Introduction to Applied Statistical Methods

Practical 10

In this practical we will look at the effect of different variables on binary outcomes. You are expected to learn the following using SPSS:

- Calculate risk ratio and odds ratios and chi-squared tests for binary outcomes with a single predictor and be able to interpret these ratios
- Use logistic regression to calculate odds ratios adjusted for possible confounding variables
- Test the fit of a logistic regression model

The dataset used to demonstrate these issues is the National Child Development Study (NCDS) data Practical_10a_data.sav

Background:

The data set consists of a sample of n=1000 sixteen years old children. Several continuous and categorical variables have been recorded:

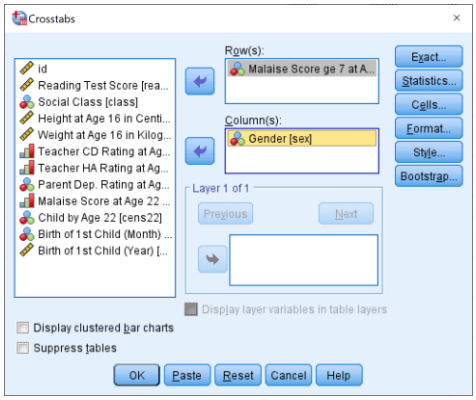
- weight – weight in kg at age 16
- sex – gender of child (0 = male, 1 = female)
- height – height in cm at age 16
- class – registrar general's classification of social class (7 categories)
- malcat – Malaise score (0=no, 1=yes)
- reading – Reading score
- scd – CD rating
- sha – Hyperactivity (HA) rating

Task 1

We want to compare the effects of a child's gender on the **risk** of depression at the age of 22 measured as the dichotomised malaise score where 0 = No malaise and 1 = Yes malaise. Calculate your answer to 3 decimal places

Use the appropriate SPSS command to create a cross tabulation of sex and malaise.

Descriptive Statistics -> Cross tabs



Malaise Score ge 7 at Age 22 0 = No, 1=Yes * Gender Crosstabulation

Count

		Gender		
		Male	Female	Total
Malaise Score ge 7 at Age 22 0 = No, 1=Yes	No	465	437	902
	Yes	26	72	98
Total		491	509	1000

Risk of having malaise in males = **26 ÷ 491 = 0·053**

Risk of having malaise in females = **72 ÷ 509 = 0·141**

We can compare the risk for each of the groups using the risk ratio.

(Risk when male) ÷ (Risk when female) = 0·053 ÷ 0·141 = 0.376

Those who were male had 0.38 times the risk of having malaise at 22 compared to those who were female.

We can also present risk as a percentage using the following formula **% decrease** = (RR - 1) x 100
Males had a 62.4% reduction in risk of malaise at 22 compared to females.

You could have also calculated this as

(Risk when female) ÷ (Risk when male) = 0.141 ÷ 0.053 = 2.660

Those who were female had 2.66 times the risk of having malaise at 22 compared to those who were male.

We can also present risk as a percentage using the following formula **% increase** = (RR - 1) x 100
Those who were female had a 166% increase in risk of having malaise at 22 compared to males.

Task 2.

We want to compare the effects of a child's gender on the odds of depression measured as the dichotomous malaise score where 0 = No malaise and 1 = Yes malaise. Calculate your answer to 3 decimal places (3dp)

Malaise Score ge 7 at Age 22 0 = No, 1=Yes * Gender Crosstabulation

Count

		Gender		
		Male	Female	Total
Malaise Score ge 7 at Age 22 0 = No, 1=Yes	No	465	437	902
	Yes	26	72	98
Total		491	509	1000

Odds of having malaise in males = **26 ÷ 465 = 0·056**

Odds of having malaise in females = **72 ÷ 437 = 0·165**

We can compare the odds for each of the groups using the odds ratio.

(Odds when male) ÷ (Odds when female) = 0·056 ÷ 0·165 = 0.339

So, the odds of having malaise at 22 when the child is male is about a 1/3 of the odds of females having malaise at 22.

