Statistical Modeling with Correlated Data

Welcome!
We're happy that you're here 🙂

Summer 2024

This workshop does not involve coding and no specific software requirements are involved, so there is no need to download anything.

Instructor: David Nichols

TA: John Lee

This workshop is brought to you by

Northwestern IT
Research Computing and Data Services

Got a programming, data, machine learning, statistics, or visualization question about your research?

We're here to help.

Go to bit.ly/rcdsconsult to request a FREE consultation.

LOGISTICS

- John will be monitoring the chat and Q&A for questions during the presentation, but I’m planning to answer questions after the presentation.
- Participants will be muted in order to keep possible noise from disturbing others.
- There are some polls and quizzes during the presentation. Please participate in these to enhance your experience and provide information to us. Individual responses will be anonymous.

Statistical Modeling with Correlated Data

Predictive statistical models such as linear regression and analysis of variance and covariance are basic workhorses of modern empirical research. The simplest and most commonly used versions of these models are designed for the analysis of independent observations. When multiple observations are gathered on the same units or units are sampled in groups, observations are no longer independent and methods that take into account correlations among dependent observations are generally required in order to make valid inferences.

This webinar will provide an introductory overview of methods for analysis of correlated data, beginning with simple tests of equality of means and touching on more extensive methods for the predictive analysis of data from repeated measurements and clustered or hierarchical/multilevel data.

Statistical Modeling with Correlated Data

Goals of the workshop:

- ➢ Demonstrate and emphasize the importance of using proper methods with correlated data
- ➢ Provide some of the vocabulary necessary to identify methods
- ➢ Provide references to examples of software packages/procedures designed to handle regression modeling with correlated data

Examples of correlated data

- Responses to survey questions from members of the same family
- Measurements taken from related animals or members of a pack or group
- Plant measurements taken from plants sharing a common ecological area
- Different types of measurements from the same object or subject, such as height and weight or opinions about different issues
- Any measurements or responses where the same thing is being observed multiple times

Outline of the presentation

- ➢ Estimation and testing: populations and samples
- ➢ Hypothesis testing and confidence intervals: Interpretation and types of errors
- ➢ The linear model for ANOVA, ANCOVA, and linear regression
- ➢ Model assumptions and their relative importance
- ➢ Within group association: intraclass correlation and its implications
- ➢ Analyzing data from complex samples

Outline of the presentation

- ➢ Some basic matrix terms and concepts
- ➢ Analyzing repeated measures data: the classical multivariate general linear model approach
- ➢ Analyzing repeated measures or other hierarchical data with linear mixed models
- ➢ Bonus mention: generalized estimating equations (GEE)
- ➢ Bonus mention: panel data models
- ➢ Bonus mention: time series models

Estimation and testing: populations and samples

Estimation and testing: Populations and samples

We care about characteristics (or *parameters*) of identifiable *populations,* which are complete sets of things (e.g., people, animals, cells), but just going out and measuring these characteristics on entire populations is often impractical or impossible, because:

- ➢ Entire populations are usually too big for us to take measurements on every unit in the population (e.g., all people in the United States), or

- ➢ The populations are theoretical rather than existing (e.g., the population of interest is all people who might receive a particular treatment or react to a particular experimental situation)

Estimation and testing: Populations and samples

We approach this problem by:

1. Selecting (*sampling*) subsets of the populations of interest
2. Collecting data from those sampled units, often following exposing those sampled units to experimental treatments or manipulations
3. Using those data to compute *estimates* of *parameters* and measures of uncertainty of these estimates
4. Making *inferences* about the parameters of entire population(s) based on the estimates from sample data

Estimation and testing: Populations and samples

In other words, statistical estimation and hypothesis testing involves making inferences about existing or possible population parameters based on data from samples

- The specified hypothesis tested is usually denoted the *null hypothesis*, and denoted by H_0
- An example of testing a specified hypothesis might be that the mean of a particular population on some characteristic is equal to 7
- Our example null hypothesis can be written as

$$H_0 \colon \mu = 7$$

Estimation and testing: Frequentist vs. Bayesian

- The most common approach to hypothesis testing, and the one discussed today, is what is known as the *frequentist* approach, which assumes that probabilities are objective values based on long-run frequencies
- An alternative approach, known as the *Bayesian* approach, also sometimes referred to as the subjective approach, where probabilities are degrees of belief, won’t be covered today
- However, we might consider offering something in the future

Interest in workshop on Bayesian methods

Please take the Zoom poll

Your responses will be anonymous

Hypothesis testing and confidence intervals: Interpretation and types of errors

Hypothesis testing and types of errors

- Hypothesis testing results in a binary decision: to reject or not to reject H_0

- The unknown actual state of affairs either corresponds to H_0 or not (*i.e.*, the null hypothesis is false)

- This produces a 2 x 2 classification of possible outcomes when testing a specified hypothesis

