

4

CHAPTER

METHODOLOGY AND EMPIRICAL RESULTS

4.1 Introduction

In chapter 3, the questions were grouped according to demographic and employment and income information, an analysis was given on the grouped questions after which the raw data obtained from both data sets were reported. From the NIDS raw data sample, it is apparent that there is a difference in income between the various degrees, while it is most apparent between an honours degree and a master's/Ph.D. degree. From the Alumni raw data sample it is evident that only one income bracket placement (R25 000 – R30 000) which seems to increase as the level of education increases. It is therefore expected that the level of education may be significant for the NIDS data set and not for the Alumni data set.

Chapter 4 examines whether the demographic and employment information represent statistically significant factors affecting income and will be estimated for each of the two data sets. To determine whether these factors are significant determinants of income, a multinomial logistic regression will be used. A multinomial logistic regression is applied to examine the relationships between a categorical dependent variable and metric or categorical independent variables (Prempeh, 2009:16; Starkweather & Moske, 2011). Multiple groups are compared through a combination of binary logistic regressions. The differences in the three income groups will be studied using the multinomial logistic regression, where the analysis will compare the low income group to the high income group and the medium income group to the high income group (Greene, 1993:720-723; Starkweather & Moske, 2011). This means that, for each independent variable, there are two comparisons and, as a result, each of the three income groups will deliver a separate equation (Field, 2009:300-312; Prempeh, 2009:16). In this study, each independent variable will have two comparisons, while the coefficients for the reference group are all zero. A variety of statistical tests will also be used to test the classification accuracy and the overall fit of the model itself.

4.2 Method

Because this study includes a single categorical dependent variable and several metric and categorical independent variables, a multinomial logistic regression will be the most suitable statistical model for both the NIDS and Alumni data sets. The objective of a multinomial logistic regression model is to predict the outcome of categorical variables, and as a result renders more than two discrete outcomes (Greene, 1993:720-723). This regression model is therefore used to predict the probabilities of the different possible outcomes of a categorical dependent variable with metric and categorical independent variables. Since the multinomial logistic regression model does not assume normality, linearity and homogeneity of variance within the independent variables, it is the preferred model for this study (Greene, 1993:720-723; Field, 2009:300-312).

The multinomial logistic regression model is reliable under a number of assumptions. This model assumes that each independent variable has a single value for each case, and assumes that the dependent variable cannot be predicted absolutely from the independent variables. The model also assumes that collinearity is relatively low and that the independent variables need not be statistically independent from each other (Greene, 1993:720-723; Prempeh, 2009; Starkweather & Moske, 2011).

This study will also consider cross-tabulation where most of the material related to cross-tabulation is addressed under Chi-square. Cross-tabulation depicts how two variables inter-relate, where the Chi-square statistic is used to see if the inter-relationship has any significant relevance. The Chi-square test for independence will be used since it considers the relationship between two categorical variables, where each variable can have two or more categories (Pallant, 2011:217). According to Pallant (2011:217), the expected frequency in any cell ought to be five or larger. If this assumption is violated, a Fisher's exact probability test should rather be used which is provided as part of the chi-square output.

Throughout, the independent variables will be referred to as predictors or determinants, however, the analysis is of cross-sectional data and can only identify relationships and not claim causation.

4.3 Multinomial logistic regression empirical results

4.3.1 NIDS data set empirical results

Table 4.1 shows the abbreviation for each variable as well as an indication of each category and its definition within the NIDS data set. This table will be of use when considering the outputs produced by SPSS, to identify each variable according to its abbreviation.

