GLM, your new BFF…

# General Linear Models

*Making friends with Regression, ANOVA, Repeated Measures, and the Linear Mixed Model*

Session 1: Tuesday, Sep 27, 4-6 pm   
Session 2: Tuesday, Oct 4, 4-6 pm   
Ballantine 118

Stephanie Dickinson  
Senior Consultant  
Department of Statistics  
Indiana Statistical Consulting Center (ISCC)  
sd3@indiana.edu

# About ISCC

Indiana Statistical Consulting Center

www.indiana.edu/~iscc  
iscc@indiana.edu

410 N Park Ave, 2nd floor

What kind of analysis should I use to answer my research questions?

How do I run the analysis in my favorite software? (also StatMath Computing) What is the statistical output telling me about me data?

How do I address the reviewer’s comments about the stats in the manuscript I submitted?

Contact us online! <http://www.indiana.edu/~iscc/contact.html>

# http://www.supplierlist.com/photo_images/139040/folding__umbrella__children_umbrella__golf__umbr.jpgWhat is a General Linear Model?

The General Linear Model (GLM) is the umbrella covering a majority of the statistical analyses used in research:

T-tests, ANOVA, ANCOVA, MANOVA, Repeated Measures,

Regression, Linear “Mixed” Models (a.k.a HLM), and more.

But how do they all fit together, and how are they used and abused?

# Terms

**Dependent Variable (DV)** = the outcome measure for the analysis (Y-axis)

**Independent Variable (IV)** = one of the predictor variables (X-axis)

Y = α + β X + ε

**Data Types:**

|  |  |  |
| --- | --- | --- |
| Interval/ Scale/ Continuous/ Metric | http://www.aperfectworld.org/clipart/academic/ruler01a.gif | Test score  Height, weight, age  Response Time  Percent correct |
| Ordinal |  | Bachelor, Masters, PhD. |
| Nominal/ Categorical (≥2)  ~ Binary (2 levels) |  | This, that, or the other;  Yes/no, right/wrong, 0/1,  Something happened or not. |

**NOTE:** It’s one thing to know the **TRUE** nature of the data (Nominal, Ordinal, Scale), but it may be a separate decision how you will **TREAT** the data in analysis.

Ex:

* Likert Scale (Ordinal) will probably need to be treated as Nominal or Scale/Interval.
* Ordinal variable with 3 levels (No, Maybe, Yes) -> treat Nominal
* Likert Scale with 9 levels (strongly disagree to agree) -> treat Continuous
* 5 levels? 4 levels?

There is not a clear “recipe” or a “one-size fits all” for these decisions…

What is meaningful for your data analysis? ...

**Independent Observations**

Most standard analyses (T-test, ANOVA, Regression, Chi-square,…) assume Independent Observations.

Each observation is a random “independent” draw from the larger population.

This assumption is built in to the calculation of the p-value.

…> p-value is the probability of seeing this effect by chance.

Only one measure of the DV for each person in the analysis

|  |  |  |
| --- | --- | --- |
| ID | Group | Test Score |
| 1 | Trt | 98 |
| 2 | Trt | 83 |
| 3 | Control | 67 |
| 4 | Control | 79 |
| etc… |  |  |

**Correlated Data**

|  |  |  |  |
| --- | --- | --- | --- |
| ID | Group | Pre | Post |
| 1 | Trt | 82 | 98 |
| 2 | Trt | 78 | 93 |
| 3 | Control | 67 | 83 |
| etc… |  |  |  |

Repeated Measures

* multiple DV’s per subject (task1, task2)
* multiple measurements of same DV (time1, time2, …)