Hypothesis testing and types of errors

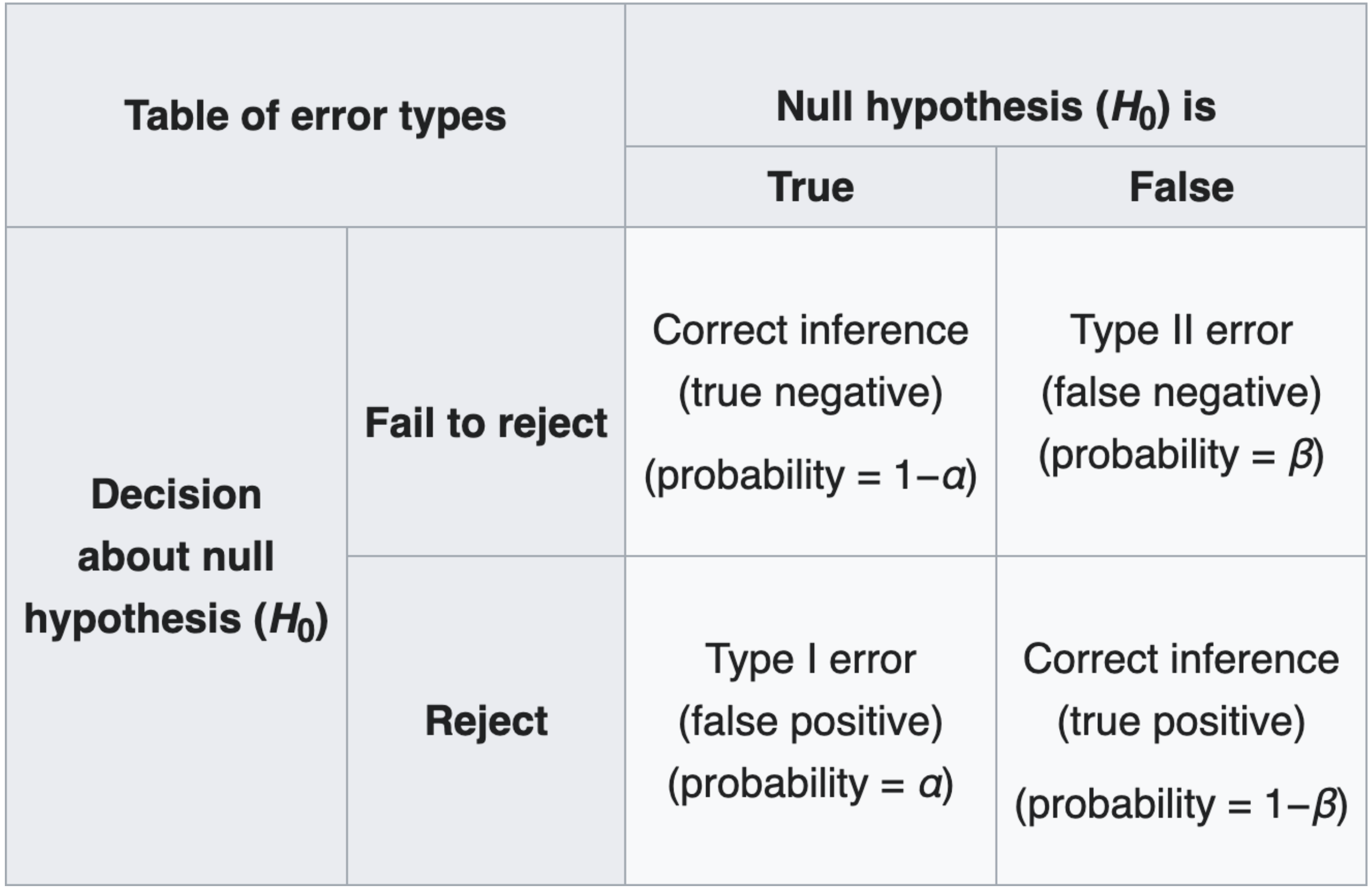


Table of error types		Null hypothesis (H_0) is	
		True	**False**
Decision about null hypothesis (H_0)	**Fail to reject**	Correct inference (true negative) (probability = $1-\alpha$)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive) (probability = $1-\beta$)

Source: Wikipedia (https://en.wikipedia.org/wiki/Type_I_and_type_II_errors)

p value interpretation: Not intuitive

- The p value is the probability in repeated sampling, assuming the model is correct, with its assumptions all met, and the null hypothesis is true, of obtaining a test statistic as large or larger in magnitude than the one observed
- For example, suppose we perform an analysis of variance to test equality of three population means
- Our null hypothesis is

$$H_0\colon \mu_1 = \mu_2 = \mu_3$$

p value interpretation: Not intuitive

- We compute an F statistic and a p value from our sample data, and

$$p = P(F \geq F_{obs} \mid H_0)$$

- $P(\quad)$ is the probability of whatever is in parentheses
- F is the theoretical F distribution with appropriate degrees of freedom
- F_{obs} is the observed F statistic we computed
- The vertical bar means conditional on, or given

p value interpretation: Not intuitive

$$p = P(F \geq F_{obs} \mid H_0)$$

- So p is the probability, given that the null hypothesis is true, of obtaining an F statistic as large or larger than the one we observed
- It's a conditional probability of data at least as discrepant from the null hypothesized state of affairs appearing given that the null is true
- It is *not* the probability that the null hypothesis is true or false, or that the alternative hypothesis is true or false

Uncertainty estimates about parameters: Confidence intervals

- In addition to tests we might perform, we want to estimate ranges of likely values of parameters
- This has become mandatory in journals in many fields
- These estimated ranges based on frequentist methods are known as confidence intervals
- In many standard cases, a $100 * (1 - \alpha)\%$ confidence interval will contain the value posited under H_0 if and only if we fail to reject H_0 at the α level

- For example, a 95% confidence interval will contain the null parameter value if and only if $p > .05$

Confidence interval interpretation: Also not intuitive

- The interpretation of a frequentist confidence interval is that in repeated sampling under the assumed model, $100 * (1 - \alpha)\%$ of such confidence intervals will contain the true population value of the estimated parameter
- It is *not* a statement that there is a given probability (e.g., .95 for a 95% confidence interval) that the population parameter is contained in that interval
- People sometimes use the term confident, as in 95% confident, that the true mean lies in the interval

The linear model for ANOVA, ANCOVA, and linear regression

Standard methods for mean comparisons: Special cases of a general linear model

- One-sample and independent-samples t tests and between-subjects analysis of variance (ANOVA), as well as analysis of covariance (ANCOVA), and linear regression models can all be formulated as instances of a single general linear model
- For a single dependent variable y, measured once per case or subject, they are all instances of a univariate general linear model
- With multiple dependent variables per case or subject, a multivariate version of the general linear model accommodates all of these models

The univariate fixed effects linear model

The univariate fixed effects linear model is

$$y_i = \beta_0 + \beta_1 X_{i1} + \ \dots + \beta_j X_{ij} + \varepsilon_i$$

- y_i is the observed dependent variable value for the i^{th} case
- β_0 is a fixed but unknown intercept or constant
- β_j is a fixed but unknown parameter for the j^{th} predictor variable
- X_{ij} is the value of the i^{th} case for the j^{th} predictor
- ε_i is a random error for the i^{th} case

The univariate fixed effects linear model

- The X_{ij} may be instances of “continuous” variables, which are also known as quantitative, interval, measured, or scale variables, or covariates
- They may also be indicators indexing the levels of categorical variables or *factors* representing separate groups of cases
- They may also be variables representing products of elemental variables (interactions between variables) or powers of a single variable (polynomial terms)

Model assumptions and their importance

The standard normal theory linear model

So far nothing has been said about the assumptions related to the random errors ε_i

For the standard model, estimated using ordinary least squares (OLS), to apply *t* and *F* tests, these assumptions can be stated as

$$\varepsilon_i \sim i.i.d.\mathcal{N}(0, \sigma^2)$$

This means the errors ε_{ij} are independently and identically distributed ("$i.i.d.$") according to a normal distribution with mean or expectation 0 and common variance σ^2

Standard linear model assumptions

These assumptions are often stated as:

1. Independence
2. Homogeneity or equality of variance
3. Normality

What are the relative importances of these assumptions for the use of standard *t* and *F* tests and associated confidence intervals?

Relative importance of assumptions quiz

Please take the Zoom quiz

Your responses will be anonymous

Relative importance of assumptions

There is generally a clear hierarchy of importance:

1. Independence

2. Homogeneity or equality of variance

3. Normality

Why independence is generally most important

- To see why independence is generally the most important of these assumptions, we need to understand the implications of errors being correlated

- An important concept for this understanding is that of intraclass correlation, which can be used to measure association of cases within groups

Within group association: Intraclass correlation and its implications

Interclass correlation

- When you think of correlation, the first type that comes to mind is likely the Pearson correlation coefficient, which measures linear association between two variables
- This is a type of *interclass* correlation, where each value within a pair belongs to a particular variable
- The classes referred to here are the two variables
- The Pearson correlation coefficient measures the extent to which pairs of values, one from each variable for the same case, resemble each other

