Logistic Regression

Content

1

What is logistic regression?

2

Types of logistic regression

3

Worked example on binary logistic regression

What is Logistic Regression?

Like the multiple regression, logistic regression is a statistical analysis used to examine relationships between independent variables (predictors) and a dependant variable (criterion)

The main difference is in logistic regression, the criterion is *nominal* (predicting group membership). For example, do age and gender predict whether one signs up for swimming lessons (yes/no)

Types of Logistic Regression

- There are primarily 2 types of logistic regression: (1) Binary and (2) Multinomial models. The difference lies in the types of the criterion variable
- Binary logistic regression is for a dichotomous criterion (i.e., 2-level variable)
- Multinomial logistic regression is for a multicategorical criterion (i.e., a variable with more than 2 levels)
- This set of slides focuses on <u>binary logistic regression</u>

Example...

A researcher would like to find out if the three predictors can predict successful enrolment into the Masters of Psychology program at JCU. The researcher recruited 30 participants who applied for the program, and asked them the following questions:

1. Interest in the Masters of Psychology program (rated 1-100)
2. Average overall score from a previous degree (scored 1-100)
3. Holding a psychology degree (yes/no)
4. Successful enrolment (yes/no)

A binary logistic regression was then conducted.

Note that sample size of 30 was used only for illustration purposes, an actual study would require larger sample size!

Location of SPSS Data Files for Practice

Example SPSS data for practice are available on **LearnJCU**:

Log in to LearnJCU -> Organisations -> Learning Centre JCU Singapore -> Learning Centre -> Statistics and Maths -> SPSS Data for Practice

Assumptions Testing

01
Multicollinearity

02
Independence of errors

03
Logit Linearity

04
Outliers

Assumptions Testing

Please refer to the SPSS guide on **Multiple Regression** of how to conduct the four assumption tests at https://www.jcu.edu.sg/current-students/student-support-services/learning-support/statistics-and-mathematics-support

Assumptions Testing

01
Multicollinearity

03
Logit Linearity

02
Independence of errors

04
Outliers

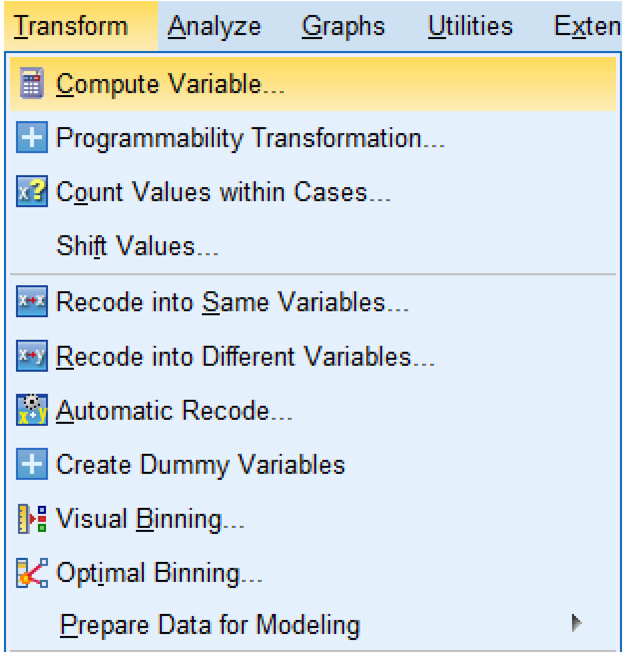
Assumptions Testing: Logit Linearity

This is an assumption that the relationship between each continuous predictor and a criterion is linear.

- *Interest* and *PreviousScore* are continuous, thus they have to be tested for this assumption.
- *PsychDegree* is categorical, hence it is not requited to be tested.

