative risk. You should take special care to avoid sources of bias and confounding in retrospective studies.

A related measure, the *risk ratio*, is defined as

$$RR = \frac{p_1}{p_2}.$$

In this case,

$$RR = \frac{\frac{17}{139}}{\frac{31}{140}} = 0.55,$$

which is interpreted as “Vitamin C users have roughly half the risk of developing a cold.” This measure of association has a simpler interpretation, but it usually cannot be estimated retrospectively because p_1 and p_2 cannot be estimated.

Specifically, with retrospective data, we cannot estimate

 $p_1 = p(\text{disease} \mid \text{exposed})$ or $p_2 = p(\text{disease} \mid \text{unexposed})$. To estimate these quantities, we typically select a group of people based on exposure status and follow them through time to see whether or not they develop disease. In a case-control study, we select patients based on disease status and then determine exposure, so that we estimate $\pi_1 = p(\text{exposed} \mid \text{disease})$ and $\pi_2 = p(\text{exposed} \mid \text{no disease})$.

So how can we estimate the odds ratio, which is a function of p_1 and p_2 , in case-control studies?

We wish to develop a statistical model for the cold data so that we can later incorporate other confounders, which might include family size or use of herbal remedies like echinacea.

A first strategy might be to fit a model using the form


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


where p is the probability of a cold, and x is the vitamin C status, where $x = 1$ for vitamin C takers and 0 otherwise. This looks like an ordinary least squares regression model, in which the response (a probability) is continuous. However, this model is problematic because p is restrained to lie in the interval $[0, 1]$, while $\beta_0 + \beta_1 x$ could technically take any value.

To ensure our estimate of p is positive, we could try fitting the model

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

but this model is also unsatisfactory because we could estimate p to be greater than 1.

To solve both problems, we fit a model of the form

$$E(y) = p = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} = \frac{1}{1 + \exp(-\beta_0 - \beta_1 x)}.$$

This *logistic function* cannot yield estimates of p that are less than 0 or greater than 1.

We are not quite finished, though. This model is a little difficult to interpret; how does taking vitamin C affect your probability of getting a cold in this model? The relationship between p and x is clearly not linear! To make interpretation easier, we will use algebra to rewrite the

model.

$$\begin{aligned} p &= \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \iff \\ \frac{p}{1-p} &= \frac{\frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}}{\frac{1}{1 + \exp(\beta_0 + \beta_1 x)}} \iff \\ \frac{p}{1-p} &= \exp(\beta_0 + \beta_1 x) \iff \\ \log\left(\frac{p}{1-p}\right) &= \beta_0 + \beta_1 x \iff \\ \text{logit}(p) &= \beta_0 + \beta_1 x. \end{aligned}$$

Thus x is linearly related to the log-odds of a cold. The parameter β_1 is interpreted as the log odds ratio of a cold, and $\frac{\exp(\beta_0+\beta_1(1))}{\exp(\beta_0+\beta_1(0))} = \exp(\beta_1)$ provides the odds ratio.



For a continuous predictor x , the odds ratio between levels i and j of the predictor is given by $\frac{\exp(\beta_0+\beta_1(i))}{\exp(\beta_0+\beta_1(j))} = \exp(\beta_1(i-j))$. Thus the effect of increasing x by the amount d is to increase the odds that $y = 1$ by a factor of $\exp(\beta_1 d)$ or to increase the log odds that $y = 1$ by an increment of $\beta_1 d$.

Logistic Regression Likelihood

Consider a study of the relationship between folic acid intake

$$x_1 = \begin{cases} 1 & \text{adequate folic acid intake} \\ 0 & \text{otherwise} \end{cases}$$

and preterm delivery

$$y = \begin{cases} 1 & \text{baby is born before 37 weeks} \\ 0 & \text{otherwise.} \end{cases}$$

Because race

$$x_2 = \begin{cases} 1 & \text{African American} \\ 0 & \text{otherwise} \end{cases}$$

is also related to the probability of preterm delivery and may be related to folic acid intake, we wish to control for it in our model.

The main question of interest is whether adequate folic acid intake is

associated with a reduced probability of preterm delivery. A secondary question may be how race is related to the probability of preterm delivery. (In the PIN study at UNC, African American women are at much lower risk of preterm delivery than are African American women in the general US population for reasons largely undetermined.)

The general logistic regression model is given by

$$\begin{aligned}\text{logit}(p_i) &= \log\left(\frac{p_i}{1-p_i}\right) \\ &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{p-1} x_{i,p-1}\end{aligned}$$

with $y_i \sim \text{Bernoulli}(p_i)$, $i = 1, \dots, n$, and the y 's independent of each other.

The likelihood is obtained as a product of the marginal distributions of

the y 's:

$$\begin{aligned} L(\mathbf{y} \mid p) &= \prod_{i=1}^n \Pr(y_i \mid p_i) \\ &= \prod_{i=1}^n [p_i^{y_i} (1 - p_i)^{1-y_i}], \end{aligned}$$

where p_i is a function of the covariates $x_{i1}, \dots, x_{i,p-1}$ and parameters $\beta_0, \dots, \beta_{p-1}$ given by

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_{p-1} x_{i,p-1})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_{p-1} x_{i,p-1})}.$$

Because this likelihood is a complex function of the elements of β , maximizing it to find the maximum likelihood estimator $\hat{\beta}$ of β requires an iterative procedure like Newton-Raphson or iteratively reweighted least squares. Computer programs (like SAS PROC LOGISTIC or SAS PROC GLM) provide maximum likelihood estimates of β as well as estimated large-sample covariances to be used for statistical inference.

Test Statistics and Model Comparisons

For the logistic regression model, a test of

$$H_0 : \beta_k = 0$$

is a test of whether the k^{th} covariate affects the probability of success, with the null hypothesis of probability of success independent of x_k .

To test this hypothesis for each variable in the model, SAS reports the Wald statistic, given by


$$Z = \frac{\widehat{\beta}_k}{\left(\widehat{\text{Var}}(\widehat{\beta})_{k+1,k+1}\right)^{\frac{1}{2}}},$$

which is asymptotically $N(0, 1)$ under H_0 in large samples. (Z^2 is often reported and follows a χ_1^2 distribution under the null.)

A likelihood ratio test statistic has somewhat better properties for

comparing nested models. We take

$$2 [\log(L(\text{larger})) - \log(L(\text{smaller}))]$$

or

$$(-2 \log(L(\text{smaller}))) - (-2 \log(L(\text{larger})))$$

and compare the resulting value to a chi-squared distribution with
degrees of freedom equal to the difference in the number of parameters
in the two models.

For small samples, these tests should not be used, and exact methods
should be used instead. You can learn more about exact methods in
BIOS665 next fall.

Example: French Skiers

Below is the SAS code used to analyze the French skier data.

```
data ski;
input vitc cold count;
cards;
1 1 17
1 0 122
0 1 31
0 0 109
;

proc logistic descending;
freq count; 
model cold=vitc;
run;
```

We use the DESCENDING option to request that the response value ordering be reversed. Thus we wish to model $\Pr(\text{cold}) = \Pr(y = 1)$ instead of $\Pr(y = 0)$. We use the FREQ COUNT statement to tell SAS that we have entered the data in a summary form. We would eliminate

this statement if we had entered 17 lines coded “1 1” for vitc and cold, 122 lines coded “1 0” for vitc and cold, 31 lines coded “0 1” for vitc and cold, and 109 lines coded “0 0” for vitc and cold. Since we do not have one line per subject in this data, we need to let SAS know this.

Selected SAS output is provided below.

The LOGISTIC Procedure

Model Information

Data Set	WORK.SKI
Response Variable	cold
Number of Response Levels	2
Number of Observations	4
Frequency Variable	count
Sum of Frequencies	279
Model	binary logit
Optimization Technique	Fishers scoring

Response Profile

Ordered Value	Total Frequency
cold	

1	1	48
2	0	231

Probability modeled is cold=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	258.184	255.312
SC	261.815	262.575
-2 Log L	256.184	251.312

The LOGISTIC Procedure
Analysis of Maximum Likelihood Estimates

Parameter	DF	Standard		Wald	
		Estimate	Error	Chi-Square	Pr > ChiSq
Intercept	1	-1.2574	0.2035	38.1575	<.0001
vitc	1	-0.7134	0.3293	4.6934	0.0303



Odds Ratio Estimates

Effect	Estimate	Point	95% Wald
		Confidence Limits	
vitc	0.490	0.257	0.934

From the Wald test and the resulting confidence limits, we see evidence that Vitamin C is protective. Note: The confidence interval for the odds ratio was computed as

$(\exp(-0.7134 - 1.96(0.3293)), \exp(-0.7134 + 1.96(0.3293)))$. Other methods for computing this confidence interval may be preferred.

Example: PIN Study

In the PIN study, we are interested in whether the probability of a preterm birth (before 37 weeks) is affected by a variety of covariates, including

- PTBANY, which equals 1 if the woman has delivered a previous preterm infant and 0 otherwise,
- RACEIND, which equals 1 if the woman is African American and 0 otherwise,
- SMOKEIND, which equals 1 for smokers and 0 otherwise,
- BMIIND, which equals 1 for underweight women and 0 otherwise, and
- EDIND, which equals 1 for women without a high school diploma and 0 otherwise.

We will fit the model

$$\begin{aligned}\text{logit}(\Pr(\text{preterm})) &= \beta_0 + \beta_1 PTBANY + \beta_2 RACEIND \\ &\quad + \beta_3 SMOKEIND + \beta_4 BMIIND \\ &\quad + \beta_5 EDIND.\end{aligned}$$

```
proc logistic descending;
model c_case1=ptbany raceind smokeind bmiind edind;
run;
*****
```

The LOGISTIC Procedure

Model Information

Data Set	WORK.NEW
Response Variable	c_case1
Number of Response Levels	2
Number of Observations	3093
Model	binary logit
Optimization Technique	Fishers scoring

Response Profile

Ordered Value	c_case1	Total Frequency
1	1	397

2

0

2696

Probability modeled is c_case1=1.

NOTE: 633 observations were deleted due to missing values for the response or explanatory variables.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	1938.403	1888.588
SC	1944.239	1923.604
-2 Log L	1936.403	1876.588

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	59.8152	5	<.0001
Score	70.2530	5	<.0001
Wald	65.3509	5	<.0001

The SAS System

2

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Wald Pr > ChiSq
Intercept	1	-2.2645	0.1073	445.0116	<.0001
ptbany	1	1.0798	0.1402	59.2982	<.0001

raceind	1	0.1944	0.1255	2.3996	0.1214
smokeind	1	0.1928	0.3201	0.3628	0.5469
bmiind	1	-0.1170	0.1694	0.4774	0.4896
edind	1	0.1346	0.1244	1.1706	0.2793



Odds Ratio Estimates

Effect	Estimate	Point	95% Wald	
			Confidence Limits	
ptbany	2.944	2.237		3.876
	1.215	0.950	1.553	
	1.213	0.648		2.271
	0.890	0.638		1.240
	1.144	0.897		1.460

1. What is the probability of preterm birth in the sample corresponding to the reference level of all categorical predictors in the model?

2. Are women with prior preterm deliveries at higher or lower risk? Is this risk significantly different from women without a history of preterm delivery? 
 3. Are African American women at higher or lower risk than women of other ethnicities? 
 4. What is the relationship between smoking and preterm risk? 

5. Is being underweight protective or possibly harmful?
 6. What is the effect of education on the outcome?
 7. What is the model-predicted probability of preterm birth for an African American woman with a PhD in biostatistics who is a non-smoker, has not had previous children, and has a normal BMI?

Now, suppose we wish to test whether a model with prior preterm delivery as the only predictor is sufficient. We fit this model below.

```
proc logistic descending;  
model c_case1=ptbany;  
run;
```



```
*****
```

The LOGISTIC Procedure

Model Information

Data Set	WORK.NEW
Response Variable	c_case1
Number of Response Levels	2
Number of Observations	3093
Model	binary logit
Optimization Technique	Fishers scoring

Response Profile

Value	c_case1	Ordered	Total
		Frequency	
1	1	397	
2	0	2696	

Probability modeled is c_case1=1.

NOTE: 70 observations were deleted due to missing values for the response or explanatory variables.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates	Intercept
			and

AIC	2372.762	2314.908
SC	2378.799	2326.982
-2 Log L	2370.762	2310.908

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	59.8536	1	<.0001
Score	71.0827	1	<.0001
Wald	66.5247	1	<.0001

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
-----------	----	----------	----------------	------------	------------

Intercept	1	-2.1093	0.0623	1146.6468	<.0001
ptbany	1	1.0421	0.1278	66.5247	<.0001

Odds Ratio Estimates

Effect	Estimate	Point	95% Wald
		Confidence	Limits
ptbany	2.835	2.207	3.642

What do you conclude about the sufficiency of the smaller model?

Interaction in Logistic Regression

Consider a study on urinary tract infections (UTI) (Koch et al., 1985). Patients were classified as having either complicated (more difficult to cure) or uncomplicated diagnosis of UTI. Because the complicated cases are more difficult to cure, investigators are interested in whether the diagnostic status of the UTI affected the effectiveness of treatment. This hypothesis corresponds to a treatment by diagnosis interaction. In this study, three treatments (A, B, and C) were provided to patients with UTI. The data are given in the table below.

Diagnosis	Treatment	Cured	Not Cured	% Cured
Complicated	A	78	28	0.74
Complicated	B	101	11	0.90
Complicated	C	68	46	0.60
Uncomplicated	A	40	5	0.89
Uncomplicated	B	54	5	0.92
Uncomplicated	C	34	6	0.85

We consider the model



$$\begin{aligned} \text{logit}(p) = & \beta_0 + \beta_1 I(\text{TRTA}) + \beta_2 I(\text{TRTB}) \\ & + \beta_3 I(\text{COMP}) + \beta_4 I(\text{TRTA}, \text{COMP}) \\ & + \beta_5 I(\text{TRTB}, \text{COMP}), \end{aligned}$$

where $I(\cdot)$ represents an indicator variable.

That is,

$$I(E) = \begin{cases} 1 & \text{if } E \text{ is true} \\ 0 & \text{otherwise} \end{cases}.$$

In this model, the reference group is patients on treatment C with uncomplicated diagnoses.

Thus the essence of this design is given by

$$\begin{bmatrix} \text{logit}(Pr(\text{Cure} | A, \text{COMP})) \\ \text{logit}(Pr(\text{Cure} | B, \text{COMP})) \\ \text{logit}(Pr(\text{Cure} | C, \text{COMP})) \\ \text{logit}(Pr(\text{Cure} | A, \text{UNCOMP})) \\ \text{logit}(Pr(\text{Cure} | B, \text{UNCOMP})) \\ \text{logit}(Pr(\text{Cure} | C, \text{UNCOMP})) \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 + \beta_3 + \beta_4 \\ \beta_0 + \beta_2 + \beta_3 + \beta_5 \\ \beta_0 + \beta_3 \\ \beta_0 + \beta_1 \quad \text{[yellow speech bubble]} \\ \beta_0 + \beta_2 \\ \beta_0 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}.$$

Below is the SAS code used to fit the interaction model as well as a reduced model with no diagnosis by treatment interaction.

```
data uti;
input diagnosis $ trt $ cure $ count;
cards;
complicated A cured 78
complicated A not 28
complicated B cured 101
complicated B not 11
complicated C cured 68
complicated C not 46
uncomplicated A cured 40
uncomplicated A not 5
uncomplicated B cured 54
uncomplicated B not 5
uncomplicated C cured 34
uncomplicated C not 6
;

proc logistic;
freq count;
```



```
class diagnosis trt/param=ref;  
model cure=diagnosis|trt;  
run;
```

```
proc logistic;  
freq count;  
class diagnosis trt/param=ref;  
model cure=diagnosis trt;  
run;
```

The SAS output is provided below.

The SAS System

1

The LOGISTIC Procedure

Model Information

Data Set	WORK.UTI
Response Variable	cure
Number of Response Levels	2
Number of Observations	12
Frequency Variable	count
Sum of Frequencies	476
Model	binary logit
Optimization Technique	Fishers scoring

Response Profile

Ordered Value	Total Frequency
cure	

1	cured	375
2	not	101

Probability modeled is cure=cured.

Class Level Information

Design Variables

Class	Value	1	2
diagnosis	complica	1	
	uncompli	0	
trt	A	1	0
	B	0	1
	C	0	0

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	494.029	459.556
SC	498.194	484.549
-2 Log L	492.029	447.556

The SAS System

2

The LOGISTIC Procedure

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
------	------------	----	------------

Likelihood Ratio	44.4726	5	<.0001
Score	44.7864	5	<.0001
Wald	39.9312	5	<.0001

Type III Analysis of Effects

Effect	DF	Wald	
		Chi-Square	Pr > ChiSq
diagnosis	1	7.7653	0.0053
trt	2	1.0069	0.6045
diagnosis*trt	2	2.6384	0.2674

Analysis of Maximum Likelihood Estimates

Parameter	DF	Standard		Wald	
		Estimate	Error	Chi-Square	Pr > ChiSq
Intercept	1	1.7346	0.4428	15.3451	<.0001
diagnosis	1	-1.3437	0.4822	7.7653	0.0053

trt	A	1	0.3448	0.6489	0.2824	0.5952
trt	B	1	0.6445	0.6438	1.0020	0.3168
diagnosis*trt	complica A	1	0.2888	0.7114	0.1649	0.6847
diagnosis*trt	complica B	1	1.1818	0.7428	2.5311	0.1116

The SAS System

3

The LOGISTIC Procedure



Model Information

Data Set	WORK.UTI
Response Variable	cure
Number of Response Levels	2
Number of Observations	12
Frequency Variable	count
Sum of Frequencies	476
Model	binary logit
Optimization Technique	Fishers scoring

Response Profile

Value	cure	Total
		Frequency
1	cured	375
2	not	101

Probability modeled is cure=cured.

