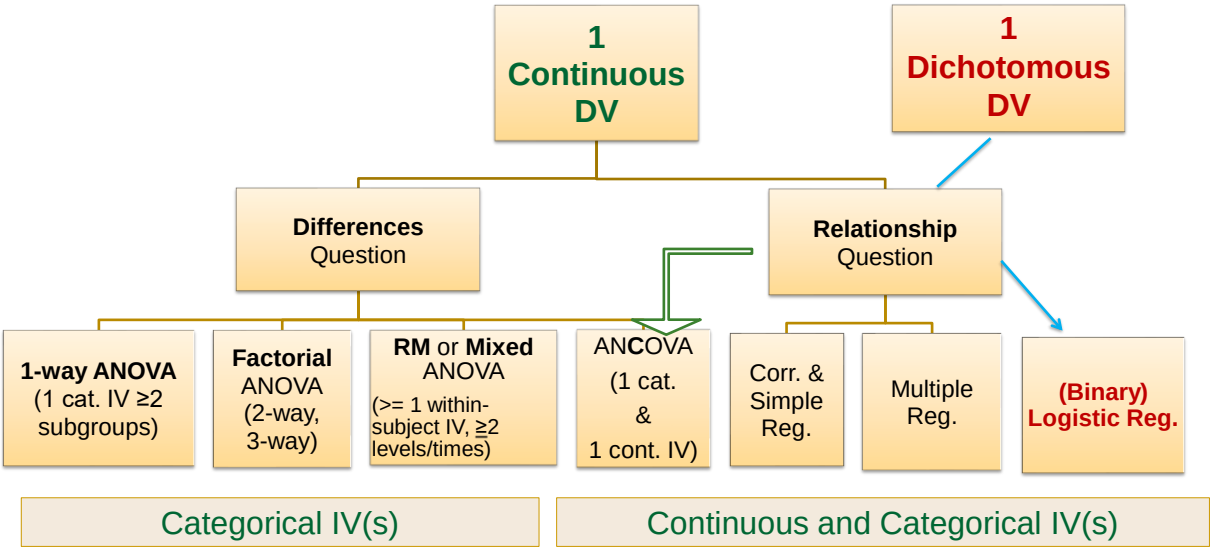
Binary Logistic Regression

Qiong (Joan) Fu, Ph.D.

- ✓ When and why do we use **logistic** regression (LR)? Why not **OLS** regression?
- ✓ Basic concepts in LR:
  - ✓ probability, odds, odds ratio, and logit (log odds)
- ✓ Example 1 with a single binary predictor
- ✓ Example 2 with a binary and a continuous predictors
  - ✓ Assessing the **overall model**
  - ✓ Assessing **individual predictors**
  - ✓ Interpreting & Reporting results
- ✓ FYI—Notes & Application in psychometrics (e.g., ROC, IRT)

For illustration of the basic concepts

1



IV = independent variable; DV = dependent variable

2

When do I consider logistic regression (LR)?

- When the **d**ependent **v**ariable (DV) is categorical (nominal or ordinal)
- ***Binary (Binomial)* LR:** DV has 2 levels/subgroups only
  - Example in your area?
  - Really dichotomous, or after combining categories into 2
- **I**ndependent **v**ariables (IVs) can be either continuous or categorical
  - No dummy coding is needed before running LR in SPSS;
  - You may select the reference/comparison subgroup of your interest.

Research Question Examples

1. Which factors predict **whether or not** a political candidate wins an election?
   - $$ and time spent on campaign, any scandal (Y/N), an incumbent (Y/N)
   - Your choices of predictors? ...
2. How well do age, gender, family history (Y/N), and exercise time predict **whether or not** an adult develops ___ (a disorder)?
   - Adapt this RQ for COVID-19?
3. To what extent are GRE, GPA, and prestige of the undergraduate program related to the **likelihood of** being admitted into the graduate school?
4. How well do college students' high school math score and getting self-efficacy (or not) predict the **likelihood of** their success on a college math course (getting A or B vs. lower grades)?