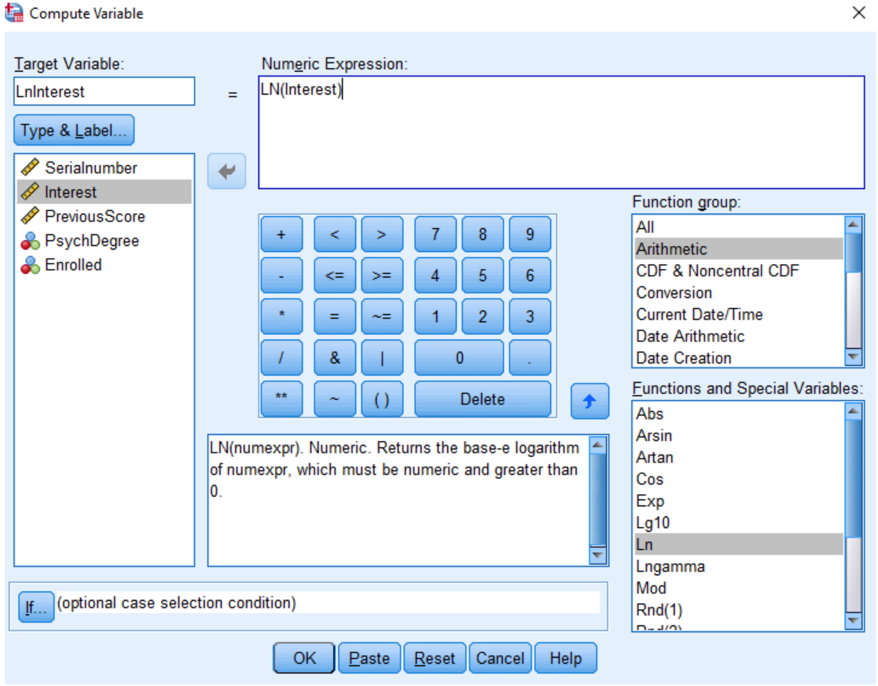
To test for this, we first need to create new variables in our dataset: Logit functions of the continuous IVs

- Transform → Compute Variable



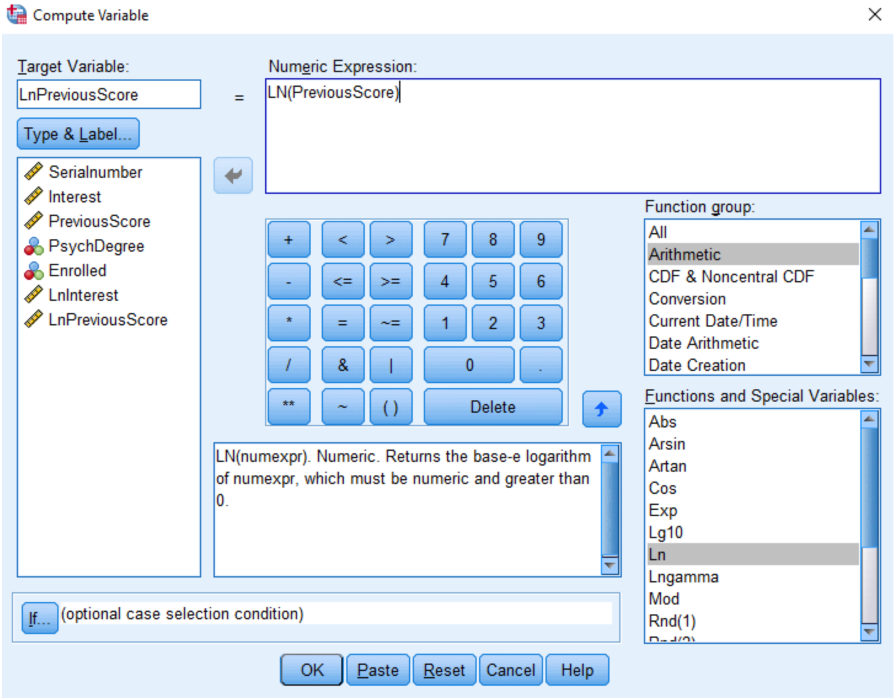
Assumptions Testing: Logit Linearity

1. Select ‘Arithmetic’ under Function group, and double click on ‘Ln’ under Functions and special variables
2. LN should appear under Numeric Expression
3. To create the logit expression of the first continuous variable (*Interest*), double click on Interest
4. Name the target variable – <u>LnInterest</u>
5. OK



