GLM Concept

- Fundamental questions to consider when trying to understand the relationship between two variables: Are the two variables related?, What is the direction of the relationship between two variables?, How strong is the relationship between two variables?
- Statistical models:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Y: dependent (outcome) variable
$\hat{Y}$: predicted score of Y
Y- $\hat{Y}$: error

  - Measuring the error of the model:
    - The basic measure of a model's error is the **Sum of Squared Errors (SSE)** where you square each individual's error then add up all the squared errors
  - **Model comparison**: make incremental changes then see if the more complex model predicts Y sufficiently better, and if it does then we conclude that adding that increased complexity (adding the new predictor X) benefited the model
  - **General Linear modeling/general multivariate regression model**: a compact way of writing several multiple linear regression models in one linear model
    - *Used when the data is normally distributed*
    - Examples: t-test, regression, ANOVA

Data = Model + Error

$$Y = \beta_0 + \varepsilon$$

Descriptive Statistics

		Notation	Definition	Formula	Excel Formula/Steps
Central Tendency	Mean	$\bar{Y}$	The average of the numbers	$\frac{\sum x}{n} = \frac{Sum}{sample\ size}$	=AVERAGE(B2:B6)
	Median	-	The middle score in a set of data arranged numerically	-	=MEDIAN(B2:B6)
	Mode	-	The most frequent score in a data set	-	=MODE(B2:B6)
Dispersion	Range	-	The difference between the max & min score	-	=MIN(B2:B6) =MAX(B2:B6)
	Standard Deviation	S	A measure of how dispersed the data is in relation to the mean	$(x_i - \bar{x})^2$= the difference between the individual value and the mean $S = \sqrt{Variance}$	1. Calculate the **Mean** 2. Calculate the **Deviation** (=Data cell – Mean cell) 3. Calculate the **Squared Deviation** (=Deviation cell^2) 4. Calculate the **Sum of Squares (SS)**- sq. dev. (=SUM(F2:F6) 5. Input the **Degrees of Freedom** (=n-1) 6. Calculate the **Variance** (=SS/df) 7. Calculate the **Standard Deviation** (=SQRT(F10)) 8. Calculate the Standard Error (=Std. Dev./(SQRT(n)))
	Variance	σ^2 / S^2	A measure of how far each number in a dataset is from the mean	$\frac{\Sigma(x_i - \bar{x})^2}{n-1} = \frac{square\ of\ deviation}{degree\ of\ freedom}$ $\frac{SSE}{df}$	

Sum of Squared Errors (SSE or SSEA): basic measure of a model's error where you square each individual's error and add them all up; Quantifies error as an area; allows us to calculate variance & standard deviation

Conceptual Formula	**Computational Formula (preferred)**
$SSE = \Sigma(Y_i - \hat{Y}_i)^2$	$SSE = \Sigma Y^2 - \frac{(\Sigma Y)^2}{N}$
1. Calculate each ind. **Deviation** (=Data cell -Mean cell) 2. Calculate each ind. **Squared Deviation** (=Deviation cell ^2) 3. **Sum** the Squared Deviations (=SUM(F2:F6))	1. Calculate each individual **Y^2** (=Data cell^2) 2. **Sum** all the Y^2 values 3. **Sum** all the Y values 4. Add values from Steps 1-3 into formula

t-Tests

Hypothesis Testing	
Definition	A method for testing a hypothesis about a parameter or population, using data measured in a sample
Purpose	To test claims or ideas about a group or population
Goal	To determine the likelihood that a population parameter, such as the mean, is likely to be true

Steps for Hypothesis Testing	

Step 1: State the null & alternative	• Null Hypothesis (H_0) - statement about the population parameter that is <u>assumed to be true</u>
	  ○ Example: Children in the US watch an average of 3 hrs. of TV a week (H_0: m = 3)
	• Alternative Hypothesis (H_1) – statement that <u>contradicts the null</u> hypothesis
	  ○ Example: Children in the US watch more or less than 3 hrs. of TV a week (H1: m ≠ 3)
	• Possible Hypotheses to Test
	  ○ Nondirectional or Two-tailed:
	    ▪ H_0: $\mu = 0$
	    ▪ H_1: $\mu \neq 0$
	  ○ Directional or One-tailed:
	    ▪ H_0: $\mu = 0$
	    ▪ H_1: $\mu > 0$ OR H1: $\mu < 0$

Step 2: Set criteria for the decision/level of significance	• Criterion of judgement upon which a decision is made regarding the value stated in the null hypothesis • Typically set at 5% (.05)

Step 3: Compute the test statistic (t-score)	$t_{obt} = \frac{M - \mu}{s_M}$
	• M: population mean
	• μ: mean of means (the number you are suspecting that your mean is different from in your hypotheses)
	• S_M : standard error (standard deviation of the mean)
	