You could have also calculated this as

(Odds when female) ÷ (Odds when male) = 0.165 ÷ 0.056 = 2.946

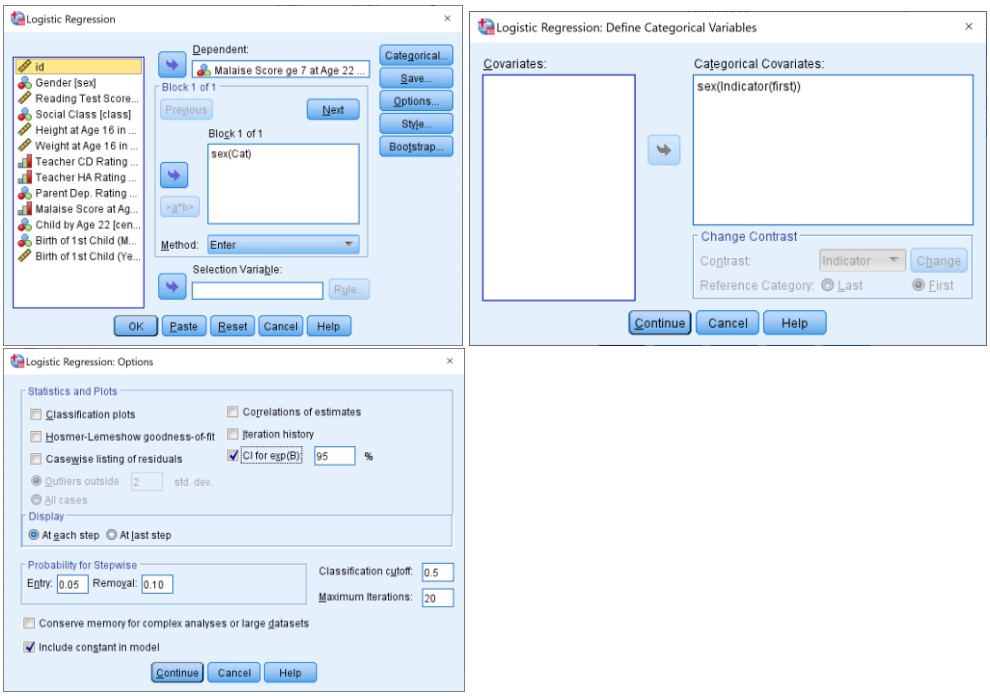
So, the odds of having malaise at 22 when the child is female is about 3 times of the odds of a male child.

Task 3

Is there an association between a childs gender and them suffering malaise at age 22?

Conduct a Binary logistic regression Regression→Binary Logistic from the Analyse menu

- Move your binary coded variable (malcat) to dependent
- Move Gender to block 1 of 1
- Identify gender as a 'categorical' variable and make the first category the reference category (i.e. 0 = male). Make sure method is set to enter and that you click on 'Change'
- Click options and check CI for Exp(B)
- Click continue
- Click OK



Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	23.009	1	.000
	Block	23.009	1	.000
	Model	23.009	1	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square

1	618.323[a]	.023	.048

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted		
			Malaise Score ge 7 at Age 22 0 = No, 1=Yes		Percentage
	Observed		No	Yes	Correct
Step 1	Malaise Score ge 7 at Age 22 0 = No, 1=Yes	No	902	0	100.0
		Yes	98	0	.0
	Overall Percentage				90.2

a. The cut value is .500

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Gender(1)	1.081	.238	20.565	1	.000	2.947	1.847	4.701
	Constant	-2.884	.202	204.794	1	.000	.056		

a. Variable(s) entered on step 1: Gender.

H_0: there is no association between gender and malaise at age 22

H_1: there is an association between gender and malaise at age 22

There was a significant improvement to the constant only model. (chi-square=23.009, df=1, p<0.001. Nagelkerke R^2 = 4.8% of the variation in malaise can be explained by the model including gender. The correct classification rate did not change from 90.2.

The odds ratio for gender is 2.95 and 1 is not included within the 95% confidence intervals, suggesting a statistically significant result (Wald=20.565, df= 1, p<0.001,).