Assumptions Testing: Logit Linearity

Repeat the procedures, this time creating the logit function of the other continuous variable (*PreviousScore*)



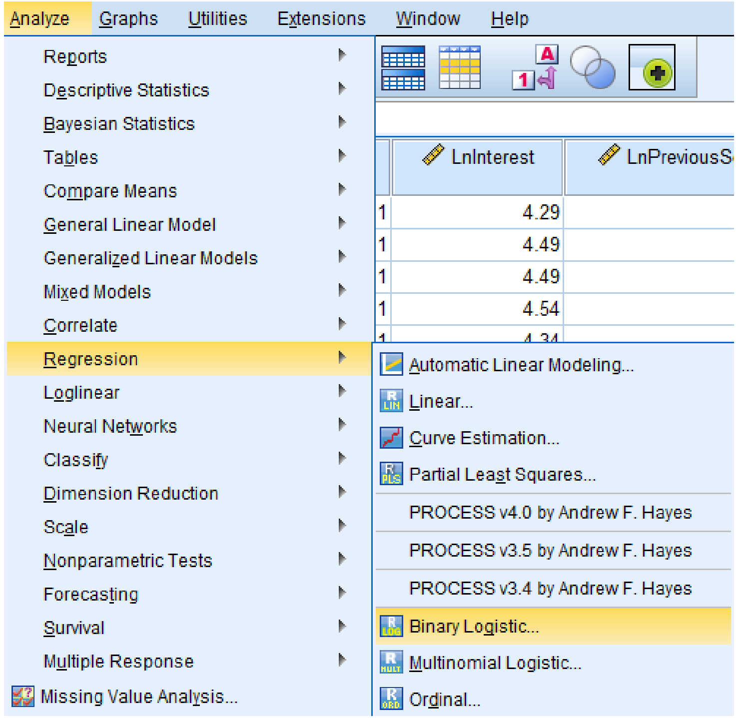
Assumptions Testing: Logit Linearity

You will see these 2 new variables in your dataset

Serialnumber	Interest	PreviousScore	PsychDegree	Enrolled	LnInterest	LnPreviousScore
1	73	50	1	1	4.29	3.91
2	89	50	1	1	4.49	3.91
3	89	50	1	1	4.49	3.91
4	94	50	1	1	4.54	3.91
5	77	50	1	1	4.34	3.91
6	65	60	1	2	4.17	4.09
7	69	60	1	2	4.23	4.09
8	55	55	1	2	4.01	4.01
9	81	60	1	2	4.39	4.09
10	75	70	1	1	4.32	4.25
11	69	70	1	1	4.23	4.25
12	70	65	1	1	4.25	4.17
13	93	68	1	1	4.53	4.22
14	79	69	1	1	4.37	4.23
15	70	70	1	1	4.25	4.25
16	90	89	1	2	4.50	4.49
17	73	75	1	2	4.29	4.32
18	80	80	1	2	4.38	4.38
19	86	79	1	2	4.45	4.37
20	78	78	1	2	4.36	4.36
21	82	77	2	2	4.41	4.34
22	81	68	2	1	4.39	4.22
23	78	70	2	1	4.36	4.25
24	76	71	2	1	4.33	4.26
25	96	80	2	2	4.56	4.38
26	72	68	2	1	4.28	4.22
27	65	75	2	2	4.17	4.32
28	66	77	2	2	4.19	4.34
29	75	80	2	2	4.32	4.38
30	70	82	2	2	4.25	4.41