Class Level Information

		Design Variables	
Class	Value	1	2
diagnosis	complica	1	
	uncompli	0	
trt	A	1	0
	B	0	1
	C	0	0

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Only Intercept	and Covariates Intercept
AIC	494.029	458.071
SC	498.194	474.733
-2 Log L	492.029	450.071

The SAS System

4

The LOGISTIC Procedure

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	41.9579	3	<.0001
Score	38.8456	3	<.0001
Wald	34.9484	3	<.0001

Type III Analysis of Effects

Effect	DF	Chi-Square	Pr > ChiSq
diagnosis	1	10.2885	0.0013
trt	2	24.6219	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq	Wald

Intercept		1	1.4184	0.2987	22.5505	<.0001
diagnosis	complica	1	-0.9616	0.2998	10.2885	0.0013
trt	A	1	0.5847	0.2641	4.9020	0.0268
trt	B	1	1.5608	0.3160	24.4010	<.0001

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits
diagnosis	complica vs uncomplica	0.382	0.212 0.688
trt	A vs C	1.795	1.069 3.011
trt	B vs C	4.762	2.564 8.847

To test the significance of the interaction effect, we conduct a likelihood ratio test of the full versus reduced models. The null hypothesis is given by $H_0 : \beta_4 = \beta_5 = 0$. The difference between 447.556 (interaction) and 450.071 (no interaction) is 2.515, which we compare to the χ^2_2 distribution because we are testing two interaction terms. The critical value for this distribution is 5.99, so we cannot reject the null hypothesis, and we conclude that the interaction between treatment and diagnosis is not significant.

We see that cure is much less likely for patients with a complicated diagnosis. In addition, both treatments A and B are superior to treatment C.

We cannot tell from the output provided whether treatment B is significantly better than treatment A. In order to test this hypothesis, we need to construct a contrast. The hypothesis of interest is $H_0 : \beta_1 = \beta_2$ and is tested using the following SAS code.

```
proc logistic;
freq count;
class diagnosis trt/param=ref;
model cure=diagnosis trt;
contrast B vs. A trt -1 1/estimate=exp;
/* ESTIMATE=EXP option requests that the OR is printed */
run;
```

The additional SAS output is provided below.

Contrast Test Results

Contrast	DF	Chi-Square	Pr > ChiSq
B vs. A	1	8.6919	0.0032

The SAS System

The LOGISTIC Procedure

Contrast Rows Estimation and Testing Results

Contrast Type	Row Estimate	Standard Error	Alpha	Lower Limit	Upper Limit
B vs. A EXP	1 2.6539	0.8786	0.05	1.3870	5.0778

Contrast Rows Estimation and Testing Results

Contrast Type	Row Chi-Square	Pr > ChiSq
B vs. A EXP	1 8.6919	0.0032

We see that treatment B is indeed superior to treatment A. Patients on treatment B have 2.65 times higher odds of cure than those on treatment A.

We may also calculate predicted probabilities and odds from the main effects model.



Diagnosis	Trt	$Pr(\text{Cure}) = p$	Odds of Cure
Comp	A	$\frac{\exp(\beta_0 + \beta_1 + \beta_3)}{1 + \exp(\beta_0 + \beta_1 + \beta_3)}$	$\exp(\beta_0 + \beta_1 + \beta_3)$
Comp	B	$\frac{\exp(\beta_0 + \beta_2 + \beta_3)}{1 + \exp(\beta_0 + \beta_2 + \beta_3)}$	$\exp(\beta_0 + \beta_2 + \beta_3)$
Comp	C	$\frac{\exp(\beta_0 + \beta_3)}{1 + \exp(\beta_0 + \beta_3)}$	$\exp(\beta_0 + \beta_3)$
Uncomp	A	$\frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}$	$\exp(\beta_0 + \beta_1)$
Uncomp	B	$\frac{\exp(\beta_0 + \beta_2)}{1 + \exp(\beta_0 + \beta_2)}$	$\exp(\beta_0 + \beta_2)$
Uncomp	C	$\frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$	$\exp(\beta_0)$

Goodness of Fit for Logistic Regression

After fitting a model, we need to know just how well that model fits the data. We measure this by the closeness of the model-predicted values to the corresponding observed values. (Is the predicted probability of cure, \hat{p} , close to 0 for those not cured and close to 1 for those cured?)

Goodness-of-fit statistics are test statistics used to assess differences between observed and predicted values, which we expect to be random. If the cell counts are sufficiently large, then these test statistics approximately follow chi-squared distributions.

Consider the following contingency table.



Diagnosis	Treatment	Cured ($j = 1$)	Not Cured ($j = 2$)	
Comp ($h = 1$)	A ($i = 1$)	n_{111}	n_{112}	
Comp ($h = 1$)	B ($i = 2$)	n_{121}	n_{122}	
Comp ($h = 1$)	C ($i = 3$)	n_{131}	n_{132}	
Unoomp ($h = 2$)	A ($i = 1$)	n_{211}	n_{212}	
Unoomp ($h = 2$)	B ($i = 2$)	n_{221}	n_{222}	
Unoomp ($h = 2$)	C ($i = 3$)	n_{231}	n_{232}	



The *Pearson chi-square goodness-of-fit test* compares observed and model-predicted cell counts by computing

$$Q_p = \sum_{h=1}^2 \sum_{i=1}^3 \sum_{j=1}^2 \frac{(n_{hij} - m_{hij})^2}{m_{hij}}, \quad \text{[!]} \quad \text{[!]}$$

where n_{hij} are the observed cell counts, and m_{hij} are the model-predicted counts.

We obtain the model-predicted counts by taking

$$m_{hij} = \begin{cases} \text{[!]} n_{hi} \cdot \hat{p}_{hi} & \text{for } j = 1 \\ n_{hi} \cdot (1 - \hat{p}_{hi}) & \text{for } j = 2 \end{cases}.$$

So to obtain the predicted number of patients cured for level h of diagnosis and level i of treatment, we multiply the total number of subjects with diagnosis h and treatment i , which is n_{hi} , by the model-predicted probability of cure for such subjects, given by \hat{p}_{hi} .

These counts may also be obtained in PROC LOGISTIC with the inclusion of the statement OUTPUT OUT=PREDICT PRED=PROB (and then printing the data in PREDICT).

For example, for subjects with a complicated diagnosis on treatment A, we observed 78 cures and 28 failures. Using the model

$$\text{logit}(p) = \beta_0 + \beta_1 I(\text{TRTA}) + \beta_2 I(\text{TRTB}) \\ + \beta_3 I(\text{COMP}),$$

we obtained $\hat{\beta} = (1.4184 \quad 0.5847 \quad 1.5608 \quad -0.9616)'.$

Thus we compute the expected number

$$m_{111} = \begin{cases} 106 \frac{\exp(1.4184+0.5847-0.9616)}{1+\exp(1.4184+0.5847-0.9616)} & \text{cured} \\ 106 \left(1 - \frac{\exp(1.4184+0.5847-0.9616)}{1+\exp(1.4184+0.5847-0.9616)}\right) & \text{not cured} \end{cases} \\ = \begin{cases} 106(0.74) = 78.35 & \text{expected patients cured} \\ 106(0.26) = 27.65 & \text{expected patients not cured} \end{cases}.$$

The Pearson chi-squared statistic can tell us whether the fit is sufficient.

We use the following SAS code to compute Pearson's chi-squared statistic for the UTI data.

```
proc logistic;
freq count;
class diagnosis trt/param=ref;
model cure=diagnosis trt/scale=none aggregate;
/* SCALE produces goodness-of-fit statistics */
/* AGGREGATE tells LOGISTIC to treat each unique combination of */
/* explanatory variables as a distinct group in computing the GOF stats */
run;
*****
```

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	DF	Value	Value/DF	Pr > ChiSq
Deviance	2	2.5147	1.2573	0.2844
Pearson	2	2.7574	1.3787	0.2519

Number of unique profiles: 6

The non-significance of this chi-squared test means that the model fits adequately.

The *deviance* or likelihood ratio chi-square, Q_L , is another traditional measure of goodness-of-fit, which is given by

$$Q_L = \sum_{h=1}^2 \sum_{i=1}^3 \sum_{j=1}^2 2n_{hij} \log \left(\frac{n_{hij}}{m_{hij}} \right).$$

We interpret this test in the same way as the Pearson chi-squared test. (Note that the ratios will be close to one, and thus their logarithms close to zero, for a model that fits well.)

These two goodness of fit tests are valid only for large enough samples. We need for each group n_{hi} to have at least 10 subjects, we need 80% of the predicted counts m_{hij} to be at least 5, and all other expected counts to be greater than 2 in order for these tests to be valid. If these conditions do not hold, exact methods for logistic regression are available.

These goodness-of-fit statistics are calculated assuming all predictors are categorical. However, continuous predictors are commonly used in logistic regression models. In such a case, sample size requirements for the validity of the above goodness-of-fit tests will rarely be met! Alternatively, fit can be assessed by fitting an appropriate expanded model with additional explanatory variables (including interactions if desired) and examining the difference in the log-likelihoods using a likelihood ratio test. Another strategy is to consider the Hosmer and Lemeshow goodness-of-fit statistic. This test places subjects into deciles based on model-predicted probabilities and then computes a Pearson chi-squared test based on the observed and expected number of subjects in the deciles. The statistic is then compared to a chi-squared distribution and is available by specifying the LACKFIT option in the MODEL statement.

Diagnostics for Logistic Regression

Although goodness-of-fit statistics tell you how well the model fits the data, they do not tell you much about where a particular model fails to fit the data, or *lack of fit*.

Suppose that you have s groups, $i = 1, \dots, s$, with n_i subjects in group i and y_i events in group i . *Pearson residuals* (the sum of their squares is Q_P , the Pearson chi-square goodness-of-fit statistic) are given by

$$\square \quad e_i = \frac{y_i - n_i \hat{p}_i}{\sqrt{n_i \hat{p}_i (1 - \hat{p}_i)}}.$$

These residuals compare the differences between observed counts and their predicted values, scaled by the observed count's standard deviation. Generally, residuals are considered indicative of lack of fit if they exceed 2 in absolute value. (Note that the e_i are formed by subtracting the mean and standardizing by the square root of the variance for binomial data.)

Deviance residuals (the sum of their squares yields the deviance statistic) are given by

$$d_i = \text{sgn}(y_i - \hat{y}_i) \left[2y_i \log \left(\frac{y_i}{\hat{y}_i} \right) + 2(n_i - y_i) \log \left(\frac{n_i - y_i}{n_i - \hat{y}_i} \right) \right]^{\frac{1}{2}},$$

where $\hat{y}_i = n\hat{p}_i$.

The SAS code and output for calculating these residuals is provided below.

```
proc logistic;
freq count;
class diagnosis trt/param=ref;
model cure=diagnosis trt/influence;
run;
*****
Regression Diagnostics
```

Case Number	Covariates		
	diagnosiscomplica	trtA	trtB
1	1.0000	1.0000	0
2	1.0000	1.0000	0
3	1.0000	0	1.0000
4	1.0000	0	1.0000
5	1.0000	0	0
6	1.0000	0	0
7	0	1.0000	0

8	0	1.0000	0
9	0	0	1.0000
10	0	0	1.0000
11	0	0	0
12	0	0	0

Regression Diagnostics

Case Number	Value	Pearson Residual							Deviance Residual						
		(1 unit = 0.55)							(1 unit = 0.31)						
		-8	-4	0	2	4	6	8	-8	-4	0	2	4	6	8
1	0.5941			*					0.7775				*		
2	-1.6833		*						-1.6394		*				
3	0.3647			*					0.4997				*		
4	-2.7422		*						-2.0700		*				
5	0.7958			*					0.9906				*		
6	-1.2566			*					-1.3765		*				
7	0.3673			*					0.5031				*		
8	-2.7226		*						-2.0638		*				
9	0.2255			*					0.3149				*		

10	-4.4352	*				-2.4612	*				
11	0.4920			*			0.6585			*	
12	-2.0324		*				-1.8084		*		

Based on the Pearson and deviance residuals, we see that the model fits poorly for quite a few groups, with the worst fit for uncured uncomplicated diagnoses with treatment B. Perhaps an interaction model (or the addition of other covariates) would improve the model fit. Although it seems that we may have conflicting results from examination of the Pearson residuals and the Pearson chi-squared statistic, we see that while the model does fit the data, there is some suggestion that we can still do better!

Example: Cirrhosis Data

The Mayo Clinic conducted a double-blinded randomized clinical trial in patients with primary biliary cirrhosis (PBC) of the liver to compare the drug D-penicillamine (DPCA) with placebo. PBC is a rare but fatal chronic liver disease (in the years since the trial was completed, the disease has become more treatable due to advances in liver transplantation).

We consider status (1=death, 0=otherwise) as the outcome of interest in comparing the two treatments. In addition, we have information about a variety of covariates, including the following.

- drug (1=DPCA, 0=placebo)
- age in days
- sex (0=male, 1=female)
- presence of ascites (0=no 1=yes)

-
- presence of hepatomegaly (0=no 1=yes)
 - presence of spiders (0=no 1=yes)
 - presence of edema (0=no edema and no diuretic therapy for edema; .5 = edema present without diuretics, or edema resolved by diuretics; 1 = edema despite diuretic therapy)
 - serum bilirubin in mg/dl
 - serum cholesterol in mg/dl
 - albumin in gm/dl
 - urine copper in ug/day
 - alkaline phosphatase in U/liter
 - SGOT in U/ml
 - triglycerides in mg/dl
 - platelets per cubic ml / 1000

-
- prothrombin time in seconds
 - histologic stage of disease

First, we do some data checking.

```
proc means;  
var age bili chol albumin copper alk_phos sgot trig platelet protime;  
run;
```

```
proc freq;  
tables status drug sex ascites hepatom spiders edema stage;  
run;
```

```
*****
```

The SAS System

1

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
age	312	18269.44	3864.81	9598.00	28650.00
bili	312	3.2560897	4.5303153	0.3000000	28.0000000
chol	284	369.5105634	231.9445450	120.0000000	1775.00
albumin	312	3.5200000	0.4198920	1.9600000	4.6400000
copper	310	97.6483871	85.6139199	4.0000000	588.0000000
alk_phos	312	1982.66	2140.39	289.0000000	13862.40



sgot	312	122.5563462	56.6995249	26.3500000	457.2500000
trig	282	124.7021277	65.1486387	33.0000000	598.0000000
platelet	308	261.9350649	95.6087423	62.0000000	563.0000000
protime	312	10.7256410	1.0043232	9.0000000	17.1000000

The SAS System

2

The FREQ Procedure

status	Frequency	Percent	Cumulative	Cumulative
			Frequency	Percent
0	187	59.94	187	59.94
1	125	40.06	312	100.00

drug	Frequency	Percent	Cumulative	Cumulative
			Frequency	Percent
0	154	49.36	154	49.36
1	158	50.64	312	100.00

sex	Frequency	Percent	Cumulative	Cumulative
			Frequency	Percent
<hr/>				
0	36	11.54	36	11.54
1	276	88.46	312	100.00

ascites	Frequency	Percent	Cumulative	Cumulative
			Frequency	Percent
<hr/>				
0	288	92.31	288	92.31
1	24	7.69	312	100.00

hepatom	Frequency	Percent	Cumulative	Cumulative
			Frequency	Percent
<hr/>				
0	152	48.72	152	48.72
1	160	51.28	312	100.00

spiders	Frequency	Percent	Cumulative Frequency	Cumulative Percent
<hr/>				
0	222	71.15	222	71.15
1	90	28.85	312	100.00

edema	Frequency	Percent	Cumulative Frequency	Cumulative Percent
<hr/>				
0	263	84.29	263	84.29
0.5	29	9.29	292	93.59
1	20	6.41	312	100.00

The SAS System

3

The FREQ Procedure

stage	Frequency	Percent	Cumulative Frequency	Cumulative Percent
-------	-----------	---------	----------------------	--------------------

1	16	5.13	16	5.13
2	67	21.47	83	26.60
3	120	38.46	203	65.06
4	109	34.94	312	100.00



Based on this checking, we will convert age into years and pseudo-center at age 50. In addition, we will convert alkaline phosphate into U/kL (divide by 1000). We will also create two edema variables: EDEMAD for those without relief from diuretics, and EDEMAND for the patients with milder cases. We will also create four variables for disease stage: STAGE1, STAGE2, STAGE3, and STAGE4. The SAS code for making these changes is presented below.

```
data pbc;  
set pbc;  
age=age/365.25-50;  
alk_phos=alk_phos/1000;  
edema_d=0;
```

```
edema_nd=0;  
if edema=1 then edema_d=1;  
if edema=.5 then edema_nd=1;  
stage1=0; stage2=0; stage3=0; stage4=0;  
if stage=1 then stage1=1;  
if stage=2 then stage2=1;  
if stage=3 then stage3=1;  
if stage=4 then stage4=1;  
run;
```

First, we fit a large model including all the predictors.

```
proc logistic descending;
model status=drug sex ascites hepatom spiders edema_d edema_nd stage2
    stage3 stage4 age bili chol albumin copper alk_phos sgot trig
    platelet protime;
run;
```