	• Calculating t_obt in Excel
	  ○ **Sum** the Y values
	  ○ Calculate each individual **Y^2** (=Data cell^2)
	  ○ **Sum** all the Y^2 values
	  ○ Calculate the **Mean**
	  ○ Calculate the **SSE**
	  ○ Calculate the **Variance** (=SSE cell/ *df* (N-1))
	  ○ Calculate the **Std. Dev.** (=SQRT(variance cell))
	  ○ Calculate the **Std. Error** (= std. dev. Cell/SQRT(N))
	  ○ Calculate **t_obt** (=(mean – hypothesis value)/std. error)
	
	• Calculate the 95% CI (M ± (t_{crit} & s_M)
	  ○ M: sample mean
	  ○ T_crit: use *t* Table
	    ▪ Need to know: df (N-1), alpha level/level of significance, & 1-tailed or 2-tailed
	  ○ s_M : standard error (std. dev./SQRT(n))
	  ○ In Excel
	    ▪ Calculate **95% CI LB** (=mean-(t_crit*std. error))
	    ▪ Calculate **95% CI UB** (=mean+(t_crit*std. error))

Step 4: Make a decision	• Compare the t_obt & t_crit values
	  ○ If t_obt > t_crit = reject the null (the sample mean is associated with a low probability of occurrence; **results are significant**)
	  ○ If t_obt < t_crit = retain/fail to reject the null (the sample mean is associated with a high probability of occurrence; **results aren't significant**)
	• Reporting/Interpretation
	  ○ Example: According to a one sample t-test, the average soda consumption of the sample is/is not statistically different from 2 cans per day (t(df) = x.xx)

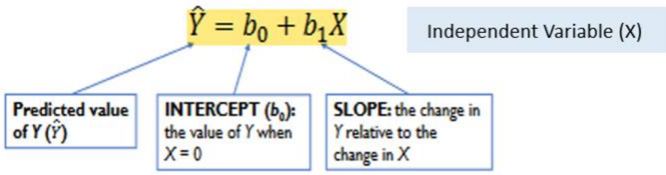
Bivariate Regression

- Simple model (mean-only model) isn't very informative since it has a lot of error – can add a predictor variable (X) to improve prediction & reduce error (provides an equation of a line that best predicts Y from X)
- Bivariate regression has two types of variables: Predictor (X; known) and Outcome/Criterion (Y; to-be-predicted)
- A population parameter (β), estimate of population parameter using sample data (*b*)
- Role of Error

Model	Source	Label	Error
$Y_i = \beta_0 + \varepsilon_i$	Model C	Total error	SST (SSE_C)
$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$	Model A	Residual error	SSE_A
	Model C- Model A	Explained error	SSR

*SST is almost always greater than SSE_A.

$SST = SSE_A + SSR$



10 Steps to Model Comparison	
Step 1: State the Augmented & Compact Models	• Need to identify the DV & IV • Compact Model: $DV = b_0 + \varepsilon$ • Augmented Model: $DV = b_0 + b_1 IV + \varepsilon$

Step 2: Identify the Null Hypothesis	• Testing to see if b1 is statistically different from 0
	  ○ $H_0: b_1 = 0$
	  ○ $H_1: b_1 \neq 0$

Step 3: Count the Number of Parameters Estimated by Each Model	• Compact Model (PC = 1)
	  ○ Parameter: b_0
	• Augmented Model (PA = 2)
	  ○ parameters: b_0 and b_1

Step 4: Calculate the Regression Equation	• Bivariate Regression Coefficient: $Y = b_0 + b_1X + \varepsilon$
	• b_1
	  ○ Conceptual Formula: $\frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2} = \frac{sum\ of\ (x\ deviation*ydeviation)}{sum\ of\ squared\ x\ deviation}$
	  ○ Computational Formula: $\frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum x)^2}{n}}$
	• b_0: $b_0 = \bar{Y} - b_1\bar{X}$
	• Calculating in Excel (NOTE: order data as Y then X)
	  ○ **Conceptual Formula (preferred)**
	    ▪ Calculate the **Mean** for X & Y
	    ▪ Calculate **Y_DEV**
	    ▪ Calculate **X_DEV**
	    ▪ Calculate **X_DEV*Y_DEV**
	    ▪ Calculate the **Sum** for X_DEV*Y_DEV
	    ▪ Calculate **X_DEV^2**
	    ▪ Calculate the **Sum** for X_DEV^2
	    ▪ Calculate **b1** (=Sum for X_DEV*Y_DEV/Sum for X_DEV^2)
	    ▪ Calculate **b0** (=Y Mean-(b1*X Mean))
	  ○ Computational Formula
	    ▪ Calculate **Sum** for X & Y
	    ▪ Calculate **X*Y**
	    ▪ Calculate the **Sum** for X*Y
	    ▪ Calculate **X^2**
	    ▪ Calculate the **Sum** for X^2
	    ▪ Calculate **b1** (=(Sum of X*Y – (Sum of X & Sum of Y/# of Cases))/(Sum of X^2 – (Sum of X^2/ # of Cases))
	    ▪ Calculate **b0** (=Y Mean-(b1*X Mean))
	• Assumptions that need to be checked – see *SPSS Output Slides*