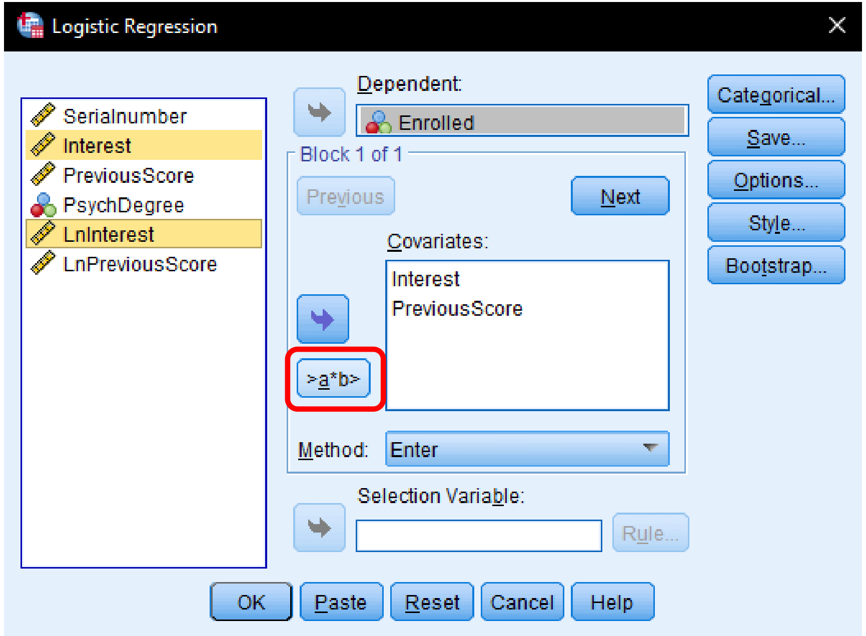
Assumptions Testing: Logit Linearity

To conduct the assumption test for logit linearity, go to **Analyze -> Regression -> Binary Logistic**



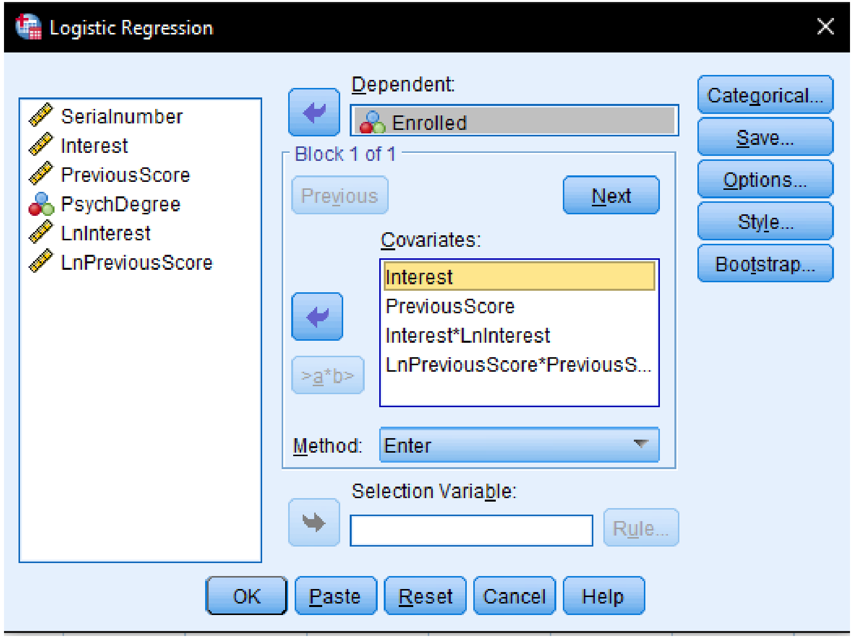
Assumptions Testing: Logit Linearity

1. Move ‘ Enrolled’ into the Dependent box
2. Move ‘Interest’ and ‘PreviousScore’ into the Covariates box
3. Holding the Ctrl key, then select ‘Interest’ and ‘LnInterest’, and click on ‘>a*b>’ to enter the interaction term into the Covariates box
4. Repeat Step 3 for ‘PreviousScore’ and ‘LnPreviousScore’