The SAS System

1

The LOGISTIC Procedure

Model Information

Data Set	WORK.PBC
Response Variable	status
Number of Response Levels	2
Number of Observations	276
Model	binary logit
Optimization Technique	Fishers scoring

Response Profile

Ordered Value	status	Total Frequency
1	1	111
2	0	165

Probability modeled is status=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Only	Intercept and Covariates
-----------	------	--------------------------------

AIC	373.984	277.117
SC	377.604	353.145
-2 Log L	371.984	235.117

 Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	136.8671	20	<.0001
Score	108.2980	20	<.0001
Wald	64.5753	20	<.0001

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard	Wald	Pr > ChiSq
			Error	Chi-Square	
Intercept	1	-12.7956	3.8279	11.1739	0.0008

drug	1	0.3091	0.3413	0.8201	0.3652
sex	1	-0.6448	0.5355	1.4500	0.2285
ascites	1	1.0535	1.3898	0.5746	0.4484
hepatom	1	0.1144	0.4000	0.0818	0.7749
spiders	1	0.3533	0.4001	0.7797	0.3772
edema_d	1	0.7758	1.4717	0.2779	0.5981
edema_nd	1	-0.0705	0.6022	0.0137	0.9068
stage2	1	2.5956	1.4906	3.0323	0.0816
stage3	1	2.8408	1.4747	3.7106	0.0541
stage4	1	2.8148	1.4855	3.5905	0.0581
age	1	0.0519	0.0183	8.0751	0.0045
bili	1	0.1532	0.0835	3.3689	0.0664
chol	1	0.000322	0.000857	0.1407	0.7076
albumin	1	-0.1408	0.5006	0.0791	0.7786
copper	1	0.00285	0.00250	1.3027	0.2537
alk_phos	1	0.2708	0.0898	9.0973	0.0026
sgot	1	0.00584	0.00322	3.2853	0.0699
trig	1	0.00248	0.00330	0.5644	0.4525
platelet	1	-0.00003	0.00199	0.0002	0.9880
protime	1	0.7345	0.2143	11.7498	0.0006

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
drug	1.362	0.698	2.659
sex	0.525	0.184	1.499
ascites	2.868	0.188	43.704
hepatom	1.121	0.512	2.455
spiders	1.424	0.650	3.119
edema_d	2.172	0.121	38.873
edema_nd	0.932	0.286	3.034
stage2	13.405	0.722	248.900
stage3	17.129	0.952	308.348
stage4	16.690	0.908	306.819
age	1.053	1.016	1.092
bili	1.166	0.990	1.373
chol	1.000	0.999	1.002
albumin	0.869	0.326	2.317
copper	1.003	0.998	1.008
alk_phos	1.311	1.099	1.563
sgot	1.006	1.000	1.012

trig	1.002	0.996	1.009
platelet	1.000	0.996	1.004
protime	2.085	1.370	3.173

Suppose that based on scientific reasoning supplied in advance by the investigator, a reduced model was fit to the data. This model did not contain the variables hepatom, edema, chol, albumin, sgot, trig, or platelet.

```
proc logistic descending;
model status=drug sex ascites spiders stage2 stage3 stage4 age bili copper alk_phos
run;
*****
```

The LOGISTIC Procedure

Model Information

Data Set	WORK.PBC
Response Variable	status
Number of Response Levels	2
Number of Observations	276
Model	binary logit

Optimization Technique

Fishers scoring

Response Profile

Ordered Value	status	Total Frequency
1	1	111
2	0	165

Probability modeled is status=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Intercept

Criterion	Intercept Only	and Covariates
AIC	373.984	265.949
SC	377.604	313.014
-2 Log L	371.984	239.949

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	132.0346	12	<.0001
Score	104.5750	12	<.0001
Wald	62.6444	12	<.0001

The SAS System

5

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald	
				Chi-Square	Pr > ChiSq
Intercept	1	-11.5823	3.0107	14.8000	0.0001
drug	1	0.2729	0.3327	0.6732	0.4119
sex	1	-0.6684	0.5141	1.6903	0.1936
ascites	1	1.3471	1.1475	1.3782	0.2404
spiders	1	0.3759	0.3805	0.9757	0.3233
stage2	1	2.6314	1.4146	3.4605	0.0629
stage3	1	2.9260	1.3958	4.3941	0.0361
stage4	1	2.8652	1.3750	4.3424	0.0372
age	1	0.0461	0.0170	7.3783	0.0066
bili	1	0.2330	0.0727	10.2842	0.0013
copper	1	0.00301	0.00247	1.4928	0.2218
alk_phos	1	0.3000	0.0874	11.7794	0.0006
protime	1	0.6613	0.1986	11.0904	0.0009

Odds Ratio Estimates

Point 95% Wald

Effect	Estimate	Confidence Limits	
drug	1.314	0.685	2.522
sex	0.513	0.187	1.404
ascites	3.846	0.406	36.461
spiders	1.456	0.691	3.070
stage2	13.894	0.868	222.276
stage3	18.652	1.209	287.655
stage4	17.553	1.186	259.844
age	1.047	1.013	1.083
bili	1.262	1.095	1.456
copper	1.003	0.998	1.008
alk_phos	1.350	1.137	1.602
protimes	1.937	1.313	2.859

To test whether the reduced model is sufficient, we may compare log-likelihoods:

239.949 (smaller) - 235.117 (larger)=4.832. Comparing this to a χ^2_8 random variable (with critical value 15.51), we conclude that the reduced model is sufficient.

Although we could reduce the model further, the investigator wanted to leave the remaining terms in the model for scientific reasons (to show he/she adjusted for the other factors).



1. What effect does treatment have on survival?
 2. What factors work to lengthen survival? How do you interpret the benefit of these factors?
 3. What factors are associated with diminished survival? How do you interpret their effects?

Lecture 19: Poisson Regression

Poisson Regression



Recall that a random variable y has Poisson distribution if

$$Pr(y = k) = \frac{\mu^k}{k!} e^{-\mu}, \mu > 0, k = 1, 2, \dots$$

We have $E(y) = var(y) = \mu > 0$. This distribution is used for counts of events occur randomly over time or space, when outcomes in disjoint periods or regions are independent. Examples are plentiful such as traffic accidents, the incidence of rare events or diseases, etc.

African Elephants Data The first of our examples deals with the number of successful matings of 41 male African elephants over a period of eight years, and to examine to what extent these numbers depend on their ages at the onset of the study.

Age Matings Age Matings Age Matings



27	0	33	3	39	1
28	1	33	3	41	3
28	1	33	3	42	4
28	1	33	2	43	0
28	3	34	1	43	2
29	0	34	1	43	3
29	0	34	2	43	4
29	0	34	3	43	9
29	2	36	5	44	3
29	2	36	6	45	5
29	2	37	1	47	7
30	1	37	1	48	2

32 2 37 6 52 9

33 4 38 2

 **Poisson Regression** In view of these count data, we will model the conditional distribution of the response Y given the explanatory variables X using the Poisson distribution whose mean is given by $\mu = \mu(\mathbf{x}) = E(y | \mathbf{x}) \geq 0$. To ensure this constraint, we assume

$$\log \mu(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \boldsymbol{\beta}^T \mathbf{x}$$

This is called a **Poisson loglinear model**. The coefficient, say β_1 of $x_1 = \text{age}$, can be interpreted as follows: if we increase age by one unit while keeping the remaining variables fixed. This affects the mean μ of the Poisson response by $100(e^{\beta_1} - 1)\%$.

Estimation

- We obtain the maximum likelihood estimate (MLE) $\hat{\beta}$ of β by maximizing the log-likelihood:

$$l(\beta) = \sum_i^n y_i \log \mu_i - \sum_i^n \mu_i \quad \text{💡}$$

Also, the MLE is asymptotically normal:

$$\hat{\beta} \sim N(\beta, \hat{V})$$

- The MLE of the mean is $\mu(\mathbf{x}) = \exp(\hat{\beta}\mathbf{x})$. In particular,
 $\hat{\mu}_i = \hat{\mu}(\mathbf{x}_i) = \exp(\hat{\beta}\mathbf{x}_i)$, $i = 1, 2, \dots, n$.
- The covariance of $\hat{\beta}$, \hat{V} is estimated by

$$\widehat{COV}(\hat{\beta}) = (\sum_i \hat{\mu}_i \mathbf{x}_i \mathbf{x}_i^T)^{-1}$$

The $se(\hat{\beta}_j)$ is the square root of the diagonal elements of the

above matrix.

- Approximate 95% CI for the mean are

$$\exp \left(\hat{\beta} \mathbf{x}_i \pm 1.96 \sqrt{\mathbf{x}_i^T \left(\sum_i \hat{\mu}_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \mathbf{x}_i} \right),$$

some other ways to calculate the CI might be better.

- The deviance is


$$D = 2 \sum_i y_i \log(y_i/\hat{\mu}_i) \approx \sum \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \sim \chi^2$$

, which is the Pearson's chi-square statistic with $n - p - 1$ df. The model for the mean is doubtful when the chi-square statistic is much greater than $n - p - 1$.

- The deviance residuals are defined by

$$d_i = sign(y_i - \hat{\mu}_i) \sqrt{2(y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i))}.$$

Pearson residuals are given by

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}.$$

Elephant Data

```
data a;  
infile elephants.txt firstobs=2;  
input age matings;  
  
proc genmod;  
model matings = age / dist = poisson  
                    lrci covb;  
  
proc genmod;  
model matings = age age*age / dist = poisson  
                    lrci covb;
```

```
run;
```

Ceriodaphnia Organisms Data This data set contains the number of Ceriodaphnia organisms that are counted in a controlled environment in which reproduction occurs among the organisms, and there are two strains involved. The data is located at ceriodaphnia.txt. Ceriodaphnia represents the number of reproduction occurred, Concentration is the chemical component in the environment that impairs the reproduction, and Strain is the strain type.

```
Cerio Conc Strain;
```

```
datalines;
```

```
82 0.00 1  
106 0.00 1  
63 0.00 1  
45 0.50 1  
34 0.50 1  
26 0.50 1  
11 1.50 1
```

```
10  1.50    1
10  1.50    1
14  1.00    2
14  1.00    2
9   1.25    2
12  1.25    2
16  1.25    2
7   1.50    2
```

...

Ceriodaphnia Organisms Data

```
proc import datafile = ceriodaphnia.csv
  out = Ceriodaphnia
  dbms = CSV
  REPLACE
;
```

```
proc genmod;
model Cerio = Conc Strain / dist = poisson
            lrci covb;
run; quit;
```

Skin Cancer Example:

- Skin cancer for women in Dallas and Minn
- 8 age group and two cities (0 = Minn, 1 = Dallas)
- y_{ij} = count (number of cases) in age i and city j
- n_{ij} = population size in age i and city j

Objective: Determine whether the risk for skin cancer adjusted for age is higher in one area than in the other.

$risk$ = probability of developing skin cancer

λ_{ij} = probability of developing skin cancer in (i, j) th group.

y	city	age	pop
1	0	1	172675
16	0	2	123065
30	0	3	96216
71	0	4	92051
102	0	5	72159
130	0	6	54722
133	0	7	32185
40	0	8	8328
4	1	1	181343
38	1	2	146207
119	1	3	121374
221	1	4	111353
259	1	5	83004
310	1	6	55932
226	1	7	29007
65	1	8	7538

Poisson Regression y_{ij} has a poisson distribution with mean

$$E(y_{ij}) = n_{ij}\lambda_{ij}.$$

Poisson regression concerns the model

$$\log(y_{ij}) = \log(n_{ij}) + \beta_0 + \sum_{i=1}^7 \beta_i age_i + \beta_8 city,$$

which is equivalent to

$$\log(\lambda_{ij}) = \beta_0 + \sum_{i=1}^7 \beta_i age_i + \beta_8 city.$$

Interpretations:

- $\beta_0 = \log(\lambda_{81}) = \log$ of the risk in age 8 group in Dallas.
- $\beta_i = \log(\lambda_{i1}) = \log$ of the risk in age i group in Dallas (for $i = 1, 2, \dots, 7$).
- $\beta_0 + \beta_8 = \log(\lambda_{80}) = \log$ of the risk in age 8 group in Minn.



-
- $\beta_i + \beta_8 = \log(\lambda_{i0}) = \log$ of the risk in age i in Minn (for $i = 1, 2, \dots, 7$).
 - $\beta_8 = \log(\lambda_{i1}) - \log(\lambda_{i0}) = \log\left(\frac{\lambda_{i1}}{\lambda_{i0}}\right)$
 - $RR_i = \frac{\lambda_{i0}}{\lambda_{i1}} = \exp(\beta_8), i = 1, \dots, 8$ where RR_i is the relative risk in age i between two cities.

SAS Code

```
*-----  
---*  
| Comparison of incidence of nonmelanoma skin cancer among |  
| women in Minneapolis St. Paul and Dallas Ft Worth. |  
| KKMN, 3e Table 24-1, p688 |  
*-----  
---*;
```

```
title LOGISTIC AND POISSON REGRESSION;
```

```
proc import datafile = skincancer.csv  
out = skin  
dbms = CSV  
REPLACE  
;
```

```
data skin;  
set skin;  
lpop = log(pop);  
run;
```

```
proc genmod;  
class age city;
```

```
model y = age city/ dist  = poisson  
offset= lpop lrci;  
run;
```

```
proc genmod;  
  class age city;  
  model y/pop = age city/ dist  = poisson  
        lrci;  
run;
```

```
proc genmod;  
  class age city;  
  model y/pop = age city/ dist  = bin  
        lrci;  
run;
```

Note: For those situations in which n_{ij} is large and λ_{ij} is very small, the Poisson distribution can be used to approximate the binomial distribution. The larger the n_{ij} and the smaller the λ_{ij} , the better is the approximation. For the skin cancer data, we expect λ_{ij} to be small, therefore, the two models (Binomial and Poisson) should give very similar conclusions.

When studying linear regression, our models were of the form

$$E[\mathbf{y}] = \beta_0 + \beta_1 \mathbf{x}_1 + \dots + \beta_{p-1} \mathbf{x}_{p-1}.$$

The response \mathbf{y} was continuous, not discrete, and we wanted to predict the mean response and explain the variability among the observed outcomes.

Lecture 19: Random Effects Model

Reading Assignment:

- Muller and Fetterman, Chapter 15: “Special Cases of Two-Way ANOVA and Random Effects” (Required)
- Littell, Milliken, Stroup, and Wolfinger. *SAS System for Mixed Models*. (Great SAS reference for longitudinal data analysis.)

Mixed effects models are an extension of the GLM for correlated data.

Applications include

- clustering by center, clinic, or primary care physician,
- responses measured on individual mice in a litter,
- meta-analysis to combine the results of many different studies, and

-
- repeated measures over time.

In addition, factors in an experiment may be considered *fixed* or *random*. We used *fixed effects* when we wish to make inferences only about the particular k levels of the factor observed in the study.

Examples include gender, treatment, or dose. We use *random effects*  when these levels are a sample from a population of levels. Examples include subjects, observers, and clinic sites. 

In order to decide whether a given factor should be treated as random or fixed, consider repetition of the experiment. If you repeated an experiment testing a new drug on patients in 5 Chapel Hill psychiatric clinics, what would have to be the same to make the experiment a repeat? Presumably, the second experiment should have the same drugs and doses as the first (fixed factors), but using 5 different psychiatric clinics may not really matter (random factor).

So far, we have considered models of the form

$$y = \text{fixed effects} + \text{error}.$$

The variance in y is partitioned into what can be explained by fixed effects and what remains unexplained. The error term is the only part of the right hand side of that equation that has any variance. In addition, recall that we assume for the `glm` that error terms are independent.

If a model contains a random factor, then we have the form

$$y = \text{fixed effects} + \text{random effects} + \text{error}.$$

The random factor also has variance, and we now partition variance into that due to fixed factors, random factors, and unexplained factors. While we assume that subjects with different levels of the random factor are independent, we allow subjects at the same level of the random factor to have correlated errors.

Introduction to Clustered Data

Multiple responses (y 's), called *repeated measures*, are often taken per subject. If the repeated measures are recorded over time, we call these *longitudinal data*. The set of one *subject's* repeated measures make up a *cluster*.

Clustered data do not have to be longitudinal. For example, clusters could be defined by members families participating in a genetic study or patients in medical clinics across North Carolina. In the family  study, we assume subjects from different families will be independent, but that subjects within a family may be correlated.

Clustered data generally signal a violation of the homogeneity of variances assumption because subjects within a cluster are typically more alike than subjects in different clusters. Regression procedures must take into account the correlation between subjects within a cluster, and assuming falsely that observations in a cluster are

independent may give invalid results.

The correlation (dependence) structure in the data takes a specific form. Clustered data imply that observations inside a cluster (family, subject) are correlated but observations from different clusters are uncorrelated.

Longitudinal Analysis

Longitudinal studies have designs in which the outcome variable is measured repeatedly over time.

Examples include

- The ARIC (Atherosclerosis Risk in Communities) Study at CSCC, in which over 15,000 subjects were examined at baseline and every three years for a number of cardiovascular endpoints.
- A study at the CPC that monitors women during pregnancy and for one year after delivery to investigate factors related to lifetime weight gain.
- HIV clinical trials in which viral load levels are monitored throughout the course of a patient's disease.

Scientific advantages of longitudinal study designs include the following.

