statstutor

community project

encouraging academics to share statistics support resources

stcp-marshall-logisticS

The following resources are associated:
'Summarising categorical variables in SPSS', 'Chi-squared in SPSS' and the 'Titanic.sav' dataset

Logistic regression in SPSS

Dependent (outcome) variable: Binary

Independent (explanatory) variables: Any

Common Applications: Logistic regression allows the effect of multiple independents on one binary dependent variable to be tested. It is predominantly used to assess relationships between the binary dependent variable and each independent variable whilst controlling for the other independent variables but also produces an equation (model) which can be used for prediction. It is similar to multiple linear regression but instead of predicting the value of the dependent variable, it can be used to calculate the probability of an event occurring. For example, when insurance companies calculate the insurance cost for a driver, they use information from the driver to calculate the probability of the driver having a crash. The higher the probability of a crash, the higher the cost of the insurance.

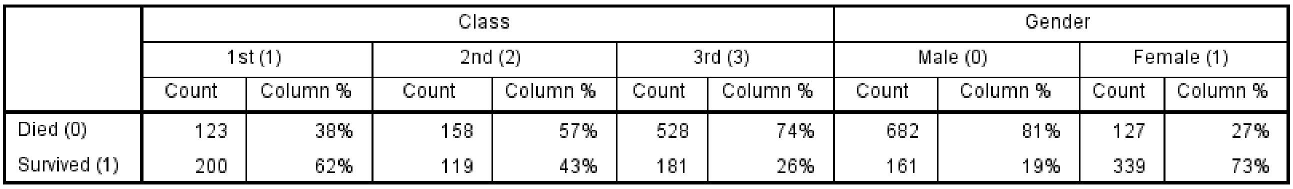
Data: The dataset *Titanic.sav* contains data on 1309 passengers and crew who were on board the ship 'Titanic' when it sank in 1912. Only 38.2% of those on board survived. Age and fare are continuous and the others are categorical. Class has three categories (1=1st, 2=2nd, 3=3rd) and passengers were grouped into three countries of residence (0=American, 1=British, 2=Other).

Variable name	pclass	survived	Residence	age	fare	Gender
Name	**Class of passenger**	**0 = died 1 = survived**	**Country of residence**	**Age**	**Price of ticket**	**0 = male 1 = female**
Abbott, Eugene	3	0	0	13	20.25	0
Appleton, Charlotte	1	1	2	53	51.48	1

Research question: Which variables affected survival? The dependent variable is survival and the other variables are the explanatory (independent) variables to be tested using logistic regression.

Summary Statistics

If you have a number of possible independent variables, look for associations between each categorical independent and the dependent variable using crosstabulations and Chi-squared tests (see the '*Summarising categorical variables in SPSS' and 'Chi-squared in SPSS'* resource) to help decide which variables to include in logistic regression. For example, the values below suggest that survival was more likely for females and those in 1st class.



