# A Re-Introduction to General Linear Models (GLM)

- Today's Class:
  - You do know the GLM
  - Estimation (where the numbers in the output come from): From least squares to residual maximum likelihood (REML)
  - Reviewing specification of fixed effects in GLMs
    - Centering continuous predictors
    - Two ways of including categorical predictors
    - Concepts in SAS: CONTAST, ESTIMATE, and LSMEANS
    - Concepts in STATA: CONTRAST, LINCOM, and MARGINS

# You do know the General Linear Model

- The **general linear model** incorporates many different labels of related single-level analyses under one unifying umbrella term:

| | Categorical Predictors | Continuous Predictors | Both Types of Predictors |
|---|---|---|---|
| Univariate (one outcome) | "ANOVA" | "Regression" | "ANCOVA" |
| Multivariate (2+ outcomes) | "MANOVA" | "Multivariate Regression" | "MANCOVA" |

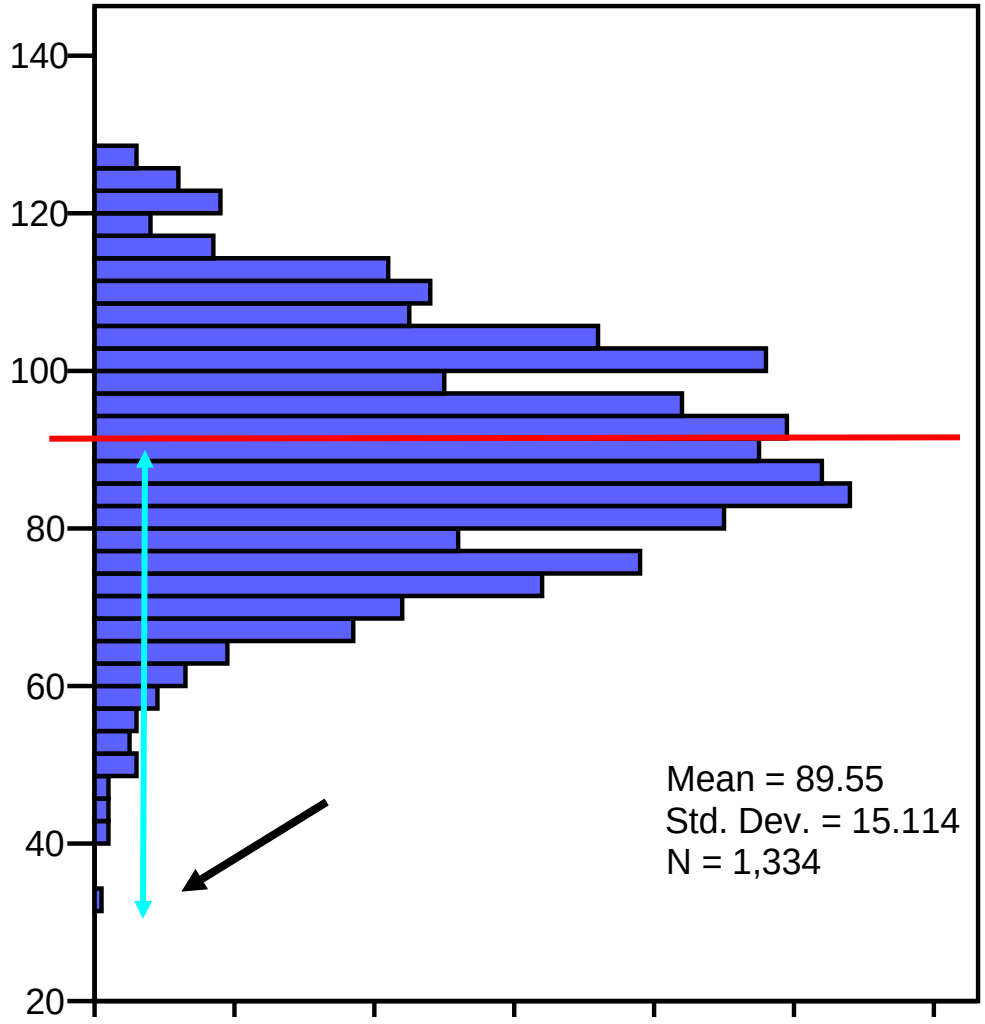
- Actually, these words are not really helpful—they create artificial distinctions among what is really just one kind of model
- What these models all have in common is the use of a **normal conditional distribution** (i.e., for the *residuals* that remain after creating conditional/expected outcomes from the model predictors)
- Note: Model predictors do NOT have distributional assumptions!

# The Two Sides of *Any* Model

- **Model for the Means:**
  - *Aka* **Fixed Effects**, Structural Part of Model
  - What you are used to **caring about for testing hypotheses**
  - How the expected outcome for a given observation varies as a function of its values on the model predictor variables

- **Model for the Variance:**
  - *Aka* **Random Effects and Residuals**, Stochastic Part of Model
  - How residuals are distributed and related across observations
  - What you are used to **making assumptions about** instead…
  - ***For the GLM***, that residuals come from a **normal** distribution, are **independent** across persons, and have **constant variance** across persons and predictors ("identically distributed")