Assumptions Testing: Logit Linearity

- You should have 4 Covariates in total
- Click OK



Assumptions Testing: Logit Linearity

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Interest	-8.430	6.888	1.498	1	.221	.000
	PreviousScore	-8.936	5.091	3.081	1	.079	.000
	Interest by LnInterest	1.550	1.282	1.463	1	.227	4.712
	LnPreviousScore by PreviousScore	1.761	.989	3.172	1	.075	5.821
	Constant	230.910	129.181	3.195	1	.074	1.918E+100

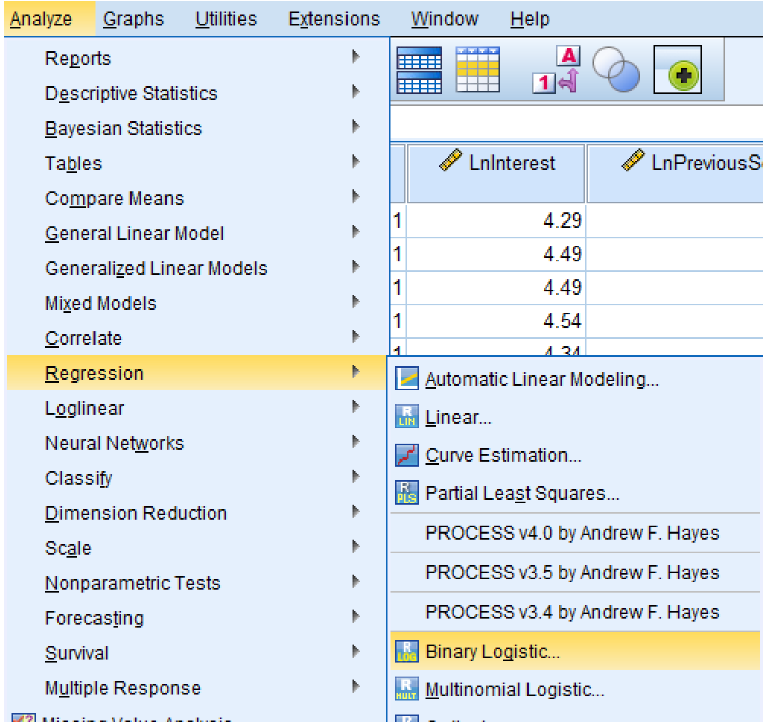
a. Variable(s) entered on step 1: Interest, PreviousScore, Interest * LnInterest , LnPreviousScore * PreviousScore .

Since the *p* values of the interaction terms are above .05, we conclude that the assumption for logit linearity is *not* violated

Logistic Regression

Now to conduct the main analysis…

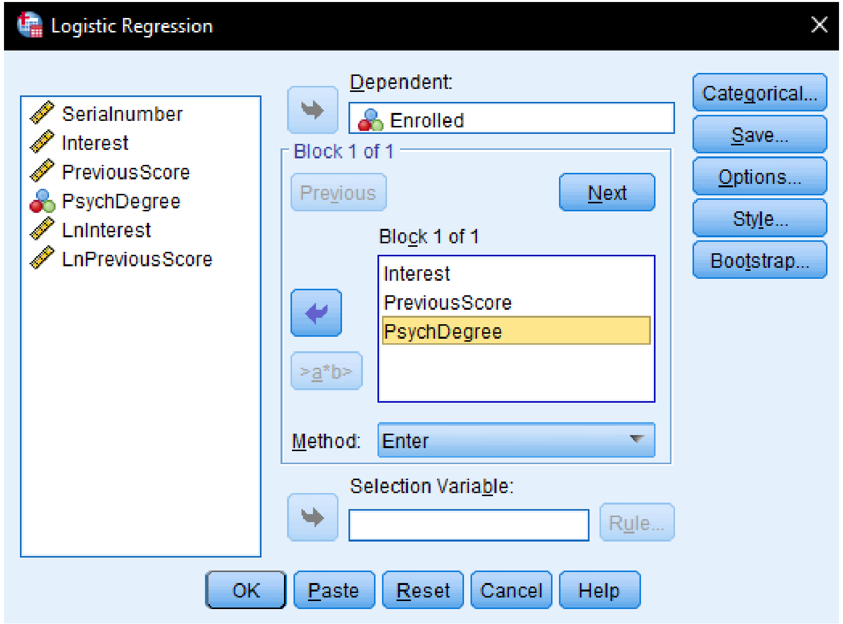
Analyze -> Regression -> Binary Logistic