Example

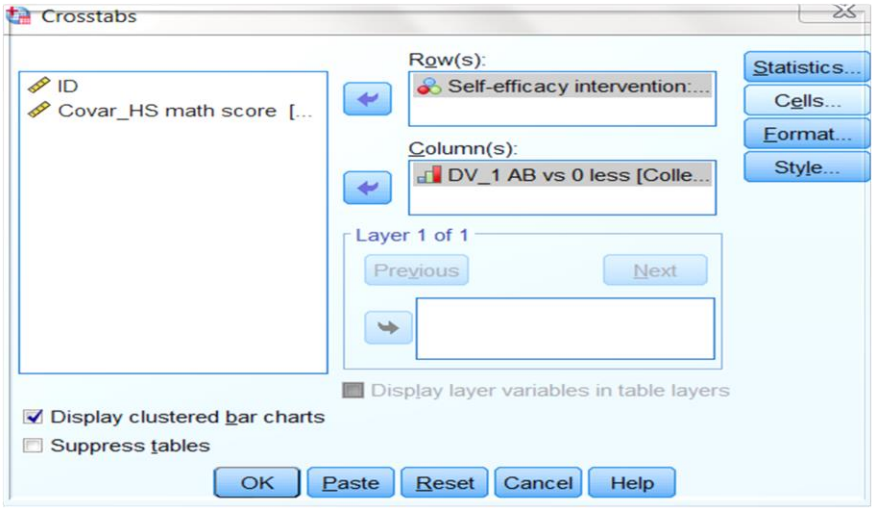
See SPSS data "Math_LR data.sav"

- Participants
  - *N* = 100 college students in a math algebra course; randomly assigned to two condition groups (self-efficacy intervention vs. control).
- **Dependent Variable** (**DV**):
  - College Algebra grades*:* ***Success* (*A* or *B*) = 1** vs. ***Less than B* = 0**
- **Independent Variable** (**IV**):
  - Self-efficacy intervention **Condition**: *Control* **= 0**, *Experimental* **= 1**
  - **HS math**: High school math scores; continuous data, range 24~98
- **Research Question** (example):

  How well do college students' high school math scores and self-efficacy conditions (Control vs. Experimental) predict their *likelihood* (or *chance/odds*) of getting high grades (A or B) versus lower grades on a college algebra course?

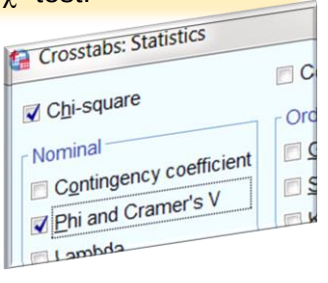
Initial Data Checking (Cross Tabulation)

- Check the **crosstab** of the categorical **IV(s) and DV**.
- **No cell** should be empty or has a particularly small frequency count before we run any inferential statistics.



In SPSS: Analyze → Descriptive Statistics → Crosstabs → Condition (Row), College Math (Column) → OK

For χ^2 test:

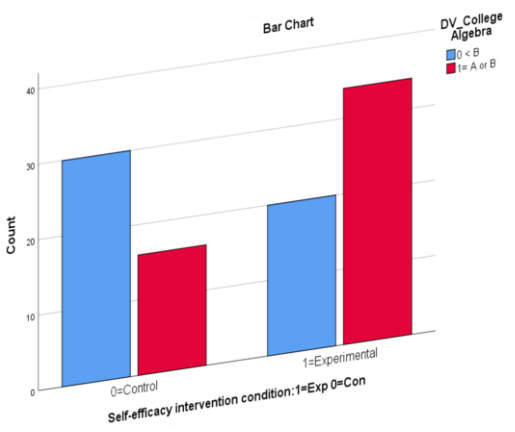


Crosstab of Categorical IV & DV

Self-efficacy condition * DV_College Algebra Cross-tabulation

Count

		DV_College Algebra		Total
		0 < B	1= A or B	
Self-efficacy condition	0=Control	30	16	46
	1=Experimental	20	34	54
Total		50	50	100



7

What if I do run OLS multiple linear regression?

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.605[a]	.366	.353	.404

a. Predictors: (Constant), Self-efficacy condition, Covar_HS math score
b. Dependent Variable: DV_College Algebra

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9.146	2	4.573	27.979	.000[b]
	Residual	15.854	97	.163		
	Total	25.000	99			

a. Dependent Variable: DV_College Algebra
b. Predictors: (Constant), Self-efficacy condition, Covar_HS math score