# The Simplest Possible Model: The "Empty Means" GLM




$$y_i = \beta_0 + e_i$$

**Filling in values:**

$$32 = \underbrace{90}_{y_{pred}} + -58$$

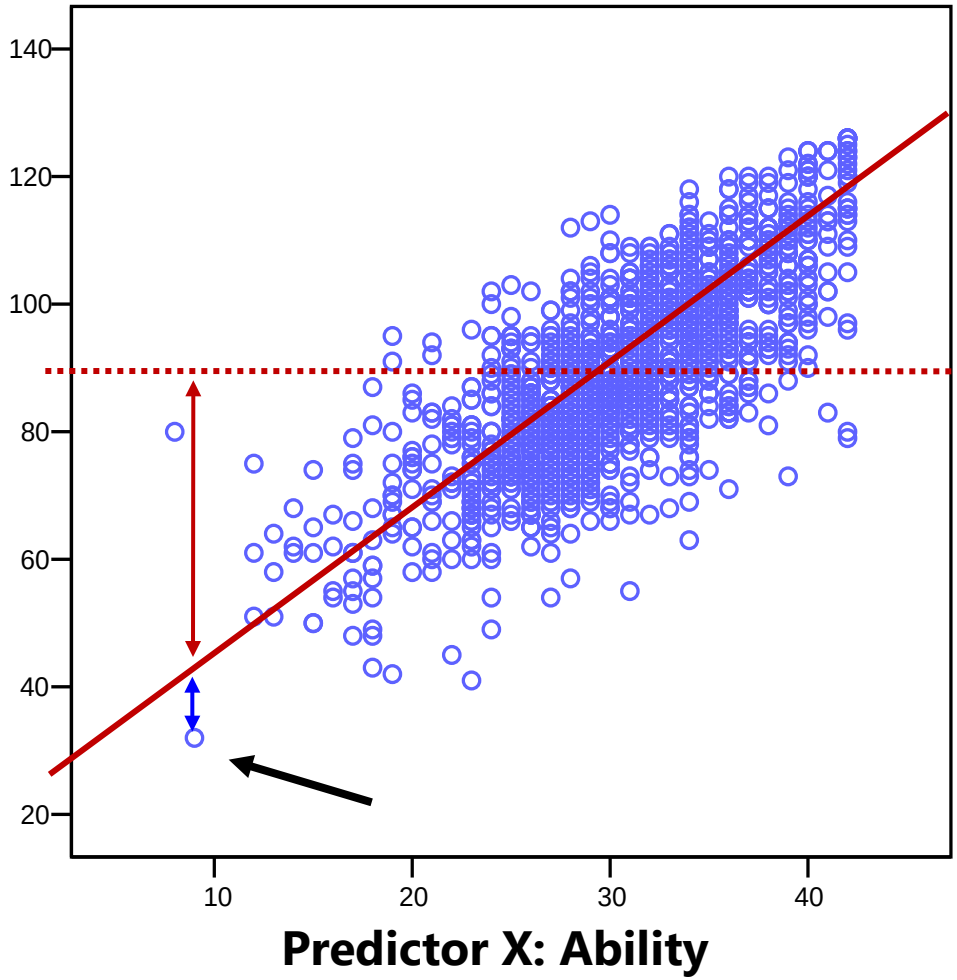
Model for the Means

$y_{pred}$ is also called $\hat{y}$ ("y-hat")