- Longitudinal studies allow investigation of events that occur over time, which is essential to the study of growth, aging, or the course of disease.
- Longitudinal studies allow us to study the order of events.
- Longitudinal studies permit more complete ascertainment of exposure histories in epidemiologic studies.
- Longitudinal studies can reduce unexplained variability in response by using the subject as his or her own control (crossover studies).

Randomized-blocks Experiment

Blocking is a powerful experimental design technique when the experimental subjects are heterogeneous with respect to certain variables that are associated with the response but are not of primary interest. Blocking consists of the following two steps:

1. grouping homogeneous experimental units together to form a block, and
2. assigning treatments at random to experimental units within a block.

Blocking prevents one type of subject from predominantly receiving one type of treatment. Suppose we conduct a national clinical trial of three treatments in 100 hospitals across the United States. Because we feel that patients within a hospital may be more alike than patients across hospitals (due to factors like population served by hospital, quality of nursing care in hospital, etc.), we wish to treat each hospital

as a block. Within each hospital (block), we will randomly assign the three treatments.

Components of Variance

Until now, we have considered only one variance, σ^2 , in our regression models. When we use random effects, we have additional variance terms to estimate. For example, we may treat hospitals as random factors in a multi-center study if we view these hospitals as random samples from a population of US hospitals (but aren't particularly interested in hospital performance). If the variance due to the hospitals is much larger than the random error, for example, then this indicates that hospitals are extremely variable, and future studies would benefit by choosing a greater number of hospitals. If hospital variance is low, then we might have saved money by recruiting more patients in fewer centers.

General Linear Mixed-Effects Model

The *general linear mixed-effects model* is often written

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i,$$

where \mathbf{Y}_i is the vector of responses for subject i , \mathbf{X}_i is the vector of fixed covariates for subject i , $\boldsymbol{\beta}$ is a p -dimensional parameter vector, \mathbf{Z}_i is a matrix of known covariates, and \mathbf{b}_i is a q -dimensional vector containing the random effects. We assume that $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$, $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \Sigma_i)$, and $\mathbf{b}_1, \dots, \mathbf{b}_n, \boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_n$ are independent. This model is also commonly called the Laird-Ware model after its developers (Laird and Ware, 1982).

In BIOS 663, we will just consider \mathbf{Z}_i to be indicator variables denoting cluster (e.g., family) membership, though more complicated structures are possible.

Randomized Block Design Example

Three drugs that were thought to affect lymphocyte production were included in an experiment along with a placebo drug (a control that has no physiological properties). To control for potential variation among experimental units (here, mice), it was decided to block together mice from the same litter (same mother and father and born at the same time). Five (5) litters of mice were used where four (4) mice were selected from each litter, then one of the four drugs was randomly assigned (without replacement) to each mouse within each litter. This resulted in a balanced randomized complete block design. A complete block is a block that receives all levels of the treatments. After a sufficient period of time, a blood sample was drawn from each mouse and the number of lymphocytes per cubic millimeter was determined. The data are represented as thousands of lymphocytes per cubic millimeter. The data are taken from Mead et al. (Mead, R., R.N. Curnow, and A.M. Hasted. 1993. Statistical Methods in Agriculture

and Experimental Biology. Chapman and Hall, London, p656).

Note that the randomization of assignments of drug levels to mice is performed independently for each litter. Since the drugs that are used in the experiment were purposefully selected and interest will be in the mean response of each drug it is reasonable to treat the drugs as fixed effects. The litters of mice, however, should be considered as a random effect. If the experiment were repeated again it is unlikely that these same litters would be used again. Further, interest would be in making inferences over all potential litters of mice rather than just the 5 litters used in this experiment. This randomized complete block model could be written as:

$$y_{ij} = \mu + \alpha_j + b_i + \varepsilon_{ij},$$

where $i = 1, \dots, 5$ and $j = 1, \dots, 4$. As usual, $\varepsilon_{ij} \sim N(0, \sigma^2)$, and we let $b_i \sim N(0, \sigma_b^2)$. y_{ij} is the lymphocyte count on the mouse in litter i receiving drug j , μ is the overall mean, α_j is the effect of the j th drug, b_i is the effect of the i th block (i.e., the i th litter), and b_i and ε_{ij} are

independent random variables. Note that we have assumed that the block effect and the treatment effect do not interact. It is the investigators responsibility to select blocks appropriate for the experiment that will not interact with the treatment. If interaction is possible, then it may be better to include the “block” in the experiment as a random treatment factor and use a factorial design so that the interaction can be investigated.

As there is a control treatment in the experiment, it may be desired to make comparisons of each treatment (drug) mean with the control level to determine separately for each drug whether or not it had an effect. This might be important in a drug screening study where interest is in determining which drugs to continue to pursue, rather than comparing among drugs to determine which is best or which ones differ in thier effects. The Dunnetts multiple comparison procedure is useful in this instance. Dunnetts procedure controls the type I error rate over the family of statements of each mean to the control. Note that comparisons of this type can also be oneided comparisons as you

may only be interested in differences of means versus the control that are positive (or negative) (drug mean is larger than the control, drug mean is less than the control). PROC MIXED permits both two and one sided comparisons. For two sided comparisons the LSMEANS statement is coded as

LSMEANS effect / DUNNETTS PDIFF=CONTROL('control level value');

while the oneided comparisons are coded as

LSMEANS effect / DUNNETTS PDIFF=CONTROLU('control level value');

where treatments are expected to be higher than the control, or

LSMEANS effect / DUNNETTS PDIFF=CONTROLL('control level value');

where treatments are expected to be below the control. In this experiment we will assume that we want the drugs to enhance (increase) the production of lymphocytes so we will use the oneided, upper comparison coding. Of course, these decisions should be made

prior to performing the data analysis and should be written in as a part of the project proposal.

Before we proceed with the analysis, let's investigate some of the ways in which blocking will affect the results and our summaries of the data, and hopefully we will also gain some insight into how blocking works and when to decide to use it. First, let's compute the variance of an individual observation y_{ij}

$$\begin{aligned}Var(y_{ij}) &= Var(\mu + \alpha_j + b_i + e_{ij}) = Var(b_i + e_{ij}) = \\&Var(b_i) + Var(e_{ij}) + Cov(b_i, e_{ij}) = \sigma_b^2 + \sigma^2.\end{aligned}$$

From this result we can compute the variance of the means of the drugs as

$$Var(\bar{y}_j) = \frac{1}{n^2} \sum_i Var(y_{ij}) = \frac{1}{n^2} \sum_i (\sigma_b^2 + \sigma^2) = \frac{1}{n} (\sigma_b^2 + \sigma^2)$$

where $n = 5$ is the number of blocks or litters in the experiment. Now let's investigate the covariance and correlation between measurements.

Note that measurements made on mice from different litters are independent as defined by our model since both random effects are independently distributed. However, mice from the same litter are not

independent as they share the common block or litter random effect. Thus the covariance between mice on treatments (drugs) j and j' from litter i is given by:

$$\begin{aligned} \text{Cov}(y_{ij}, y_{ij'}) &= \text{Cov}(\mu + \alpha_j + b_i + e_{ij}, \mu + \alpha'_{j'} + b_i + e_{ij'}) \\ &= \text{Cov}(b_i + e_{ij}, b_i + e_{ij'}) = \text{Var}(b_j) = \sigma_b^2. \end{aligned}$$



```
*****
* lympho.sas *;
* *;
* To study the effects of four drugs (A, B, C, and D, where D is a
* placebo), four (4) mice from each of five (5) litters were used,
* where litters were treated as blocks. Drugs were randomly *;
* assigned to the mice within each litter. Lymphocyte counts *;
* (thousands per cubic millimeter) were measured on each mouse. *;
* *;
* Source: Mead, R., R.N. Curnow, and A.M. Hasted. 1993. Statistical
* Methods in Agriculture and Experimental Biology. Chapman *;
```

* and Hall, London, p656. S540.S7.M4.1993 *;

Title1 Drug Effects on Lymphocyte Counts;

options ps=55 ls=80 pageno=1 nodate;

data lympho;

input drug \$ litter lcytes;

label drug=drug litter=litter lcytes=lymphocyte counts;

datalines;

A 1 7.1

B 1 6.7

C 1 7.1

D 1 6.7

A 2 6.1

B 2 5.1

C 2 5.8

D 2 5.4

A 3 6.9

```
B 3 5.9
C 3 6.2
D 3 5.7
A 4 5.6
B 4 5.1
C 4 5.0
D 4 5.2
A 5 6.4
B 5 5.8
C 5 6.2
D 5 5.3
;
```

```
proc print data=lympho label;
run;
```

```
proc plot data=lympho;
```

```
plot lcytes*litter=* $ drug;  
run;  
  
/* Fit an RBD Model */  
proc mixed data=lympho method=REML covtest;  
class litter drug;  
model lcytes = drug;  
random litter;  
lsmeans drug / adjust=dunnett pdiff=control(D);  
/* compare drugs to the placebo level, oneailed test */  
lsmeans drug / adjust=dunnett pdiff=controlu(D);  
run;
```

Drug Effects on Lymphocyte Counts

20

The Mixed Procedure

Model Information

Data Set WORK.LYMPHO

Dependent Variable LCytes

Covariance Structure	Variance Components
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Class Level Information

Class	Levels	Values
Litter	5	1 2 3 4 5
Drug	4	A B C D

Dimensions

Covariance Parameters	2
Columns in X	5
Columns in Z	5
Subjects	1

Max Obs Per Subject	20
Number of Observations	
Number of Observations Read	20
Number of Observations Used	20
Number of Observations Not Used	0

Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	38.70809588	
1	1	18.49496651	0.00000000

Convergence criteria met.

Drug Effects on Lymphocyte Counts 21

The Mixed Procedure

Covariance Parameter Estimates

Standard	Z			
Cov Parm	Estimate	Error	Value	Pr Z
Litter	0.3869	0.2830	1.37	0.0858
Residual	0.05308	0.02167	2.45	0.0072

Fit Statistics

-2 Res Log Likelihood	18.5
AIC (smaller is better)	22.5
AICC (smaller is better)	23.4
BIC (smaller is better)	21.7

Type 3 Tests of Fixed Effects

Num	Den			
Effect	DF	DF	F Value	Pr > F

Drug	3	12	11.59	0.0007
------	---	----	-------	--------

Least Squares Means

Standard

Effect	Drug	Estimate	Error	DF	t Value	Pr > t
Drug	A	6.4200	0.2966	12	21.64	<.0001
Drug	B	5.7200	0.2966	12	19.28	<.0001
Drug	C	6.0600	0.2966	12	20.43	<.0001
Drug	D	5.6600	0.2966	12	19.08	<.0001
Drug	A	6.4200	0.2966	12	21.64	<.0001
Drug	B	5.7200	0.2966	12	19.28	<.0001
Drug	C	6.0600	0.2966	12	20.43	<.0001
Drug	D	5.6600	0.2966	12	19.08	<.0001

Differences of Least Squares Means

Standard

Effect	Drug	Drug	Estimate	Error	DF	t Value	Tails
Drug	A	D	0.7600	0.1457	12	5.22	Both
Drug	B	D	0.06000	0.1457	12	0.41	Both
Drug	C	D	0.4000	0.1457	12	2.75	Both
Drug	A	D	0.7600	0.1457	12	5.22	Upper
Drug	B	D	0.06000	0.1457	12	0.41	Upper
Drug	C	D	0.4000	0.1457	12	2.75	Upper

Differences of Least Squares Means

Effect	Drug	Drug	Adjustment	Adj P
Drug	A	D	Dunnett-Hsu	0.0006
Drug	B	D	Dunnett-Hsu	0.9543
Drug	C	D	Dunnett-Hsu	0.0448
Drug	A	D	Dunnett-Hsu	0.0003
Drug	B	D	Dunnett-Hsu	0.5844
Drug	C	D	Dunnett-Hsu	0.0224

“Screamer” Study

A study was designed to investigate different strategies for reducing disruptive vocalizations among nursing home residents with Alzheimer’s disease. Each resident in the study received four treatments (3 interventions and a control). This *crossover study* (patients received all the treatments) was designed with a “washout” period between treatments so that there would be no “carryover” effects.

Study subjects were nursing home residents, and the observational units were resident responses on each treatment. There were twelve residents (clusters) aged 71-87, 11 of whom were female. We have 4 observations (treatment responses) in each cluster. We call treatment the *crossover factor*.

This study design, in which each subject serves as his/her own control, is more powerful than a study that randomizes one-fourth (or 3)

residents to one of the four treatments. The ability to make within-subject comparisons allows us to control for much person-to-person variability.

The response of interest is the percentage of 120 fifteen-second audiotape samples with vocalizations above a fixed decibel level. The four treatments (administered on different days) were a control, a stuffed teddy bear, headphones with music, and both the teddy bear and headphones with music.

We wish to know if there are any differences in the effectiveness of the treatments.

Consider the following hypothetical subset of the data.

Subject	Control	Bear	Music	Both
1	45	33	8	37
5	23	16	13	9
6	22	28	31	12
7	46	46	30	16
8	68	58	68	48
10	13	13	7	8
11	18	27	17	9
13	35	32	51	4
15	49	44	21	27
16	57	57	74	16
19	36	30	46	30
21	33	21	21	21

Using PROC MEANS in SAS, we obtain the following summary statistics.

Treatment	Mean	Std Dev
Control	37.08	16.59
Bear	33.75	14.73
Music	32.25	22.66
Both	19.75	13.35

We will again use the model



$$y_{ij} = \mu + \alpha_j + b_i + \varepsilon_{ij},$$

where $i = 1, \dots, 12$ and $j = 1, \dots, 4$. As usual, $\varepsilon_{ij} \sim N(0, \sigma^2)$. But how about the distribution of b_i ?

Correlation

Two variables, Y_1 and Y_2 , are *independent* if the conditional distribution of $Y_1 | Y_2$ does not depend on Y_2 .

Two variables, Y_1 and Y_2 , are *uncorrelated* (i.e., not *linearly* dependent) if $E[(Y_1 - \mu_{Y_1})(Y_2 - \mu_{Y_2})] = 0$. Note that variables can be uncorrelated without being independent (e.g., $Y_1 = (-1, 0, 1)$ and $Y_2 = (1, 0, 1)$).

Two variables are *correlated* if $E[(Y_1 - \mu_{Y_1})(Y_2 - \mu_{Y_2})] \neq 0$. The *covariance* between Y_1 and Y_2 is given by $E[(Y_1 - \mu_{Y_1})(Y_2 - \mu_{Y_2})]$. Covariance can take any value and will be dependent on the units of the variables of interest. To make it independent of the units, we can divide by the standard deviations of the two variables to obtain the

correlation, which must lie between -1 and 1:

$$\text{Corr}(Y_1, Y_2) = \frac{E[(Y_1 - \mu_{Y_1})(Y_2 - \mu_{Y_2})]}{\sigma_{Y_1} \sigma_{Y_2}}.$$

Repeated measures from the same person are usually *positively* correlated.

We use SAS to calculate the covariance and correlation matrices for the screamer study below.

```
data new;
input subject control bear music both;
cards;
1 45 33 8 37
5 23 16 13 9
6 22 28 31 12
7 46 46 30 16
8 68 58 68 48
10 13 13 7 8
11 18 27 17 9
13 35 32 51 4
```

```
15   49   44   21   27
16   57   57   74   16
19   36   30   46   30
21   33   21   21   21
; 
```

```
proc corr data=new cov;
var control bear music both;
run;
*****
```

The CORR Procedure

4 Variables: control bear music both



Covariance Matrix, DF = 11

	control	bear	music	both
control	275.3	223.9	251.9	160.1
bear	223.9	217.1	249.7	103.4
music	251.9	249.7	513.6	79.1

both	160.1	103.4	79.1	178.2
------	-------	-------	------	-------

Simple Statistics					
Variable	N	Mean	Std Dev	Sum	Minimum
control	12	37.08333	16.59386	445.00000	13.00000
bear	12	33.75000	14.73478	405.00000	13.00000
music	12	32.25000	22.66405	387.00000	7.00000
both	12	19.75000	13.34933	237.00000	4.00000

Pearson Correlation Coefficients, N = 12

Prob > |r| under H0: Rho=0

	control	bear	music
control	1.00000	0.91585 <.0001	0.67000 0.0171
bear	0.91585 <.0001	1.00000	0.74800 0.0051
music	0.67000 0.0171	0.74800 0.0051	1.00000
both	0.72281 0.0079	0.52607 0.0789	0.26164 0.4114

We notice that the independence assumption is not valid.



Given the responses for subject i , we define the *covariance matrix* as the following symmetric matrix of variances and covariances:

$$\text{Cov} \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{ik} \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1k} \\ \sigma_{12} & \sigma_{22} & & \sigma_{2k} \\ \vdots & & \ddots & \vdots \\ \sigma_{1k} & \sigma_{2k} & \dots & \sigma_{kk} \end{bmatrix}.$$

Note that this matrix contains $\frac{k(k+1)}{2}$ unique elements. We sometimes refer to this as an *unstructured* covariance matrix. Sometimes it is advantageous to model the covariance structure more parsimoniously (especially when there are many repeated measurements). A popular covariance matrix that allows correlation to deteriorate over time is an **autoregressive covariance matrix**. For example, the first-order autoregressive or AR(1) covariance matrix is given by





$$\text{Cov} \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ y_{ik} \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{k-1} \\ \rho & 1 & \rho & & \rho^{k-2} \\ \vdots & & & \ddots & \vdots \\ \rho^{k-1} & \rho^{k-2} & \dots & \rho & 1 \end{bmatrix}$$

and contains 2 parameters.

