

# General Linear Model

A starting point

# General Linear Model

- Recall in regression a person's score may be represented as follows

$$y = a + bX + e$$

- There is a constant, the effect of predictor variable, plus error
- In terms of prediction  $\hat{y} = a + bX$
- However, there are errors in prediction, i.e. residual variance ( $y - \hat{y}$ )



# General Linear Model

- It is important to understand the linear model since it is the basis for most of the science of psychology
- You can have neat ideas all you want but until the theory can be stated in some specific model form, i.e. a mathematical representation of the idea, then all you have is an opinion, and not a scientific one<sup>1</sup>
- Furthermore, in reading research if you understand the model, even if only in a general sense, you are able to come to an informed interpretation regarding the results and draw your own conclusions

# ANOVA and Regression

- As an example, ANOVA can be seen as a special case of Multiple Regression, basically MR with categorical variables
- In ANOVA the general linear model can be represented as

$$y = \mu + \tau + e$$

- There is a constant, in this case a grand mean, the treatment effect, and error
- However the treatment effect here will need a special coding scheme to produce the desired result
  - We'll also see why I didn't put  $\tau X$

# Effects coding

- We will conduct an MR and ANOVA on the same data to illustrate their equivalence
- The coding scheme used will be 'effects coding' so as to be in keeping with the typical null hypothesis for ANOVA
  - The effects coding will allow for comparisons to the grand mean
- The 45+ group is the reference group
  - Note that if we had their specific ages it would be silly to group them, we only do this for demonstrative purposes here

	group	effectsyoun g	effectsmid	rating
1	18-25	1	0	7
2	18-25	1	0	4
3	18-25	1	0	6
4	18-25	1	0	8
5	18-25	1	0	6
6	18-25	1	0	6
7	18-25	1	0	2
8	18-25	1	0	9
9	25-45	0	1	5
10	25-45	0	1	5
11	25-45	0	1	3
12	25-45	0	1	4
13	25-45	0	1	4
14	25-45	0	1	7
15	25-45	0	1	2
16	25-45	0	1	2
17	45+	-1	-1	2
18	45+	-1	-1	3
19	45+	-1	-1	2
20	45+	-1	-1	1
21	45+	-1	-1	2
22	45+	-1	-1	1
23	45+	-1	-1	3
24	45+	-1	-1	2

# ANOVA

- Anova output
- 1 way analysis of variance with 'Group' as the grouping variable

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
group	64.000 <sup>a</sup>	2	32.000	11.586	.000
Error	58.000	21	2.762		
Corrected Total	122.000	23			

group	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
18-25	6.000	.588	4.778	7.222
25-45	4.000	.588	2.778	5.222
45+	2.000	.588	.778	3.222

# Regression

- With regression we will use the two effects-coded variables as predictors
- What do we find?
- The exact same results
- How is this so?

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.724 <sup>a</sup>	.525	.479	1.662

a. Predictors: (Constant), effectsmid, effectsyoung

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	64.000	2	32.000	11.586	.000 <sup>a</sup>
	Residual	58.000	21	2.762		
	Total	122.000	23			

a. Predictors: (Constant), effectsmid, effectsyoung  
b. Dependent Variable: rating

# Regression

- Let's apply the coefficients to see what we'd come up with for predicted values
- With effects coding our constant, i.e. the intercept, is the grand mean and the coefficients tell us how far away the other groups are from that
- Young is 2 points above with a mean of 6
- Mid is the same as the grand mean, 4
- Old is the only value left that could produce a grand mean of 4 i.e. 2
  - This is why you only use 2 coded variables, as that's all that's necessary to determine the third given the grand mean
    - Lost a degree of freedom