**$y_i$ error variance:**

$$\frac{\sum(y_i - y_{pred})^2}{N - 1}$$

# "Linear Regression" Model with a Continuous Predictor ($X_i$ = Ability)




$$y_i = \beta_0 + \beta_1 X_i + e_i$$

**Empty Model:**

**32 = 90 + −58**

**Ability (X) Model:**

**32 = 29 + 2*9 + −15**

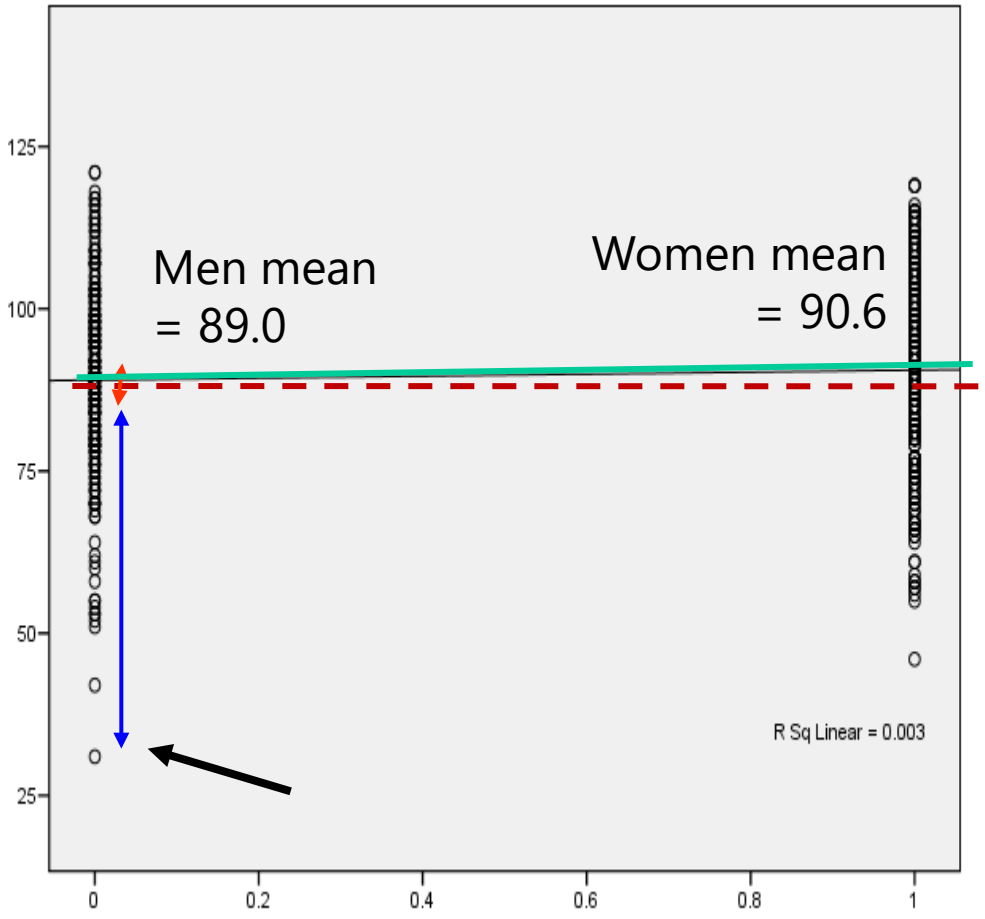
$y_{pred}$

**Model for the Means**

**$y_i$ error variance:**

$$\frac{\sum(y_i - y_{pred})^2}{N - 2}$$

# "ANOVA" Model with a Categorical Predictor ($Z_i$ = Sex)




**Predictor Z: Sex (0=M, 1=W)**

$$y_i = \beta_0 + \beta_2 Z_i + e_i$$

**Empty Model:**

**32 = 90 + −58**

**Sex (Z) Model:**

**32 = 89 + 1.6*0 + −57**

$y_{pred}$

Model for the Means

**$y_i$ error variance:**

$$\frac{\sum(y_i - y_{pred})^2}{N - 2}$$

# The Two Sides of a General Linear Model

$$y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \cdots + e_i$$

Our focus today

- **Model for the Means (Predicted Values):**
  - Each person's expected (predicted) outcome is a weighted linear function of his/her values on $X$ and $Z$ (and any other predictors), each measured once per person (i.e., this is a univariate model)
  - **Estimated parameters are called fixed effects** (here, $\beta_0$, $\beta_1$, and $\beta_2$)
  - The number of fixed effects will show up in formulas as ***k*** (so ***k* = *3*** here)

- **Model for the Variance:**
  - $e_i \sim N(0, \sigma_e^2)$ → ONE source of residual (unexplained) deviation
  - $e_i$ has a mean of 0 with some estimated constant variance $\sigma_e^2$, is normally distributed, is unrelated to $X$ and $Z$, and is unrelated across all observations (which is just one per person here)
  - **Estimated parameter is the residual variance only** (not each $e_i$),
  - Proportion of variance reduced relative to *empty means* model = **$R^2$**

# See? You do know the GLM

- The **general linear model** incorporates many different labels of related single-level analyses under one unifying term:

| | Categorical Predictors | Continuous Predictors | Both Types of Predictors |
|---|---|---|---|
| Univariate (one outcome) | "ANOVA" | "Regression" | "ANCOVA" |
| Multivariate (2+ outcomes) | "MANOVA" | "Multivariate Regression" | "MANCOVA" |

- What these models all have in common is the use of a normal conditional distribution (for the *residuals* that remain after creating conditional outcomes from the model predictors)
- The use of these words **almost always implies estimation using** "**least squares**" (LS), *aka* "**ordinary least squares**" (OLS)

# How Estimation Works (In Brief)

- Most statistical estimation routines do one of three things:

- Minimize Something: Typically found with names that have "least" in the title. Forms of **least squares** include "Generalized", "Ordinary", "Weighted", "Diagonally Weighted", "WLSMV", and "Iteratively Reweighted." **Typically the estimator of last resort**…

- Maximize Something: Typically found with names that have "maximum" in the title. Forms include "**Maximum likelihood**", "ML", "Residual Maximum Likelihood" (REML), "Robust ML". Typically the gold standard of estimators, and what we will use this semester. **REML is the same thing as least squares for complete data.**

- Use Simulation to Sample from Something: more recent advances in simulation use resampling techniques. Names include "**Bayesian** Markov Chain Monte Carlo", "Gibbs Sampling", "Metropolis Hastings", "Metropolis Algorithm", and "Monte Carlo". Used for complex models in which ML is not available or feasible.

# Least Squares (LS) Estimation

| Source | Sum of Squares (SS) | Degrees of Freedom (DF) | Mean Square (MS) | F-ratio |
|---|---|---|---|---|
| Model (from predictor model) | $SS_{model} = \sum(\beta_0 - y_{pred})^2$ | $DF_{num}$ = #fixed effects – 1 (for $\beta_0$) | $MS_{model} = SS_{model} / DF_{num}$ | F-ratio = $MS_{model} / MS_{error}$ |
| Error (from empty model) | $SS_{error} = \sum(y_i - y_{pred})^2$ | $DF_{denom}$ = #people – #fixed effects – 1 (for $\beta_0$) | $MS_{error} = SS_{model} / DF_{num}$ | |
| Total | $SS_{total} = SS_{model} + SS_{error}$ | $DF_{total} = DF_{num} + DF_{denom}$ | | |

- $MS_{model}$ = how much error you reduced per added fixed effect
- $MS_{error}$ = how much error is left, per possible new fixed effect (otherwise known as "residual" or "error" variance)
- Compare F-ratio to critical value given $DF_{num}$ and $DF_{denom}$ to get *p*-value for **model $R^2$** (proportion reduction in error, PRE)

# Least Squares (LS) Estimation

- Uses fixed effect estimates that minimize: $\sum_{i=1}^{N}(e_i^2)$
  - (Sum of squared residuals across persons)
  - Invented c. 1840, can be done via matrix algebra, so it will always work

- Has "closed form" solution (=easy formula) when used for general linear models (GLM) for single outcomes given:
  - $e_i \sim N(0, \sigma_e^2)$ → normal, independent, constant variance ONLY

- For GLM for multiple outcomes, LS quickly becomes useless…
  - Cannot handle missing outcomes (listwise-deletes entire person instead)
  - Only two options for modeling covariance between outcomes
  - Then why do it this way? Dogma + lack of awareness of alternatives…

- For non-normal outcomes, LS is no longer used at all…

# Maximum Likelihood to the Rescue

- Maximum likelihood estimation is the better way of finding the model estimates using all the data, and it comes in 2 flavors:
- **"Residual (or restricted) maximum likelihood"**
  - Only available for general linear models or general linear mixed models (key: based on normally distributed residuals at all levels of analysis)
  - **REML = LS** given complete outcomes, but it doesn't require them
  - Estimates variances the same way as in LS (accurate) → $\frac{\sum(y_i - y_{pred})^2}{N - k}$
- **"Maximum likelihood" (ML; also called FIML*)**
  - Is more general, is available for all of the above, as well as for non-normal outcomes and models with latent variables (CFA/SEM/IRT/DCM)
  - Is NOT the same as LS: it under-estimates variances by not accounting for the # of estimated fixed effects → $\frac{\sum(y_i - y_{pred})^2}{N}$
- **FI = Full information → it uses all original data (they both do)*