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity
		B	Std. Error	Beta			Tolerance
1	(Constant)	-.504	.142		-3.555	.001	
	Covar_HS math score	.015	.002	.551	6.625	.000	.945
	Self-efficacy condition	.152	.083	.152	1.825	.071	.945

a. Dependent Variable: DV_College Algebra

The whole model looks great based on the R^2 ☺

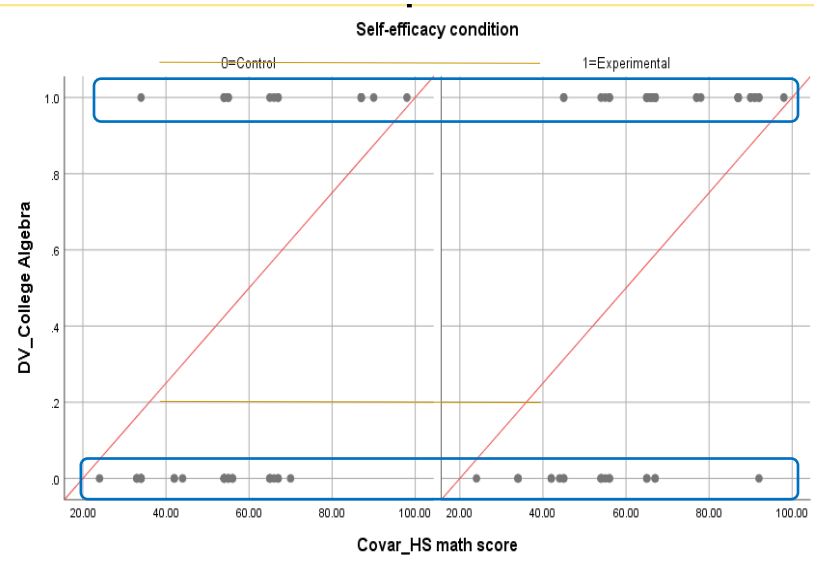
But… what about the **assumptions**?

8

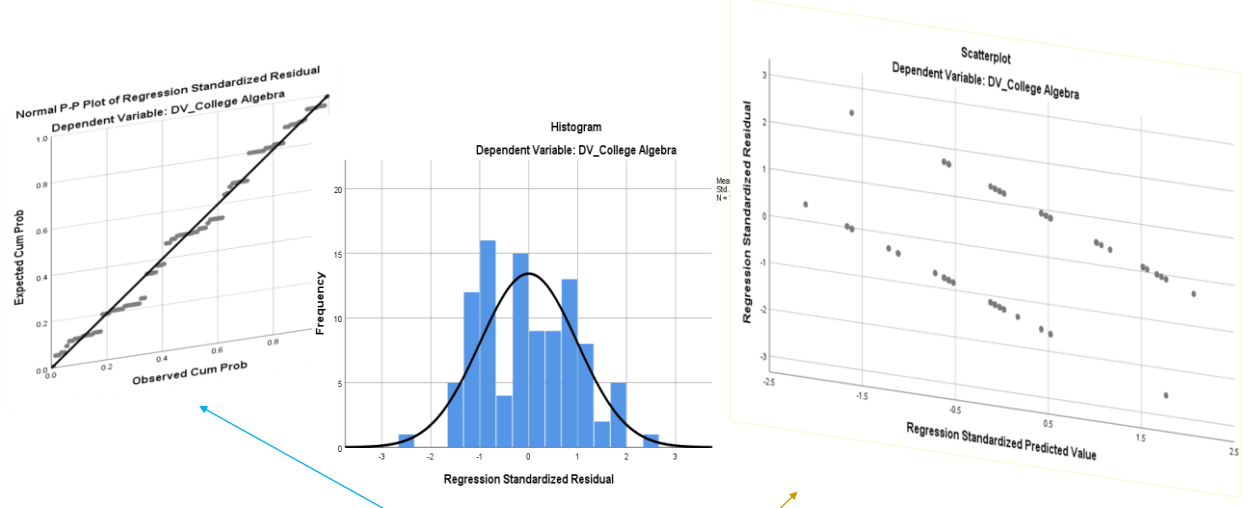
What if I do run OLS multiple regression?

Using OLS regression, SPSS would oblige and create a line of best fit.

But how about the **linearity assumption,** first of all?



- No matter how the HS math score changes, the practical values for the outcome only have “1” and “0”.
- So the model using OLS regression is **meaningless**!

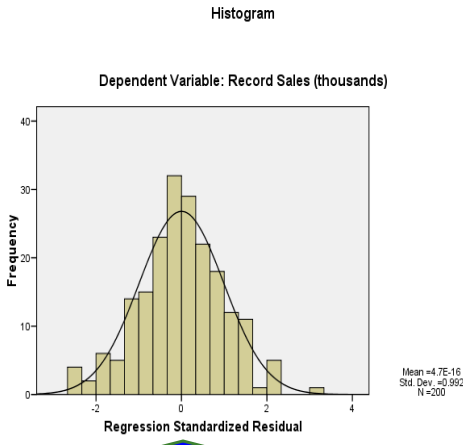


Two more **assumptions**:

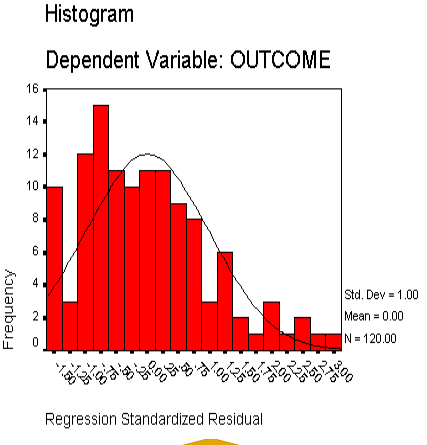
- Normality of error (Z residual)—**met?**
- Homogeneity of error (Z residual)—**met??**

Rev. OLS Regression: Normality of Residuals

--- Check whether or how much the histogram of standard residuals is similar to the normal curve



Good



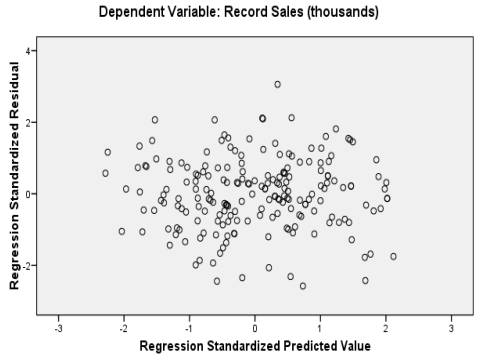
Bad

Andy Field

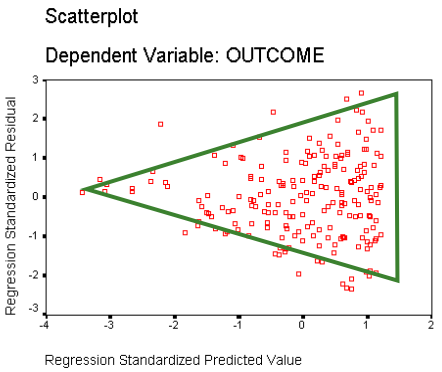
11

Rev. OLS Regression: Homoscedasticity

- **Standardized Residual Plot:**
  - ✓ ZRESID (Y axis) vs. ZPRED (X axis)
- **A random pattern** is expected.



Good

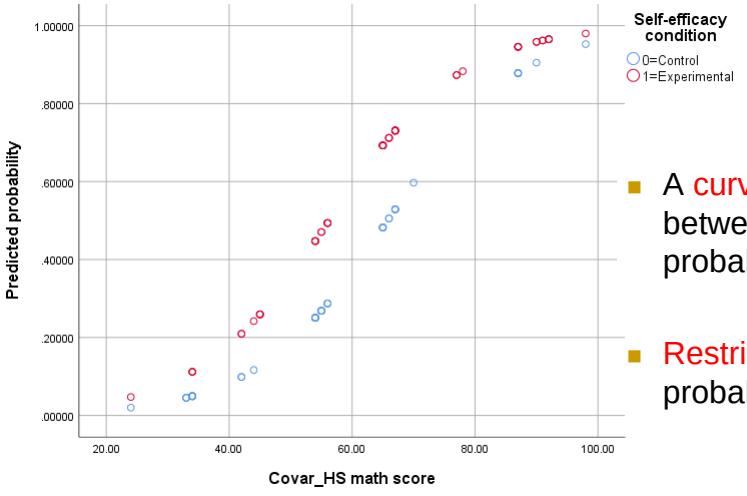


Bad

Andy Field

12

So…Can we transform the outcome into **probability** of success as the outcome?

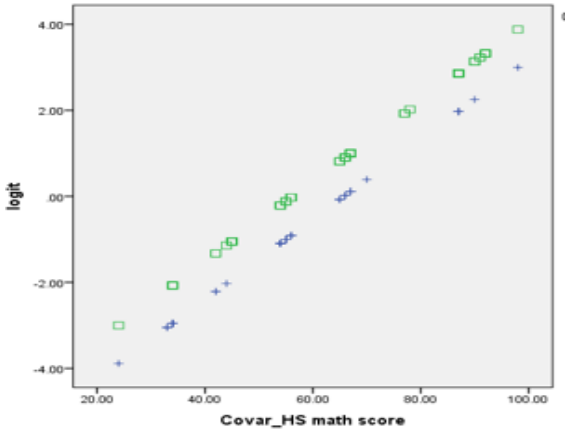


As the HS math score increases, the probability of success (A or B) increases.