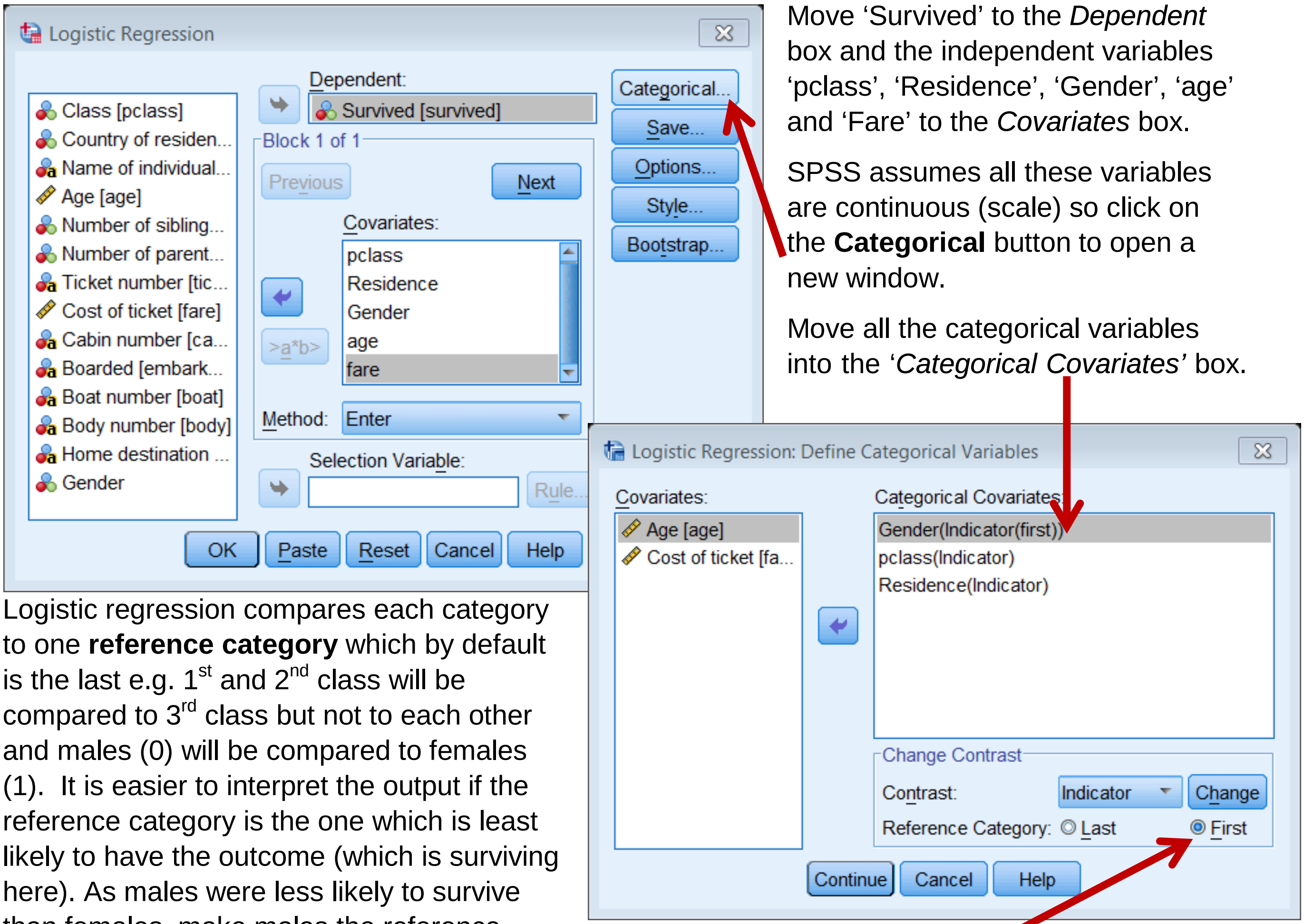
	Class						Gender			
	1st (1)		2nd (2)		3rd (3)		Male (0)		Female (1)	
	Count	Column %	Count	Column %	Count	Column %	Count	Column %	Count	Column %
Died (0)	123	38%	158	57%	528	74%	682	81%	127	27%
Survived (1)	200	62%	119	43%	181	26%	161	19%	339	73%

It is also important to look for categories with small frequencies and check whether pairs of independent variables are strongly related as both situations can cause problems.
In the associated '*Chi-squared in SPSS'* resource' there was significant evidence of an association, between nationality and survival. 56% of Americans survived compared to 32% of the British passengers and 35% of other nationalities. Logistic regression allows other values to be controlled for when assessing the relationship between nationality and survival.

Carrying out the analysis in SPSS

As the highest number (1) for the dependent variable 'Survived' indicates surviving, the output from the logistic regression procedure will compare the likelihood of survival between groups.

To run a logistic regression, go to *Analyze → Regression → Binary Logistic*



Logistic Regression

Class [pclass]
Country of residen...
Name of individual...
Age [age]
Number of sibling...
Number of parent...
Ticket number [tic...
Cost of ticket [fare]
Cabin number [ca...
Boarded [embark...
Boat number [boat]
Body number [body]
Home destination ...
Gender

Dependent: Survived [survived]

Block 1 of 1

Covariates: pclass, Residence, Gender, age, fare

Method: Enter

Selection Variable:

Logistic Regression: Define Categorical Variables

Covariates: Age [age], Cost of ticket [fa...

Categorical Covariates: Gender(Indicator(first)), pclass(Indicator), Residence(Indicator)

Change Contrast

Contrast: Indicator

Reference Category: ○ Last ● First

Move 'Survived' to the *Dependent* box and the independent variables 'pclass', 'Residence', 'Gender', 'age' and 'Fare' to the *Covariates* box.