# Maximum Likelihood to the Rescue

- Even though REML = LS for complete outcomes, we will begin by using software based in REML instead of LS
  - In SPSS, SAS, or STATA: one routine called **"MIXED"** instead of separate routines for GLM, REGRESSION, or ANOVA (or *t*-tests)
  - So "sums of squares" and "mean squares" are no longer relevant

- Why use MIXED (even when it is the same as LS)?
  - **Big-time convenience**: MIXED has options to produce fixed effects that are model-implied, but not directly given (e.g., pairwise comparisons, simple slopes of interactions)
  - **Model comparisons** (*F*-test for change in $R^2$ from new effects) can be requested for any combinations of fixed effects
  - **Generalizability**: We can estimate **univariate or multivariate** models for normal outcomes using the same MIXED routine
  - For **non-normal outcomes**, there are parallel routines in SAS (GLIMMIX) and STATA (several), but not in SPSS ("pseudo-ML")

# Intermediate Summary

- **What is *not* new:**
  - We will be starting with the same kind of univariate general linear models for single outcomes per person you already know (regression, ANOVA, ANCOVA) …
  - We will practice interpreting main effects and interaction terms among all kinds of predictors (as needed for MLM)

- **What *is* new:**
  - Rather than finding the fixed effects and residual variance through least squares (which yields sums of squares, mean squares, and so forth), the program will find them using residual maximum likelihood, of which least squares is a special case with limited applicability for real data…

# Testing Significance of Fixed Effects (of Predictors) in the Model for the Means

- Any single-df **fixed effect** has 4-5 relevant pieces of output:
  - **Estimate** = best guess for the fixed effect from our data
  - **Standard Error** = index of the precision of fixed effect estimate (i.e., quality of the "most likely" estimate)
  - ***t*-value or *z*-value** = Estimate / Standard Error
  - ***p*-value** = probability that fixed effect estimate is ≠ 0
  - **95% Confidence Interval** = Estimate ± 1.96*SE = range in which true (population) value of estimate is expected to fall 95% of the time

- Compare test statistic (*t* or *z*) to critical value at chosen level of significance (known as alpha): this is a "**univariate Wald test**"
- Whether the *p*-value is based on *t* or *z* varies by program…

# Evaluating Significance of Fixed Effects

Fixed effects can be tested via **Wald** tests: the ratio of its estimate/SE forms a statistic we compare to a distribution

| | Denominator DF is infinite (Proper Wald test) | Denominator DF is estimated instead ("Modified" Wald test) |
|---|---|---|
| Numerator DF = 1 *(test one fixed effect)* is **Univariate Wald Test** | use ***z*** distribution (Mplus, STATA) | use ***t*** distribution (SAS, SPSS) |
| Numerator DF > 1 *(test 2+ fixed effects)* is **Multivariate Wald Test** | use $\boldsymbol{\chi^2}$ distribution (Mplus, STATA) | use ***F*** distribution (SAS, SPSS) |
| Denominator DF options (in Stata use "small", not a default) | not applicable, so DDF is not given | SAS, STATA 14: BW, KR<br>SAS, STATA 14, SPSS: Satterthwaite |

# Standard Errors for Fixed Effects

- Standard Error (SE) for fixed effect estimate $\beta_X$ in a one-predictor model (remember, SE is like the SD of the estimated parameter):

$$SE_{\beta X} = \sqrt{\frac{\text{residual variance of Y}}{\text{variance of X} * (N-k)}}$$

- When more than one predictor is included, SE turns into:

$$SE_{\beta X} = \sqrt{\frac{\text{residual variance of Y}}{Var(X) * (1-R_X^2) * (N-k)}}$$

$R_X^2$ = X variance accounted for by other predictors, so $1-R_X^2$ = unique X variance

- So all things being equal, SE is smaller when:
  - More of the outcome variance has been reduced (better predictive model)
    - This means fixed effects can become significant later if $R^2$ is higher then
  - The predictor has less covariance with other predictors (less collinearity)
    - Best case scenario: X is uncorrelated with all other predictors

- If SE is smaller → *t*-value or *z*-value is bigger→ *p*-value is smaller

# Multivariate Wald Tests of Fixed Effects

- General test for significance of **multiple fixed effects** at once
- Special cases of this you have already seen:
  - "Omnibus" *F*-test for the effect of a grouping variable
  - *F*-Test of Model $R^2$ or change in $R^2$ in hierarchical regression
  - Implies "numerator DF" > 1

- Available for sets of fixed effects via CONTRAST (in SAS or STATA) or TEST statements (in STATA, SPSS, or Mplus)
  - SAS CONTRAST: Separate each fixed effect by commas to indicate separate numerator DF
  - STATA TEST: List fixed effects in separate sets of parentheses to indicate separate numerator DF

# Multivariate Wald Tests of Fixed Effects

- For example with 3 continuous predictors:

$$y_i = \beta_0 + \beta_1(X_{1i}) + \beta_2(X_{2i}) + \beta_3(X_{3i}) + e_i$$

- The df=3 test of the model $R^2$ is given by default in STATA MIXED, but not in SAS
  - Predictors are treated as continuous by default in SAS

```
PROC MIXED DATA=work.dataname METHOD=REML;
     MODEL y = x1 x2 x3 / SOLUTION;
     CONTRAST "Model R2 F-Test with df=3" x1 1, x2 1, x3 1;
RUN;
```

- Here is how you'd ask for it in Stata (c. indicates "continuous"):

```
mixed y c.x1 c.x2 c.x3, ///
    variance reml dfmethod(residual)
      test (c.x1=0) (c.x2=0) (c.x3=0), small
```

# Multivariate Wald Tests of Fixed Effects

- What if you wanted to test the $R^2$ change after including predictors X2 and X3?