Perhaps the most popular choice for mixed-model analysis of variance assumes that the correlation between repeated measurements is due to each subject's underlying level of response. This *subject effect* is treated as a random variable in mixed-model ANOVA. For example, if the expected response to treatment is given by $E(Y_{ij}) = \mathbf{X}_{ij}\boldsymbol{\beta}$, then the response for subject i differs from the population mean by a subject effect, b_i , and a within-subject error, w_{ij} , leading to the model

$$Y_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + b_i + w_{ij},$$

where $b_i \sim N(0, \sigma_b^2)$ and $w_{ij} \sim N(0, \sigma_w^2)$. The covariance matrix for this model has the *compound symmetry* form:

$$\text{Cov} \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{ik} \end{bmatrix} = \begin{bmatrix} \sigma_b^2 + \sigma_w^2 & \sigma_b^2 & \dots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma_w^2 & & \sigma_b^2 \\ \vdots & & \ddots & \\ \sigma_b^2 & \sigma_b^2 & \dots & \sigma_b^2 + \sigma_w^2 \end{bmatrix}$$

and contains only 2 parameters. Although this is not valid under as wide a set of circumstances as the unrestricted covariance, at times the unrestricted covariance can contain too many parameters to estimate from the data.

It can be shown that the correlation between two subjects in the same group under the CS structure is given by $\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}$. This is known as an exchangeable correlation. In this case, the α 's are fixed treatment

effects, and the b 's represent each subject's own response tendency (some people may be more vocal than others, and thus their responses may tend to be high regardless of "treatment").

One option for our analysis is to use SAS PROC MIXED to extend PROC GLM and allow for our clusters of correlated observations.

Here, we define the *cluster* to be the study subject. Responses across clusters, or of different subjects, are assumed to be independent.

Responses within a cluster, or on an individual subject, are assumed to be correlated.

The following SAS code was used to input the data and fit the mixed effects model to control for correlation among members in a cluster.

```
data new2(keep=subject trt scream);
set new;
trt=control; scream=control; output;
trt=bear; scream=bear; output;
trt=music; scream=music; output;
trt=both; scream=both; output;
run;
```

```
proc mixed data=new2;
  class subject trt;
  model scream=trt/s;
  repeated/type=cs subject=subject r;
  contrast bear-both trt 1 -1 0 0;
  contrast bear-control trt 1 0 -1 0;
  contrast bear-music trt 1 0 0 -1;
  contrast both-control trt 0 1 -1 0;
  contrast both-music trt 0 1 0 -1;
  contrast control-music trt 0 0 1 -1;
run;
```

The output is provided below.

The Mixed Procedure

Model Information

Data Set	WORK.NEW2
Dependent Variable	scream

Covariance Structure	Compound Symmetry
Subject Effect	subject
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Class Level Information

Class	Levels	Values
subject	12	1 5 6 7 8 10 11 13 15 16 19 21
trt	4	bear both control music

Dimensions

Covariance Parameters	2
Columns in X	5
Columns in Z	0
Subjects	12

Max Obs Per Subject	4
Observations Used	48
Observations Not Used	0
Total Observations	48

Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	385.19441916	
1	1	366.18072711	0.00000000

Convergence criteria met.

Estimated R Matrix for subject 1

Row	Col1	Col2	Col3	Col4
1	296.08	178.08	178.08	178.08

2	178.08	296.08	178.08	178.08
3	178.08	178.08	296.08	178.08
4	178.08	178.08	178.08	296.08

The Mixed Procedure

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
CS	subject	178.08
Residual		118.01

Fit Statistics

-2 Res Log Likelihood	366.2
AIC (smaller is better)	370.2
AICC (smaller is better)	370.5
BIC (smaller is better)	371.2

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	19.01	<.0001

Solution for Fixed Effects

Effect	trt	Standard					
		Estimate	Error	DF	t Value	Pr > t	
Intercept		32.2500	4.9673	11	6.49	<.0001	
trt	bear	1.5000	4.4349	33	0.34	0.7373	
trt	both	-12.5000	4.4349	33	-2.82	0.0081	
trt	control	4.8333	4.4349	33	1.09	0.2837	
trt	music	0	

Type 3 Tests of Fixed Effects

Num	Den
-----	-----

Effect	DF	DF	F Value	Pr > F
trt	3	33	5.84	0.0026
Contrasts				
Label	Num DF	Den DF	F Value	Pr > F
bear-both	1	33	9.97	0.0034
bear-control	1	33	0.56	0.4576
bear-music	1	33	0.11	0.7373
both-control	1	33	15.28	0.0004
both-music	1	33	7.94	0.0081
control-music	1	33	1.19	0.2837

We see that the treatments are indeed different. Specifically, the teddy bear and music together are significantly better than all the other treatments.

SAS code and output for PROC MIXED with the UN structure are given below.

```
proc mixed data=new2 dfbw noclprint;
  class subject trt;
  model sbp=trt/s;
  repeated/type=un subject=subject r;
  contrast bear-both trt 1 -1 0 0;
  contrast bear-control trt 1 0 -1 0;
  contrast bear-music trt 1 0 0 -1;
  contrast both-control trt 0 1 -1 0;
  contrast both-music trt 0 1 0 -1;
  contrast control-music trt 0 0 1 -1;
  run;
*****
```

The Mixed Procedure

Model Information

Data Set	WORK.NEW2
Dependent Variable	scream
Covariance Structure	Unstructured

Subject Effect	subject
Estimation Method	REML
Residual Variance Method	None
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Dimensions

Covariance Parameters	10
Columns in X	5
Columns in Z	0
Subjects	12
Max Obs Per Subject	4

Number of Observations

Number of Observations Read	48
Number of Observations Used	48
Number of Observations Not Used	0

Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	385.19441916	
1	1	340.32444257	0.00000000

Convergence criteria met.

Estimated R Matrix for subject 1

Row	Col1	Col2	Col3	Col4
1	275.36	223.93	251.98	160.11
2	223.93	217.11	249.80	103.48
3	251.98	249.80	513.66	79.1591
4	160.11	103.48	79.1591	178.20

The Mixed Procedure

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
UN(1,1)	subject	275.36
UN(2,1)	subject	223.93

UN(2,2)	subject	217.11
UN(3,1)	subject	251.98
UN(3,2)	subject	249.80
UN(3,3)	subject	513.66
UN(4,1)	subject	160.11
UN(4,2)	subject	103.48
UN(4,3)	subject	79.1591
UN(4,4)	subject	178.20

Fit Statistics

-2 Res Log Likelihood	340.3
AIC (smaller is better)	360.3
AICC (smaller is better)	367.0
BIC (smaller is better)	365.2

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
9	44.87	<.0001

Solution for Fixed Effects

Effect	trt	Standard					
		Estimate	Error	DF	t Value	Pr > t	
Intercept		32.25	6.54	11	4.93	0.0005	
trt	bear	1.50	4.38	11	0.34	0.7390	
trt	both	-12.50	6.66	11	-1.87	0.0876	
trt	control	4.83	4.87	11	0.99	0.3427	
trt	music	0	

Type 3 Tests of Fixed Effects

Effect	Num	Den	Pr > F	
	DF	DF		
trt	3	11	11.70	0.0010

The Mixed Procedure

Contrasts

Label	Num	Den	F Value	Pr > F
	DF	DF		
bear-both	1	11	12.49	0.0047
bear-control	1	11	2.99	0.1118
bear-music	1	11	0.12	0.7390
both-control	1	11	27.04	0.0003
both-music	1	11	3.51	0.0876
control-music	1	11	0.98	0.3427

We can check whether the compound symmetry assumption is valid for the screamer data by comparing the residual log likelihoods from the compound symmetry and unrestricted covariance models.

To test the adequacy of the compound symmetry assumption, we consider a chi-squared test of the difference between the two residual log likelihoods. Recall that the compound symmetric covariance estimates 2 parameters, while the unstructured covariance estimates $\frac{4(5)}{2} = 10$ parameters.

	UN	CS
-2 Res Log L	340.3	366.2
df	1	9

Thus we compare $-2 \log \text{likelihood ratio} = (366.2 - 340.3) = 25.9$ to the chi-squared distribution with 8 df (which has critical value 15.5) and therefore reject the null hypothesis that the smaller model (compound

symmetry) is valid. We conclude that the compound symmetry assumption is not valid for the data.

Parallel Groups Repeated Measures Design

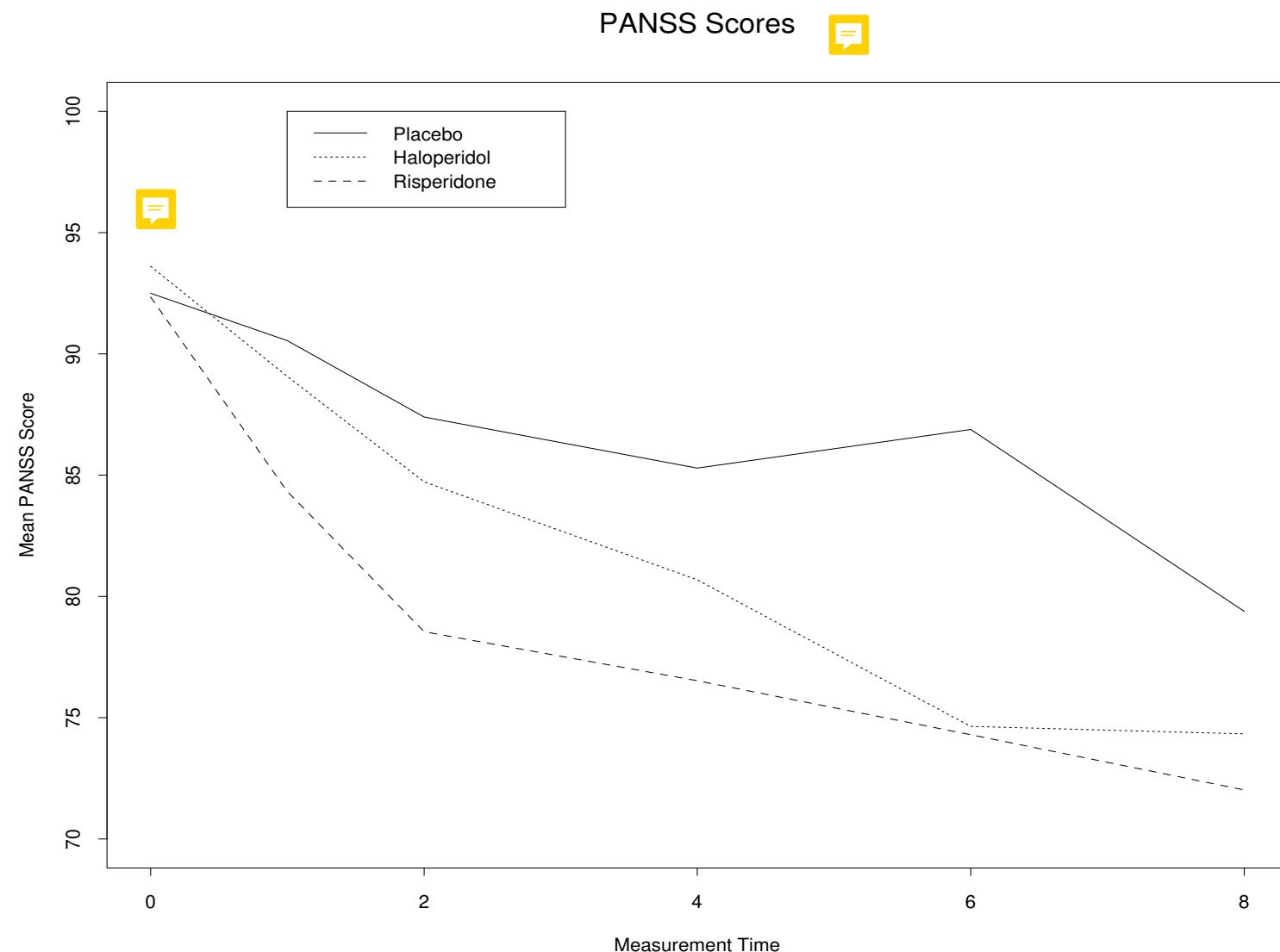
In the parallel groups design, two or more groups of subjects are measured repeatedly over time. For example, groups may be defined by *randomization* to one of several treatments.

We consider designs in which all subjects are meant to be measured at the same set of follow-up times, but these methods will allow for certain types of missing data.

The analysis goal is to characterize the patterns of change over time in the groups and to determine whether the patterns differ in the groups.

Example: Schizophrenia Clinical Trial

A multi-center study was conducted to assess the effect of two treatments and a placebo on decreasing the PANSS (positive and negative symptom scale) score. Patients were randomized to receive only one of the two treatments (haloperidol and risperidone) or a placebo. Measurements of the PANSS were taken at baseline and after 1, 2, 4, 6, and 8 weeks of treatment. Consider the following plot of PANSS score means at each measurement time. What do you think about the patterns of change over time?



Notice that since this is a randomized study, the group means should be the same at baseline, so any test of treatment effect is essentially a test of treatment by time interaction (we cannot have parallel lines if there really is a treatment effect, since parallel lines would imply coincident lines in this case).

In the analysis of parallel groups data, we may have one or more of the following hypotheses.

- H_0 : Are the profiles of the means similar in the three groups? (That is, are the line segments between adjacent measurement occasions parallel?) This hypothesis is the hypothesis of no group by time interaction.
- H_0 : If the profiles are parallel, are they also at the same level? This is the hypothesis of no group effect. (In a randomized study, the profiles should all start at the same level.)
- H_0 : If the profiles are parallel, are the means constant over time? This is the hypothesis of no time effect.

The appropriate hypotheses for any given dataset must be derived from the relevant scientific issues in that study. For example, we are interested only in the

first and last hypotheses for this randomized study. (Why?)

Notation for Parallel Groups Repeated Measures

Changing notation slightly, let n be the number of subjects (clusters) and N be the total number of observations. We consider the univariate representation (one row for each observation of the outcome) of the multivariate data.

If there are k measurement occasions, we define $k - 1$ indicator variables. For the i^{th} observation, we let $x_{ij} = 1$ if the observation was taken at time j and 0 otherwise, $j = 1, \dots, k - 1$. In addition, let x_{ik} be the indicator variable for the haloperidol group and $x_{i,k+1}$ be the indicator variable for the placebo group. (Risperidone is the reference group for treatment.)

We also define $2(k - 1)$ interaction terms as products of the time and group indicator variables. With these terms in the model, our model has $1 + (k - 1) + (2) + 2(k - 1) = 3k$ parameters for the mean model. If we use an unrestricted covariance matrix, we have $\frac{6(7)}{2} = 21$ additional parameters for the covariance.



If we let i index all the responses, $i = 1, \dots, N$, then the model is given by

$$\begin{aligned}y_i = & \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} \\& + \beta_6 x_{i6} + \beta_7 x_{i7} + \beta_8 x_{i8} + \beta_9 x_{i9} + \beta_{10} x_{i10} \\& + \beta_{11} x_{i11} + \beta_{12} x_{i12} + \beta_{13} x_{i13} + \beta_{14} x_{i14} \\& + \beta_{15} x_{i15} + \beta_{16} x_{i16} + \beta_{17} x_{i17} + \varepsilon_i,\end{aligned}$$

where



- x_{i1}, \dots, x_{i5} represent measurement times 0, 1, 2, 4, and 6 (with time 8 as the reference),

- x_{i6} and x_{i7} are the indicator variables for haloperidol and placebo, respectively (risperidone is the reference),

- x_{i8}, \dots, x_{i12} are the interactions between haloperidol and time, and
- x_{i13}, \dots, x_{i17} are the interactions between placebo and time.

Assumptions of the Parallel Groups Model

1. The subjects are random samples from each of the groups.
2. Observations from different individuals are independent, while repeated measurements on the same individual are not assumed to be independent.
3. The vector of outcomes for a given subject has a multivariate normal distribution.
4. The expected values of the individual observations are given by the linear regression model.
5. If observations are missing, they are missing at random (MAR).

Example: Schizophrenia Data

The following SAS code may be used to fit this model to the schizophrenia data (assuming the data are already in univariate format). Included are statements necessary to estimate differences in cell means for various time and treatment combinations. Note the ordering of the contrast coefficients. The first six coefficients represent haloperidol (TRT=1) at the six times (in numerical order). The next six coefficients represent placebo (TRT=2) at the six times, and the last six coefficients represent risperidone. The “e” option instructs SAS to print information about the contrast coefficients so that you can check to be sure you coded them properly (always a good idea!).

```
proc mixed data=new2 noclprint;
class id trtnew time t;
model y=trtnew time time*trtnew/s;
repeated t/type=un subject=id r;
estimate time 1 vs. baseline haloperidol1 - diff for placebo
      time*trtnew -1 1 0 0 0 0 1 -1 0 0 0 0 0 0 0 0 0/e;
estimate time 1 vs. baseline placebo - diff for risperidone
      time*trtnew 0 0 0 0 0 0 1 -1 0 0 0 0 -1 1 0 0 0 0/e;
estimate time 2 vs. baseline haloperidol1 - diff for placebo
      time*trtnew -1 0 1 0 0 0 1 0 -1 0 0 0 0 0 0 0 0/e;
```

```
estimate time 2 vs. baseline placebo - diff for risperidone
    time*trtnew 0 0 0 0 0 0 1 0 -1 0 0 0 -1 0 1 0 0 0/e;
estimate time 4 vs. baseline haloperidol1 - diff for placebo
    time*trtnew -1 0 0 1 0 0 1 0 0 -1 0 0 0 0 0 0 0/e;
estimate time 4 vs. baseline placebo - diff for risperidone
    time*trtnew 0 0 0 0 0 0 1 0 0 -1 0 0 -1 0 0 1 0 0/e;
estimate time 6 vs. baseline haloperidol1 - diff for placebo
    time*trtnew -1 0 0 0 1 0 1 0 0 0 -1 0 0 0 0 0 0 0/e;
estimate time 6 vs. baseline placebo - diff for risperidone
    time*trtnew 0 0 0 0 0 0 1 0 0 0 -1 0 -1 0 0 0 1 0/e;
estimate time 8 vs. baseline haloperidol1 - diff for placebo
    time*trtnew -1 0 0 0 0 1 1 0 0 0 0 -1 0 0 0 0 0 0/e;
estimate time 8 vs. baseline placebo - diff for risperidone
    time*trtnew 0 0 0 0 0 0 1 0 0 0 0 -1 -1 0 0 0 0 1/ e;
run;
```

Selected SAS output is provided below.