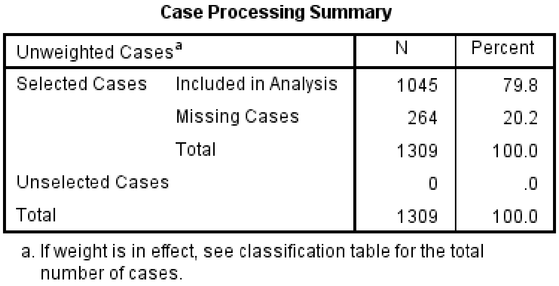
SPSS assumes all these variables are continuous (scale) so click on the **Categorical** button to open a new window.

Move all the categorical variables into the '*Categorical Covariates'* box.

Logistic regression compares each category to one **reference category** which by default is the last e.g. 1st and 2nd class will be compared to 3rd class but not to each other and males (0) will be compared to females (1). It is easier to interpret the output if the reference category is the one which is least likely to have the outcome (which is surviving here). As males were less likely to survive than females, make males the reference category by selecting 'Gender, then changing the '*Reference category*' to **First** and clicking **Change** and then click **Continue**. Click on the **Options** button menu, select the '*CI for Exp(B)*' option and **Continue**. Then run the procedure by clicking **OK.**

Interpreting the output

Whilst there are no distributional assumptions for logistic regression, it is preferable to have a decent sample size particularly if there are several independent variables. Cases are only included if there are values for every independent variable so choose variables to be included carefully. Here, 264 cases (20.2%) have been excluded as they have missing values on at least one variable.

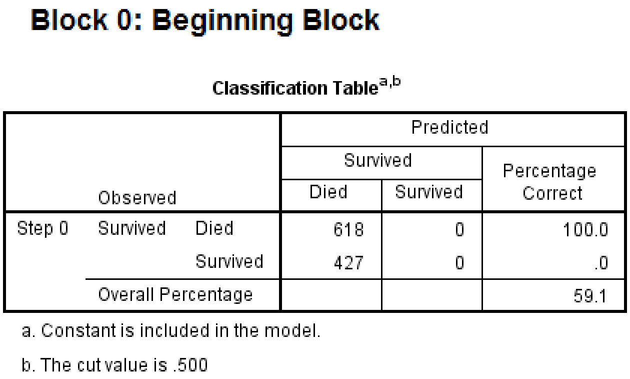


Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	1045	79.8
	Missing Cases	264	20.2
	Total	1309	100.0
Unselected Cases		0	.0
Total		1309	100.0

a. If weight is in effect, see classification table for the total number of cases.

The logistic regression process will produce an equation (model) which can be used to estimate the probability of survival for an individual using the values of the independent variables. In a basic logistic regression, two models will be compared. The full model (Block 1) contains all the selected independent variables and null model (Block 0) contains no independent variables so every individual is given the same probability of survival. In order to assess whether a model is reliable, a cut-off (usually 0.5) is used to allocate individuals to one of the outcome groups (died or survived). Here, an individual with a predicted probability of survival of more than 0.5 will be allocated to the survival group. The '**Classification Table'** compares the actual and predicted groups to assess how many would be correctly classified.



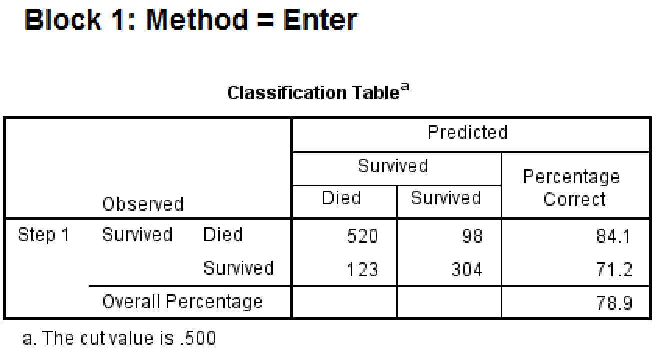
Block 0: Beginning Block

Classification Table[a,b]

			Predicted		
			Survived		Percentage Correct
Observed			Died	Survived	
Step 0	Survived	Died	618	0	100.0
		Survived	427	0	.0
	Overall Percentage				59.1

a. Constant is included in the model.