$$y_i = \beta_0 + \beta_1(X_{1i}) + \beta_2(X_{2i}) + \beta_3(X_{3i}) + e_i$$

- Here is change in $R^2$ in SAS:

```
PROC MIXED DATA=work.dataname METHOD=REML;
     MODEL y = x1 x2 x3 / SOLUTION;
     CONTRAST "Change in R2 F-Test with df=2" x2 1, x3 1;
RUN;
```

- Here is change in $R^2$ in Stata (NOT provided by default):

```
mixed y c.x1 c.x2 c.x3, ///
    variance reml dfmethod(residual),
      test (c.x2=0) (c.x3=0), small
```

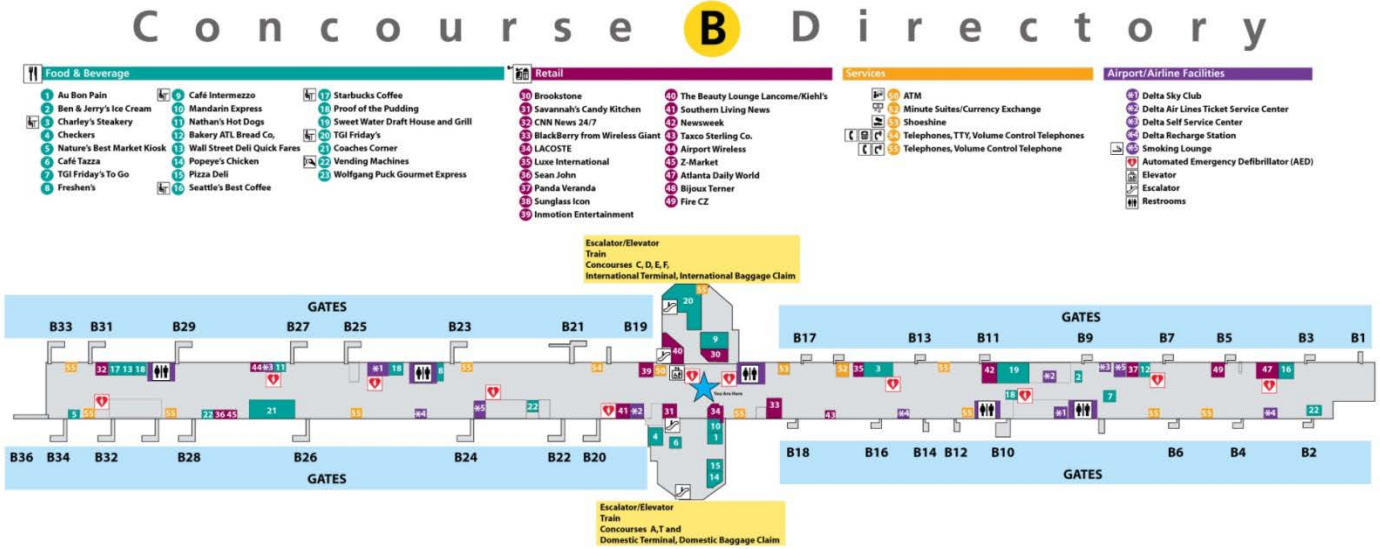
# Specifying the Effects of Predictors

- From now on, we will think carefully about exactly **how** the **predictor variables** are entered into the **model for the means** (i.e., by which a predicted outcome is created for each person)

- Why don't people always care? Because the scale of predictors:
  - Does NOT affect the amount of outcome variance accounted for ($R^2$)
  - Does NOT affect the outcomes values predicted by the model for the means (so long as the same predictor fixed effects are included)

- Why should this matter to us?
  - ***Because the Intercept = expected outcome when all predictors = 0***
  - Can end up with nonsense values for intercept if X = 0 isn't in the data, so we need to change the scale of the predictors to include 0
  - Scaling becomes more important once interactions are included or once random intercepts are included (i.e., variability around fixed intercept)

# Why the Intercept $\beta_0$ *Should* Be Meaningful…




**This is a very detailed map…**
**But what do we need to know**
**to be able to use the map at all?**

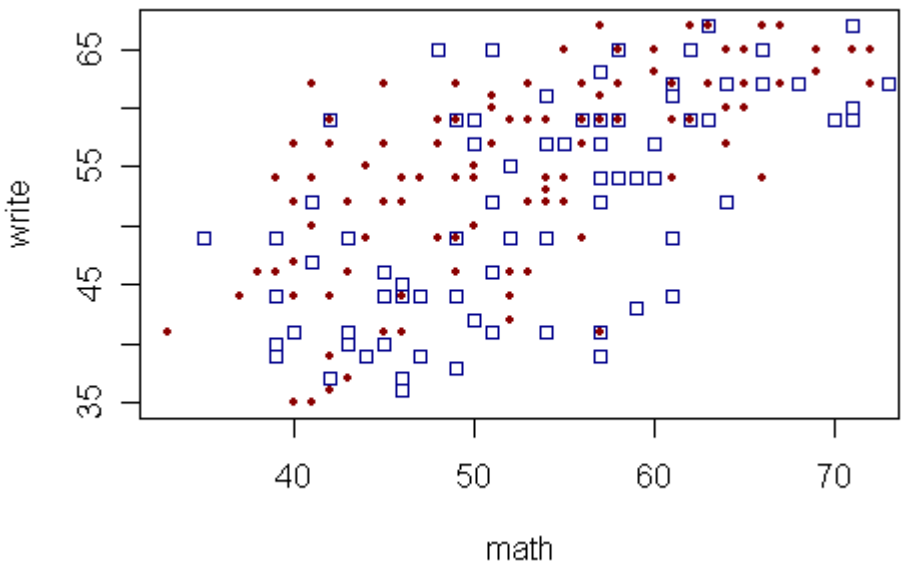
# What the Intercept $\beta_0$ *Should* Mean to You…

The model for the means will describe what happens to the predicted outcome Y

"as X increases" or

"as Z increases"

and so forth…




But you won't know what the predicted outcome is supposed to be unless you know where the predictor variables are starting from!

Therefore, the **intercept** is the "**YOU ARE HERE**" sign in the map of your data… so it should be somewhere in the map*!