Table 4-1: Variable abbreviation and definition list for NIDS data set

Nr.	Abbreviation	Variable	Definition
1	PID	Personal identifier	N/A
2	POG	Population group	1 = African 2 = Coloured, Indian & Asian 3 = White
3	GEN	Gender	1 = Male 2 = Female
4	AGE	Age in years	1 < 36 2 = 36 – 45 3 = 46 – 55 4 > 55
5	MST	Marital status	1 = Married 2 = Never married 3 = Other
6	HID	Highest degree obtained	1 = Bachelor's degree 2 = Bachelor's & diploma 3 = Honours degree 4 = Master's degree/Ph.D.
7	OCC	Primary occupation	1 = Managers 2 = Professionals 3 = Service & Sales workers 4 = Other
8	AWH	Average weekly work hours	1 < 41 hours 2 = 41 – 45 hours 3 > 45 hours
10	INC	Monthly Income bracket before tax and general deductions	1 = R1 – R13 000 2 = R13 001 – R22 000 3 > R22 000

Source: Author's (2013)

Firstly, it should be noted that the variables age, population group, occupation, and average weekly work hours were eliminated from the multinomial logistic regression model for either of one of two reasons, or both. The first reason for the elimination of the variables was because missing cases were reported within specific income categories,

while the second reason for the elimination was as a result of variables suffering from multicollinearity. The variables suffering from multicollinearity were identified by means of the standard error column from the multinomial logistic regression model, where a standard error greater than 2 indicates that the variable suffers from multicollinearity.

The second point of interest is the overall fit of the model, since the elimination of some of the variables affects the accuracy and the predictive ability of the model itself. The SPSS output model renders a model fitting information table (Table 4-2), estimating whether or not the relationship between the independent variables and the dependent variable is supported. The significance of the likelihood ratio chi-square statistic at the 5% level would indicate that there is a relationship between the dependent variable and independent variables (Morgan & Teachman, 1988). The null hypothesis of this statistic states that there is no difference between the model with, and the model without, the independent variables, the rejection of the null hypothesis would imply that there is some meaningful relationship between the dependent and independent variables (Morgan & Teachman, 1988). According to Table 4-2, the likelihood ratio chi-square statistic is significant at the 5% level and it can therefore be concluded that the null hypothesis can be rejected. The presence of a relationship between income and the independent variables (GEN, MTS and HID) was supported. The prominent decrease in the -2 log likelihood between the model without (Intercept only) and the model with (Final) the independent variables also substantiates the notion that there is some meaningful relationship.

The likelihood ratio chi-square test is an important test of the significance of the relationship within the model, yet it does not render an outcome of the accuracy and errors of the model. The accuracy and errors associated with the model are important since they could assist in detecting whether the model is likely to underestimate or overestimate the predicted values. The results obtained from the model also provide a classification accuracy rate found in the classification table (Table 4-3) which should be compared to a calculated probability accuracy criterion to assess the practicability of the model. Should the classification accuracy rate be greater than the probability accuracy criterion, a conclusion can be drawn that the model is feasible and useful for the specific purpose. According to Table 4-3, the classification accuracy rate (56.0%) is larger than the probability accuracy criterion (43.6%), indicating that the model is feasible and useful for this specific purpose. What should be noted and accounted for is that,

according to the values provided in the classification table (Table 4-3), the model tends to overestimate the predicted values.

Table 4-2: Model fitting information for the NIDS data set

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	148.339			
Final	90.053	58.287	12	.000

South African labour and development research unit; author's

Table 4-3: Classification table for the NIDS data set

Observed	Predicted			
	1.00	2.00	3.00	Percent Correct
1.00	35	24	2	57.4%
2.00	12	32	10	59.3%
3.00	4	14	17	48.6%
Overall Percentage	34.0%	46.7%	19.3%	56.0%

South African labour and development research unit; author's

Table 4-4 provides the parameter estimates for the multinomial logistic regression model of the NIDS data set. The significance of the variables are provided in the "Sig." column, while the parameter estimates are given in the "Exp(B)" column. The table can be interpreted as follows; firstly the last category (e.g. female) within each variable (e.g. gender) is used as the reference group while, secondly, the last income category is used as the reference group for all the variables within the first and second income category. If the Exp(B) value is smaller than unity, it is an indication that the category under question (e.g. male) is less likely to earn the associated income within that income category than the reference group (e.g. female), with reference to the last income category.

