Introduction to the General Linear Model

GradQuant Workshop

Fundamental Issues of the GLM

- Three basic questions asked:
  - Is there a relationship between variables?
  - What direction is this relationship?
  - What is the size of this relationship
- Modeling the data
- Assessment of error
- Model comparisons

Terminology of the GLM

- “General” refers to the many tests encompassed by GLM
- Our Y variable is the outcome, predicted, or dependent variable
- Our X variable(s) is the regressor, predictor, or covariate
- More loose terms
  - Typically called regression with continuous predictors
  - ANOVA with categorical predictors
  - ANCOVA with at least one of each
  - But really, they’re the same thing

Predicting scores

- Data = Model + error
- Modeling begins with a very simple value- $Y = \bar{Y} + e_i$
- Model fit is judged according to the ordinary least squares estimation
  - $\frac{\sum(\check{Y}-\bar{Y})^2}{N}$ = variance in the residuals
- Relationship between accuracy, correlation, and residuals
- This model is used for most common statistical techniques

What is the best predictor?

- Imagine we had no predictor variables…what would our best guess be?
- $\frac{\sum(\check{Y}-\bar{Y})^2}{N}$ is at a minimum when using the mean
- When using predictors we would use the “conditional” mean
- With perfect prediction, observed and expected values are the same- $\frac{\sum(\check{Y}-\bar{Y})^2}{N}$ is zero

Bivariate regression

- Regression equation forms basis of the GLM
- Variables can be included to reduce the residual error
- $Y_i = b_0 + b_1 X_{i1} + e_i$
- b_0 represents the expected values when X = 0
- b_1 is the expected change in Y for a one unit change in Y
- e_i = the error after taking model prediction into account
- This regression equation represents the best fit line