* There is no *wrong* way to center (or not), only *weird*…

# What if I want a different intercept?

- Choosing a location for your model-estimated intercept does not lock you into only that reference location…

- **ESTIMATE statements (in SAS) to the rescue!**
  - **TEST in SPSS, LINCOM in STATA, NEW in Mplus**

- These statements allow to you to request model-predicted fixed effects for any values of your predictors
(i.e., new intercept values = conditional/predicted outcomes)

- Rules for ESTIMATE-type statements:
  - If you want a predicted outcome, you MUST include the intercept
  - Variable names sometimes refer to their predictor values, and sometimes to their model fixed effects, depending on what is being estimated
  - The default value for continuous predictors is 0
  - The default value for categorical predictors varies by program
    - Sometimes the mean across groups (SAS), or requires an input explicitly (SPSS)

# Continuous Predictors

- For **continuous** (quantitative) predictors, **we** will make the intercept interpretable by **centering** (so new variable goes in the equation and the model):
  - **Centering** = subtract a constant from each person's variable value so that **the 0 value** falls within the range of the new centered predictor
  - Typical → Center around predictor's mean: $\text{Centered } X_1 = X_1 - \overline{X_1}$
    - Intercept is then expected outcome for "average $X_1$ person"
  - Better → Center around meaningful constant $C$: $\text{Centered } X_1 = X_1 - C$
    - Intercept is then expected outcome for person with that constant (even 0 may be ok)

- These statements can be used to request predicted outcomes (i.e., intercepts) for specific combinations of predictor values
  - SAS ESTIMATE: Must write a separate statement per prediction
  - STATA MARGINS: Can specify a range of predictor values

# Continuous Predictors

- For example: $y_i = \beta_0 + \beta_1(X_{1i} - 10) + \beta_2(X_{2i} - 5) + e_i$
- In SAS, requires separate statement per prediction

```
PROC MIXED DATA=work.dataname METHOD=REML;
  MODEL y = x1c x2c / SOLUTION;
  CONTRAST "Model R2 F-Test with df=2" x1c 1, x2c 1;
  ESTIMATE "Pred Y if X1=10, X2=5" intercept 1 x1c  0 x2c  0;
  ESTIMATE "Pred Y if X1= 8, X2=4" intercept 1 x1c -2 x2c -1;
  ESTIMATE "Pred Y if X1= 8, X2=6" intercept 1 x1c -2 x2c  1;
  ESTIMATE "Pred Y if X1=12, X2=4" intercept 1 x1c  2 x2c -1;
  ESTIMATE "Pred Y if X1=12, X2=6" intercept 1 x1c  2 x2c  1;
  RUN;
```

- In STATA, margins at(variable=(start(by)end)

```
mixed y c.x1c c.x2c, ///
      variance reml dfmethod(residual),
      test (c.x2=0) (c.x3=0), small
      margins, at(xlc=(-2(2)2) x2c(-1(1)1)) vsquish
```

# Categorical (Grouping) Predictors

- For **categorical** predictors, **<u>either we or the program</u>** will make the intercept interpretable by making a **reference group**
  - Which is more convenient depends on what's in the rest of the model
- **If you create your own reference group via the following, the program treats the new predictor variables as "continuous" even if they represent group differences!**
  - **To do it yourself: Denote a reference group** by giving it a 0 value on all predictor variables created from the original grouping variable, then $\beta_0$ = expected outcome for that reference group specifically
  - Accomplished via "dummy coding" (aka, "reference group coding")
    → Two-group example using *Gender*: 0 = Men, 1 = Women
    (or 0 = Women, 1 = Men)
  - Alternative approach I usually do not like to use:
    → "Contrast/effect coding"→ *Gender*: −0.5 = Men, 0.5 = Women

# Categorical Predictors Modeled as Continuous

- For 2+groups, we need: ***dummy codes = #groups – 1***
  - "Group" variable: Control=0, Treat1=1, Treat2=2, Treat3=3
  - Variables: *d1*= 0, 1, 0, 0 → difference between Control and T1
    *d2*= 0, 0, 1, 0 → difference between Control and T2
    *d3*= 0, 0, 0, 1 → difference between Control and T3
- **d1, d2, and d3 are then continuous variables** as far as the program is concerned, which implies the following:
  - **All predictors** that distinguish the groups (e.g., d1, d2, d3) MUST be in the model to get these specific group-difference interpretations!
    - e.g., `MODEL y = d1` → d1 = difference between T1 and mean of C,T2,T3
  - Fixed effects for these dummy codes will not *directly* tell you about differences among non-reference groups…
    - e.g., `MODEL y = d1 d2 d3` → won't give differences among T1,T2,T3
    - … But you can still get them: **ESTIMATE** statements to the rescue!

# Categorical Predictors Modeled as Continuous

- Model: $y_i = \beta_0 + \beta_1 d1_i + \beta_2 d2_i + \beta_3 d3_i + e_i$
  - "Group" variable: Control=0, Treat1=1, Treat2=2, Treat3=3
  - New variables to be created for the model: *d1*= 0, 1, 0, 0 → difference between Control and T1; *d2*= 0, 0, 1, 0 → difference between Control and T2; *d3*= 0, 0, 0, 1 → difference between Control and T3
- How does the model give us **all possible group differences**? By determining each group's mean, and then the difference…

| Control Mean (Reference) | Treatment 1 Mean | Treatment 2 Mean | Treatment 3 Mean |
|---|---|---|---|
| $\beta_0$ | $\beta_0 + \beta_1 d1_i$ | $\beta_0 + \beta_2 d2_i$ | $\beta_0 + \beta_3 d3_i$ |

- The model for the 4 groups directly provides 3 differences (control vs. each treatment), and indirectly provides the other 3 differences (differences between non-reference treatments)