- A curvilinear relationship between covariate (X) and probability of success.
- Restricted range with probabilities (0-1)

13

How about transformation with **ln(odds)**?



- **Linear** relationship between **logits for DV** and HS math score (X1) by condition group (X2)?

  - **logits = ln(odds)**
  - **ln = natural log**
- **No limits in range** for logits of DV.

Note. The lines would be **flat** if the predictor (HS math score; X1) is not significantly related with the logit of the outcome.

14

Binary Logistic Regression (LR)

- LR estimates the odds of a certain event occurring.
  - This is the category of primary interest in the outcome (e.g., *success*);
  - **coded 1 in SPSS—Double check the coding table in SPSS output.**
  - The other category (e.g., *failure*) is the **reference, coded 0 in SPSS**
- LR applies *maximum likelihood estimation* (MLE) after transforming the dichotomous DV into a logit variable
  - Logit **= log odds** = ln(odds) = natural log of the odds

$$\text{logit}(Y) = \ln(odds) = \ln\left(\frac{\text{Probability of Success}}{\text{Probability of Failure}}\right)$$
$$= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_j X_j$$

Basic concepts coming first….

FYI—Math Refresher

Common log (log)

$10^1=10 \rightarrow \log(10)=1;$

$10^2=100 \rightarrow \log(100)=2$

$10^3=\_\_\_ \rightarrow \log(\_\_\_)=\_\_\_$

Natural log (ln)

$e^1 = 2.718^1 = 2.718 \rightarrow \ln(2.718)=1;$

$e^2 = 2.718^2 = 7.38 \rightarrow \ln(7.38) = 2;$

$e^3 = 19.947 \rightarrow \ln(19.947) = \_\_$

2 Probabilities
→ odds (at one level/value of X),

2 odds (from 2 adjacent levels/values of X)
→ Odds Ratio,

Odds Ratio ↔
logit [ln(odds)]

$$\text{logit} = \log \text{odds} = \ln(odds) = \ln\left(\frac{\text{Prob. of Success}}{\text{Prob. of Failure}}\right)$$
$$= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_j X_j$$
$$Odds = \frac{\text{Prob. of Success}}{\text{Prob. of Failure}} = Exp(\text{logit})$$
$$\text{Prob. of Success} = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_j X_j)}}$$
$$= \frac{1}{1+e^{-(\text{logit})}} = \frac{\text{odds}}{1+\text{odds}}$$

Example 1. LR with One Dichotomous IV only

Manual computation for illustration of concepts:

Probabilities
→ Odds
→ Odds Ratio
↔ Logit [ln(odds)]

DV: College_Math	IV: SE Intervention		Total
	Control (0)	Exp(1)	
A or B =1	16	34	50
< B =0	30	20	50
Total	48	54	100

	Control **(0)**	Experimental **(1)**	**Notes**
Probability for higher grades(**P1**)	16/46 = 0.35	34/54 = 0.63	**Within** each subgroup
Probability of lower grades (**P2**)	30/46 = 0.65	20/54 = 0.37	
Odds (= P1/P2)	0.35/0.65 = **0.53**	0.63/0.37 = **1.70**	
Odds Ratio (OR) = Odds_E/Odds_C	/	1.70/0.53 = **3.19**	**Between groups**: As the predictor increases by 1 (Group changes from control to the experimental)
logit = ln(odds)	/	ln(3.19) = **1.16**	
Double check	/	Exp(1.16) = 3.19	

Ex.1: **SPSS Output**

Go to SPSS: Analyze → Regression → Binary Logistic…

DV: College_Math; **IV: Condition** for SE (1 = Experimental vs. 0 = Control)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Self-efficacy condition(1)	1.159	.419	7.668	1	.006	3.187	1.403	7.241
	Constant	-.629	.310	4.123	1	.042	.533		

a. Variable(s) entered on step 1: Self-efficacy condition.

Odds Ratio (OR) = Exp(B) = Exp(1.159) = 3.187

$$\ln\left(\frac{P\ of\ A\&B}{P\ of\ less}\right) = -0.63 + \mathbf{1.16}^{*}\ \text{Group}$$

The logit equation is **not** of interest. No need to report or interpret.

- Check the prior slide to make sense of the **odds 0.53** for the Control group (=0), and then the **OR 3.187** for the Experimental group (=1).