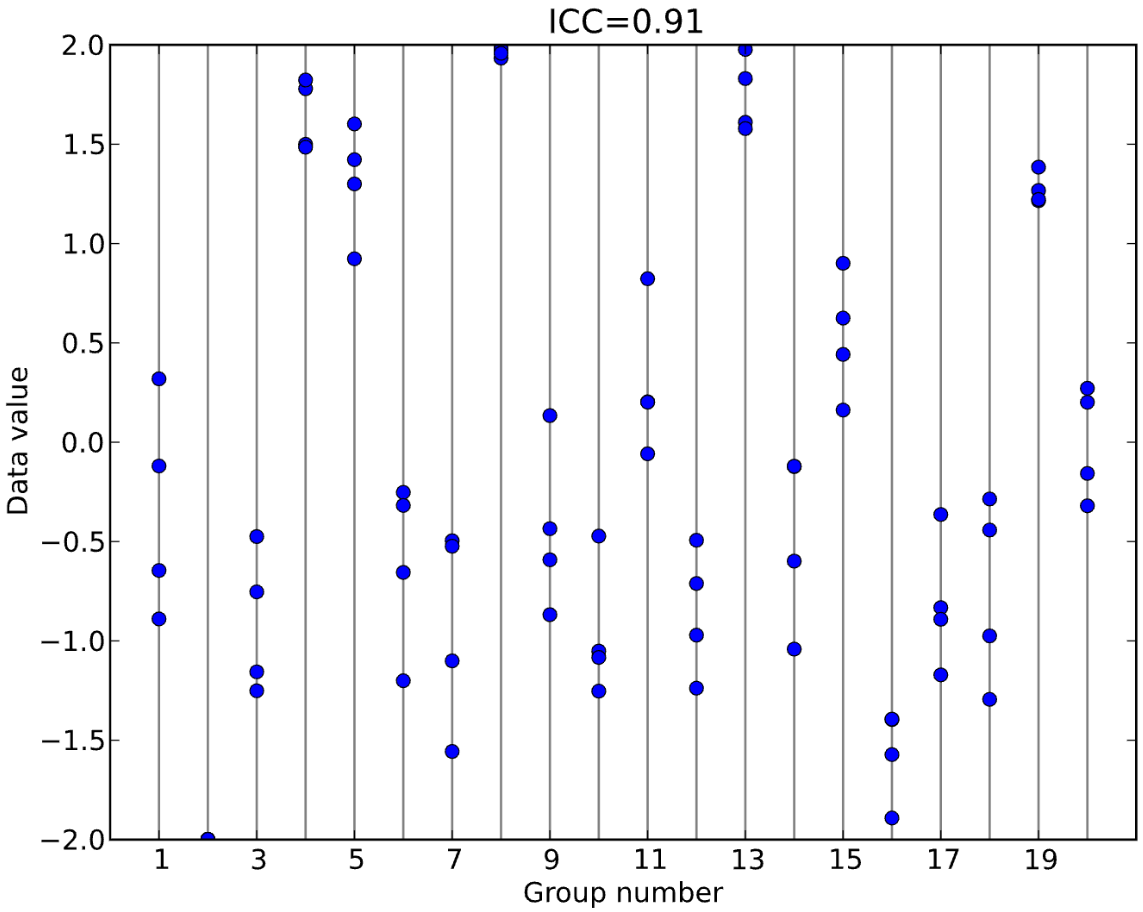
Interclass vs. intraclass correlation

- An *intraclass* correlation (ICC) measures the extent to which values within a group resemble each other relative to the extent to which values in different groups resemble each other

- Each group may have an arbitrary number of cases, and the group sizes need not be equal

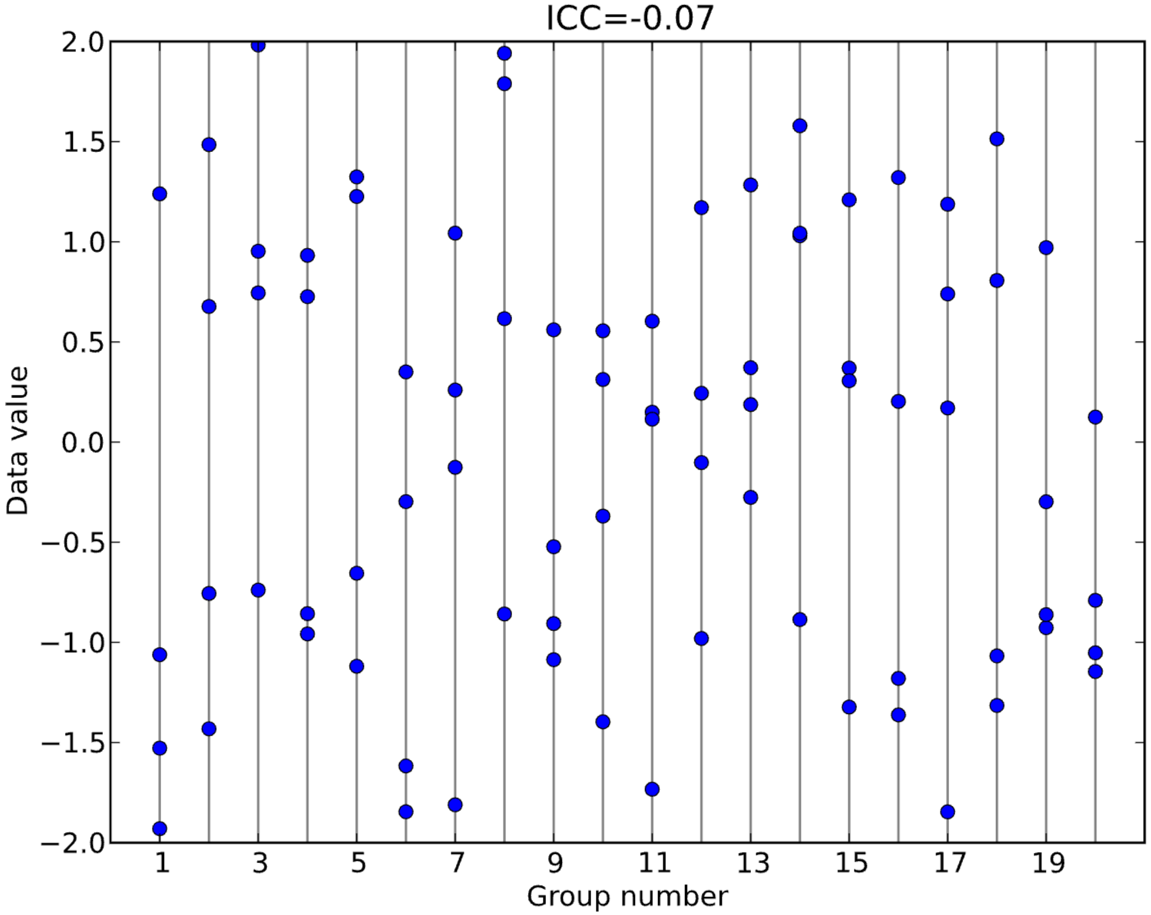
- Values within a class are interchangeable

Example data with a high ICC



By Skbkekas - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=6982852

Example data with a small negative ICC



By Skbkekas - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=6982834

Association within groups: intraclass correlation

- There are a number of types of intraclass correlation coefficients (*ICCs*) and multiple approaches to estimating them

- There is plenty of material involving ICCs and their estimation to be the focus of a lengthy workshop

- Here we're concerned with ICCs that measure the extent to which observations within sampled groups or clusters are more (or less) similar than observations in other sampled groups or clusters

Implications of positive intraclass correlation

In a standard two-sample t test or one-way independent samples analysis of variance, we assume that errors from cases within groups are uncorrelated as well as between errors from cases in different groups

Let's consider conducting a one-way analysis of variance where instead of being independent, the errors within groups are positively correlated with common intraclass correlation ρ, while errors for observations in different groups remain independent