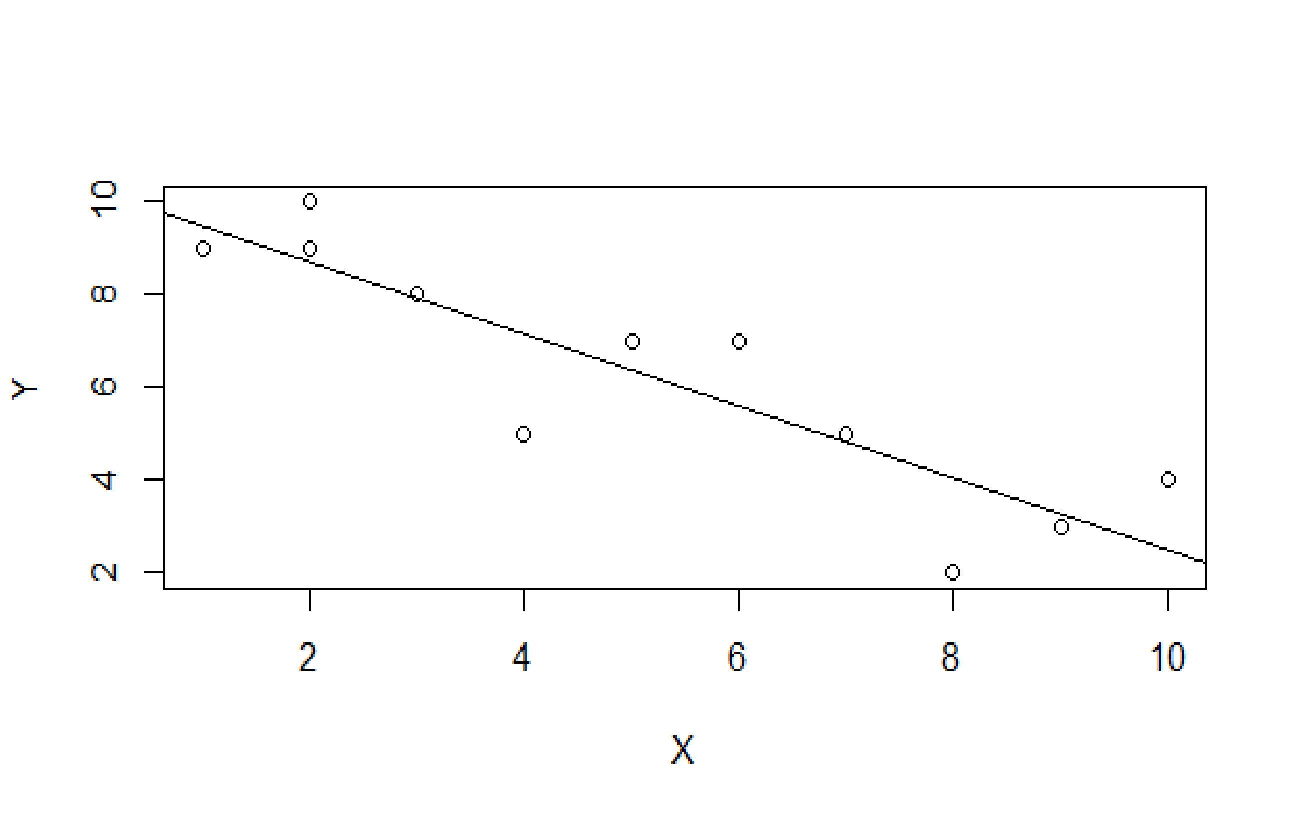
The best fit line

Data

X	Y
8	2
9	3
9	3
10	4
6	7
7	5
4	5
5	7
3	8
1	9
2	9
2	10

The model and graph

$$\check{Y} = 10.27 - .78X$$



Correlation

- Measure of linear association between X and Y, addresses the three questions of the GLM
- Regression parameters can be used to calculate correlation
  - $r_{xy} = b_1 \frac{s_y}{s_x}$
- Standardized regression equation:
  - $\breve{Z}_y = rZ_x$
- Correlation of previous data is *r* = -.90
  - $\breve{Z}_y = -.90Z_x$
- With one predictor, r is also equal to correlation between predicted and observed Y's.

Variance explained

- We can make one modification to our model
  - $Var(data) = Var(model) + Var(error)$
- Our model will tell us the proportion of variance explained
  - $R^2 = 1 - \frac{Var(error)}{Var\ (data)}$
  - $r^2 =$ .80
- This is applied to the multivariate case, and used to evaluate overall model fit

Traditional t-test

- A more specialized form of the regression
- $r = \sqrt{\frac{t^2}{t^2+df}}$
- Equivalent to $Y_i = b_0 + b_1 X_{i1} + e_i$
  - Where b_1 is equal the mean difference between groups
  - e_i is the within group variations
- The goal of a t-test is the same goal as that of OLS regression
- All information from a t-test can be gained from regression and vice versa

3 types of t tests

One sample t-test	Test whether the population mean is different from a constant 1 distribution
Paired Samples t-test	Test whether the population mean of differences between paired scores is equal to 0 2 distributions Correlation/relationship exists
Independent Samples t-test	Test the relationship between 2 categories and a quantitative variables 2 distributions NO relationship exists

Sample Data

X	Y
0	3.0
0	2.0
0	1.0
0	2.0
0	3.0
0	4.0
0	4.0
0	5.0
1	3.0
1	2.0
1	3.0
1	7.0
1	8.0
1	6.9
1	10.0
1	11.0
1	9.0

```
t.test(Y~X,data)
Welch Two Sample t-test data: Y by X
t = -3.215, df = 10.347, p-value = 0.008871
0.95 percent confidence interval: -6.365351 -1.167982
```

```
Call: lm(formula = Y ~ X, data = data)
Coefficients:
        Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.8889 0.8284 3.487 0.00305 **
X            3.7667 1.1716 3.215 0.008871
Residual standard error: 2.485 on 16 degrees of freedom
Multiple R-squared: 0.3925, Adjusted R-squared: 0.3545
F-statistic: 10.34 on 1 and 16 DF, p-value: 0.008871
```