How to Interpret Odds Ratio [i.e., Exp(*B*)]

- The predictor variables in LR are assessed using **Odds Ratio** [**OR**; i.e., **Exp(*B*)**], rather than slopes (*B*).
- **Compare Exp(*B*) with 1**:
- For a **dichotomous** predictor: Compared with the *reference* group (coded 0; e.g., control), the odds for the focus group (coded 1; female, experimental) …
- For a **continuous** predictor: As the predictor increases by one unit, the odds …

  - **Exp(B) = 1**, then the predictor has no effect on the odds of success (1 = even chance for success vs. not. As exp(0) = 1, then B ≈ 0]
  - **Exp(B) > 1**: e.g., *Exp(B)* = 1.10 → then the odds of of success **increase by** 10% [b/c 1.10 – 1 = 0.10].
  - **Exp(B) < 1**: e.g., *Exp(B)* = 0.70 → then the odds of of success **decrease by** 30% [b/c 1 – 0.70 = 0.30].

  - For a **significant** predictor, the 95% confidence interval (CI) of the OR **does not cross 1** (not zero)!

Ex.1: Interpretation of Odds Ratio for Gender

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Self-efficacy condition(1)	1.159	.419	7.668	1	.006	3.187	1.403	7.241
	Constant	-.629	.310	4.123	1	.042	.533		

a. Variable(s) entered on step 1: Self-efficacy condition.

- The **predicted odds (or likelihood)** for students in the **control group** (coded 0) to get A or B (coded 1) versus lower grades (coded 0) is 53%, odds ratio (OR) = 0.53.
- The predicted **odds** for students in the **experimental group** (coded 1 for SPSS data) to get higher versus lower grades is **3.19 times as high as** that for students in the control group , OR = 3.19 (95% CI: 1.40-7.24), *p* = .006.
  - Alternatively,
    - …is **2.19 *times higher than* that**…;
    - … **increases *by a factor of* 3.19** …

Practice: Interpret the Odds Ratio [Exp(B)]

Variables	Odds ratio [Exp(B)]	Interpret
Gender (1=F, 0=M)	• 1.80 in Study 1 • 0.60 in Study 2 • 2.50 in Study 3	
Confidence (0-10 points)	• 1.80 in Study 1 • 0.60 in Study 2 • 2.50 in Study 3	

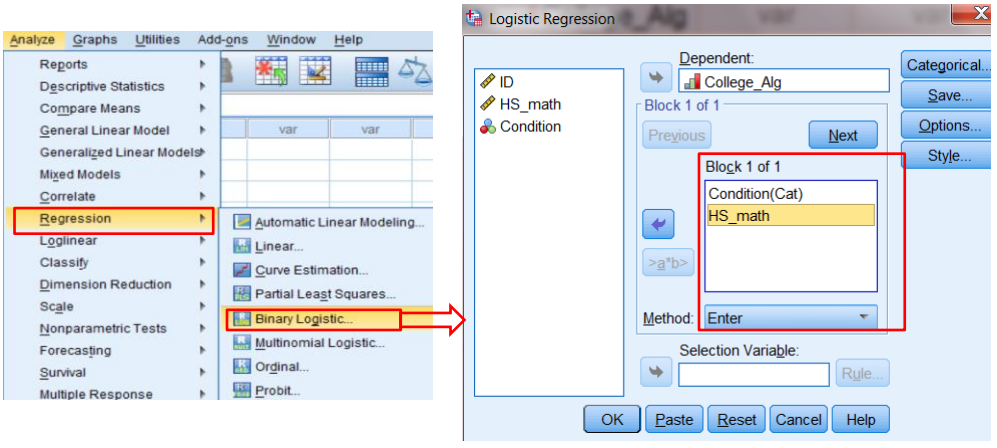
Ex2. Multiple Logistic Regression

See SPSS data "Math_LR data.sav"

- Participants
  - *N* = 100 college students taking a math course; randomly assigned to two conditions of self-efficacy intervention (0 = Control, 1 = Experimental).
- **Outcome (DV):**
  - College_Algebra grades: **Success (*A or B*) = 1** vs. ***Less than B* = 0**
- **Predictors** :
  - **Group—**SE intervention **Condition**: 0 = Control, 1 = Experimental.
  - **HS math**: High school math scores; continuous data, range 24~98
- **Research Question (example)**:

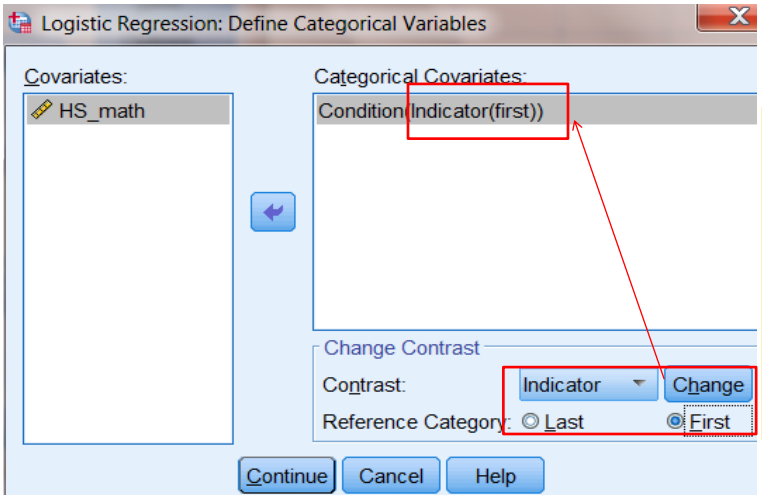
  How well do college students' high school math scores and self-efficacy condition groups (Control vs. Experimental) predict their *likelihood* (or *chance/odds*) *of* getting higher grades (A or B) versus lower grades on a college math course?

Multiple Logistic Regression in SPSS (I)



- If you need an interaction term, you would **not need to compute an interaction term** in the SPSS dataset as for the OLS regression.
  - Instead, you would only need to highlight both IVs and click >a*b>
- However, no interaction term is included in the current example.

Logistic Regression in SPSS (II)



Group—SE intervention
Condition:

0 = Control,

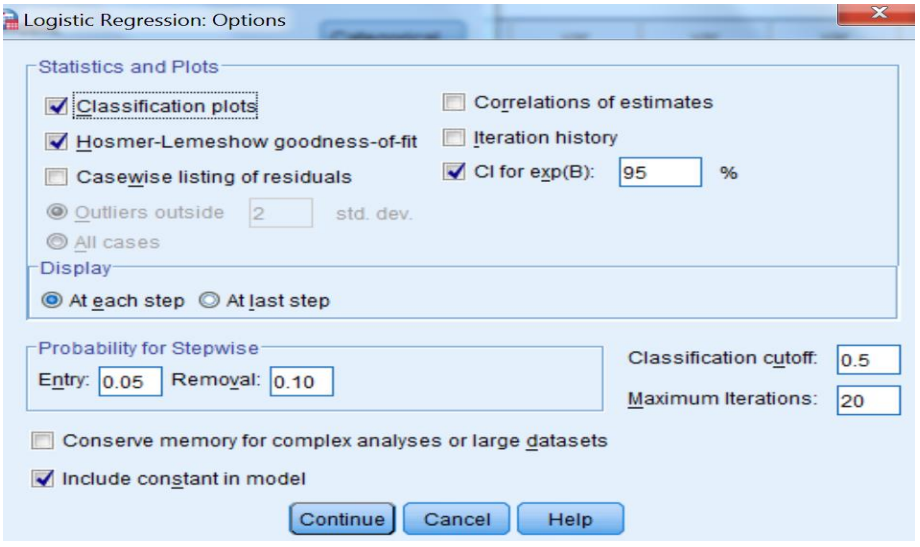
1 = Experimental

Do **not** choose **Simple** (1) here—that would give you different coding.

- Click "**Categorical**" button to define categorical variable(s)
- Send the categorical IV to the right box →check "First" to select lower category (0; Control in this dataset) as reference category →click "Change" → click "Continue"

Logistic Regression in SPSS (III)

- click “Options” → check the choices to your interest → click “Continue”