Considering the variable gender, the male category is significant at the 5% level within the first income category, while the male category is significant at the 10% level within the second income category. The significance of the female category is not given, since it is used as the reference category for the variable gender. According to Table 4-4, when considering the first income category (R1 – R13 000) with reference to the last

income category (> R22 000), males are less likely to earn within the first income category than females; while in the case of the second income category (R13 001 – R22 000), males are also less likely to earn within this category compared to females in relation to the last income category. What this means is that males are more likely to earn within higher income categories than females, for this sample. This finding is also in line with the relevant literature, stating that males earn significantly higher wages compared to females (Rospabe, 2001:4-7; Psacgarpoulos & Patrinos, 2004:129).

The second variable considered in the table is marital status (MTS), where the married category is significant at the 5% level within the first income category and insignificant for the never married category, while both the married and never married categories are insignificant within the second income category compared to the last income category. Table 4-4 indicates that married individuals are less likely to earn a wage within the first income category compared to individuals within the reference category (other – widow/widower, divorced or separated, etc.), with reference to the last income category. Considering the parameter estimates, it is evident that married individuals are more likely to earn a wage within higher income categories compared to individuals who are within the never married and other MTS category. This finding is in line with the relevant literature, indicating that married individuals are more significantly associated with higher earnings than other marital status categories (Rospabe, 2001:7).

From Table 4-4, it is evident that the degree obtained (HID) has a significant impact on an individual's associated income category. All the levels of education are significant at the 5% level apart from the honours degree category within the second income category, which is significant at the 10% level. It is apparent that lower levels of education are associated with lower income categories, while higher levels of education are associated with higher income categories. This can be seen from the parameter estimates, where those individuals with a bachelor's degree are more likely to fall within the first income category compared to those with a master's degree or Ph.D.; while those individuals with a bachelor's degree and a diploma are also more likely to fall within the first income category compared to those with a master's degree or Ph.D., but less so than those with only a bachelor's degree. The same result is obtained for those individuals with an honours degree compared to those with a master's or Ph.D., yet those individuals with an honours degree have lower parameter estimates than those

with a bachelor's degree and those with an additional diploma; indicating that lower levels of education are associated with lower income categories for this sample.

Table 4-4: Multinomial logistic regression model parameter estimates for the NIDS data set

INC	B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
							Lower Bound	Upper Bound
INC = 1	Intercept	1.121	1.119	1	.317			
	GEN = 1	-1.460	.542	1	.007	.232	.080	.672
	GEN = 2	0 ^b		0				
	MTS = 1	-2.544	.838	1	.002	.079	.015	.406
	MTS = 2	-1.672	1.061	1	.115	.188	.024	1.502
	MTS = 3	0 ^b		0				
	HID = 1	2.522	.911	1	.006	12.457	2.091	74.213
	HID = 2	2.349	.943	1	.013	10.479	1.651	66.486
	HID = 3	2.000	.968	1	.039	7.389	1.109	49.237
	HID = 4	0 ^b		0				
INC = 2	Intercept	.107	1.045	1	.918			
	GEN = 1	-.972	.501	1	.053	.378	.142	1.011
	GEN = 2	0 ^b		0				
	MTS = 1	-.797	.875	1	.362	.451	.081	2.503
	MTS = 2	-.299	1.090	1	.784	.742	.088	6.279
	MTS = 3	0 ^b		0				
	HID = 1	2.045	.725	1	.005	7.726	1.865	32.000
	HID = 2	2.056	.745	1	.006	7.816	1.816	33.639
	HID = 3	1.352	.786	1	.086	3.865	.828	18.049
	HID = 4	0 ^b		0				

a. The reference category is: 3.00.

b. This parameter is set to zero because it is redundant.