# Categorical Predictors Modeled as Continuous

- Model: $y_i = \beta_0 + \beta_1 d1_i + \beta_2 d2_i + \beta_3 d3_i + e_i$

| Control Mean (Reference) | Treatment 1 Mean | Treatment 2 Mean | Treatment 3 Mean |
|---|---|---|---|
| $\boldsymbol{\beta_0}$ | $\boldsymbol{\beta_0 + \beta_1 d1_i}$ | $\boldsymbol{\beta_0 + \beta_2 d2_i}$ | $\boldsymbol{\beta_0 + \beta_3 d3_i}$ |

| | Alt Group | Ref Group | Difference |
|---|---|---|---|
| • Control vs. T1 = | $(\beta_0+\beta_1)$ | $-\ (\beta_0)$ | $= \beta_1$ |
| • Control vs. T2 = | $(\beta_0+\beta_2)$ | $-\ (\beta_0)$ | $= \beta_2$ |
| • Control vs. T3 = | $(\beta_0+\beta_3)$ | $-\ (\beta_0)$ | $= \beta_3$ |
| • T1 vs. T2 = | $(\beta_0+\beta_2)$ | $-\ (\beta_0+\beta_1)$ | $= \beta_2 - \beta_1$ |
| • T1 vs. T3 = | $(\beta_0+\beta_3)$ | $-\ (\beta_0+\beta_1)$ | $= \beta_3 - \beta_1$ |
| • T2 vs. T3 = | $(\beta_0+\beta_3)$ | $-\ (\beta_0+\beta_2)$ | $= \beta_3 - \beta_2$ |

# ESTIMATEs with manual dummy codes

| | Alt Group | Ref Group | Difference |
|---|---|---|---|
| • Control vs. T1 = | $(\beta_0+\beta_1)$ | $-\ (\beta_0)$ | $= \beta_1$ |
| • Control vs. T2 = | $(\beta_0+\beta_2)$ | $-\ (\beta_0)$ | $= \beta_2$ |
| • Control vs. T3 = | $(\beta_0+\beta_3)$ | $-\ (\beta_0)$ | $= \beta_3$ |
| • T1 vs. T2 = | $(\beta_0+\beta_2)$ | $-\ (\beta_0+\beta_1)$ | $= \beta_2 - \beta_1$ |
| • T1 vs. T3 = | $(\beta_0+\beta_3)$ | $-\ (\beta_0+\beta_1)$ | $= \beta_3 - \beta_1$ |
| • T2 vs. T3 = | $(\beta_0+\beta_3)$ | $-\ (\beta_0+\beta_2)$ | $= \beta_3 - \beta_2$ |

Note the order of the equations: the reference group mean *is subtracted from* the alternative group mean.

In SAS ESTIMATE statements (or SPSS TEST or STATA LINCOM), the variables refer to their fixed effects; the numbers refer to the operations of their fixed effects.

```
TITLE "Manual Contrasts for 4-Group Diffs";
PROC MIXED DATA=work.dataname METHOD=REML;
MODEL y = d1 d2 d3 / SOLUTION;
CONTRAST "Omnibus F-test df=3 group main effect" d1 1, d2 1, d3 1;
ESTIMATE "Control Mean" intercept 1 d1 0 d2 0 d3 0;
ESTIMATE "T1 Mean"      intercept 1 d1 1 d2 0 d3 0;
ESTIMATE "T2 Mean"      intercept 1 d1 0 d2 1 d3 0;
ESTIMATE "T3 Mean"      intercept 1 d1 0 d2 0 d3 1;

ESTIMATE "Control vs. T1"     d1  1 d2  0 d3 0;
ESTIMATE "Control vs. T2"     d1  0 d2  1 d3 0;
ESTIMATE "Control vs. T3"     d1  0 d2  0 d3 1;
ESTIMATE "T1 vs. T2"          d1 -1 d2  1 d3 0;
ESTIMATE "T1 vs. T3"          d1 -1 d2  0 d3 1;
ESTIMATE "T2 vs. T3"          d1  0 d2 -1 d3 1;  RUN;
```

Intercepts are used only in predicted outcomes.

Positive values indicate addition; negative values indicate subtraction.

# LINCOMs with manual dummy codes

| | Alt Group | Ref Group | Difference |
|---|---|---|---|
| • Control vs. T1 = | $(\beta_0+\beta_1)$ | $-\ (\beta_0)$ | $= \beta_1$ |
| • Control vs. T2 = | $(\beta_0+\beta_2)$ | $-\ (\beta_0)$ | $= \beta_2$ |
| • Control vs. T3 = | $(\beta_0+\beta_3)$ | $-\ (\beta_0)$ | $= \beta_3$ |
| • T1 vs. T2 = | $(\beta_0+\beta_2)$ | $-\ (\beta_0+\beta_1)$ | $= \beta_2 - \beta_1$ |
| • T1 vs. T3 = | $(\beta_0+\beta_3)$ | $-\ (\beta_0+\beta_1)$ | $= \beta_3 - \beta_1$ |
| • T2 vs. T3 = | $(\beta_0+\beta_3)$ | $-\ (\beta_0+\beta_2)$ | $= \beta_3 - \beta_2$ |

Note the order of the equations: the reference group mean *is subtracted from* the alternative group mean.

In SAS ESTIMATE statements (or SPSS TEST or STATA LINCOM), the variables refer to their fixed effects; the numbers refer to the operations of their fixed effects.

```
display as result "Manual Contrasts for 4-Group Diffs"
mixed y c.d1 c.d2 c.d3, /// variance reml dfmethod(residual),
  test (c.d1=0) (c.d2=0) (c.d3=0), small // Omnibus F-test df=3 group main effect
lincom _cons*1 + c.d1*0 + c.d2*0 + c.d3*0, small // Control Mean
lincom _cons*1 + c.d1*1 + c.d2*0 + c.d3*0, small // T1 Mean
lincom _cons*1 + c.d1*0 + c.d2*1 + c.d3*0, small // T2 Mean
lincom _cons*1 + c.d1*0 + c.d2*0 + c.d3*1, small // T3 Mean
lincom           c.d1*1 + c.d2*0 + c.d3*0, small // Control vs T1
lincom           c.d1*0 + c.d2*1 + c.d3*0, small // Control vs T2
lincom           c.d1*0 + c.d2*0 + c.d3*1, small // Control vs T3
lincom           c.d1*-1 + c.d2*1  + c.d3*0, small // T1 vs T2
lincom           c.d1*-1 + c.d2*0  + c.d3*1, small // T1 vs T3
lincom           c.d1*0  + c.d2*-1 + c.d3*1, small // T2 vs T3
```