Random Effects

* multiple observations within subject
* Hierarchical Data
* Unit of analysis is clustered in larger groups
* Students within class, class within school

|  |  |  |  |
| --- | --- | --- | --- |
| ID | School | Group | Test Score |
| 1 | 1 | Trt | 98 |
| 2 | 1 | Trt | 93 |
| 3 | 1 | Control | 67 |
| 4 | 2 | Control | 79 |
| etc… |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| ID | Group | Item | Response time |
| 1 | Trt | 1 | 3.2 |
| 1 | Trt | 2 | 2.5 |
| 1 | Trt | 3 | 3.7 |
| 2 | Trt | 1 | 4.2 |
| etc… |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | DV | Data structure | Analyses |
| **General Linear Model** | DV is Continuous  http://www.aperfectworld.org/clipart/academic/ruler01a.gif | Independent Observations | T-test  ANOVA, ANCOVA Linear Regression |
| **General Linear Mixed Model** | DV is Continuous  http://www.aperfectworld.org/clipart/academic/ruler01a.gif | Correlated Data | “Mixed” Models  Repeated Measures  Random Effects (HLM) |
| **Generalized Linear Model** | DV is Categorical | Independent Observations | Pearson Chi-square  Logistic Regression  (VARBRUL) |
| **Generalized Linear Mixed Model** | DV is Categorical | Correlated Data | Repeated Measures Logistic Regression  GEE, GLIMMIX |

|  |  |  |
| --- | --- | --- |
| **General Linear Model**, Independent Observations | | |
| DV is Continuous http://www.aperfectworld.org/clipart/academic/ruler01a.gif | IV is Categorical | T-test (1 IV: 2 groups (Binary)),  One way ANOVA (1 IV: >2 groups),  Two-way ANOVA (2 IV’s) Factorial ANOVA (>2 IV’s) |
| IV is Continuoushttp://www.aperfectworld.org/clipart/academic/ruler01a.gif | Pearson Correlation (1 IV)  Linear Regression (1 IV)  Multiple Linear Regression (>1 IV) |
| IV’s are Both | ANCOVA, Multiple Linear Regression |
| Multiple DV’s, Continuous |  | Paired T-test (1 IV, 2 levels)  Repeated Measures ANOVA (≥2 levels)  MANOVA (≥2 DV’s) |
| **General Linear Mixed Model**, Correlated Data | | |
| DV is Continuous http://www.aperfectworld.org/clipart/academic/ruler01a.gif | IV’s are Both/Either | Linear Mixed Models  (Repeated Measures, Random Effects) (HLM) |

# What is the difference between ANOVA and Regression?

Nothing! Well…not exactly nothing…

* More about a matter of perspective - experimental design vs observational data
* Mathematically, the model for an ANCOVA (1 categorical IV with 1 continuous “covariate”) is identical to a Regression with 1 categorical IV and 1 continuous IV.
* A “covariate” is just another Independent Variable.
* ANOVA usually includes factorial interactions between IV “factors”, but it’s really up to you…
* Software for “ANOVA” and “Regression” usually have different defaults, but you can always get one from the other. And “GLM” does both!
* ANOVA often includes interactions, shows sums of squares with F tests
* Regression will not include interactions by default, shows beta parameter coefficients

The plan:

1. Assumptions (and Violations) of GLMs
2. The Steps of Analysis

**General Linear Models** (today)

Independent Observations

1. T-test
2. ANOVA
3. ANCOVA
4. Simple Linear Regression
5. Multiple Regression

Correlated Data

1. Paired t-test
2. Repeated Measures ANOVA

**General Linear Mixed Models** (next time)

# Assumptions of General Linear Models

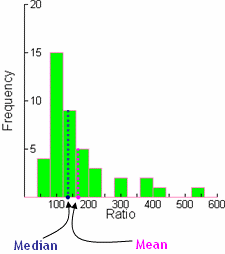
1. Data are a representative sample of some larger population which I’m making inferences about.
2. Observations are independent (or else modeled appropriately in a Repeated Measures or “Mixed” model)
3. The continuous predictor variables (IV)’s have a “linear” relationship with the outcome (DV).
4. There are equal variances between the groups (or across values of continuous predictor variables).
5. Residuals are Normally distributed.  
   *note:* Normally distributed DV is only a ‘proxy’ for this…

~ Note that multicollinearity is not an assumption of the model…just something important to consider.