South African labour and development research unit; author's

Furthermore, the highest degree obtained within the second income category shows that lower levels of education have higher parameter estimates compared to higher levels of education, thus indicating that higher levels of education are associated with higher income categories. What should be noted is that those with only a bachelor's degree and those with an additional diploma, have similar likelihoods of falling within the second income category compared to those with a master's degree or Ph.D., with reference to the last income category. The finding that higher levels of education are associated with higher income categories, while lower levels of income are associated with lower income categories supports the finding within the literature considered in this study (Bhorat, 2000:3; Rospabe, 2001:21; Keswell & Poswell, 2004:849). Since higher levels of education are associated with higher income categories for this sample, it can

be said that the rate of return is higher for higher levels of education than for lower levels of education; considering that the rate of return to education only regards the income component and not the cost of education. Education can therefore be seen as an important determinant of income and is positively associated, where an increase in the level of education could render higher earnings.

4.3.2 Alumni data set empirical results

Table 4-5 comes to show the abbreviation for each variable as well as an indication of each category and its definition within the Alumni data set. Similar to Table 4.1, this table will be of use when considering the outputs produced by SPSS, in order to identify each variable according to its abbreviation.

Table 4-5: Variable abbreviation and variable list for the Alumni data set

Nr.	Abbreviation	Variable	Definition
1	PID	Personal identifier	N/A
2	GEN	Gender	1 = Male 2 = Female
3	AGE	Age in years	1 = 22 and 23 years 2 = 24 and 25 years 3 > 25 years
4	MST	Marital status	1 = Never married 2 = Living with partner 3 = Married
5	HID	Highest degree obtained	1 = Bachelor's degree 2 = Honours degree 3 = Master's degree/Ph.D.
6	OCC	Primary occupation	1 = Insurance, Banking, and Finance 2 = General management, Operations, Accounting, and Analyst 3 = Logistics and Marketing 4 = Other
7	AWH	Average weekly work hours	1 < 41 hours 2 = 41 – 45 hours 3 > 45 hours
8	YOM	Year of matriculation	1 = 2009 2 = 2008 3 = 2007 4 = 2006 5 < 2006

9	FOS	Field of study	1 = Economics 2 = Risk Management 3 = International Trade 4 = Economics & Risk Management 5 = Economics & International Trade
10	EMPS	Employed while studying	1 = Yes 2 = No
11	PROV	Province of primary employment	1 = Gauteng 2 = North-West 3 = Other
12	YEXP	Total years of work experience	1 = 1 year 2 = 2 years 3 = 3 years 4 > 3 years
13	INC	Monthly Income bracket before tax and general deductions	1 = R1 – R13 000 2 = R13 001 – R22 000 3 > R22 000

Source: Author's (2013)

Variables, average weekly work hours, year of matriculation, field of study, employed while studying, and total years of work experience were eliminated from the Multinomial logistic regression model for either one of two reasons, or both, namely due to missing cases being reported and secondly due to multicollinearity.

According to Table 4-6, the likelihood ratio chi-square statistic is significant at the 5% level and it can therefore be concluded that the null hypothesis can be rejected that there is no difference between the model with and the model without the independent variables. The presence of a relationship between income and the independent variables (GEN, AGE, MTS, HID, OCC and PROV) was supported. The decrease in the -2 log likelihood between the model without the independent variables (Intercept only) and the model with (Final) the independent variables also confirms the impression that there is some meaningful relationship between the dependent and independent variables.

As described previously, the accuracy and errors associated with the model is important, since they assist in detecting whether the model is likely to underestimate or overestimate the predicted values. The classification accuracy rate, which can be found in Table 4-7, shows that the model has an accuracy rate of 53.8% and that the model also tends to overestimate the predicted values. The classification accuracy rate (53.8%) is larger than the calculated probability accuracy criterion (36.78%), indicating that the model is feasible.