Implications of positive intraclass correlation quiz

Please take the Zoom quiz

Your responses will be anonymous

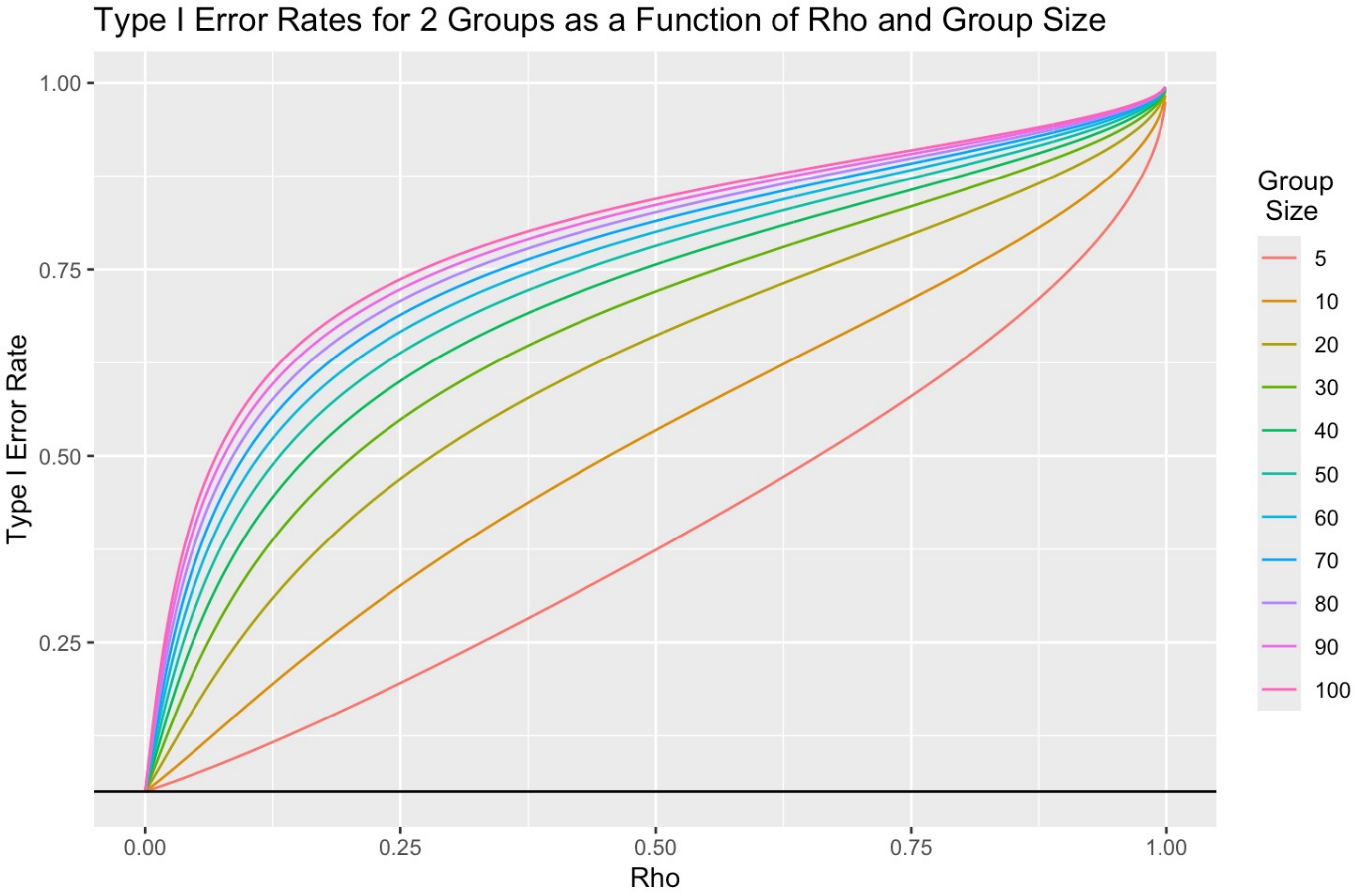
Answers to quiz questions

- With $\rho = .1$ and 20 cases evenly divided into two groups, the standard t or F test has a Type I error rate of about .17, more than three times the nominal .05 level

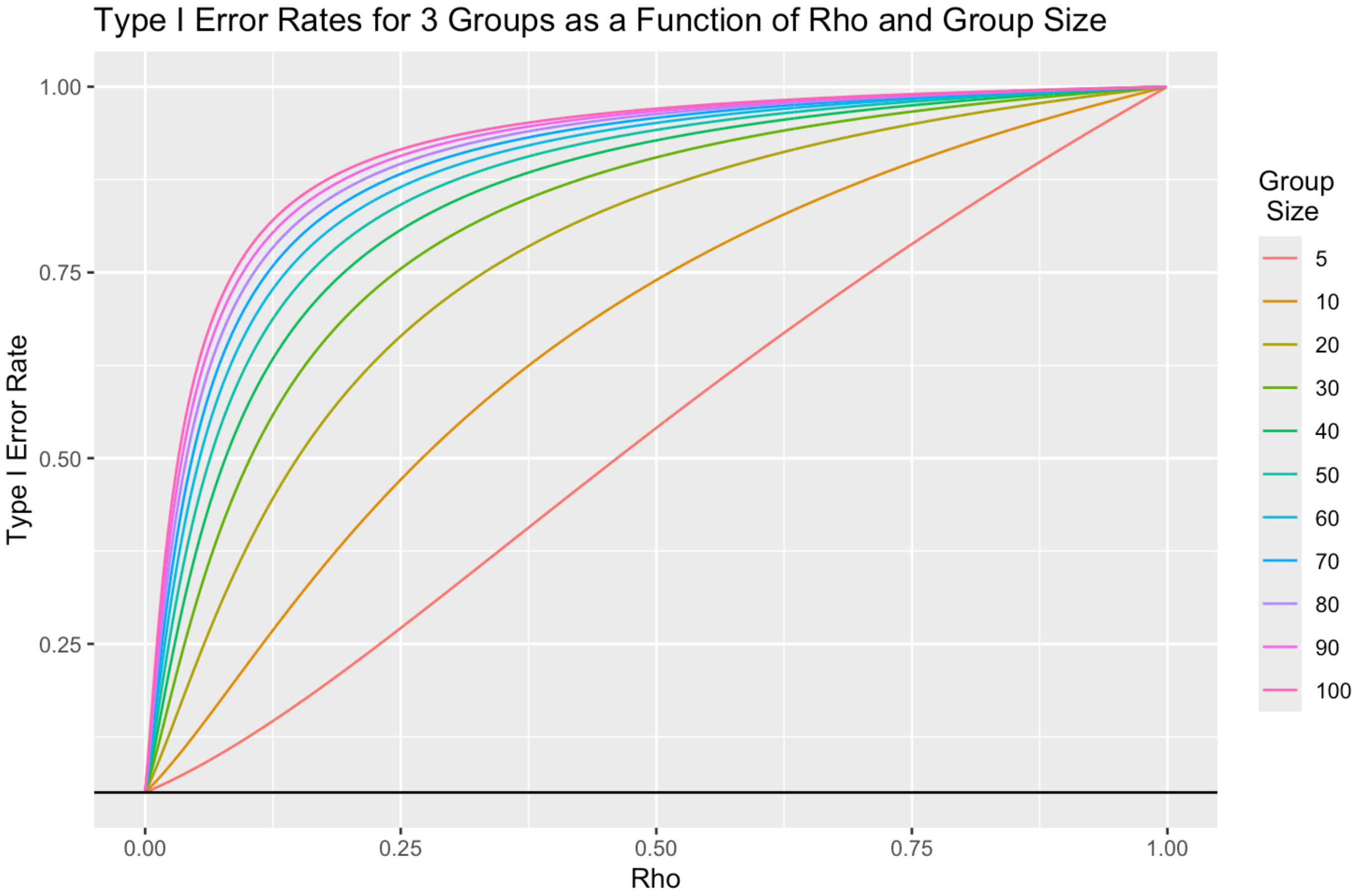
- If we triple the sample size to 30 cases per group and 60 overall, the Type I error rate is about .34