Logistic Regression

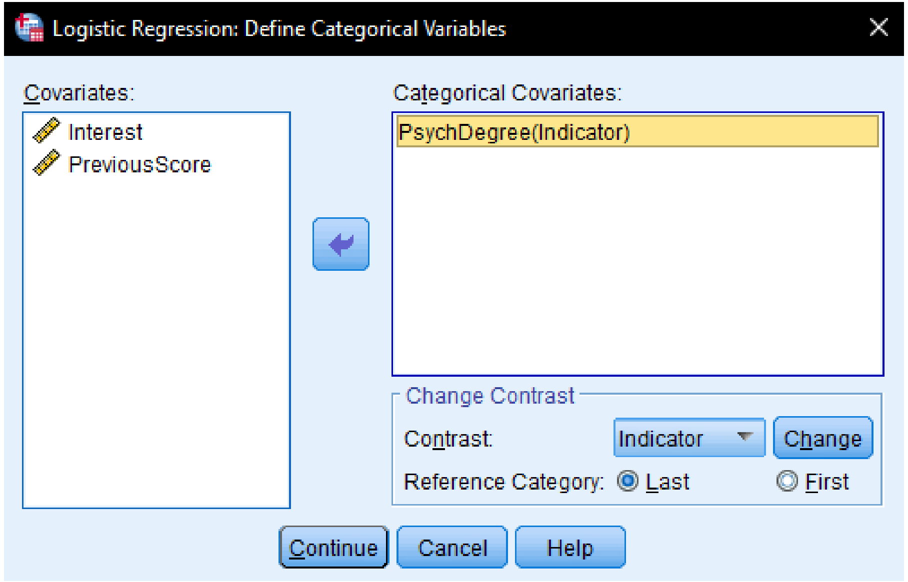
1. Move ‘Enrolled’ into the Dependent box
2. Move ‘Interest’, ‘PreviousScore’ and ‘PsychDegree’ into the Covariates box

Note that ‘PsychDegree’ is a categorical variable.