Effect Size

- Independent Sample Equation – Use Total N

$$d = \frac{\overline{X}_1 - \overline{X}_2}{\hat{S}} \qquad \hat{S} = \sqrt{\frac{N_1}{2}}\left(S_{\overline{X}_1 - \overline{X}_2}\right)$$

- Paired Sample Equation – N is number of pairs

$$d = \frac{\overline{X} - \overline{Y}}{\hat{S}_D} \qquad \hat{S}_D = \sqrt{N}\left(S_{\overline{D}}\right)$$

We also get an correlation!

$$r = \sqrt{\frac{t^2}{t^2 + df}}$$

Confidence Intervals

- Independent Samples Equation

$$LL = \left(\overline{X}_1 - \overline{X}_2\right) - t_\alpha\left(S_{\overline{X}_1 - \overline{X}_2}\right)$$

$$UL = \left(\overline{X}_1 - \overline{X}_2\right) + t_\alpha\left(S_{\overline{X}_1 - \overline{X}_2}\right)$$

- Paired Samples Equation

$$LL = \left(\overline{X} - \overline{Y}\right) - t_\alpha\left(S_{\overline{D}}\right)$$

$$UL = \left(\overline{X} - \overline{Y}\right) + t_\alpha\left(S_{\overline{D}}\right)$$

ANOVA

- ANOVA is another specific form of regression
- Assesses the relationship between outcome and multiple categories
- Capable of doing everything a t-test can do, $F = t^2$ with 1 df in numerator
- Most parts of ANOVA have direct analogs in regression
- The $\eta^2 = \left(\frac{SS_{model}}{SS_{total}}\right)$ statistic used in ANOVA is the value of R^2

Hypothesis testing with ANOVA

T test	ANOVA
Research question: the effect of Drug X on depression - Give 1 group a dosage of drug X and another gets zero dosage	Research question: the effect of Drug X on depression - You give 1 group high dosage of Drug X, a 2nd group low dosage, and a 3rd group gets zero dosage
State IVs and DVs	State IVs and DVs
State hypotheses	State your hypothesis
Calculate t statistic	Calculate F ratio
Compare to sampling distribution for t	Compare to sampling distribution for F
Reject of retain H0	Reject or retain H0
	Follow up multiple comparison test

Sample data from ACT and education

Education level	Mean	SD
Less than high school	27.48	5.21
High school	27.49	6.06
Some college	26.98	5.81
Completed college	28.29	4.85
Some graduate work	29.26	4.35
Completed graduate degree	29.60	3.95

ANOVA vs. Regression

```
Call: lm(formula = ACT ~ as.factor(education))
Residuals: Min 1Q Median 3Q Max
-23.9773 -3.2945 0.5263 3.7055 9.0227
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)   27.4737 0.6319       43.480  < 2e-16 ***
Education1     0.0152  0.9513       0.016   0.98725
education2    -0.4964  0.9573      -0.519   0.60425
Education3     0.8209  0.6943       1.182   0.23748
education4     1.7872  0.7511       2.379   0.01761 *
education5     2.1292  0.7488       2.844   0.00459 **
Residual standard error: 4.771 on 694 degrees of freedom
Multiple R-squared: 0.02887, Adjusted R-squared: 0.02187
F-statistic: 4.126 on 5 and 694 DF, p-value: 0.001063
```

```
summary(aov(ACT~as.factor(education)))
           Df  SumSq     MeanSq      F value Pr(>F)
           5    470      93.90   4.126 0.00106 **
Residuals 694 15794 22.76
```

$$\eta^2 = \left(\frac{SS_{model}}{SS_{total}}\right) = \left(\frac{470}{470+15794}\right) = .0288$$

What have we seen so far?