- If we maintain the group sizes of 30 per group but have three groups, for 90 total cases, the Type I error rate is about .69

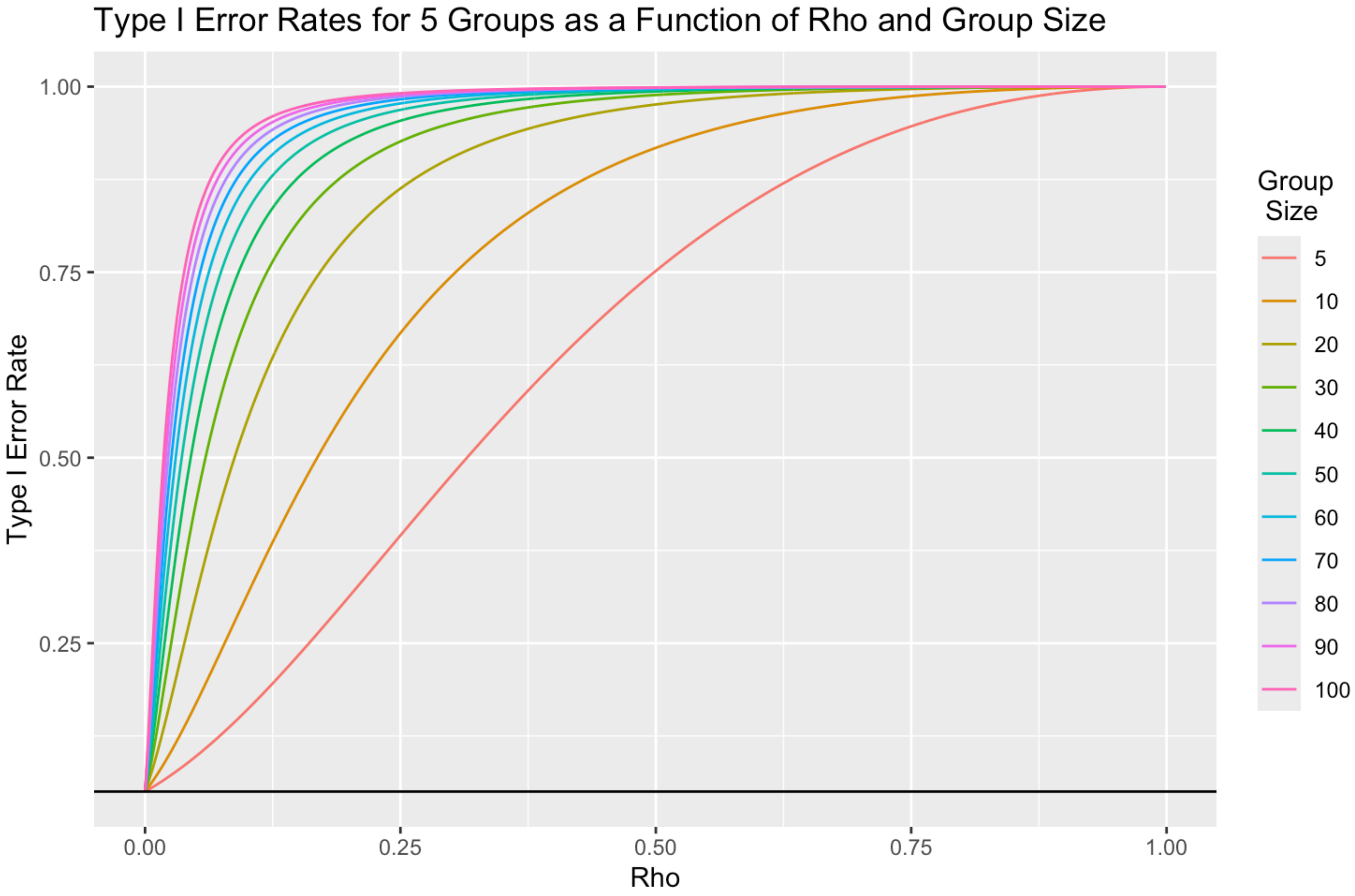
Type I error rates with 2 groups



Type I error rates with 3 groups



Type I error rates with 5 groups



Implications of positive ICC

- Effects on Type I error can be dramatic
- For a given number of groups and observations per group, as ρ increases, the Type I error rate increases and can approach 1
- For a given ρ, increasing the sample size either by adding more groups of the same size or increasing the sample size within groups results in increases in the Type I error rate
- Effects on confidence interval coverage levels are equally bad: as the true α level increases the percentage of $100 * (1 - \alpha)\%$ confidence intervals that cover the true 0 value decreases and can approach 0%

Implications of positive ICC

- Effects on Type I error can be dramatic and are typically much more severe than those associated with inequality of variance or non-normality

- Unlike heterogeneity of variance or non-normality, *increasing sample sizes does not help – it makes things worse!*

- The central limit theorem doesn’t help here – it assumes independence of observations

Implications of negative ICC

- As might be expected based on what happens with positive ρ, when ρ is negative and we assume independence, Type I error rates are reduced below nominal levels, and coverage rates for confidence intervals exceed nominal coverage levels

- Positive ICCs are much more common than negative ones

- Some common estimators of ρ are based on ratios of variance estimates, which can only be non-negative

Common sources of correlated errors

The most common reasons for correlated errors are:

- Units being sampled in groups rather than independently

- Units sampled independently being exposed to common conditions not included in the assumed model

- Repeated measurements taken on the same cases or units

Analyzing data from complex samples

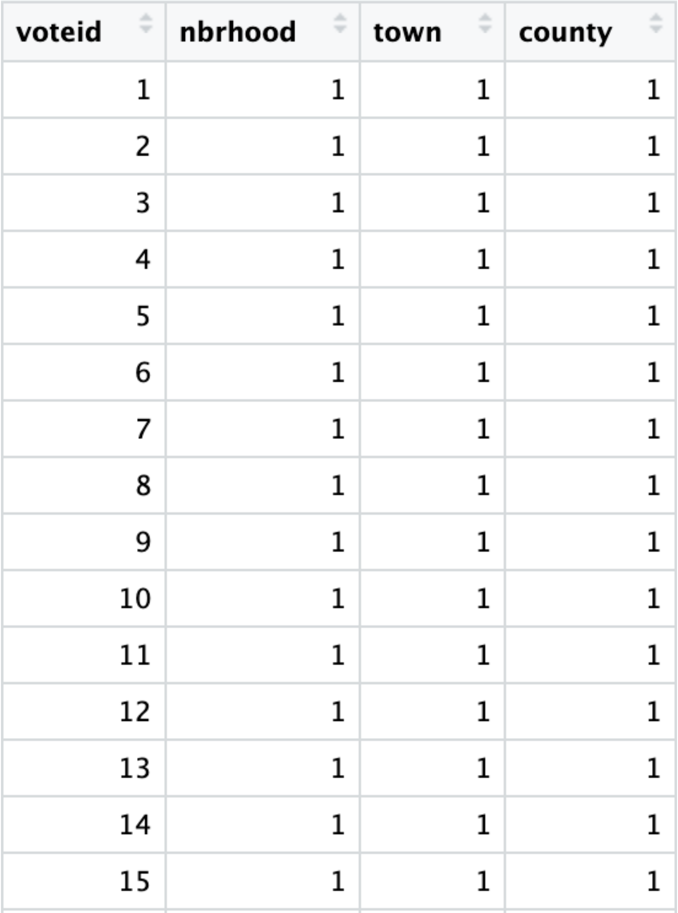
Correlation due to sampling in groups or clusters: Complex sampling