Step 5: Compute SSE_T	• Also SST; represents Total Error
	• Comes from the compact model: $SSE_T = \Sigma(Y_i - \bar{Y})^2$
	• Calculating in Excel (*NOTES:* order data as Y then X; will need to move over b1 and b0 values or calculate them)
	  ○ Calculate **Mean** of Y
	  ○ Calculate **Y_DEV**
	  ○ Calculate **Y_DEV^2**
	  ○ Calculate **SST** (=SUM for Y_DEV^2 data column)

Step 6: Compute SSE_A	• Residual Error; Error for the Augmented Model
	• $SSE_A = \Sigma(Y_i - \hat{Y}_i)^2$ ($\hat{Y}_i$ is the score predicted by the regression equation)
	• Calculating in Excel
	  ○ Calculate **Y_HAT or Y Predicted** (=b0+(b1*X value)
	  ○ Calculate **DEV** (=Y Value – Y Predicted Value)
	  ○ Calculate **DEV^2**
	  ○ Calculate **SSE_A** (=SUM for DEV^2 data column)

Step 7: Compute SSR	• Explained Error; Reduced Sum of Squares • Calculating in Excel ○ Calculate **SSR** (=SSE_T-SSE_A)

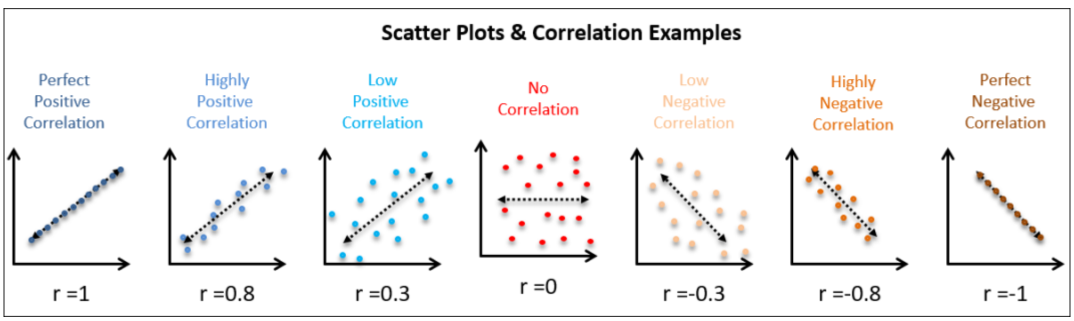
Step 8: Compute PRE/R^2	• Proportion of Reduced Error Variance; Percentage of Reduced Sum of Squares of Error • Referred to as the "variance explained" • Calculating in Excel ○ Calculate **R^2** (=SSR/SST)
Step 9: Complete Summary Table	(see table below)

Step 10: Make Decision Regarding H_0	• Look at the *F* statistic (calculated) to make a decision
	  ○ F critical value can be obtained from F table
	• When you have F_{obt} and F_{crit}
	  ○ If $F_{obt} < F_{crit}$, we do not reject H_0
	  ○ If $F_{obt} > F_{crit}$, we reject H_0
	• If you have SPSS output, you are given a p value – compare to alpha level of .05
	  ○ If *p* > .05, we do not reject H_0
	  ○ If *p* < .05, we reject H_0

Source	SS	df	MS	F	R^2
Model Comparison	SSR	PA - PC	$MS_{model} = \frac{SSR}{PA - PC}$	$F = \frac{MS_{model}}{MS_{residual}}$	$R^2 = \frac{SSR}{SSE_T}$
Residual	SSE_A	N - PA	$MS_{residual} = \frac{SSE_A}{N - PA}$		
Total	SSE_T	N - PC			

Correlation

- Statistical procedure used to describe the strength and direction of the linear relationship between 2 variables (degree to which two variables are associated)
  - Values range from -1, to 1, with both of those values insinuating a perfect correlation (rare). Values of r = 0 doesn't mean the variables aren't related, it just means they aren't LINEARLY related.
  - Correlation doesn't equal causation
  - Key Descriptors