- Models that look like competitors really are not
- Even comparisons of means are using OLS
- Better models are those that reduce residual error
- Effect sizes are analogous across different methods as well

Comparing models

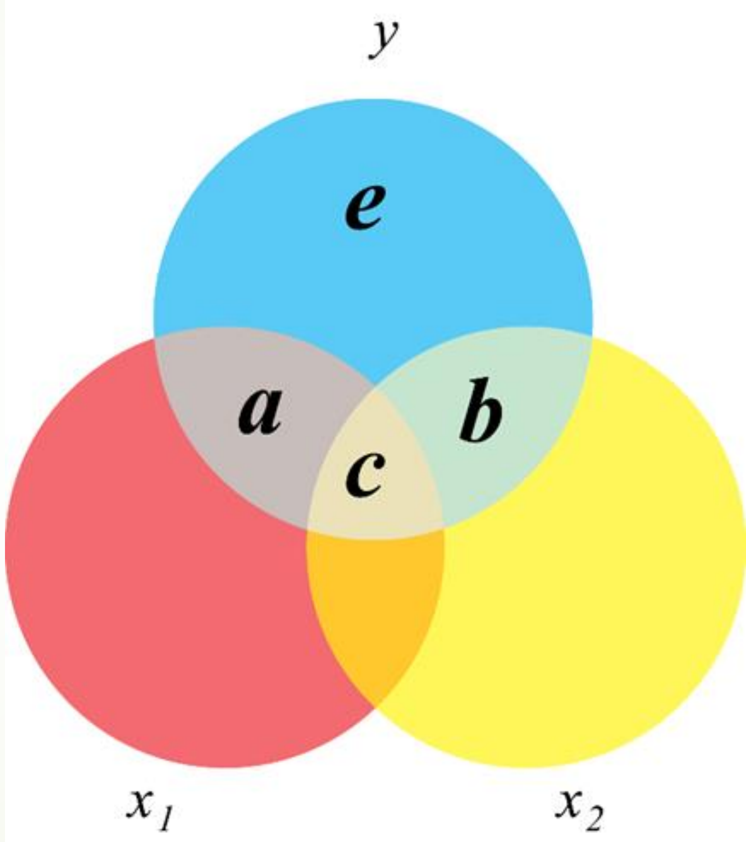
- Remember:
  - Data = Model + error
- The goal of adding predictors should be to reduce the error
  - $\Delta R^2 = R^2_{model\ 2}$- $R^2_{model\ 1}$
- If additional predictors reduce error, they should be included
- Parsimonious models should be preferred

Multiple Regression

- $Y_i = b_0 + b_1X_{i1} + b_2X_{i2} + \ldots b_kX_{ik} + e_i$
- The regression equation has no limit on predictors
- Each of these coefficients represents partial coefficients
- b_0 is now the predicted value when all X's are zero
- Can build models simultaneously or hierarchically

Partial coefficients

- We are often interested in knowing *partial* relationships
- Tells us unique relationship or contribution
- Necessary for making causal inferences
- Several different measures of partial coefficients