# When Assumptions are Violated

1. Not a representative sample? Not good. Redefine the “population”?
2. Not a linear relationship? Add polynomial terms (square, cubic,..)?
3. Not independent observations?
   * Aggregate data to the individual level? (esp. binary data!)
   * Model the correlation structure in Linear Mixed Models
4. Equal variances?
   * Levene’s test is only one diagnostic measure… (careful with p-values)
   * What is Std. Dev. in each group? How different are they?
   * Is one more than twice as big as the other?
   * If sample sizes between groups are equal, ANOVA is Robust
   * Log transformations often help with variances too

“Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance”, Glass, Peckham and Sanders. 1972 42: 237 *REVIEW OF EDUCATIONAL RESEARCH*

1. Not Normally distributed residuals?
   * Be skeptical of tests of normality…p-value is more significant with larger sample size, but…
   * But large sample size means you don’t have to worry as much (Central Limit Theorem means ANOVA is Robust)
   * Try taking the log?

Non-parametric tests also an option:

* Wilcoxon Rank-sum (T-test)
* Kruskal-Wallis (ANOVA)
* Less powerful.
* Still assume independent observations.

# Steps of Analysis

1. Formulate Research Question and Hypothesis Test
2. Identify appropriate analysis

* Identify IV’s and DV’s, and data types for each
* Consider data structure (independent or correlated data) and possibly restructure (aggregate?)

1. Perform descriptive statistics and univariate plots
   * Frequency tables for all categorical variables
   * Means, Std.Dev., Min, Med, Max for continuous variables
   * Histograms for continuous variables, especially DV’s
     + If DV not (approximately) normal, consider transformation for linear model
2. Plot relationship of DV on IV’s
   * Box plots for DV on categorical variables
   * Scatterplots for DV on continuous variables
3. Run Analysis
4. Check assumptions
   * Normal residuals
   * Equal variance
5. Interpret Results

* Post Hoc tests
* Estimated Marginal Means
* Beta “slope” coefficients

# Software notes

|  |  |  |
| --- | --- | --- |
|  | SPSS | SAS |
| T-test | Compare means > Independent T-test… or GLM | Proc ttest… or Proc GLM |
| ANOVA | Compare means > One-way ANOVA… or GLM | Proc ANOVA…or GLM |
| ANCOVA | GLM | Proc GLM |
| Regression | Regression… or GLM | Proc Reg…or GLM |
| Multiple Regression | Regression\*… or GLM | Proc Reg\*…or GLM |

* Regression procedures have some advantages like Stepwise selection,…
* But they often make categorical IV variables more difficult since you have to create your own “dummy” variables for each categorical level…(see below)

# “Dummy Variables”

* 0/1 Indicator variables for each level of the categorical variable
* You need one less Dummy variable than the number of levels

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ID | State | Age | Sex | Response | Illinois | Indiana | Male |
| 1 | Illinois | 18 | M | 82 | 1 | 0 | 1 |
| 2 | Illinois | 20 | F | 95 | 1 | 0 | 0 |
| 3 | Indiana | 22 | M | 78 | 0 | 1 | 1 |
| 4 | Ohio | 25 | M | 79 | 0 | 0 | 1 |
| … |  |  |  |  |  |  |  |

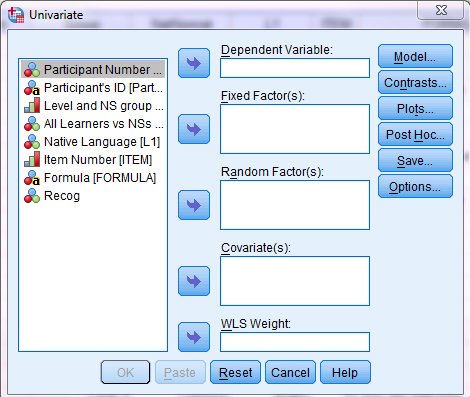
Now you can do a Regression (which assumes you’re giving it “sort-of” continuous variables).