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	
	B	Std. Error	Beta			
1	(Constant)	4.000	.339		11.791	.000
	effectsyoun	2.000	.480	.724	4.169	.000
	effectsmid	.000	.480	.000	.000	1.000

a. Dependent Variable: rating

# Regression

- So our grand mean is four, and our regression equation is:
- Applying the equation to case 1 we'd get
- $4 + 2 * 1 + 0 = 6$
- We'd predict a value of 6, which just so happens to be the group mean
- In fact, for everyone in group one (18-25) we'd predict 6 with this coding scheme

$$\hat{y} = 4 + 2X_{effectyoung} + 0_{effectold}$$

	group	effectsyoun g	effectsmid	rating
1	18-25	1	0	7
2	18-25	1	0	4
3	18-25	1	0	6
4	18-25	1	0	8
5	18-25	1	0	6
6	18-25	1	0	6
7	18-25	1	0	2

# Regression

- So our predicted values will be the group mean, making the model SS, the sum of the squares of the deviations of group means from the grand mean
- The residual variance is the sum of squares for each score minus its group mean
- Adding those gives our total variance for the DV

.44009209000003E-010

group	rating	PRE_1	RES_1
18-25	4	6.00	-2.00
18-25	6	6.00	.00
18-25	8	6.00	2.00
18-25	6	6.00	.00
18-25	6	6.00	.00
18-25	2	6.00	-4.00
18-25	9	6.00	3.00
25-45	5	4.00	1.00
25-45	5	4.00	1.00
25-45	3	4.00	-1.00
25-45	4	4.00	.00
25-45	4	4.00	.00
25-45	7	4.00	3.00
25-45	2	4.00	-2.00
25-45	2	4.00	-2.00
45+	2	2.00	.00
45+	3	2.00	1.00
45+	2	2.00	.00
45+	1	2.00	-1.00
45+	2	2.00	.00
45+	1	2.00	-1.00
45+	3	2.00	1.00
45+	2	2.00	.00

$$\sum n(\bar{X}_j - \bar{X}_{..})^2$$

$$\sum \sum (X_{ij} - \bar{X}_j)^2$$

$$\sum (X_{ij} - \bar{X}_{..})^2$$

# Back to Anova

- So again for any score we have the basic model

$$y = \mu + \tau + e$$

- A score is simply a combination of
  - The constant
    - The grand mean
  - The treatment effect
    - difference in their group mean – the grand mean (times 1)
  - The residual
- Note also that this is the exact same model for the independent samples t-test

# More

- ANCOVA = sequential (hierarchical) MR
- Factorial between groups adds another treatment effect and interaction term
- Repeated measures<sup>1</sup>
  - $\mu$  = grand mean
  - $\pi$  = constant associated with the  $i$ th individual, how much their mean differs from the average person
  - $\tau$  = constant associated with the  $j$ th treatment, how its mean differs from the average treatment mean
  - $e$  = error
- Logistic Regression
  - The first is in terms of log odds, the second is in terms of probability of group membership
    - $e$  is not error (implicit) but the natural logarithm
- Common Factor Model
- The general linear model that incorporates both multivariate and univariate situations where  $Y$  represents a vector or matrix of DVs

$$Y = b_0 + b_1 X_{cov} + e$$

$$Y = b_0 + b_1 X_{cov} + b_2 X_2 + e$$

$$Y_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ij}$$

$$Y_{ij} = \mu + \pi_i + \tau_j + e_{ij}$$

$$\text{Log Odds} = b_0 + b_1 X + e$$

$$\Pr(Y = 1) = \frac{e^{b_0 + b_1 X}}{1 + e^{b_0 + b_1 X}}$$

$$X_i = \lambda_i \xi + \delta_i$$

$$Y = Xb + e$$

# The Statistical Language

- Statistics is a language used for communicating research ideas and findings
- We have various dialects with which to speak it and of course pick freely of the words available
- Sometimes we prefer to do regression and talk about amount of variance to be accounted for
- Sometimes we prefer to talk about mean differences and how large those are
  - In both cases we are interested in the effect size
- Which tool we use reflects how we have chosen our measures and want to talk about our results