- ➢ Deliberate sampling of groups or clusters of units is common in designed survey sampling, often referred to as complex sampling

- ➢ Full scale design and analysis of complex survey data generally involves construction of a sampling frame consisting of the entire finite population of interest

- ➢ A sampling frame is just a list of the units in a population, typically in the form of a data set containing variables used to select the sampled units

Example sampling frame



voteid	nbrhood	town	county
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5	1	1	1
6	1	1	1
7	1	1	1
8	1	1	1
9	1	1	1
10	1	1	1
11	1	1	1
12	1	1	1
13	1	1	1
14	1	1	1
15	1	1	1

Complex sampling

- If we were to just randomly sample at the individual voter ID level, with each individual having an equal chance of being sampled, this would be a simple random sample (SRS)
- Cluster sampling often involves multiple stages of sampling, and may also involve sampling clusters within strata of other characteristics
- For example, with our polling data, we might stratify by county and sample only some towns, or we might stratify by both county and town and sample neighborhoods
- An additional stage of sampling typically would then be selecting individual voters from the sampled neighborhoods

Complex sampling

- At each stage of the sampling process, weights are created that are accumulated (multiplied over stages) so that at the end, they can be used to weight sample data to accurately represent the population and compute variance estimates and standard errors that accurately represent the uncertainty involved in using the samples for estimation of population quantities
- There are a number of types of designs used under different circumstances, and properly handling data from a given type involves specific adjustments to methods for estimating variances
- In most cases, especially when cluster sampling is involved, failing to apply proper variance estimation methods results in estimates that are too small, resulting in increases in Type I errors and optimistically narrow confidence intervals

Complex sampling

- Complex sampling is essentially its own subfield of statistics with a long and extensive history of methods that are applied to all types of models where inference is involved
- Government statistical agencies such as the U. S. Census Bureau and Statistics Canada make extensive use of such methods, and generally will publish guides to analyses of their survey products that help researchers produce appropriate estimates
- This typically involves applying the provided weights and specifying in survey analysis software the particulars of the design

Complex sampling

- Two general approaches to computing appropriate variance estimates are used:

  1. Linearization, or Taylor series linearization, also known as delta method approximation

  2. Replication or resampling methods, such as jackknife, balanced repeated replication (BRR), and bootstrap

- Adjustments are also sometimes use to align sample characteristics to known population values (terms include poststratification, raking, and calibration)

Complex sampling variance estimation: Linearization methods

- Linearization methods have the advantage that they do not involve resampling or iterative calculations for variance estimation
- Some statisticians consider these to be the “gold standard” for estimating variances in complex samples
- A disadvantage is that they require derivations of formulas for particular designs, and it can be very complicated to derive them for complex estimates such as nonlinear combinations of means
- They also require access to sample design variables (stratification and clustering variables), and in some cases these may not be available due to privacy considerations

Complex sampling variance estimation: Replication and resampling methods

- ➢ Replication or resampling methods are in some ways more general and flexible
- ➢ They can be applied to situations where deriving linearization estimates is difficult
- ➢ They can be applied without access to design variables
- ➢ However, they do require more computational overhead and time
- ➢ Fortunately, most computers today are sufficiently powerful that replication and resampling methods for most problems are practical

Analyzing complex sample data

- Major statistical packages available at Northwestern all support linearization methods
- R, SAS, and Stata also support replication and resampling methods – SPSS does not
- Main package in R is survey
- SAS procedures begin with SURVEY (e.g., SURVEYREG)
- SPSS procedures begin with CS (e.g., CSGLM)
- Stata facilities are listed under svy and are very extensive

Complex sampling bottom lines

- Complex sampling is a big specialized area with its own terminology and lots of complicated issues to consider
- The most important thing to do as a researcher is to understand how sampling is done and to match analysis methods to that sampling design – don’t just apply standard methods that assume simple random sampling
- If full design information is available and estimation of standard quantities is all that’s desired, linearization methods are probably an easier way to go
- If full design information is not available and/or complicated estimators are involved, replication or resampling methods might be preferred or required

Some basic matrix terms and concepts

Some matrix concepts and terms

Statistics makes a great deal of use of matrix algebra, so a few basic terms are helpful (and will be mentioned later):

- A *matrix* is a two-dimensional array of numbers, such as

$$\begin{bmatrix} 7 & 5 & 8 \\ 3 & 9 & 2 \end{bmatrix}$$

- The dimensions of the matrix are described by the number of rows and columns, as in $r \times c$ or $i \times j$
- This is a 2×3 matrix (2 rows, 3 columns)

Some matrix concepts and terms

- The individual numbers of elements of the matrix are often referred to using subscript notation
- M_{ij} refers to the element in the i^{th} row and j^{th} of the matrix M
- In our example matrix

$$\begin{bmatrix} 7 & 5 & 8 \\ 3 & 9 & 2 \end{bmatrix}$$

$$M_{13} = 8$$

and

$$M_{22} = 9$$

Some matrix concepts and terms

- A one-dimensional array is referred to as a *vector*
- There are two types of vectors

- row vectors

$$\begin{bmatrix} 3 & 7 & -1 \end{bmatrix}$$

- column vectors

$$\begin{bmatrix} 9 \\ 13 \end{bmatrix}$$

Some matrix concepts and terms

- A matrix with the same number of rows and columns is a square matrix

$$\begin{bmatrix} -2 & 7 \\ 13 & 5 \end{bmatrix}$$

- A square matrix M where $M_{ij} = M_{ji}$ for all i, j is a *symmetric* matrix

$$\begin{bmatrix} 2.5 & 4.5 & 7 \\ 4.5 & 10 & 12 \\ 7 & 12 & 40 \end{bmatrix}$$

Some matrix concepts and terms

- Probably the most common symmetric matrices you’ll see are covariance and correlation matrices

- Our example symmetric matrix is a covariance matrix (often denoted by C)

$$C = \begin{bmatrix} 2.5 & 4.5 & 7 \\ 4.5 & 10 & 12 \\ 7 & 12 & 40 \end{bmatrix}$$

Some matrix concepts and terms

- This covariance matrix C in standardized form yields the correlation matrix R

$$R = \begin{bmatrix} 1 & .9 & .7 \\ .9 & 1 & .6 \\ .7 & .6 & 1 \end{bmatrix}$$

Some matrix concepts and terms

- Flipping a matrix so that the rows become the columns and the columns become the rows is known as transposing it

- The resulting matrix is known as the transpose of the original matrix

- A matrix transpose of the matrix M might be denoted as M' or M^T

- The transpose of symmetric matrix is the same as the original

Some matrix concepts and terms

- A square matrix where $M_{ij} = 0$ for all $i \neq j$ is a *diagonal* matrix

$$\begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

Some matrix concepts and terms

- A diagonal matrix where $M_{ii} = 1$ for all i is an *identity* matrix

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Some matrix concepts and terms