b. The cut value is .500



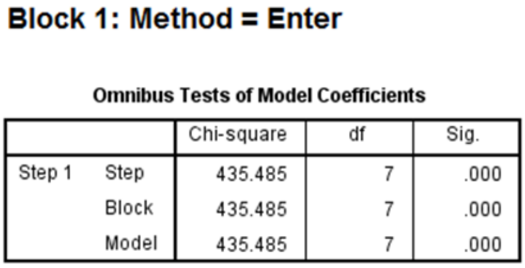
Block 1: Method = Enter

Classification Table[a]

			Predicted		
			Survived		Percentage Correct
Observed			Died	Survived	
Step 1	Survived	Died	520	98	84.1
		Survived	123	304	71.2
	Overall Percentage				78.9

a. The cut value is .500

59.1% of individuals were correctly classified using the null model (everyone was classified as dying) and 78.9% were correctly classified using the full model which is a large improvement.

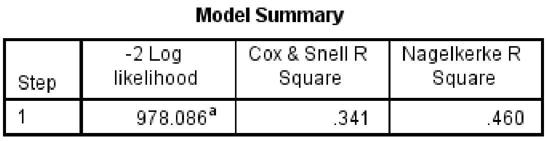


Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	435.485	7	.000
	Block	435.485	7	.000
	Model	435.485	7	.000

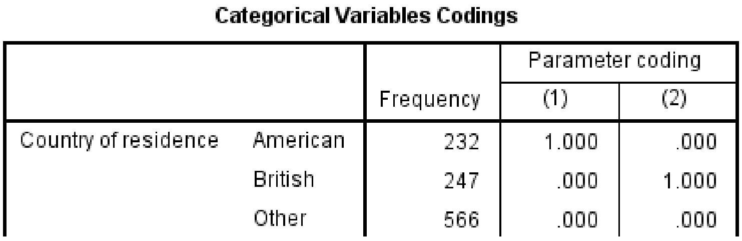
As well as assessing the individual significance of each independent variable, the full model is tested using a Likelihood Ratio (LR) test to see if it is a significant improvement (p-value < 0.05) on the null model in the 'Model' row of the '**Omnibus Tests of Model Coefficients'** table. For a basic logistic regression, all the independent variables are entered in the same block (and step) so the three rows of the table are the same. The model was statistically significant when compared to the null model, $\chi^2(7) = 435.485, p < 0.001.$



Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	978.086[a]	.341	.460

R^2 value in linear regression gives a measure of the proportion of variation of the dependent variable explained by the model. Whilst this cannot be calculated for logistic regression, 'pseudo R^2' values appear in the '**Model Summary'** table. The Nagelkerke R Square value is 0.46 so 46% of the variation in survival can be explained by the full model suggesting that predictions are fairly reliable.



Categorical Variables Codings

		Frequency	Parameter coding	
			(1)	(2)
Country of residence	American	232	1.000	.000
	British	247	.000	1.000
	Other	566	.000	.000

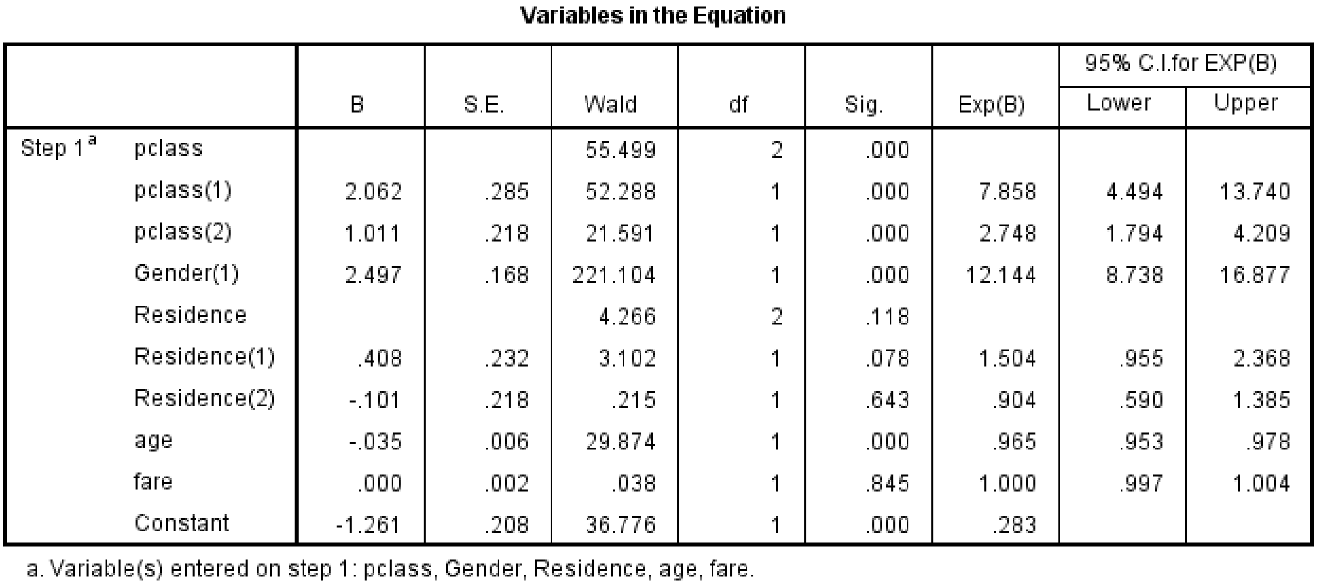
Before interpreting the '**Variables in the Equation**' table, check the '**Categorical Variables Codings'** table which clarifies the reference category and the categories labelled 1 and 2 for each categorical independent variable. Here, the reference category is 'Other', Residence(1) is American and Residence (2) is British.