# Using BY/CLASS/i. statements instead

- Designate as "**categorical**" predictor in program syntax
  - If you let **SAS**/SPSS do the dummy coding via **CLASS**/BY, then the **highest/last group is default reference**
    - In SAS 9.4 you can change reference group: REF='level' | FIRST | LAST but it changes that group to be last in the data (→ confusing)
    - "Type III test of fixed effects" provide omnibus tests by default
    - **LSMEANS**/EMMEANS can be used to get all means and comparisons without specifying each individual contrast
  - If you let STATA do the dummy coding via i.group, then the **lowest/first group is default reference**
    - Can change reference group, e.g., last = ref → ib(last).group
    - CONTRAST used to get omnibus tests (not provided by default)
    - MARGINS can be used to get all means and comparisons with much less code than describing each individual contrast
  - No such thing as "categorical" predictors in Mplus ☹
    - You must create contrasts manually for all grouping variables

# SAS Main effects of **Categorical** Predictors

```
TITLE "Program-Created Contrasts for 4-Group Diffs via CLASS";
PROC MIXED DATA=work.dataname METHOD=REML;
CLASS group;
MODEL y = group / SOLUTION;
LSMEANS group / DIFF=ALL;
```

**CLASS** statement means "make my dummy codes for me"

**The LSMEANS line above gives you ALL of the following… note that one value has to be given for each possible level of the categorical predictor in *data* order**

```
ESTIMATE "Control Mean" intercept 1 group 1 0 0 0;
ESTIMATE "T1 Mean"      intercept 1 group 0 1 0 0;
ESTIMATE "T2 Mean"      intercept 1 group 0 0 1 0;
ESTIMATE "T3 Mean"      intercept 1 group 0 0 0 1;

ESTIMATE "Control vs. T1"  group -1  1  0  0;
ESTIMATE "Control vs. T2"  group -1  0  1  0;
ESTIMATE "Control vs. T3"  group -1  0  0  1;
ESTIMATE "T1 vs. T2"       group  0 -1  1  0;
ESTIMATE "T1 vs. T3"       group  0 -1  0  1;
ESTIMATE "T2 vs. T3"       group  0  0 -1  1;

CONTRAST "Omnibus df=3 main effect F-test" group -1 1 0 0,
                                           group -1 0 1 0,
                                           group -1 0 0 1;
```

When predicting intercepts, 1 means "for that group only"

When predicting group differences, contrasts must sum to 0; here −1 = ref, 1 = alt, and 0 = ignore

**CLASS also gives this contrast by default**

**Can also make up whatever contrasts you feel like using DIVISOR option:**

```
ESTIMATE "Mean of Treat groups" intercept 1 group  0 1 1 1 / DIVISOR=3;
ESTIMATE "Control vs. Mean of Treat groups" group -3 1 1 1 / DIVISOR=3;
RUN;
```

# STATA Main effects of **Categorical** Predictors

```
display as result "Program-Created Contrasts for 4-Group Diffs"
display as result "i. means make my dummy codes for me (factor var)"
mixed y ib(last).group, /// variance reml dfmethod(residual),
contrast i.group, small // Omnibus F-test
margins i.group, pwcompare(pveffects) df(#)// Means per group and mean diffs
```

**The MARGINS line above gives you ALL of the following… note that one value has to be given for each possible level of the categorical predictor in *data* order**

```
lincom _cons*1 + 1.group*1 + 2.group*0 + 3.group*0 + 4.group*0, small // Control Mean
lincom _cons*1 + 1.group*0 + 2.group*1 + 3.group*0 + 4.group*0, small // T1 Mean
lincom _cons*1 + 1.group*0 + 2.group*0 + 3.group*1 + 4.group*0, small // T2 Mean
lincom _cons*1 + 1.group*0 + 2.group*0 + 3.group*3 + 4.group*1, small // T3 Mean
lincom           1.group*-1 + 2.group*1  + 3.group*0  + 4.group*0, small // Control vs T1
lincom           1.group*-1 + 2.group*0  + 3.group*1  + 4.group*0, small // Control vs T2
lincom           1.group*-1 + 2.group*0  + 3.group*0  + 4.group*1, small // Control vs T3
lincom           1.group*0  + 2.group*-1 + 3.group*1  + 4.group*0, small // T1 vs T2
lincom           1.group*0  + 2.group*-1 + 3.group*0  + 4.group*1, small // T1 vs T3
lincom           1.group*0  + 2.group*0  + 3.group*-1 + 4.group*1, small // T2 vs T3
```

**Can also make up whatever contrasts you feel like (no DIVISOR option?) :**

```
lincom _cons*1 + 1.group*0  + 2.group*.33 + 3.group*.33 + 4.group*.34, small // Mean of Treat
lincom           1.group*-1 + 2.group*.33 + 3.group*.33 + 4.group*.33, small // Cont v Treat
```

# Summary

- Today was about fixed effects in the model for the means
  - Within the context of GLM as a unifying starting point, but these same concepts will readily apply to MLM, SEM, etc
  - Output will result from (residual) ML instead of least squares

- Key points to take with you:
  - Fixed effects are tested for significance using univariate or multivariate Wald tests (*t*- or *z*-value from ratio of estimate / SE)
  - All predictors should always have a meaningful 0 value (adjusting predictor scales is called "centering" or "recoding")
  - When you manually code one categorical predictor variable into separate dummy coded variables, they are then "continuous"
  - You can write code to request *model-implied* fixed effects (such as predicted outcomes or non-reference group differences)