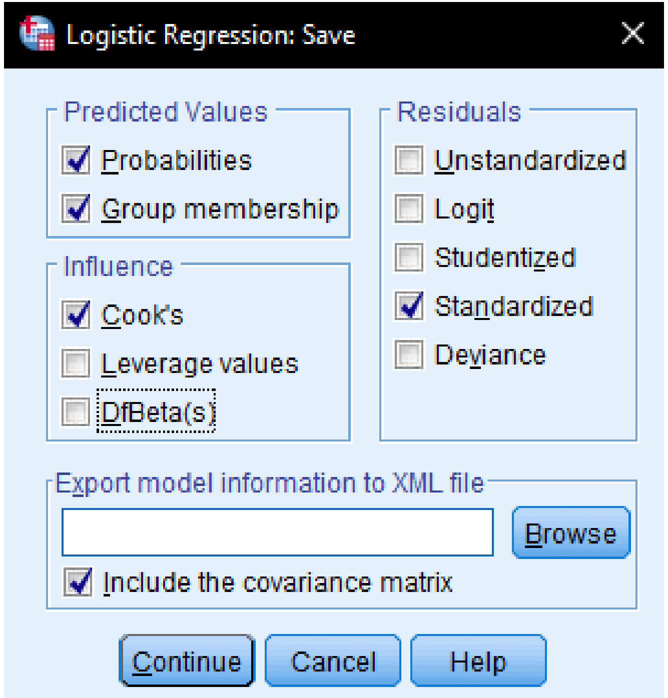
Logistic Regression

3. Click on ***Categorical***
4. Select ‘PsychDegree’ as a categorical covariate
5. Continue



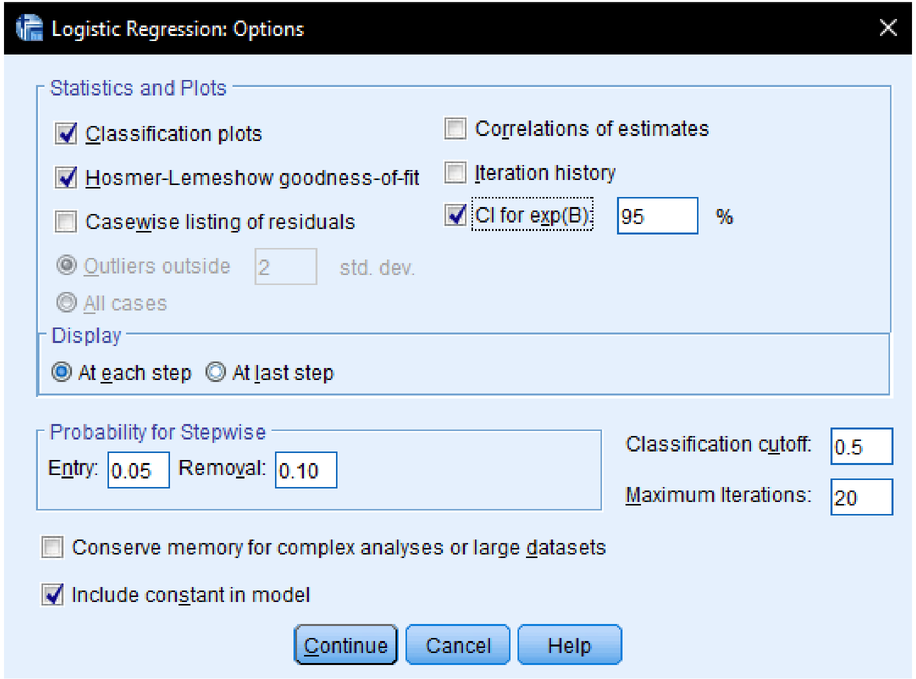
Logistic Regression

6. Click on **Save**
7. Select *Probabilities*, *Group membership*, *Cook's* (this can be used to screen for outliers), and *Standardized Residuals*
8. Continue



Logistic Regression

9. Click on **Options**
10. Select *Classification plots*, *Hosmer-Lemeshow goodness-of-fit*, and *CI for exp(B)*
11. Continue, and OK



Assumptions Testing: Outliers

01
Multicollinearity

03
Logit Linearity

02
Independence of errors

04
Outliers

Assumptions Testing: Outliers

- Outliers can be tested together with the main analysis
- Looking at the dataset, *Cook's distance* is added as a new variable
- Since all the values are < 1, we conclude that there are no outliers