The Mixed Procedure

Model Information

Data Set

WORK.NEW2

Dependent Variable	y
Covariance Structure	Unstructured
Subject Effect	id
Estimation Method	REML
Residual Variance Method	None
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Dimensions

Covariance Parameters	21
Columns in X	28
Columns in Z	0
Subjects	523
Max Obs Per Subject	6
Observations Used	2468
Observations Not Used	670
Total Observations	3138

Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	21961.16171710	
1	2	20151.85618890	0.00610058
2	1	20096.23405069	0.00167392
3	1	20080.65977798	0.00027233
4	1	20078.27196287	0.00001373
5	1	20078.16116305	0.00000004
6	1	20078.16081868	0.00000000

Convergence criteria met.

Estimated R Matrix for id 1

Row	Col1	Col2
1	356.09	238.69
2	238.69	507.39

The Mixed Procedure

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
UN(1,1)	id	356.09
UN(2,1)	id	250.92
UN(2,2)	id	429.79
UN(3,1)	id	238.69
UN(3,2)	id	382.01
UN(3,3)	id	507.39
UN(4,1)	id	229.26
UN(4,2)	id	361.56
UN(4,3)	id	451.91
UN(4,4)	id	571.64
UN(5,1)	id	216.73
UN(5,2)	id	340.27
UN(5,3)	id	432.90
UN(5,4)	id	515.10
UN(5,5)	id	630.37
UN(6,1)	id	193.48
UN(6,2)	id	317.84

UN(6,3)	id	396.08
UN(6,4)	id	484.38
UN(6,5)	id	562.16
UN(6,6)	id	625.87

Fit Statistics

-2 Res Log Likelihood	20078.2
AIC (smaller is better)	20120.2
AICC (smaller is better)	20120.5
BIC (smaller is better)	20209.6

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
20	1883.00	<.0001

Solution for Fixed Effects

Effect	trtnew	time	Standard				
			Estimate	Error	DF	t Value	Pr > t
Intercept			78.6367	1.5171	519	51.83	<.0001
trtnew	1		6.9346	3.5402	519	1.96	0.0507
trtnew	2		17.0886	3.7912	519	4.51	<.0001
trtnew	3		0

The Mixed Procedure

Solution for Fixed Effects

Effect	trtnew	time	Standard				
			Estimate	Error	DF	t Value	Pr > t
time		0	13.8419	1.4886	519	9.30	<.0001
time		1	5.6702	1.3052	519	4.34	<.0001
time		2	1.8247	1.2063	519	1.51	0.1310
time		4	0.9563	1.0284	519	0.93	0.3529
time		6	0.1495	0.8046	519	0.19	0.8527
time		8	0
trtnew*time	1	0	-5.8040	3.4774	519	-1.67	0.0957

trtnew*time	1	1	-2.1313	3.0885	519	-0.69	0.4905
trtnew*time	1	2	-0.09359	2.8842	519	-0.03	0.9741
trtnew*time	1	4	-0.2524	2.4959	519	-0.10	0.9195
trtnew*time	1	6	-0.2067	1.9499	519	-0.11	0.9156
trtnew*time	1	8	0
trtnew*time	2	0	-17.0786	3.7332	519	-4.57	<.0001
trtnew*time	2	1	-10.6730	3.3797	519	-3.16	0.0017
trtnew*time	2	2	-6.6133	3.1757	519	-2.08	0.0378
trtnew*time	2	4	-3.9971	2.7871	519	-1.43	0.1521
trtnew*time	2	6	2.4792	2.2450	519	1.10	0.2700
trtnew*time	2	8	0
trtnew*time	3	0	0
trtnew*time	3	1	0
trtnew*time	3	2	0
trtnew*time	3	4	0
trtnew*time	3	6	0
trtnew*time	3	8	0

Type 3 Tests of Fixed Effects

Effect	Num	Den	F Value	Pr > F
	DF	DF		
trtnew	2	519	10.96	<.0001
time	5	519	8.20	<.0001
trtnew*time	10	519	3.68	<.0001

Coefficients for time 1 vs. baseline
haloperidol1 - diff for placebo

Effect	trtnew	time	Row1
Intercept			
trtnew	1		
trtnew	2		
trtnew	3		
time		0	

The Mixed Procedure

Coefficients for time 1 vs. baseline
haloperidol1 - diff for placebo

Effect	trtnew	time	Row1
time		1	
time		2	
time		4	
time		6	
time		8	
trtnew*time	1	0	-1
trtnew*time	1	1	1
trtnew*time	1	2	
trtnew*time	1	4	
trtnew*time	1	6	
trtnew*time	1	8	
trtnew*time	2	0	1
trtnew*time	2	1	-1
trtnew*time	2	2	
trtnew*time	2	4	
trtnew*time	2	6	
trtnew*time	2	8	
trtnew*time	3	0	

trtnew*time	3	1
trtnew*time	3	2
trtnew*time	3	4
trtnew*time	3	6
trtnew*time	3	8

Coefficients for time 1 vs. baseline
placebo - diff for risperidone

Effect	trtnew	time	Row1
Intercept			
trtnew	1		
trtnew	2		
trtnew	3		
time		0	
time		1	
time		2	
time		4	
time		6	
time		8	
trtnew*time	1	0	

trtnew*time	1	1
trtnew*time	1	2
trtnew*time	1	4
trtnew*time	1	6
trtnew*time	1	8

The Mixed Procedure

Coefficients for time 1 vs. baseline
placebo - diff for risperidone

Effect	trtnew	time	Row1
trtnew*time	2	0	1
trtnew*time	2	1	-1
trtnew*time	2	2	
trtnew*time	2	4	
trtnew*time	2	6	
trtnew*time	2	8	
trtnew*time	3	0	-1
trtnew*time	3	1	1
trtnew*time	3	2	
trtnew*time	3	4	

trtnew*time	3	6
trtnew*time	3	8

.....some output omitted.....

Estimates

Label	Estimate
time 1 vs. baseline haloperidol1 - diff for placebo	-2.7329
time 1 vs. baseline placebo - diff for risperidone	-6.4056
time 2 vs. baseline haloperidol1 - diff for placebo	-4.7549

Estimates

Label	Standard Error	DF
time 1 vs. baseline haloperidol1 - diff for placebo	2.5776	519
time 1 vs. baseline placebo - diff for risperidone	2.0356	519
time 2 vs. baseline haloperidol1 - diff for placebo	3.1184	519

Estimates

Label	t Value	Pr > t
time 1 vs. baseline haloperidol1 - diff for placebo	-1.06	0.2895

time 1 vs. baseline placebo - diff for risperidone	-3.15	0.0017
time 2 vs. baseline haloperidol1 - diff for placebo	-1.52	0.1279

The Mixed Procedure

Estimates

Label	Estimate
time 2 vs. baseline placebo - diff for risperidone	-10.4653
time 4 vs. baseline haloperidol1 - diff for placebo	-7.5298
time 4 vs. baseline placebo - diff for risperidone	-13.0814
time 6 vs. baseline haloperidol1 - diff for placebo	-13.9604
time 6 vs. baseline placebo - diff for risperidone	-19.5578
time 8 vs. baseline haloperidol1 - diff for placebo	-11.2746
time 8 vs. baseline placebo - diff for risperidone	-17.0786

Estimates

Label	Standard Error	DF
-------	----------------	----

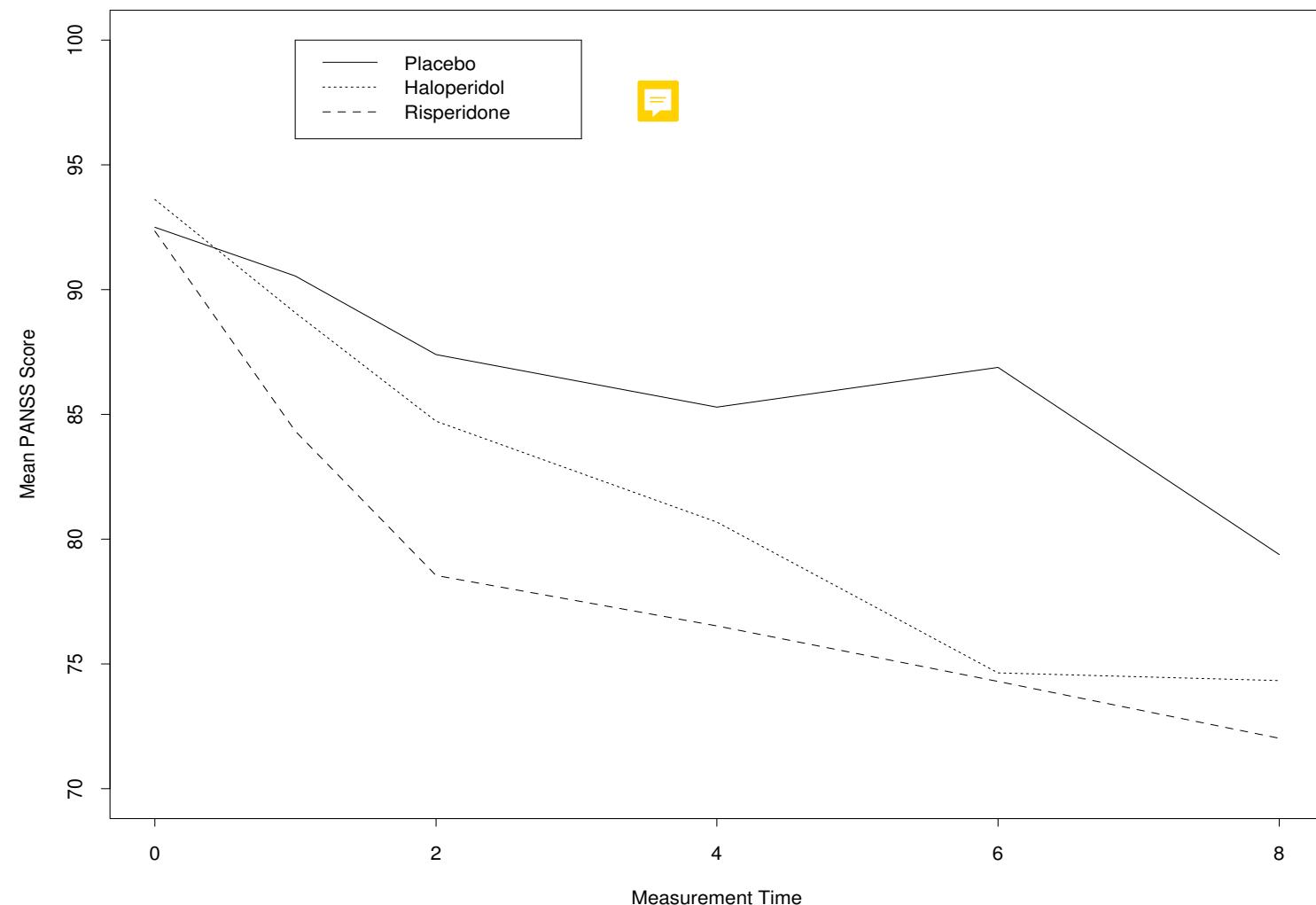
time 2 vs. baseline placebo - diff for risperidone	2.4715	519
time 4 vs. baseline haloperidol1 - diff for placebo	3.5989	519
time 4 vs. baseline placebo - diff for risperidone	2.8744	519
time 6 vs. baseline haloperidol1 - diff for placebo	4.2460	519
time 6 vs. baseline placebo - diff for risperidone	3.3765	519
time 8 vs. baseline haloperidol1 - diff for placebo	4.6473	519
time 8 vs. baseline placebo - diff for risperidone	3.7332	519

Estimates

Label	t Value	Pr > t
time 2 vs. baseline placebo - diff for risperidone	-4.23	<.0001
time 4 vs. baseline haloperidol1 - diff for placebo	-2.09	0.0369
time 4 vs. baseline placebo - diff for risperidone	-4.55	<.0001
time 6 vs. baseline haloperidol1 - diff for placebo	-3.29	0.0011
time 6 vs. baseline placebo - diff for risperidone	-5.79	<.0001
time 8 vs. baseline haloperidol1 - diff for placebo	-2.43	0.0156
time 8 vs. baseline placebo - diff for risperidone	-4.57	<.0001

What do we conclude based on the hypothesis tests given in the SAS output?

PANSS Scores



Using the following SAS code and partial output, test whether the compound symmetry covariance is defensible for these data.



```
proc mixed data=new2 noclprint;
class id trtnew time t;
model y=trtnew time time*trtnew/s;
repeated t/type=cs subject=id r;
run;
*****
```

The Mixed Procedure

Model Information

Data Set	WORK.NEW2
Dependent Variable	y
Covariance Structure	Compound Symmetry
Subject Effect	id
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Dimensions

Covariance Parameters	2
Columns in X	28
Columns in Z	0
Subjects	523
Max Obs Per Subject	6
Observations Used	2468
Observations Not Used	670
Total Observations	3138

Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	21961.16171710	
1	2	20548.58902724	0.00002967
2	1	20548.34199238	0.00000011
3	1	20548.34114692	0.00000000

Convergence criteria met.

Estimated R Matrix for id 1

Row	Col1	Col2
1	455.62	302.14
2	302.14	455.62

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
CS	id	302.14
Residual		153.48

The Mixed Procedure

Fit Statistics

-2 Res Log Likelihood	20548.3
AIC (smaller is better)	20552.3

AICC (smaller is better)	20552.3
BIC (smaller is better)	20560.9

Treating Time as a Continuous Variable

If the means tend to change linearly over time, we may wish to treat time as a continuous variable rather than as a categorical variable. If means do change linearly with time, then the treatment effect in a continuous-time model is captured in one single parameter, leading to more powerful tests.

If time is treated as continuous, the model is given by

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i,$$

where

-  X_{i1} represents measurement time as a continuous variable taking the values 0, 1, 2, 4, 6, or 8,
- X_{i2} and X_{i3} are the indicator variables for haloperidol and placebo, respectively,
- X_{i4} is the interaction between haloperidol and time, and

- X_{i5} is the interaction between placebo and time.

To fit this model, we remove TIME from the class statement and use the following SAS code.

```
proc mixed data=new2 noclprint;  
class id trtnew t;  
model y=trtnew time time*trtnew/s;  
repeated t/type=un subject=id r;  
run;
```

Note that we still need the variable "t" (which takes the same values as "time") declared as a class variable so that we can tell SAS how the repeated measures are ordered within a subject. This is why we define two variables, "t" and "time," with the same values (if we wish to treat time as continuous, we still need a version of that variable that can be treated as a class variable).

For subjects on haloperidol, we have

$$E(Y_i) = (\beta_0 + \beta_2) + (\beta_1 + \beta_4)TIME_i,$$

for subjects on placebo, we have

$$E(Y_i) = (\beta_0 + \beta_3) + (\beta_1 + \beta_5)TIME_i,$$

and for subjects on risperidone we have

$$E(Y_i) = \beta_0 + \beta_1 TIME_i.$$

Clearly, the model assumes each group's mean changes linearly over time.

Selecting a Method for Parameterizing Time

Because fitting time as a class variable is equivalent to fitting a polynomial of order 5 in time, we can view the continuous time parameterization as a model nested in the categorical time parameterization. Thus a likelihood ratio test is appropriate for comparing the two models.

Using the following SAS code and selected output, test whether the continuous time model is appropriate for the schizophrenia data.

```
proc mixed data=new2 noclprint method=ml;
class id trtnew time t;
model y=trtnew time time*trtnew/s;
repeated t/type=un subject=id r;
run;
```

```
proc mixed data=new2 noclprint method=ml;
class id trtnew t;
model y=trtnew time time*trtnew/s;
repeated t/type=un subject=id r;
run;
```

```
*****
```

The Mixed Procedure

Model Information

Data Set	WORK.NEW2
Dependent Variable	y
Covariance Structure	Unstructured
Subject Effect	id
Estimation Method	ML
Residual Variance Method	None
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Dimensions

Covariance Parameters	21
Columns in X	28
Columns in Z	0
Subjects	523
Max Obs Per Subject	6
Observations Used	2468
Observations Not Used	670
Total Observations	3138

Iteration History

Iteration	Evaluations	-2 Log Like	Criterion
0	1	22020.87344360	
1	3	20159.01391755	0.00575612
2	2	20128.62705395	0.00065058
3	1	20123.02179920	0.00003529
4	1	20122.73459148	0.00000018
5	1	20122.73319055	0.00000000

Convergence criteria met.