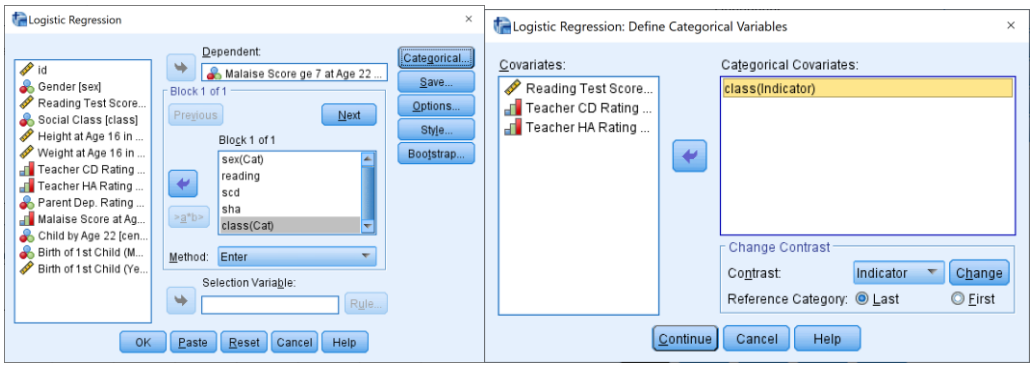
The odds of being of having malaise at 22 for females is 2.947 (95% CI 1.847, 4.701) times that of Males.

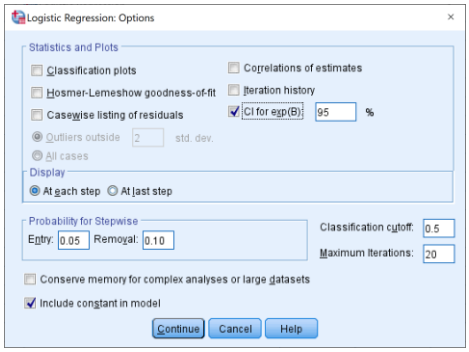
Task 4

Estimate the effect of gender on depression, as measured by the dichotomised malaise score, controlling for reading score, CD rating, HA rating and class.

Conduct a Binary logistic regression Regression→Binary Logistic from the Analyse menu

- Move your binary coded variable (malcat) to dependent box
- Move gender, reading, scd, sha and class to covariates
- Make sure method is set to enter
- Identify gender and class as a 'categorical' variable and make genders reference category 'first' (i.e. 0 = male) and class leave as 'last'. Make sure method is set to enter and that you click on 'Change'
- Click continue
- Click options and check CI for Exp(B)
- Click continue
- Click OK





Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	51.047	10	.000
	Block	51.047	10	.000
	Model	51.047	10	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	590.285[a]	.050	.105

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

Observed			Predicted: Malaise Score ge 7 at Age 22 0 = No, 1=Yes		Percentage Correct
			No	Yes	
Step 1	Malaise Score ge 7 at Age 22 0 = No, 1=Yes	No	901	1	99.9
		Yes	97	1	1.0
	Overall Percentage				90.2

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Gender	1.278	.252	25.685	1	.000	3.590	2.190	5.885
	Reading Test Score	-.035	.017	4.376	1	.036	.966	.935	.998
	Teacher CD Rating at Age 11	.110	.086	1.630	1	.202	1.116	.943	1.322
	Teacher HA Rating at Age 11	.197	.108	3.321	1	.068	1.218	.985	1.505
	Social Class			6.606	6	.359			
	Social Class(1)	-1.561	.857	3.319	1	.068	.210	.039	1.126
	Social Class(2)	-.837	.520	2.584	1	.108	.433	.156	1.201
	Social Class(3)	-1.194	.619	3.721	1	.054	.303	.090	1.019
	Social Class(4)	-.780	.475	2.696	1	.101	.459	.181	1.163
	Social Class(5)	-.773	.514	2.262	1	.133	.462	.169	1.264
	Social Class(6)	-.218	.621	.124	1	.725	.804	.238	2.714
	Constant	-2.069	.666	9.653	1	.002	.126		

a. Variable(s) entered on step 1: Gender, Reading Test Score, Teacher CD Rating at Age 11, Teacher HA Rating at Age 11, Social Class.

H_0: there is no association between gender and malaise at age 22 adjusting for reading score, scd, sha and class

H_1: there is an association between gender and malaise at age 22 adjusting for reading score, scd, sha and class

There was a significant improvement to the constant only model. (chi-square=51.047, df=10, p<0.001. Nagelkerke R^2 = 10.5% of the variation in malaise can be explained by the model including gender, reading score, scd, sha and class. The correct classification rate did not change from 90.2.

The odds ratio for gender is 3.59 and 1 is not included within the 95% confidence intervals suggesting a statistically significant result (Wald=25.685, df= 1, p<0.001).

This means adjusting for potential confounders, that the odds of having malaise at 22 for females is 3.590 (95% CI 2.190, 5.885) times that of Males.

Task 5

The researcher would like to understand the effect of reading score on depression, as measured by malaise at 22 in the adjusted logistic model. What is the relationship between the odds ratios and the coefficients in the model?