Serialnumber	Interest	PreviousScore	PsychDegree	Enrolled	LnInterest	LnPreviousScore	PRE_1	PGR_1	COO_1	ZRE_1
1	73	50	1	1	4.29	3.91	.07239	1	.00892	-.27935
2	89	50	1	1	4.49	3.91	.02224	1	.00157	-.15083
3	89	50	1	1	4.49	3.91	.02224	1	.00157	-.15083
4	94	50	1	1	4.54	3.91	.01524	1	.00092	-.12441
5	77	50	1	1	4.34	3.91	.05423	1	.00565	-.23946
6	65	60	1	2	4.17	4.09	.41803	1	.23694	1.17990
7	69	60	1	2	4.23	4.09	.34547	1	.24950	1.37645
8	55	55	1	2	4.01	4.01	.41044	1	.68618	1.19849
9	81	60	1	2	4.39	4.09	.17314	1	.63564	2.18530
10	75	70	1	1	4.32	4.25	.62296	2	.15271	-1.28539
11	69	70	1	1	4.23	4.25	.72399	2	.28544	-1.61960
12	70	65	1	1	4.25	4.17	.52140	2	.11212	-1.04375
13	93	68	1	1	4.53	4.22	.23053	1	.08954	-.54736
14	79	69	1	1	4.37	4.23	.50840	2	.10924	-1.01695
15	70	70	1	1	4.25	4.25	.70834	2	.25187	-1.55840
16	90	89	1	2	4.50	4.49	.91629	2	.01731	.30226
17	73	75	1	2	4.29	4.32	.81121	2	.02494	.48242
18	80	80	1	2	4.38	4.38	.84817	2	.02401	.42310
19	86	79	1	2	4.45	4.37	.74983	2	.06942	.57761
20	78	78	1	2	4.36	4.36	.82545	2	.02594	.45985
21	82	77	2	2	4.41	4.34	.60281	2	.10282	.81172
22	81	68	2	1	4.39	4.22	.27912	1	.07000	-.62226
23	78	70	2	1	4.36	4.25	.40203	1	.10982	-.81996
24	76	71	2	1	4.33	4.26	.47936	1	.14367	-.95955
25	96	80	2	2	4.56	4.38	.45503	1	.57961	1.09438
26	72	68	2	1	4.28	4.22	.43649	1	.14852	-.88010
27	65	75	2	2	4.17	4.32	.80317	2	.04636	.49504
28	66	77	2	2	4.19	4.34	.83887	2	.03130	.43827
29	75	80	2	2	4.32	4.38	.80807	2	.03018	.48736
30	70	82	2	2	4.25	4.41	.89505	2	.01390	.34243

Logistic Regression: Results

Classification Table[a,b]

Observed			Predicted: Enrolled No	Predicted: Enrolled Yes	Percentage Correct
Step 0	Enrolled	No	0	15	.0
		Yes	0	15	100.0
	Overall Percentage				50.0

a. Constant is included in the model.

b. The cut value is .500

- The purpose of logistic regression is thus to find out if the *prediction accuracy* of the model can be improved by predictor variables
- This table shows the regression model with no predictors involved (block 0). **This model (at Step 0) can correctly predict if someone successfully enrolled 50% of the time.**

Logistic Regression: Results

In block 1, all the predictors were entered simultaneously

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	11.598	3	.009
	Block	11.598	3	.009
	Model	11.598	3	.009

A *p* value <.05 suggests that, overall, the predictors significantly improved the prediction accuracy of the model

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	29.991[a]	.321	.428

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

R square values of the regression model

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	11.809	8	.160

Measure of model fit. A *p* value >.05 suggests a good model fit

Logistic Regression: Results

Classification Table[a]

Observed			Predicted: Enrolled 0	Predicted: Enrolled No	Percentage Correct
Step 1	Enrolled	0	10	5	66.7
		No	5	10	66.7
	Overall Percentage				66.7

a. The cut value is .500

In Step 1, the addition of the predictors resulted in the model being able to predict successful enrolment 66.7% of the time (compared to 50% in block 0; 16.7% improvement !)

Logistic Regression: Results

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Interest	-.077	.054	2.051	1	.152	.926	.833	1.029
	PreviousScore	.160	.066	5.972	1	.015	1.174	1.032	1.335
	PsychDegree(1)	.668	1.003	.443	1	.506	1.950	.273	13.938
	Constant	-5.611	4.931	1.295	1	.255	.004		

a. Variable(s) entered on step 1: Interest, PreviousScore, PsychDegree.

- This table tells us which predictors are significant. Only Previous score is a significant predictor ($p < .05$)
- In logistic regression, Exp(B) is commonly used to interpret results, and is expressed as **an odds ratio**
- In other words, an increase of 1 unit in Previous score results in a 17.4% more chance of enrolling in the masters program (1.174 – 1 = .174, meaning .174 above 1)
- The other statistics (e.g., B, Wald, 95% CI) can also be reported in the writeup

Write-Up

An example write-up can be found on page 228 in

Allen, P., Bennett, K., & Heritage, B. (2019). *SPSS Statistics: A Practical Guide* (4th ed.). Cengage Learning.

Questions?

learningcentre-singapore@jcu.edu.au