Perfectly Correlated • Either -1 or 1
Strongly Correlated • r > |.75|
Moderately Correlated • |.5| < r > |.75|
Weakly Correlated • |.25| < r > |.5|
Uncorrelated • 0 < r > |.25|
Sign • Positive or Negative

- Significance

Two-tailed Testing	One-tailed Testing
H_0: r = 0; there is no association between the two variables H_1: r ≠ 0; there is an association between the two variables	H_0: r = 0; there is no association between the two variables H_1: r < 0 or r > 0; there is a positive/negative association between the two variables

- Pearson's r (Most common, 2 continuous variables)

Calculating in Excel (Note: Will have data for X & Y)	Interpretation

• Calculate the **Mean** for X & Y
• Calculate **X_DEV**
• Calculate **X_DEV^2**
• Calculate the **Sum** for X_DEV^2
• Calculate **Y_DEV**
• Calculate the **Y_DEV^2**
• Calculate the **Sum** for Y_DEV^2
• Calculate **X_DEV*Y_DEV**
• Calculate the **Sum** for X_DEV*Y_DEV
• Calculate **Pearson's r** (=Sum for X_DEV*Y_DEV/SQRT(X_DEV^2*Y_DEV^2))
• Calculate **R^2** (=Pearson's r^2)
	• Obtain Critical Value
	  ○ Need to know: df (n-2), level of significance, & one- or two-tailed
	• Compare Obtained to Critical Values
	  ○ Reject the Null (statistically significant)
	    ▪ Obtained > Critical
	  ○ Retain the Null (not significant)
	    ▪ Obtained < Critical
	• Write Up
	  ○ Example: Using the Pearson correlation coefficient, there is a statistically significant relationship between X and Y, r=-.897,p<.05.

- Point Biserial (1 continuous & 1 categorical variable) – see *SPSS Output Slides*
- Spearman's p (rho) (have ordinal variables)
  - Relationship must be monotonic (as 1 variable increases, the other either increases or decreases)
  - Can't handle normal data, not sensitive to outliers
- Kendall's tau-b

- Alternative to Spearman's when sample size is small – therefore, assumptions are the same

One-Way ANOVA

- <u>**Analysis of Variance (ANOVA):**</u> omnibus test used to determine whether at least 1 of 3+ group means are different from one another
  - DV (Y) is continuous
  - IV (X) is categorical (the grouping variable)
- Equivalent to a t-test but can have more levels (groups) tested at once
- If the means of k different groups are all the same, then the variance between them is 0
- Hypothesis
  - H_0: $\mu_1 = \mu_2$… [μ represents the population mean for each group]
  - H_1: At least one mean is different
- *F* statistic: Ratio of Variances (=Between/Within)
  - Should be a positive value.
  - F = 1 means no mean difference because the between and within group variabilities would be the same. The higher your F the greater chance of it being significant (would mean there is a much higher variability between groups than you see within groups)
  - Calculating in Excel

Mean	SS_{BG}	SS_{WG}	SST	Complete Summary Table
• Calculate the **Mean** for each group • Calculate the **Grand Mean** (includes all subjects regardless of group)	• Calculate the **DEV.** • Calculate the **SQ. DEV.** • Calculate the **Sum** for SQ. DEV. • Calculate **SSR (Between)** (=(G1 n*(G1 mean-grand mean)^2) + (G2 n*(G2 mean-grand mean)^2)	• Calculate **SSE (Within)** (=G1 Sum for SQ. DEV. + G2 Sum for SQ. DEV.)	• Calculate **SST** (=SSR+SSE)	• df Column ○ BG: # of groups -1 ○ WG: Total # of subjects - # of groups ○ Total: Total # of subjects – 1 • MS column ○ BG: SS_{BG} – BG df ○ WG: SS_{WG} – WG df • F Statistic ○ MS_{BG}/MS_{WG} • eta-squared (η^2) ○ SS_{BG}/SST ○ Small Effect: .01 < η^2 < .06 ○ Medium Effect: .06 < η^2 < .14 ○ Large Effect: .14 < η^2

Source	SS	df	MS	F
Between Groups	SS_{BG}	k - 1	$MS_{BG} = \frac{SS_{BG}}{k-1}$	$F = \frac{MS_{BG}}{MS_{WG}}$
Within Groups	SS_{WG}	N - k	$MS_{WG} = \frac{SS_{WG}}{N-k}$	
Total	SST	N - 1		

**Where N is total number of subjects and k is number of groups.

- Interpretation
  - Obtain F crit value (from F table)
    - df between – horizontal
    - df within - vertical
  - If F value is GREATER than F crit then we can say there is a significant difference between the means of the groups (reject the null)

Miscellaneous

- If a test result is **NOT SIGNIFICANT**, then the alpha level should be written in the following way in a reporting/write up: p>.05
  - If result is significant: p<.05