Table 4-6: Model fitting information for the Alumni data set

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	194.328			
Final	141.893	52.436	36	.038

Source: Alumni survey data set; author's

Table 4-7: Classification table for the Alumni data set

Observed	Predicted				Percent Correct
	1.00	3.00	4.00	5.00	
1	6	7	0	1	42.9%
2	4	22	2	5	66.7%
3	0	6	3	3	25.0%
4	0	6	2	11	57.9%
Overall Percentage	12.8%	52.6%	9.0%	25.6%	53.8%

Source: Alumni survey data set; author's

Table 4-8 shows the parameter estimates for the multinomial logistic regression model of the Alumni data set. As can be seen from the table, few variables have significant categories within the model. As a result, only those variables with categories that are significant will be interpreted.

Considering the first income category (R1 – R10 000), only the first category of age and the first category of province of primary employment is significant; both are significant at the 5% level. When studying the parameter estimate (36.223) of those individuals who are 22 and 23 year of age (AGE = 1), it is evident that those individuals are more likely to earn within the first income category compared to individuals who are older than 25 years of age (AGE= 3), with reference to the last income category (> R20 000). What this finding indicates is that younger individuals are more likely to earn within the first income category compared to older individuals, while referring to the last income category. The second variable with a significant category is that of the province of primary employment; the Gauteng province (PROV = 1) has a parameter estimate (0.012) which is less than unity, indicating that those individuals located within the Gauteng province are less likely to earn within the first income category compared to individuals from other provinces, with reference to the last income category. The finding that Gauteng is the only significant province and is also less likely to fall within the first income category compared to other provinces, provides reason to assume that those

individuals located in Gauteng are more likely to earn higher incomes which is in line with the findings of Bhorat (2000:7).

The third category of occupation (logistics and marketing) as well as the first category of province of primary employment (Gauteng) are the only significant categories within the second income category (R10 001 – R15 000); both are significant at the 5% level. From Table 4-8, it is evident that the third category of province of primary employment has a high parameter estimate (15.415) which is above unity, indicating that those whose occupation is listed under logistics and marketing are more likely to be associated with earning a wage within the second income category than those individuals who have listed their occupation as other, with reference to the third income category. Those who consider their occupation as either within the industry of logistics or marketing, are therefore more likely to earn an income between R10 001 and R15 000, than other occupational categories which are more likely to earn within higher income categories, for this sample. From Table 4-8, it is apparent that the Gauteng province is the only significant category (PROV = 1) within the province of primary employment category for the second income category. According to the table, those who are employed in the Gauteng province are less likely to earn within the second income category than those from other provinces, but less so than within the first income category when referring to the third income category. An assumption can therefore be made that those who are employed in the Gauteng province are more likely to earn within higher income categories than any of the other provinces. This result is expected since the geographic location of employment is an important factor of influence because large industries tend to be strategically placed within a particular country (Coe, Hess, Yeung, Dicken, & Henderson, 2004). It therefore stands to reason that those workers located near large industry will most probably earn a greater wage than those located in regions where unemployment is high and industry is small, Bhorat (2000:7) also supports these findings.

Considering the third income category (R15 001 – R20 000), only the second category of occupation (general management, operations, accounting, and analyst) is significant at the 5% level, while the third category of occupation (logistics and marketing) and the second category of age (24 and 25 years) are significant at the 10% level. Both the second (15.850) and third (11.062) occupational categories have parameter estimates that are higher than unity, indicating that those individuals who listed their occupations