Estimated R Matrix for id 1

Row	Col1	Col2
1	354.04	237.31
2	237.31	504.31

The Mixed Procedure

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
UN(1,1)	id	354.04
UN(2,1)	id	249.48
UN(2,2)	id	427.29
UN(3,1)	id	237.31
UN(3,2)	id	379.79
UN(3,3)	id	504.31
UN(4,1)	id	227.95
UN(4,2)	id	359.46
UN(4,3)	id	449.18
UN(4,4)	id	567.96
UN(5,1)	id	215.48
UN(5,2)	id	338.29
UN(5,3)	id	430.28
UN(5,4)	id	511.81

UN(5,5)	id	625.86
UN(6,1)	id	192.37
UN(6,2)	id	315.99
UN(6,3)	id	393.69
UN(6,4)	id	481.27
UN(6,5)	id	558.20
UN(6,6)	id	621.01

Fit Statistics

-2 Log Likelihood	20122.7
AIC (smaller is better)	20200.7
AICC (smaller is better)	20202.0
BIC (smaller is better)	20366.9

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
20	1898.14	<.0001

Solution for Fixed Effects

Effect	trtnew	time	Standard				
			Estimate	Error	DF	t Value	Pr > t
Intercept			78.6367	1.5109	519	52.05	<.0001
trtnew	1		6.9346	3.5253	519	1.97	0.0497
trtnew	2		17.0886	3.7749	519	4.53	<.0001
trtnew	3		0

The Mixed Procedure

Solution for Fixed Effects

Effect	trtnew	time	Standard				
			Estimate	Error	DF	t Value	Pr > t
time		0	13.8420	1.4824	519	9.34	<.0001
time		1	5.6703	1.2993	519	4.36	<.0001
time		2	1.8247	1.2006	519	1.52	0.1292
time		4	0.9563	1.0231	519	0.93	0.3504
time		6	0.1495	0.8002	519	0.19	0.8518

time		8	0
trtnew*time	1	0	-5.8040	3.4627	519	-1.68	0.0943
trtnew*time	1	1	-2.1313	3.0743	519	-0.69	0.4885
trtnew*time	1	2	-0.09356	2.8704	519	-0.03	0.9740
trtnew*time	1	4	-0.2524	2.4830	519	-0.10	0.9191
trtnew*time	1	6	-0.2067	1.9392	519	-0.11	0.9152
trtnew*time	1	8	0
trtnew*time	2	0	-17.0786	3.7169	519	-4.59	<.0001
trtnew*time	2	1	-10.6729	3.3639	519	-3.17	0.0016
trtnew*time	2	2	-6.6133	3.1602	519	-2.09	0.0369
trtnew*time	2	4	-3.9971	2.7726	519	-1.44	0.1500
trtnew*time	2	6	2.4792	2.2327	519	1.11	0.2673
trtnew*time	2	8	0
trtnew*time	3	0	0
trtnew*time	3	1	0
trtnew*time	3	2	0
trtnew*time	3	4	0
trtnew*time	3	6	0
trtnew*time	3	8	0

Type 3 Tests of Fixed Effects

Effect	Num	Den	F Value	Pr > F
	DF	DF		
trtnew	2	519	11.03	<.0001
time	5	519	8.25	<.0001
trtnew*time	10	519	3.71	<.0001

The Mixed Procedure

Model Information

Data Set	WORK.NEW2
Dependent Variable	y
Covariance Structure	Unstructured
Subject Effect	id
Estimation Method	ML
Residual Variance Method	None
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Dimensions

Covariance Parameters	21
Columns in X	8
Columns in Z	0
Subjects	523
Max Obs Per Subject	6
Observations Used	2468
Observations Not Used	670
Total Observations	3138

Iteration History

Iteration	Evaluations	-2 Log Like	Criterion
0	1	22063.13468076	
1	3	20255.08219146	0.00579452
2	2	20224.45097439	0.00070989
3	1	20218.22109780	0.00004441
4	1	20217.85586897	0.00000028
5	1	20217.85369080	0.00000000

Convergence criteria met.



Estimated R Matrix for id 1

Row	Col1	Col2
1	357.07	227.20
2	227.20	553.82

The Mixed Procedure

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
UN(1,1)	id	357.07
UN(2,1)	id	240.70
UN(2,2)	id	442.31
UN(3,1)	id	227.20
UN(3,2)	id	410.52

UN(3,3)	id	553.82
UN(4,1)	id	218.99
UN(4,2)	id	380.37
UN(4,3)	id	485.04
UN(4,4)	id	592.86
UN(5,1)	id	206.63
UN(5,2)	id	342.41
UN(5,3)	id	443.17
UN(5,4)	id	514.05
UN(5,5)	id	613.21
UN(6,1)	id	184.15
UN(6,2)	id	307.33
UN(6,3)	id	387.60
UN(6,4)	id	465.04
UN(6,5)	id	530.13
UN(6,6)	id	584.45

Fit Statistics

-2 Log Likelihood	20217.9
AIC (smaller is better)	20271.9
AICC (smaller is better)	20272.5

BIC (smaller is better) 20386.9

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
20	1845.28	<.0001

Solution for Fixed Effects

Effect	trtnew	Standard					
		Estimate	Error	DF	t Value	Pr > t	
Intercept		90.0941	0.9663	519	93.24	<.0001	
trtnew	1	2.3082	2.1578	519	1.07	0.2852	
trtnew	2	1.5762	2.1491	519	0.73	0.4636	
trtnew	3	0	
time		-1.2618	0.1713	519	-7.37	<.0001	
time*trtnew	1	0.4097	0.4057	519	1.01	0.3130	
time*trtnew	2	1.7698	0.4404	519	4.02	<.0001	
time*trtnew	3	0	

The Mixed Procedure

Type 3 Tests of Fixed Effects

Effect	Num	Den	F Value	Pr > F
	DF	DF		
trtnew	2	519	0.71	0.4911
time	1	519	7.83	0.0053
time*trtnew	2	519	8.14	0.0003

Summary

When analyzing data from the parallel groups repeated measures design, consider the following strategy.

1. Choose a working covariance structure. (Note that the choice of the mean model and the covariance structure are interdependent.) Use the REML (default) log-likelihood as the criterion for selecting a covariance structure. As a general rule of thumb, the unstructured model should be used unless a simpler covariance structure is clearly satisfactory. **(Restricted (residual) maximum likelihood (REML))**: REML is an alternative to full maximum likelihood estimation and is typically the default method in most statistical packages. Rather than maximizing the likelihood of the data, it maximizes the likelihood of the observed residuals. REML obtains initial estimates of the fixed effects using ordinary least squares and then using these estimates it maximizes the likelihood of the residuals (in which the fixed effects are subtracted off) to obtain estimates of the variance parameters. Finally the estimated variance parameters are used to obtain generalized least squares estimates of the fixed effect parameters. **REML is a good alternative to**



-
- ML when the sole focus is on estimation of the variance components. The variance components obtained via ML are biased when the samples are small while REML estimates are unbiased. The problem with REML for model building is that the "likelihoods" obtained for models with different fixed effects are not comparable. Hence it is not valid to compare models with different fixed effects using a likelihood ratio test or AIC when REML is used to estimate the model. For this reason we have not used REML in this course. You must be aware of the distinction between ML and REML because REML is the default for most software packages. Thus you will generally need to explicitly specify maximum likelihood estimation if that's what you desire.)
2. Decide in advance whether to model the effect of treatment on patterns of change by
 - (a) time by treatment interaction with time coded as a categorical variable,
 - (b) time by treatment interaction with time coded as a continuous variable,
or

-
- (c) a treatment main effect ANCOVA model with baseline treated as a covariate and time treated as a categorical variable

The ML log likelihood (PROC MIXED METHOD=ML) may be used to compare any nested models differing by more than one degree of freedom.

- 3. Determine the final form of the regression (mean) model
- 4. Fit the final model using REML
- 5. Only estimation and test of covariate effects are considered here. It is also of great interest in testing and estimating covariance matrix.

Lecture 20: Power & Sample Size Calculation

Reading Assignment:

- Muller and Fetterman, Chapter 17: “Understanding and Computing Power for the GLM” (for more background)

Motivation

One of the most common questions asked of a statistician about study design is the number of patients to include.

It is an important question, because if a study is too small it will not be able to answer the question posed, and would be a waste of time and money. It could also be deemed unethical because patients may be put at risk with no apparent benefit.

However, it is also undesirable for studies to be too large because resources would be wasted if fewer patients would have sufficed, and it is unethical to expose more patients to the less effective treatment.

Next we will discuss the power calculation for continuous response variables followed by categorical response variables.

Continuous Response Variable

Model Statement

$$\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times 1} + \mathbf{e}_{n \times 1}$$

with $n >> p$. Assume $\mathbf{e} \sim N(0, \sigma^2 \mathbf{I})$ and that $\boldsymbol{\beta}$ contains fixed and unknown constants. Also assume that \mathbf{X} contains fixed values, known without appreciable error, conditional upon having collected the sample. Any power computed with the methods described in this chapter applies only to the particular choice of \mathbf{X} used in the calculations. Changing \mathbf{X} changes the design.

Estimation

Assume \mathbf{X} is of full rank. Let

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}.$$

Then the predicted values can be computed as

$$\hat{\mathbf{y}} = \mathbf{X} \hat{\boldsymbol{\beta}},$$

the sum of squares error equals

$$SSE = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{y}'[I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y},$$

and the variance estimate,

$$\hat{\sigma}^2 = \frac{SSE}{n - p}.$$

Let β be the primary parameters and $\theta = \mathbf{C}\beta$ be the secondary parameters. Assume \mathbf{C} an $a \times p$ matrix ($a \leq p$) with full (row) rank of a . Therefore

$$\mathbf{M} = \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'$$

is full (row) rank of a . Thus θ is both estimable and testable.

The General Linear Hypothesis

$$H_0 : \theta = \theta_0$$

$$\text{versus } H_A : \theta \neq \theta_0.$$

Estimability of θ and full rank of \mathbf{M} ensures uniqueness of the sum of squares for the hypothesis,

$$SSH = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathbf{M}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

and also ensures uniqueness and existence of the likelihood ratio test.

Test Statistic, Null Case

Under the assumptions described above, the maximum likelihood approach provides a test with type I error rate exactly equal to the

target, α . The likelihood ratio test statistic has many equivalent forms:

$$\begin{aligned} F_{obs} &= \frac{SSH/a}{SSE/(n-p)} \\ &= \frac{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathbf{M}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)/a}{\hat{\sigma}^2} \\ &= \frac{\hat{\rho}_*^2/a}{(1 - \hat{\rho}_*^2)/(n-p)}, \end{aligned}$$

where $\hat{\rho}_*^2$ represents a generalized squared correlation. When the model spans an intercept and the hypothesis does not span the intercept, then the generalized correlation reduces to the usual multiple correlation ("corrected" for the intercept). Under H_0 , $F \sim F(a, n-p)$.



Test Statistic, Non-Null Case

Define *power* as the probability of rejecting H_0 , whether or not H_0 is

true or not. This definition differs slightly from the traditional definition. But this new definition greatly simplifies the discussion of power. With the current definition, if H_0 holds, then the power of the GLH equals the type I error rate. A test is *unbiased* if the expected rejection rate is no more than α for null cases and no less than α for alternative cases. Among all tests unbiased and invariant to location and scale, the likelihood ratio test represents the uniformly most powerful test.

Define

$$f_A = \frac{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \mathbf{M}^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) / a}{\sigma^2}$$



$$= \frac{\rho_*^2 / a}{(1 - \rho_*^2) / n}$$

$$\approx \frac{\rho_*^2 / a}{(1 - \rho_*^2) / (n - p)}.$$

The last equation has no "hats" compared to the last equation in F_{obs} , which uses estimates of parameters. Note that f_A is a parameter, a constant, while F_{obs} is a random variable.

Under H_A ,

$$F_{obs} \sim F(a, n - p, \omega),$$

with

$$\begin{aligned}\omega = a \times f_A &= \frac{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \mathbf{M}^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)}{\sigma^2} \\ &= \frac{\rho_*^2}{(1 - \rho_*^2)/n} \quad (*)\end{aligned}$$

Refer ω as the noncentral parameter because it captures the amount by which the model deviates from the central case.

Properties of ω in the GLM

First note that $0 \leq \omega < \infty$. Having $\omega = 0$ implies H_0 holds and power equals α . Also changing the scale of the data (such as from meters to centimeters) does not change ω .

For the independent groups T test of equality of means, assuming



equal cell sizes,



$$\omega = \frac{n}{4} \frac{(\mu_1 - \mu_2)^2}{\sigma^2}.$$

For a multiple regression model that includes an intercept, ω associated with the overall test of (corrected) regression, the usual test of all slopes equal to zero is

$$\begin{aligned}\omega &= \frac{\sigma_Y^2 \rho^2}{\sigma_Y^2 (1 - \rho^2)/n} &= \frac{\rho_*^2}{(1 - \rho_*^2)/n} \\ &= \frac{n \rho_*^2}{(1 - \rho_*^2)}\end{aligned}$$

Here $\sigma^2 = \sigma_Y^2 (1 - \rho^2)$ represents the usual residual variance for the model; while σ_Y^2 represents the variance of \mathbf{Y} . In turn, ρ^2 , the usual squared multiple correlation coefficient, provides a scale free measure of effect and $(1 - \rho^2)$ provides a scale free measure of residual variance.

In both special cases, ω increases with sample size, decreases with error variance, and increases with amount of effect. Equation (*) allows generalizing this statement to all cases of GLM. In the general case, $(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ captures the size of the effect, while sample size n , hides in $\mathbf{M} = \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'$ through \mathbf{X} . Consider a balanced ANOVA design with cell mean coding, what is $(\mathbf{X}'\mathbf{X})^{-1}$? In summary, ω varies only as function of sample size, mean difference, and error variance.

Computing Power

1. Specify $\alpha, \sigma^2, \mathbf{X}, \boldsymbol{\beta}, \mathbf{C}$ and $\boldsymbol{\theta}_0$.
2. Find the critical value, $f_{crit} = F_F^{-1}(1 - \alpha; a, n - p);$.
3. Compute the noncentral parameter, ω , as defined in equation (*).
4. Compute $Power = 1 - F_F(f_{crit}; a, n - p, \omega).$

Factors in Choosing a Design

1. Test size α , which changes when using a Bonferroni correction for multiple analyses
2. The size of σ^2 , which must often be varied due to uncertainty about the variable, the population or the study design
3. Varying \mathbf{X} includes changing total sample size, cell size ratios, and the distribution of control variables. Typically balanced designs maximize power.
4. Varying β evaluates the impact of the strength and pattern of effects. Some choices of β are equivalent in terms of power: for example, for an overall test of equality of means in a three-group ANOVA balanced design, the following choices for β are equivalent: $[200, 210, 220]'$, $[0, 10, 20]'$, $[210, 200, 220]'$.
5. The choice of \mathbf{C} plays the primary role in specifying the hypothesis.

-
6. The choice of θ_0 completes specification of the hypothesis. In most cases, $\theta_0 = \mathbf{0}$.

Using Parameter Estimates in Power Analysis

Unfortunately, several quantities are required before we can do any calculations (and the argument is a little circular!) Speculation drives power analysis not data. Despite that, in some cases data from an earlier study fuel the speculation. Estimation of σ^2 , β or both may be used to compute ω . As a function of one or more parameter estimates, the noncentral value estimate becomes a random variable, as does the corresponding power. The process adds additional source of uncertainty.

Taylor and Muller (1995) proposed the following methods to construct confidence intervals for both noncentrality and power when using $\hat{\sigma}^2$:

let

$$\hat{\omega} = \frac{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \mathbf{M}^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) / a}{\hat{\sigma}^2}$$

with $\hat{\sigma}^2$ based on v degrees of freedom. Also let $c_{crit} = F_{\chi^2}^{-1}(\alpha_c; v)$, the α_c quantile of a χ^2 random variable. Compute the α_c quantile for ω as

$$\hat{\omega}_c = \hat{\omega} \frac{c_{crit}}{v}.$$

Using $\hat{\omega}_c$ to compute power yields an α_c quantile for power. They reported similar results based on estimated σ^2 and β .

Example1: Kidney Disease

Falk et al (1992) randomly assigned 24 participants to one of two treatments intended to slow the worsening of kidney disease. Higher levels of creatinine indicate worse function. Using the reciprocal of serum creatinine level as the dependent variable allowed the investigators to meet the Gaussian assumption. The scientists considered an increase of 0.50dL/mg a clinically important improvement.

If we state the model as

$$\begin{aligned}\mathbf{y}_{24} &= \mathbf{X}_{24 \times 2} \boldsymbol{\beta}_{2 \times 1} + \mathbf{e} \\ &= \begin{bmatrix} \mathbf{1}_{12} & 0_{12} \\ 0_{12} & \mathbf{1}_{12} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \mathbf{e}\end{aligned}$$

The hypothesis of interest uses $\mathbf{C} = [-1, 1]$ and $\theta_0 = 0$. Let $\delta = \mu_2 - \mu_1$. Then $\theta = \delta = 0.5$, which reflects the investigators' interest. Assuming $\sigma^2 = 0.068$ that estimated from the study, which led to a $\hat{\omega} = 22.06$ and a power of 0.96 under $\alpha = 0.01$. Taylor and Muller computed the 95% CI for ω as [11.01, 36.880] and for power as [0.688, 0.999]. The asymmetry of the χ^2 leads to asymmetric confidence intervals. Taylor and Muller recommended using one-sided confidence intervals for power, since we usually wish to make statements of inequality, such as ensuring power no less than P . The one sided interval is [0.75, 1].