- If each diagonal element in a diagonal matrix is replaced by a square matrix and all off-diagonal elements are replaced by square matrices with every element 0, we have a *block diagonal* matrix

Some matrix concepts and terms

An example of a block diagonal matrix:

$$\begin{bmatrix} 1 & .5 & .3 & 0 & 0 & 0 & 0 & 0 & 0 \\ .5 & 1 & .7 & 0 & 0 & 0 & 0 & 0 & 0 \\ .3 & .7 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & .5 & .3 & 0 & 0 & 0 \\ 0 & 0 & 0 & .5 & 1 & .7 & 0 & 0 & 0 \\ 0 & 0 & 0 & .3 & .7 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & .5 & .3 \\ 0 & 0 & 0 & 0 & 0 & 0 & .5 & 1 & .7 \\ 0 & 0 & 0 & 0 & 0 & 0 & .3 & .7 & 1 \end{bmatrix}$$

Matrix representation of the univariate fixed effects linear model

The matrix representation of the model presents a more compact form:

$$\mathrm{y} = X\beta + \varepsilon$$

- y is a column vector of observed values with N rows
- β is a fixed but unknown column vector of parameters for the k parameters in the model ($k = j + 1$ if there is an intercept)
- X is an $N \times k$ matrix of predictor variable values
- ε is a column vector of random errors with N rows

Analyzing repeated measures data: the classical multivariate general linear model approach

Analyzing repeated measures: The multivariate general linear model

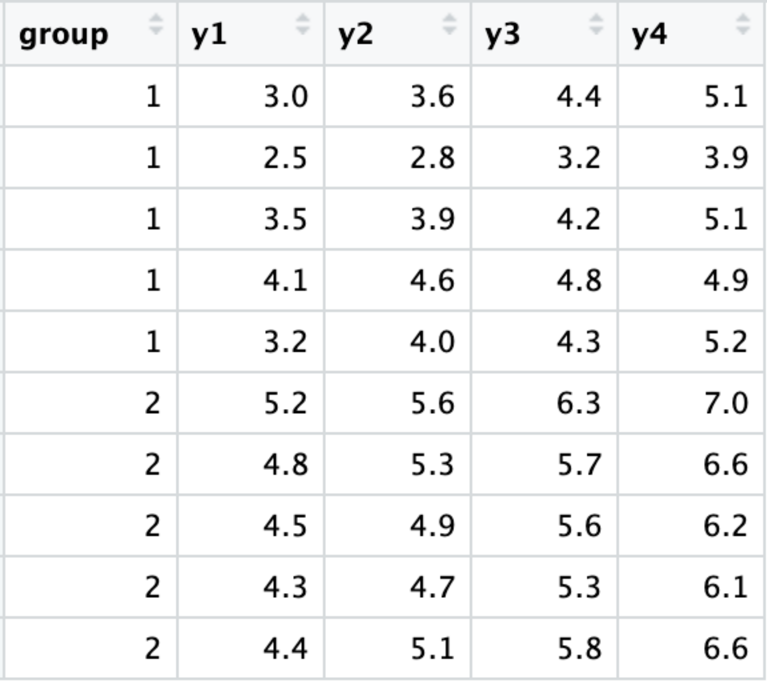
With more than one dependent variable per case or subject, such as with repeated measures data, a multivariate version of the general linear model can be used:

$$\mathrm{Y} = X\beta + \mathrm{E}$$

- Y is an $N \times p$ matrix where the p columns are the p dependent variables
- β is a fixed but unknown $k \times p$ matrix of parameters for the k parameters in the model for each of the p dependent variables
- X is the same $N \times k$ matrix of predictor variable values
- E is an $N \times p$ matrix of random errors

Data structure for the multivariate general linear model

- Data for use with the multivariate general linear model are structured with one case per subject
- An ID variable is optional
- Repeated measurements are held in different variables on the same case
- This structure is referred to as the multivariate approach setup, or wide data



group	y1	y2	y3	y4
1	3.0	3.6	4.4	5.1
1	2.5	2.8	3.2	3.9
1	3.5	3.9	4.2	5.1
1	4.1	4.6	4.8	4.9
1	3.2	4.0	4.3	5.2
2	5.2	5.6	6.3	7.0
2	4.8	5.3	5.7	6.6
2	4.5	4.9	5.6	6.2
2	4.3	4.7	5.3	6.1
2	4.4	5.1	5.8	6.6

The multivariate general linear model for repeated measures

- Errors are still assumed to be independent between cases or subjects, but are allowed to be correlated within a given subject

- Univariate F and t tests as well as an array of multivariate tests can be formulated from the estimates of the multivariate model

- Tests for the within subjects or repeated measures part of the model can be based on very general assumptions about the covariances of the repeated measures within subjects

The multivariate general linear model for repeated measures

- This model is available in general purpose statistical software packages available to Northwestern researchers, including R, SAS, SPSS, and Stata

- SAS and SPSS offer GLM procedures with extensive built-in options that make such analyses particularly easy

The multivariate general linear model for repeated measures

This model is does have some limitations:

- It requires complete data on all variables for a subject in order to be able to include that case in an analysis
- This complete data requirement makes it less than useful in situations where not all repeated measurements are available for some subjects or where subjects are not measured at a consistent set of time points
- The modeling of the within-subjects part of the model automatically includes a full-factorial design on the within-subjects factors
- Covariates for a given subject affect all time points for the repeated measures, so time-varying covariates are not accommodated

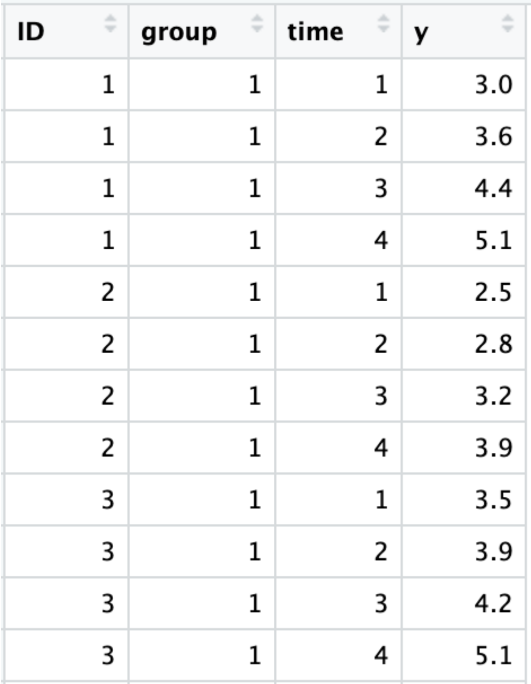
Analyzing repeated measures or other hierarchical data with linear mixed models

Linear mixed models for repeated measures and other hierarchical data structures

- Multiple measurements on the same subject can be thought of as hierarchically structured, where observations are nested within subjects
- An alternative approach to analyzing repeated measures, which can also be applied to data from other hierarchical structures, involves representing data with each measurement for a subject as a separate case of data, with cases for a given subject linked by a common value of a subject identifier (ID) variable, and levels of within-subjects variables indexed by one or more variables

Data structure for mixed models

- An ID variable is required to link cases for a subject
- Repeated measurements are held in the same variable over different cases with the same ID
- An index variable is used to identify repeated measures
- This structure is referred to as the univariate or mixed models approach setup, or as long or narrow data