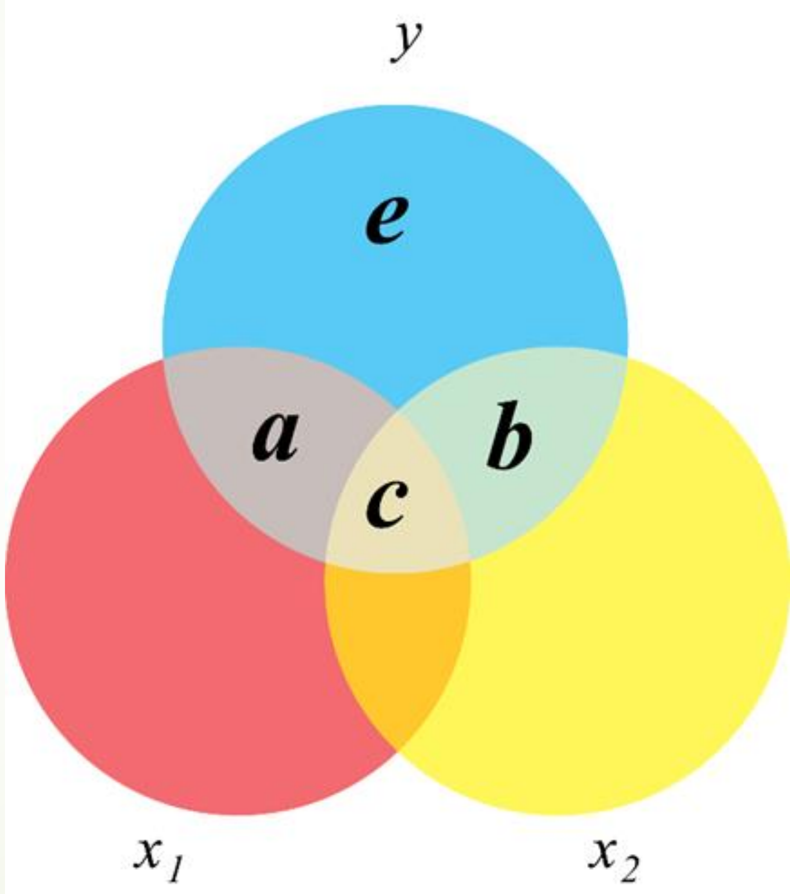


Partial coefficients

$$\frac{B+C}{A+C+E+B}=r_{y2}^{2}$$

$$\frac{B}{A+C+E+B}=r_{y(2.1)}^{2}=sr_{2}^{2}$$

$$\frac{B}{B+E}=r_{y2.1}^{2}=pr_{2}^{2}$$



Standardized Regression

- Often our units don't have substantive meaning
- We can z-score our variables to give more meaning
- Standardized slopes include a special meaning
- Denoted as β
- Inferences and model fit will remain the same as unstandardized

More multiple regression

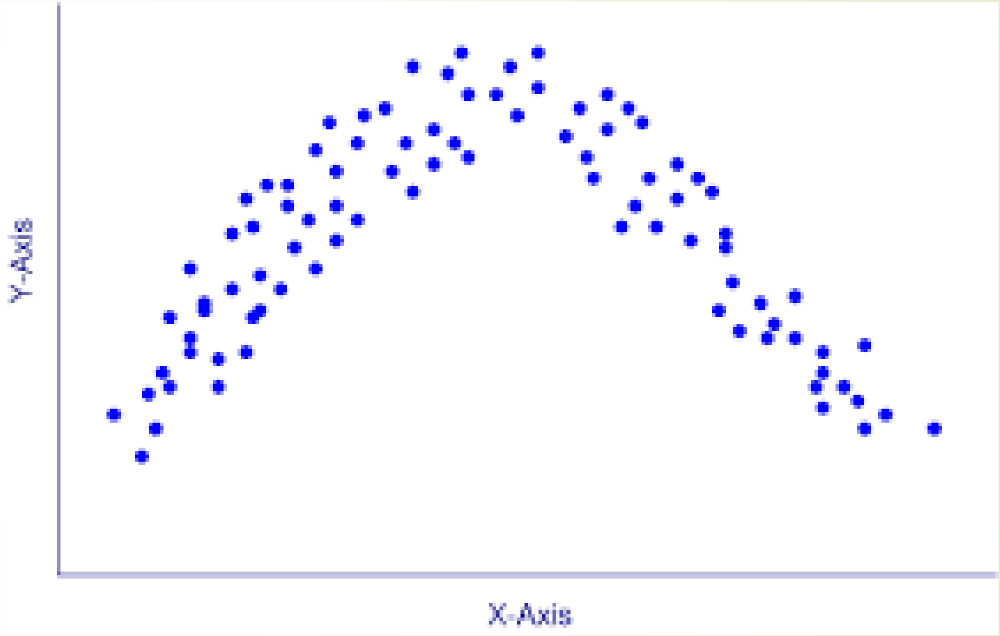
- No statistical difference between covariate and predictor in regression
- Predictors can be either continuous or categorical
- Typically $r > \beta$
- But $r < \beta$ can happen
- Can proceed hierarchically or simultaneously, depending on research question

Modeling Interactions

- Typically modeled as a product of predictors
- $Y_i = b_0 + b_1X_{i1} + b_2X_{i2} + b_3X_{i1}X_{i2} \ldots b_kX_{ik} + e_i$
- Indicate the extent to which the effect of one variable relies on another variable
- Positive sign represents synergy, negative represents dampening