Example2: Medical Cost Suppose we want to design a study to

determine whether there is a linear relationship between nursing home patients' ages and their annual costs for medication. It would be considered unimportant from an economic and medical standpoint if age explained less than .04 of the variability in medication cost. We want a significance level of $\alpha = 0.05$. Calculate the power when sample size $n=200$.

$$\text{Solution: } f_{crit} = F_F^{-1}(1 - \alpha; a, n - p) = F_F^{-1}(0.95; 1, 198) = 3.889$$

$$\omega = \frac{n\rho^2}{(1-\rho^2)} = 200 * 0.04 / (1 - 0.04) = 8.33$$

$$power = 1 - F_F(f_{crit}; 1, n-p, \omega) = 1 - F_F(3.889; 1, 198, 8.33) = 0.82.$$

Power Reporting

Single power values rarely suffice to inform the scientist. As statisticians, the authors use tables and plots of power values to return the choices of sample size to the principal investigator. A typical table involves varying two or three of the factors controlling power: mean difference, variance and sample size.

 Power for comparing two means

Sample Size per group	Detectable Effect Size	
	80% power	90% power
10	1.3	1.5
20	0.9	1.1
50	0.6	0.7
100	0.4	0.5
200	0.28	0.33

where effect size = $\frac{|\mu_1 - \mu_2|}{\sigma}$.

Dichotomous Responses

For a binary outcome we need to specify type I error, and proportions P_1 and P_2 where P_1 is the expected outcome under the control intervention and $P_1 - P_2$ is the minimum clinical difference which it is worthwhile detecting.

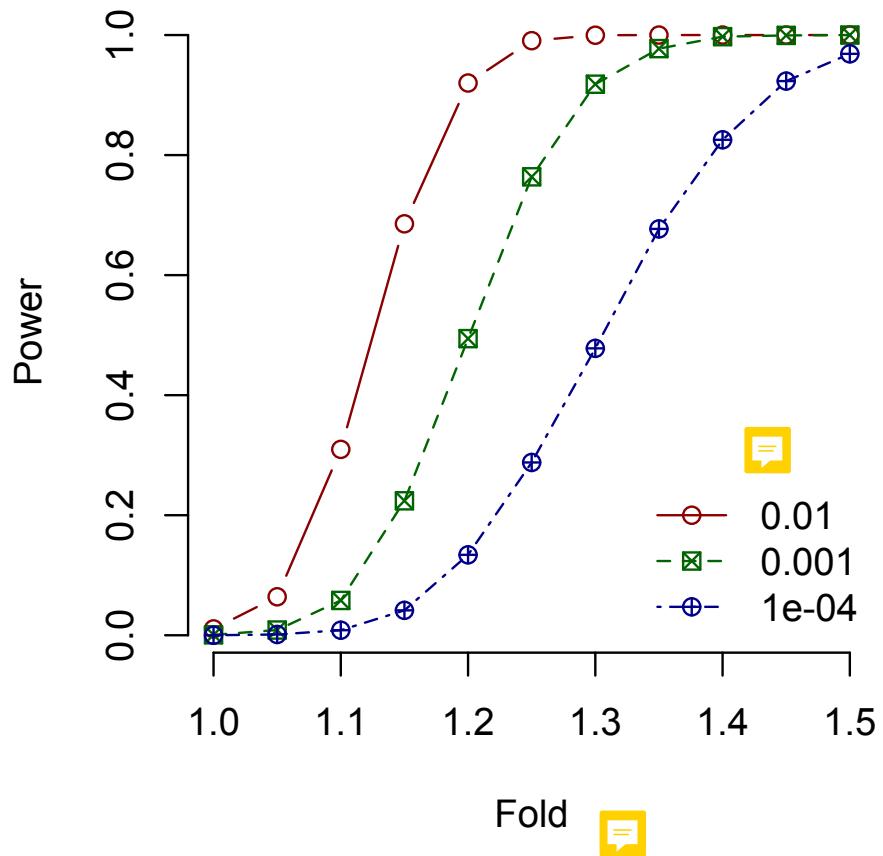
1. Specify α, P_1 and P_2 .
2. Find the critical value, $z_{crit} = N^{-1}(1 - \alpha/2)$;
3. Calculate $\omega = \frac{P_1 - P_2}{\sqrt{P_1(1-P_1)/n_1 + P_2(1-P_2)/n_2}}$
4. Compute $Power = 1 - N(z_{crit}; \omega, 1)$.

Power for comparing two proportions

P1	P2	n1	n2	Power
0.2	0.3	100	100	0.38
0.2	0.3	200	200	0.64
0.15	0.3	100	100	0.73
0.15	0.3	200	200	0.95

Illustrate Power by Figure

Power of differential expression study for three different p-value cutoffs.



Software:



- Commercial one: nquery
<https://uncapps.its.unc.edu/vpn/index.html>
- SAS and R have various functions for power analysis
- online power calculation: <http://powerandsamplesize.com/>

How Much to Do?

In practice the sample size is often fixed by other criteria, such as finance or resources, and the formula is used to determine a realistic effect size. If this is too large, then the study will have to be abandoned or increased in size.

Five key questions regarding sample size:

1. What is the main purpose of the study?
2. What is the principal measure of patient outcome?
3. How will the data be analyzed to detect a treatment difference?
4. What type of results does one anticipate with standard treatment?

5. How small a treatment difference is it important to detect and with what degree of certainty?

Thus in order to calculate the sample size for a study it is first necessary to decide upon what your outcome is. If your outcome variable is continuous you will need to have some measure of what you would expect its mean value to be in the control group together with an estimate of its standard deviation. You will also need to know what size of effect you expect or is desirable (be realistic with this). If your outcome variable is binary you will need to have an idea of the proportions falling into the two outcome categories, and what change in these proportions can be expected or is desirable.

After deciding on the purpose of the study and the principle outcome measure, the investigator must decide how the data are to be summarized and analyzed to detect a treatment difference. Thus, the

investigator must choose an appropriate summary measure of this outcome and then calculate a sample size based on the smallest treatment difference in this summary measure that is of such clinical value that it would be very undesirable to fail to detect. Given answers to all of the five questions above, we can then calculate a sample size.

Summary

Finally, the end of the semester! We made it!

Topic 1: Introduction and Overview



1. Why linear regression, why not t-test?
2. Basic concepts: Population, Sample, parameter, statistic
3. Statistical Activities: Parameter Estimation, Inference

Topic 2: Linear Algebra Review

1. Matrix operation, matrix addition, matrix multiplication ...
2. An *orthogonal matrix* is a **square matrix** with $\mathbf{A}' = \mathbf{A}^{-1}$.
3. Rules of Matrix Operation.
4. Linear Dependence and Rank, matrix determinant
5. Positive Definite and Semi-positive Definite Matrices 
6. Inverse and Generalized Inverse
7. Eigenvalues, Eigenvectors. Suppose \mathbf{A} is an symmetric matrix.
Then there exists an orthogonal (column orthonormal) matrix \mathbf{V} such that $\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}'$.

8. Random Vectors and Matrices

$$E(\mathbf{A}\mathbf{Y} + \mathbf{b}) = \mathbf{A}E(\mathbf{Y}) + \mathbf{b} = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$$

$$\text{Cov}(\mathbf{A}\mathbf{Y} + \mathbf{b}) = \mathbf{A}\text{Cov}(\mathbf{Y})\mathbf{A}' = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'.$$

9. Important Distributions for Linear Models. If $Z \sim N(0, 1)$,
 $X_1 \sim \chi^2(n_1)$ and $X_2 \sim \chi^2(n_2)$, and X_1 and X_2 are independent.
Construct random variables following t-distribution and F  distribution.
10. Maximum Likelihood Estimates (MLE)

Topics 3 and 4: Simple Linear Regression and the General Linear Model: Estimation and Testing

1. $\mathbf{y}_{n \times 1} = \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$
2. Least Squares Estimation: $\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$.
 $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $\operatorname{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.
3. HILE Gauss
 - Existence Assumption 
 - Linearity Assumption
 - Independence Assumption 
 - Homogeneity Assumption
 - Gaussian Errors Assumption 
4. $\boldsymbol{\beta}$ is the vector of primary parameters, and $\boldsymbol{\theta}_{a \times 1} = \mathbf{C}_{a \times p} \boldsymbol{\beta}_{p \times 1}$ is

a vector of secondary parameters, defined by \mathbf{C} , the *contrast matrix*. Each row of \mathbf{C} defines a new scalar parameter in terms of the β 's, e.g., $\beta_1 - \beta_2$. The general linear hypothesis is

$$H_0 : \boldsymbol{\theta}_{a \times 1} = \boldsymbol{\theta}_0$$

$$H_A : \boldsymbol{\theta}_{a \times 1} \neq \boldsymbol{\theta}_0.$$

5. Estimability and Testability of a Parameter. If \mathbf{X} is full rank , then $\hat{\boldsymbol{\beta}}$ exists (uniquely), $\boldsymbol{\beta}$ is estimable, and any (nonzero) \mathbf{C} gives estimable $\boldsymbol{\theta}$. If \mathbf{C} is full rank, $\boldsymbol{\beta}$ is testable.



6. Computation of Test Statistic and p-value. Let

$$\mathbf{M}_{a \times a} = \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}' \text{ and } SSH_{1 \times 1} = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathbf{M}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

The test-statistic is

$$F_{obs} = \frac{SSH/a}{SSE/(n-p)} = \frac{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \mathbf{M}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)/a}{\hat{\sigma}^2} = \frac{MSH}{MSE}$$

Topic 5: Some Distributional Results for the GLM

- If \mathbf{X} is full rank, $\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$
- $\boldsymbol{\theta} = \mathbf{C}_{a \times p}\boldsymbol{\beta}$, then $\widehat{\boldsymbol{\theta}} \sim N_a(\boldsymbol{\theta}, \sigma^2\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')$.
- Predicted Values: Conditional Means and Future Observations
 - $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = [\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y} = \mathbf{H}\mathbf{y}$,
 - $E(\widehat{\mathbf{y}}) = \mathbf{X}\boldsymbol{\beta}$,
 - $\text{cov}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.
- Definitions and Properties of Residuals
- Residual Variance $\widehat{\sigma}^2 = \frac{SSE}{n-p} = \frac{\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}}{n-p} = \frac{\mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}}{n-p}$

Topic 6: Multiple Regression: General Consideration

- Basic Sum Squares:

$$USS(\text{total}) = USS(\text{model}) + SSE, \quad \mathbf{y}'\mathbf{y} = \mathbf{y}'\mathbf{H}\mathbf{y} + \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}.$$

$$\begin{aligned} CSS(\text{total}) &= CSS(\text{model}) + SSE \\ \mathbf{y}' \left[\mathbf{I} - \frac{1}{n} \mathbf{J}_n \mathbf{J}'_n \right] \mathbf{y} &= \mathbf{y}' \left[\mathbf{H} - \frac{1}{n} \mathbf{J}_n \mathbf{J}'_n \right] \mathbf{y} + \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}. \end{aligned}$$

- $F_{obs} = \frac{MS(\text{hypothesis})}{MSE} = \frac{SSH/dfH}{SSE/dfE}$
 $\equiv \frac{[SSE(\text{reduced}) - SSE(\text{full})]/[dfE(\text{reduced}) - dfE(\text{full})]}{SSE(\text{full})/dfE(\text{full})}$
 $= \frac{CSS(\text{Regression})/(p-1)}{SSE(\text{full})/(n-p)}.$

Reject the hypothesis if $F_{obs} \geq F_F^{-1}(1 - \alpha, p - 1, n - p) = f_{crit}$.

The usual test of overall regression assumes model spans an intercept and excludes the intercept from the test.

- ANOVA table.

- Usual “Corrected” R^2 :

$R_c^2 = \frac{CSS(\text{Regression})}{CSS(\text{Regression}) + SSE(\text{full})} = \frac{CSS(\text{Regression})}{CSS(\text{total})}$. R_c^2 estimates ρ_c^2 , the population ratio of model to total variance, with $0 \leq \rho_c^2 \leq 1$ and $0 \leq R_c^2 \leq 1$.

- The corrected test for overall regression,

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$$

holds if and only if $H_0 : \rho_c^2 = 0$

Topic 7: Testing Hypotheses in Multiple Regression

- All tests compare two models: the full model and the reduced model (this is the basic idea of likelihood ratio tests, called the *likelihood ratio principle*).
- Overall test: $F_{obs} = \frac{CSS(\beta_1, \dots, \beta_{p-1})/(p-1)}{SSE(\beta_0, \dots, \beta_{p-1})/(n-p)}$.
- Added-Last Test: the *added-last test* seeks to assess the usefulness of one predictor, above and beyond all others.
Coefficient Estimates/t-test table, Type III table. The F statistic is

$$F_{obs} = \frac{\frac{SSE(\text{reduced}) - SSE(\text{full})}{dfE(\text{reduced}) - dfE(\text{full})}}{\frac{SSE(\text{full})/dfE(\text{full})}{}} = \frac{(\hat{\theta} - \theta_0)' \mathbf{M}^{-1} (\hat{\theta} - \theta_0) / dfH}{\mathbf{y}' (\mathbf{I} - \mathbf{H}) \mathbf{y} / dfE},$$

where $\mathbf{C} = \begin{bmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 \end{bmatrix}_{1 \times p}$

- Added-in-Order Test: the *added-in-order test* seeks to assess the

contribution of predictor j above and beyond all of the preceding $j - 1$ predictors (without the $j + 1$, $j + 2$, etc. predictors in the model).

- Group Added-Last Tests
- Group Added-in-order Tests

Topic 8: Correlations

$$\begin{aligned}\rho &= \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \\ R &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left(\sum_{i=1}^n (X_i - \bar{X})^2\right) \left(\sum_{i=1}^n (Y_i - \bar{Y})^2\right)}}.\end{aligned}$$

Partial correlations describe the strength of the linear relationship between two variables, Y and X , after controlling for the effects of other variables \mathbf{Z} .

Topic 9: GLM Assumption Diagnostics

- The First Step: Get to Know Your Data
- Homogeneity: violations seen in the pattern of residuals.
- Independence: assessed through logic of sampling scheme.
- Linearity: examine pattern of residuals.
- Existence: (finite sample...).
- Gaussian distribution: distributional assessment involves box plot of residuals, histogram of residuals, and test of Gaussian distribution of residuals. (The discrepancy between T and Gaussian random variables somewhat inflates the probability of rejecting the null...why?)
- Outliers: leverage, Influence: Cook's Distance

Topic 10: Computation Diagnostics

- Colinearity
- Eigenanalysis
- Condition Number and Condition Index: the *condition index* for the k th eigenvalue equals $\sqrt{\lambda_1/\lambda_k}$. The maximum condition index, called the *condition number*
- R_j^2 , Tolerance, and VIF

$$R_j^2 = R^2(X_j, \{X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_{p-1}\})$$

$$\text{VIF}_j = \frac{1}{1 - R_j^2} = \frac{1}{\text{tolerance}}.$$

- Leverage
- Cook's distance

Topic 11: Selecting the Best Model

1. Specify the maximum model under consideration.
2. Specify a criterion for model selection.
3. Specify a strategy for applying the criterion.
4. Conduct the analysis.

Topic 12: ANOVA

- Coding schemes

$$\text{Es}(\mathbf{X}_{ref}) = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}_{3 \times 3} \quad \text{Es}(\mathbf{X}_{cell}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}_{3 \times 3}$$

$$\text{Es}(\mathbf{X}_{anova}) = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}_{3 \times 4}$$

$$\text{Es}(\mathbf{X}_{effect}) = \begin{bmatrix} 1 & -1 & -1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}_{3 \times 3}$$



- Step down test

Topic 13: Coding Schemes for Regression

- (Regression)
$$\mathbf{y} = \begin{bmatrix} 1 & \mathbf{x} \\ 1 & \mathbf{x} \\ 1 & \mathbf{x} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \boldsymbol{\varepsilon}$$

- (ANOVA)
$$\mathbf{y} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \boldsymbol{\varepsilon}$$

- (Intercept Only)
$$\mathbf{y} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} [\beta_0] + \boldsymbol{\varepsilon}$$
- (Null)
$$\mathbf{y} = \boldsymbol{\varepsilon}$$

Topic 14: Logistic Regression

- Definition of odds, and odds ratio
- The general logistic regression model is given by

$$\begin{aligned}\text{logit}(p_i) &= \log \left(\frac{p_i}{1 - p_i} \right) \\ &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{p-1} x_{i,p-1}\end{aligned}$$

with $y_i \sim \text{Bernoulli}(p_i)$, $i = 1, \dots, n$, and the y 's independent of each other.

- Interpretation of regression coefficients in terms of odds ratio.
- Model comparison by likelihood ratio test
- Logistic regression with categorical covariates and their interactions.
- Goodness of fit test

Topic 15: Mixed Effects Model

- When data are correlated and the independence assumption does not hold, mixed effects models are one way to adjust for the non-independence of observations 
- Random effects may be introduced to account for the fact that observations within one subject (or more generally, within one cluster) may be more alike than observations from different clusters
- Forms of covariance matrices for clustered and repeated measurements
- Parameter interpretation of models for longitudinal data