ID	group	time	y
1	1	1	3.0
1	1	2	3.6
1	1	3	4.4
1	1	4	5.1
2	1	1	2.5
2	1	2	2.8
2	1	3	3.2
2	1	4	3.9
3	1	1	3.5
3	1	2	3.9
3	1	3	4.2
3	1	4	5.1

Linear mixed models for repeated measures and other hierarchical data structures

- This structure allows us to represent situations where each subject might have measurements at only a subset of time points, perhaps even unique ones

- We need only complete data for each represented measurement occasion for that subject

- It also allows for predictor variables to vary over time for a given subject (time-varying covariates)

Clustered data: hierarchical structures

Aside from formal survey sampling and repeated measures, many situations involve data that are hierarchically structured and where treatments may be applied to individual units as members of groups

- ➢ Education: individual students might share the same teacher and classroom, and teachers as well as students are in the same school
- ➢ Biology: individual plants might share immediate environments or animals may be relatives or share important ecosystems
- ➢ Medicine: Patients might share physicians, hospitals or medical centers and/or groups, while physicians might share hospitals or medical centers and/or groups

Clustered data: hierarchical structures

- ➢ Shared characteristics of these groupings often introduce correlation among units in the same groups or clusters

- ➢ Failure to separate variability due to sharing things other than a treatment level often leads to biased estimates of treatment effects

- ➢ Assuming that these other shared experiences are not completely confounded with treatments, we can model their influences and try to separate these from effects of treatments

Analysis of hierarchical data via mixed models

- The univariate linear model discussed earlier treats predictor variables as what are known as fixed effects

- It assumes that the levels of grouping factors in our model that are represented in the data constitute the entire set of levels of each factor, or at least the set about which we want to make inferences

- The systematic effects are all modeled as functions of these fixed effects, and ε represents the only source of random variation in the model

Analysis of hierarchical data via mixed models

- With hierarchical data, the levels of grouping variables included in data may be a sample of a wider population of levels
- This introduces an additional source of random variation into the data
- We can model this additional variation by treating these as what are called random effects
- Modeling random effects can account for correlations among grouped observations
- Models with a combination of fixed and random effects are known as mixed models

Analysis of hierarchical data via mixed models

The linear mixed model equation is

$$y = X\beta + Zu + \varepsilon$$

- u is a column vector of random effects parameters of length q, the number of random effects parameters in the model

- Z is an $N \times q$ design matrix for the random effects parameters (N is the total number of observations over all subjects)

Analysis of hierarchical data via mixed models

The random effects u are assumed to distributed as

$$u \sim \mathcal{N}(0, G)$$

and the errors ε are assumed to be distributed as

$$\varepsilon \sim \mathcal{N}(0, R)$$

G and R are block-diagonal matrices, where each subject’s block is independent of that for every other subject, but relationships within blocks can be of various structures

Analysis of hierarchical data via mixed models

- This model allows us to model certain correlation or covariance structures for cases that are in the same groups that define the random effects subjects, while maintaining independence across groups or subjects
- The simplest such model has a separate random intercept for each subject measuring that subject’s systematic departure from the overall fixed effect mean or intercept
- This single random effect allows us to estimate a common non-negative intraclass correlation within subjects in the G matrix
- The R matrix is modeled as a diagonal matrix with the residual variance in each diagonal element

Analysis of hierarchical data via mixed models

- A variety of more complicated models can be formulated, allowing a great deal of flexibility in effects and covariance structures

- These structures include ones where time is measured flexibly instead of at fixed points and where predictor variables vary over time within a given subject

- The available structures also can accommodate unequal variances across time and/or levels of predictors

Mixed model complexities

The power and flexibility of mixed models is accompanied by some important and substantial complexities

- Unlike the general linear model, estimation of mixed model equations requires iterative techniques based on likelihood methods: maximum likelihood (ML) or restricted/residual maximum likelihood (REML)
- Such models may take much longer to estimate than those handled via least squares methods
- Convergence to stable and globally correct values of parameter estimates can sometimes be difficult or impossible to achieve in cases where the model is not consistent with the data or the data are sparse

Mixed model complexities

- Sample sizes generally need to be substantially larger in mixed models in order to obtain good results
- Model selection is more complicated than with only fixed effects
- Denominator degrees of freedom associated with F statistics in unbalanced mixed models are not simple values and have to be estimated
- Some controversy exists about the validity of the assumed null distributions for these statistics

Primary software options for linear mixed models

- All of the major general purpose statistical software packages available at Northwestern offer mixed models capabilities
- The most popular package in R for such methods is lme4 and its lmer function
- SAS has a MIXED procedure that offers linear mixed models in a very general framework with a huge array of options
- SPSS has a MIXED procedure in a similar general framework with somewhat fewer options
- Stata's offers a mixed procedure for mixed/hierarchical/multilevel linear models

Bonus mention: generalized estimating equations (GEE)

Generalized Estimating Equations (GEE)

- Another approach for correlated data regression modeling that shares similarities in terms of data structure and specifications is the use of generalized estimating equations (GEE)
- Only fixed effects are involved, but repeated measures can be specified for subjects, allowing linked observations to be modeled as correlated
- Less strict assumptions about errors and covariance structures are made than with likelihood-based methods, but more stringent assumptions about missing time points or measurements are required

Generalized Estimating Equations (GEE)

- GEE variance estimation can produce results that are identical to complex samples methods in some situations
- All four of the software packages discussed offer GEE methods, typically in the context of generalized linear models, with linear models being a special case

Bonus mention: panel data models

Panel data models

- Another type of longitudinal or repeated data is what is known as panel data
- Panel data is obtained by measuring the same subjects, objects, or entities over a set of common time points
- It is sometimes referred to as cross-sectional time series data
- It is common and especially popular in economics, sociology, and political science
- Many tests and diagnostics have been created for these models by econometricians

Panel data models

- A number of models and approaches to estimation are available
- A common distinction is between models treating subjects or entities as fixed vs. random
- Modeling approaches often involve combining OLS estimation with various adjustments or as one in a set of steps, though more complicated methods are also used
- R, SAS, and Stata all offer extensive facilities for panel models
- Capabilities in SPSS are much more limited, restricted to overlaps with mixed models and other kinds of regression models available for data more generally

Bonus mention: time series models

Time series models

- Time series modeling, sometimes referred to as forecasting, has a substantial history and literature
- Contributions have been made from various perspectives, but econometricians and statisticians associated with business and commercial interests have produced the bulk of the literature and methods
- For a single extended series of values from measurements at sequential time points with common intervals, methods known as exponential smoothing and autoregressive integrated moving average (ARIMA) models are the most widely used

Time series models

- For cyclical data seasonal models that accommodate regular periodic effects of time are widely used
- Exponential smoothing and ARIMA models both include methods for seasonal or cyclical data
- Seasonal ARIMA models are sometimes referred to as SARIMA models
- A plethora of types of models with more general capabilities are available, many referred to with acronyms

Time series models

- A particularly popular type of generalization of autoregressive models are known as ARCH, or autoregressive conditional heteroskedasticity models, which model the error variance as autoregressive (AR) functions of prior variances
- GARCH, or generalized autoregressive conditional heteroskedasticity models, model error variances as autoregressive moving average (ARMA) functions of prior variances
- All four major packages offer a number of methods, though those in SPSS are more limited than in the other three

Questions?