The adjusted odds ratio for the reading score is 0.966 and 1 is not included within the 95% confidence intervals, suggesting a statistically significant event, (Wald=4.376, df= 1, p=0.036).

As reading score increases by one unit the odds of depression is expected to decrease by 0.966 (95% CI 0.935, 0.998), keeping all other variables in the model constant. (The coefficient β can be described as the change in log odds on dichotomous malaise score for an increase of one unit in the reading score).

Task 6

What does the Hosmer-Lemeshow test tell us about the fit of the adjusted logistic model?

- Click options and check Hosmer and Lemeshow

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	10.075	8	.260

P value =0.260 of Hosmer and Lemeshow Test is higher than the 0.05 level of significance thus is not significant and indicates a good fit for the model.

We are now going to use the Practical_10b_data.sav

- To make predictions from a logistic regression model
- To understand classification

The file contains two variables. The first diagnosis is a binary outcome, which tells us whether each patient has developed a disease (coded 0 for no and 1 for yes). Health score is a continuous variable measuring the patients fitness.

Task 7

We wish to investigate the relationship between health score and having the disease.

Run a binary logistic regression to investigate this relationship.

Analyse→Regression→Binary logistic

- Move Diagnosis to dependent
- Move Health Score to covariates
- Under Options select Hosmer and Lemeshow Goodness of fit and C.I. for exp(β)
- Click Continue
- Under save check probabilities
- Click continue
- Click OK

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	94.769	1	.000
	Block	94.769	1	.000
	Model	94.769	1	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	271.750[a]	.271	.384

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

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	HealthScore	-.708	.092	59.550	1	.000	.492	.411	.590
	Constant	1.785	.339	27.786	1	.000	5.959		

a. Variable(s) entered on step 1: HealthScore.

The odds ratio for health score is 0.492 and 1 is not included within the 95% confidence intervals, indicating a statistically significant result, (Wald=59.550, df= 1, p<0.001) .

This means the odds of developing the disease decreases by 0.492 (95% CI 0.411, 0.590) as the Health Score increases by one unit (The coefficient value of -0.708 is the change in log odds on diagnosis for an increase of one unit in the Health score).

The risk of developing disease almost halves as health score increases by one unit.

Task 8

What does the Hosmer-Lemeshow test tell us about the fit of the adjusted logistic model?

- Click options and check Hosmer and Lemeshow

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	7.173	6	.305

P value =0.305 of Hosmer and Lemeshow Test is higher than the 0.05 level of significance thus is not significant and indicates a good fit for the model.

Task 9

Suppose a clinician wishes to use Health Score as a screening tool, so that those who are likely to have the disease based on their health score will be sent for further tests. They are interested in making sure that most of those with the disease will receive the additional tests.

Looking at the classification table comment on how effective a screening tool the model is for the clinician's purposes.

Classification Table[a]

			Predicted		
			diagnosis		Percentage
	Observed		0	1	Correct
Step 1	diagnosis	0	186	24	88.6
		1	45	45	50.0

Overall Percentage			77.0

a. The cut value is .500

The cut off is the probability of having the disease (i.e. the probability of success), the 0.5 here represents a 50% chance to get the disease and 50% chance to not get the disease, based on this probability the cut-off threshold probability of 0.5 is used (default), 88.6% are correctly predicted to not to develop the disease and 50% are correctly predicted to be positive to develop the disease.

We would need to either think about changing the cut-off so we can better predict the number of positives for the disease or improve the model to give better predictive capability. The probability we have chosen here is arbitrary. Generally, where you set the cut-off will depend on the relative importance of the probability of detecting true event cases (sensitivity) and the probability of misclassifying non-events as events (false positive rate).

Task 10

What is the probability of patient being diagnosed with the disease if their health score is 8 compared to a patient with a health score of 2?

L=1.785 - 0.708Healthscore

L=1.785 – 0.708 x 8

L=-3.879

P=exp(L)/1 +exp(L)

P=exp(-3.879)/ 1 + exp(-3.879)

P=0.02067/1.02067

P=0.02

L=1.785 – 0.708 x 2

L=0.369

P=exp(0.369)/ 1 + exp(0.369)

P=1.44629/2.44629

P=0.59

There is 2% probability that a patient with a Health score of 8 will be diagnosed with the diseased compared to 59% if the patients' health score was 2.