DV = Response

IV = Age, Illinois, Indiana, Male

Y = α + β1 Age + β2 Illinois(0/1) + β3 Indiana(0/1) + β4 Male(0/1) + ε

# “Factors” vs “Covariates”

* GLM creates these dummy variables for you (behind the scenes) when you put them in as “factors” in SPSS (or “class” variables in SAS.)
* In SPSS, “Factors” are any categorical IV. “Covariates” are any continuous IV.
* In SAS, the “Class” statement is for any categorical IV. Others are continuous.
* Your model will be “wonky” to say the least if you mix them up...

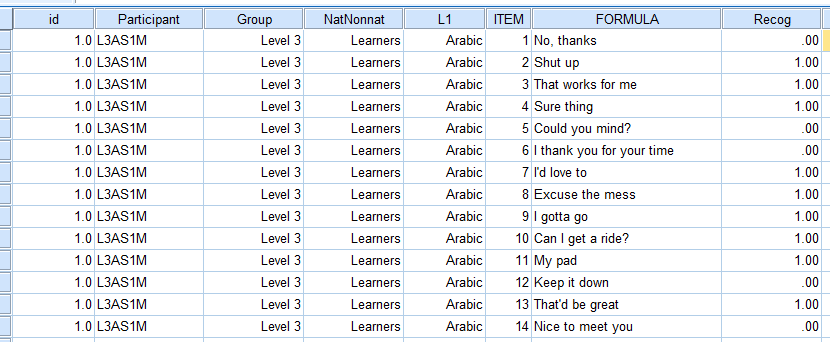
**Continuous variables**

**DON’T USE**

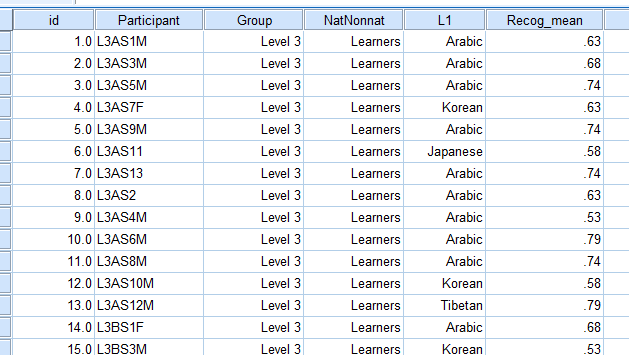
**Nominal variables**

**English language learners and conventional expressions**(Inspired by and adapted from Kathleen Bardovi-Harlig <http://www.indiana.edu/~dsls/publications/KathleenAAAL2011.pdf>)

English language learners (n=75) (levels 3, 4, 5) and Native English speakers (n=25) each answer a questionnaire listing 30 conventional expressions (ex: *No thanks I’m full, Sorry I’m late, and No problem)* and indicate whether they recognize the expressions or not. We count the % recognized for each participant.



The raw DV is binary 0/1 (1=Recognized, 0=Not Recognized). To make a nice pretty continuous outcomes (and remove the correlation between subjects), we can aggregate to get the % Recognized for each subject.



DV: #(or %) expressions Recognized - Continuous

IV: Group (Learner or Native) - Categorical, Binary (2 levels)

Analysis is at participant-level, one observation per person - independent observations

## = Independent Samples T-test

IV: Group (Level 3, Level 4, Level 5 or Native) - Categorical (4 levels)

## = One-way ANOVA

IV: Group (Level 3, Level 4, Level 5 or Native)  
IV: Sex

## = Two-way ANOVA? Or is it ANCOVA?

Just depends if you want to include interaction between sex and group.

Even if you consider sex a “covariate” need to include it in the “factor” box.

IV: Age

## = Regression (Correlation)

IV: Group (Level 3, Level 4, Level 5 or Native)  
IV: Sex  
IV: Age

## = Multiple Regression? or GLM

Run all “Steps of Analysis” from above…

(see SPSS…)

Resources

UCLA Stats Consulting - <http://www.ats.ucla.edu/stat/>   
What Statistical Analysis should I use? <http://www.ats.ucla.edu/stat/mult_pkg/whatstat/default.htm>

*Discovering Statistics Using SPSS*  
by Andy Field

*A Guide to Doing Statistics in Second Language Research Using SPSS*   
by Jenifer Larson-Hall