SPSS Syntax

```
CROSSTABS
  /TABLES=Condition BY College_Alg
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ PHI
  /CELLS=COUNT
  /COUNT ROUND CELL
  /BARCHART.
```

```
**LR multiple regression, using Control = 0 as reference (vs 1= Exp) .

LOGISTIC REGRESSION VARIABLES College_Alg
  /METHOD=ENTER HS_math Condition
  /CONTRAST (Condition)=Indicator(1)
  /CLASSPLOT
  /PRINT=GOODFIT CI(95)
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

DV: College_Alg

IV1: Group for SE intervention

IV2: HS_math

- **Crosstab** for the categorical IV (Condition Group) and the DV.
- Make sure that the categorical variable has the correct **reference category** if you have selected the **first** group (with lowest coding) as reference.
  - **0 = Control**
  - 1 = Experimental

Output: Coding in SPSS for DV and Group IV

This is an opportunity to double check whether the variables were coded in SPSS as you planned for?

Dependent Variable Encoding

Original Value	Internal Value
0 < B	0
1= A or B	1

Categorical Variables Codings

		Frequency	Parameter coding (1)
Self-efficacy condition	0=Control	46	.000
	1=Experimental	54	1.000

Ignore Block 0 output (empty model).

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	44.204	2	.000
	Block	44.204	2	.000
	Model	44.204	2	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	94.426[a]	.357	.476

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Overall Fit of the Model

- The **model including the two predictors** fits the data significantly better than the empty model with the intercept only, $\chi^2_{(2)} = 44.20$, $p < .001$.
- *df* = 2 is due to adding the 2 predictors

Significance Tests

- We **can report but cannot interpret** these "pseudo" R^2 values as "% of variance in the DV explained".
- The variance of a dichotomous or categorical DV depends on the frequency distribution of that variable.

Goodness-of-fit Test for Overall Model

- Some researchers prefer to use the Hosmer and Lemeshow (H-L) chi-square (χ^2) test of goodness of fit, particularly if
  - **continuous** covariates are in the model, or
  - the sample size is small.
  - http://www.statsdirect.com/help/regression_and_correlation/logi.htm
  - The H-L chi-square statistic is **expected to have *p* > .05**, indicating that the model adequately fits the data
- In SPSS, Options → "Hosmer-Lemeshow goodness of fit".
  - Using the Entry method,

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.477	7	.485

Reporting the LR Results in a Table

Predictors	B	S.E.	*p*	Exp(B)	[95% CI]
Self-efficacy condition	0.88	0.51	.083	2.42	[0.89, 6.56]
Covar_HS math score	0.09	0.02	< .001	1.10	[1.05, 1.14]
Intercept	-6.12	1.28	< .001	.002	
Test	χ^2	*df*	*p*		
Overall model evaluation					
Likelihood-ratio test	44.20	2	< .001		
Goodness-of-fit test					
Hosmer & Lemeshow	6.48	7	.485		

Notes. Pseudo R^2 = 0.36 (Cox & Snell), 0.48 (Nagelkerke).

Independent variables: Self-efficacy condition group (1= Experimental, 0 = Control); High school math scores.

Dependent variable: College algebra grades (1 = A or B, 0 = Less than B).

Reporting Results: Assessing Predictors Using Odds Ratio

- Based on the SPSS results for the logistic regression using students' Self-efficacy condition group (1 = Experimental, 0 = Control) and high school math scores as predictors,
- Controlling for *high school math scores*, the predicted odds for students in the Self-efficacy (SE) *experimental* group (coded 1) to get A or B (coded 1; vs. "< B" = 0) on the college math course are 2.42 times as high as (or **1**.42 times high**er** than) that for students in the SE *control* group (coded 0) . The group difference was non- or marginally significant, p = .083, Odds Ratio (OR) = 2.42, 95% CI = 0.89-6.56.
- Controlling for SE *condition group*, as his/her high school math score increases by one point, the predicted odds of getting A or B (vs. "< B") increases significantly by 10%, p < .001, OR = 1.10, 95% CI = 1.05-1.14.

Notes on LR

- Popular in education, health, or business areas.
  - Used for item response theory (IRT) measurement models.
- (Optional) Validating the LR model with classification rates (possibly with additional ROC).
- FYI: ROC Curve:

  https://case.truman.edu/files/2015/06/SPSS-Logistic-Regression.pdf
  - **sensitivity** (True positive) vs. **specificity** (True negative)

(FYI only) Reporting Classification Rate

- ❑ The classification rate is an indicator of the predictive success by the model.
  - The model correctly classifies 37 participants who got *grades A or B*, but misclassifies 13 others.
    - ❖ Namely, the model correctly classifies 74% of the cases (*true positive*).
  - The model correctly classifies 41 participants who got *grades less than B*, but misclassifies 9 others.
    - ❖ The model correctly classifies 82% of the cases (*true negative*).
  - The overall accuracy of classification is 78%.
  - By contrast, the overall accuracy of classification based on the empty model using the intercept (constant) is lower—50%.

Classification Table[a]

			Predicted		
			DV_College Algebra		Percentage Correct
	Observed		0 < B	1= A or B	
Step 1	DV_College Algebra	0 < B	41	9	82.0
		1= A or B	13	37	74.0
	Overall Percentage				78.0

a. The cut value is .500

Including two predictors

Classification Table[a,b]

			Predicted		
			DV_1 AB vs 0 less		Percentage Correct
	Observed		Less than B	A or B	
Step 0	DV_1 AB vs 0 less	Less than B	0	50	.0
		A or B	0	50	100.0
	Overall Percentage				50.0

a. Constant is included in the model.

b. The cut value is .500

Including the intercept only