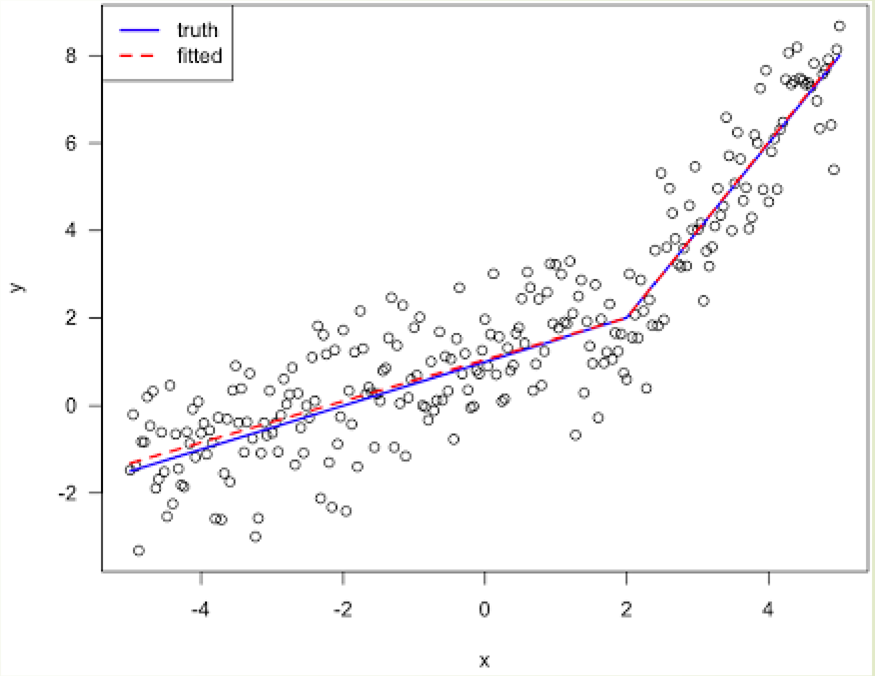
The GLM can handle non-linearity

- If need a non-linear model, we add one more term
- $Y_i = b_0 + b_1 X_{i1} + b_2 X_{i1}^2 + e_i$
- Can easily interpret sign of b_2
- Important to center
- Parameters also become more interpretable with centered variables



Nonlinearity may not be a curve

- Sometimes data fit two curves together
- $\check{Y}$ = -2 + .5X When $x \leq 2$
- $\check{Y}$ = -2 + 1.5X When $x > 2$
- Need to be careful about overfitting



Analysis of Covariance

- Often used when dealing with continuous and categorical predictors
- Long history of figuring out effect of condition at constant levels of other variable– HA!
- Begin by adjusting outcome based on level of covariate (continuous)
- Test association with remaining categorical variable
- Reduces error variance and clarifies relationship
- Regression doesn't care, all variables are welcome

Regression to the mean

- Extreme scores on X are associated with less extreme scores on Y
- This doesn't mean there is less variability
- Occurs whenever $r < 1.0$
- Can deceive us into thinking effects exist when they really don't
- Indicates the importance of controlling for a previous time point

Assumptions of the GLM

- Normality of residuals
- Outcome must be continuous
- Independence of observations
- Homoscedasticity
- No measurement error

Normality of residuals

- Likely indicates misspecified model
- Curve might be more appropriate than a line
- Could mean violation of our second assumption
- Best course is to figure source of non-normality

Discrete outcomes

- Entire family of models for outcomes that are not continuous
  - Logistic regression
  - Multinomial logistic regression
  - Ordinal logistic regression
  - Poisson regression or negative binomial for counts
- All rely on maximum likelihood estimation
- Often have the other assumptions as well

Nonindependence

- Disaggregate variables (known as the atomistic fallacy)
- Aggregate up to the group level (ecological fallacy)
- Two stage least squares
- Cluster robust standard errors
- Multilevel models