The '**Variables in the Equation**' table contains the Wald test results for each independent variable after controlling for the others. The **Sig.** column contains the p-values for each variable so first look

for significant values (p-value < 0.05). If you wish to use the logistic regression model for prediction, the **B** column contains the coefficients for the model but for interpretation of significant effects, use the **Exp(B)** column which gives odds ratios.
The odds of an event happening in one group is calculated as

$$\frac{probability\ event\ happens}{probability\ event\ DOES\ NOT\ happen} = \frac{p}{1-p} \quad e.g.\ odds(dying\ for\ males) = \frac{probability\ of\ male\ dying}{probability\ of\ male\ not\ dying}$$

An odds ratio is used to compare the odds of an event happening in two groups with the reference category on the bottom e.g. $\frac{odds\ (dying\ for\ males)}{odds(dying\ for\ females)} = 12.144$. This means that the odds of a male dying are 12.144 times the odds for a female dying. An odds ratio of 1 means that both groups are equally likely to die.



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	pclass			55.499	2	.000			
	pclass(1)	2.062	.285	52.288	1	.000	7.858	4.494	13.740
	pclass(2)	1.011	.218	21.591	1	.000	2.748	1.794	4.209
	Gender(1)	2.497	.168	221.104	1	.000	12.144	8.738	16.877
	Residence			4.266	2	.118			
	Residence(1)	.408	.232	3.102	1	.078	1.504	.955	2.368
	Residence(2)	-.101	.218	.215	1	.643	.904	.590	1.385
	age	-.035	.006	29.874	1	.000	.965	.953	.978
	fare	.000	.002	.038	1	.845	1.000	.997	1.004
	Constant	-1.261	.208	36.776	1	.000	.283		

a. Variable(s) entered on step 1: pclass, Gender, Residence, age, fare.

For the categorical variables, the first p-value indicates whether there is an association between that independent variable and the dependent and the other rows compare the individual categories with the reference category. For example, class is significant as a whole and the other rows compare 1st or 2nd class to 3rd class. For significant results, next go to the **Exp(B)** column which gives the odds ratio compared to the reference category e.g. the odds of survival for those in 1st class [pclass(1)] were 7.858 times the odds for those in 3rd class. For continuous variables, it is often easier to explain the odds ratio as a percentage increase or decrease by calculating the Odds ratio - 1 For age, the odds ratio is 0.965 calculate the change in odds for a one unit increase in age by e.g. age is negative so the probability of dying decreases with age. The odds ratio represents The continuous variable age is significant and the the odds ratio os very close

Reporting logistic regression

When there are several independent variables, it is a good idea to present the detailed results in a table and a summary of the key significant results in the write up.
A logistic regression was carried out to assess the effect of class, gender, nationality, age and price of ticket on the likelihood of dying on the Titanic. The overall model was statistically significant when compared to the null model, (χ^2(7) = 435.485, p < 0.001), explained 46% of the variation of survival (Nagelkerke R^2) and correctly predicted 78.9% of cases. Class (p<0.001), gender (p<0.001) and age (p<0.001) were significant but nationality (p=0.118) and price of ticket (p=0.845) were not. The odds of dying were 12.