Heteroscedasticity

- Often a byproduct of violations
- Check for subgroup differences
- Transform variables
- Adjustment of the standard errors
- Weighted least squares
- But really, problem is not that large

Measurement error

- In the bivariate case, will attenuate relationships
- In the multivariate case ??????????
- Could correct for unreliability (but need proper reliability estimates!)
- Could always try to get more reliable measures
- Latent variable modeling will correct this issue

Orthogonality

- ANOVA assumes uncorrelated factors
  - Also model must be balanced
- If predictors are correlated, model is not orthogonal
- Regression easily handles correlate X's
- If unbalanced, or factors are correlated, then advantages of regression become more pronounced

Coding schemes for regression

- Dummy coding
- Effects coding
- Contrast coding

Dummy coding

- Require the use of a "reference group"
- Reference group = 0, all others = 1
- G -1 variables are needed
- Intercept is the mean of reference group
- Slopes represent means of other groups

Example Dummy coding

- $\check{Y} = b_O + b_1 Dog + b_2 Cat$
- Because bird is zero, b_O is the mean for birds
- The equation for Dog:
- $Mean\ Dog = b_O + b_1(1) + b_2(0)$
- The equation for Cat:
- $Mean\ Cat = b_O + b_1(0) + b_2(1)$
- b's represent mean differences from the bird group

Variable	X_1	X_2
Dog	1	0
Cat	0	1
Bird	0	0

Effects Coding

- Requires a "throw away" group
- This group = -1, all others 1
- Still need G -1 variables
- Intercept and slopes change meaning
- Closest to what ANOVA is doing
- Most information is redundant with dummy coding

Example Effects Coding

Variable	X_1	X_2
Dog	1	0
Cat	0	1
Bird	-1	-1

- $\check{Y} = b_O + b_1 Dog + b_2 Cat$
- $Mean\ Bird = b_O + b_1(-1) + b_2(-1)$
- $Mean\ Bird = b_O - b_1 - b_2$
- $Mean\ Dog = b_O + b_1(1) + b_2(0)$
- $Mean\ Dog = b_O + b_1$
- $Mean\ Cat = b_O + b_1(0) + b_2(-1)$
- $Mean\ Cat = b_O + b_2$

- Grand mean:
- $\frac{Bird+Dog+Cat}{3}$
- $\frac{(b_O - b_1 - b_2) + b_O + b_1 + b_O + b_2}{3}$
- $\frac{3b_O + b_1 - b_1 + b_2 - b_2}{3}$
- $\frac{3b_O}{3} = b_O$
- Intercept is grand mean, slopes are deviations from GM

Contrast Coding

- Requires more specific hypotheses about data
- Several necessary or desirable properties
  - Contrasts must sum to zero
  - Distances of 1 preferable (for interpretable coefficients)
  - Sum of the product of contrasts should equal 0 (this ensures orthogonality)
  - Still need G-1 variables for orthogonality
- Parameters are now differences between contrast groups
- Intercept is more difficult to interpret
- Can give different results from dummy and effects coding

Example Contrast Coding

Variable	X_1	X_2
Dog	1/3	-1/2
Cat	1/3	1/2
Bird	-2/3	0

- Look at what we are predicting here and if we satisfy our requirements
- $Mean\ Bird = b_O + b_1(-2/3) + b_2(0)$
- $Mean\ Bird = b_O - \frac{2}{3b_1}$
- $Mean\ Cat = b_O + b_1(1/3) + b_2(1/2)$
- $Mean\ Cat = b_O + \frac{1}{3b_1} + \frac{1}{2b_2}$
- $Mean\ Dog = b_O + b_1\left(\frac{1}{3}\right) - b_2(1/2)$
- $Mean\ Dog = b_O + \frac{1}{3b_1}$ - $\frac{1}{2b_2}$
- What do our parameters mean in this context?

Causality

- $Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots b_k X_{ik} + e_i$
  - This *implies* we know causal relationship
- Statistical control is better than nothing
- But model misspecification is a major issue
- Causality is in design, not analysis
- In some contexts, this may not matter