under general management, operations, accounting, analyst, logistics, and marketing, are more likely to earn within the third income category than individuals who have listed their occupations as other, with reference to the third income category. What should be noted is that those within the industry of logistics and marketing are less likely to earn within the third income category than those listed under general management, operations, accounting, and analyst, for this sample. An assumption for this sample can therefore be made that those individuals who listed their occupations under general management, operations, accounting, analyst, logistics, and marketing, are more likely to earn within lower income categories compared to other occupational categories, when referring to the last income category. When considering the second category of age for the third income category, it is evident that those individuals who are 24 and 25 years of age are less likely to earn between R15 001 – R20 000 than those who are older than 25 years of age, referring to the last income category. This finding can be seen as somewhat differing from the norm, since it is expected that older individuals will earn within higher wage categories than younger individuals, according to the relevant literature (Rospabe, 2001:7; Kabubo-Mariara, 2003:15). According to Chang & Huang (2005:2103), age as a determinant of income is more important for higher job levels than for lower job levels, and that age was not a significant determinant of competency at any of the considered job levels, thus indicating that age did not affect an individual's physical or intellectual calibre. The reason as to why this finding was encountered can be disputed, these results could therefore require further investigation within future studies.

The majority of the results obtained from the multinomial logistic regression was expected, and are in line with the relevant literature. The results obtained from the multinomial logistic regression indicates that younger individuals are more likely to earn within lower income categories, while older individuals are more likely to earn within higher income categories. A second finding was that individuals who are employed in the Gauteng province are more likely to earn within higher income categories than individuals from other provinces, while it was also found that occupations including general management, operations, accounting, and analyst were also found to be associated with higher income categories. What should be noted is that education (highest degree obtained) was not significant within the Alumni data set, this could signify that the level of education does not significantly affect an individual's income within the first few years of employment (only persons who have graduated between

2009 and 2012 were considered in the Alumni sample). Further investigation is therefore required. The following section considers the cross-tabulation estimates between income and the highest degree obtained to determine whether or not statistically significant differences occur between the various income categories with regards to the highest degree obtained.

Table 4-8: Multinomial logistic regression model parameter estimates for the Alumni data set

INC		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
INC = 1	Intercept	1.381	2.050	.454	1	.501			
	OCC = 1	-.753	1.155	.425	1	.514	.471	.049	4.527
	OCC = 2	-.243	1.351	.032	1	.857	.784	.056	11.076
	OCC = 3	2.007	1.604	1.565	1	.211	7.439	.321	172.596
	OCC = 4	0 ^b			0				
	GEN = 1	-1.244	.993	1.571	1	.210	.288	.041	2.017
	GEN = 2	0 ^b			0				
	AGE = 1	3.590	1.767	4.125	1	.042	36.223	1.134	1157.356
	AGE = 2	1.709	1.502	1.294	1	.255	5.523	.291	104.911
	AGE = 3	0 ^b			0				
	MTS = 1	.197	1.316	.022	1	.881	1.217	.092	16.071
	MTS = 2	-2.167	1.870	1.344	1	.246	.114	.003	4.470
	MTS = 3	0 ^b			0				
	HID = 1	-1.286	2.191	.345	1	.557	.276	.004	20.241
	HID = 2	.481	1.880	.065	1	.798	1.618	.041	64.449
	HID = 3	0 ^b			0				
	PROV = 1	-4.436	1.444	9.442	1	.002	.012	.001	.201
	PROV = 2	-1.798	1.443	1.552	1	.213	.166	.010	2.803
	PROV = 3	0 ^b			0				
INC = 2	Intercept	1.051	1.452	.525	1	.469			
	OCC = 1	.555	.842	.434	1	.510	1.741	.335	9.063
	OCC = 2	1.240	1.086	1.304	1	.253	3.456	.411	29.023
	OCC = 3	2.735	1.280	4.566	1	.033	15.415	1.254	189.479
	OCC = 4	0 ^b			0				
	GEN = 1	-.428	.744	.331	1	.565	.652	.152	2.803
	GEN = 2	0 ^b			0				
	AGE = 1	1.441	1.318	1.196	1	.274	4.225	.319	55.886
	AGE = 2	.664	1.051	.399	1	.527	1.943	.248	15.241
	AGE = 3	0 ^b			0				
	MTS = 1	.763	1.093	.488	1	.485	2.145	.252	18.256
	MTS = 2	-.388	1.301	.089	1	.765	.678	.053	8.687
	MTS = 3	0 ^b			0				
	HID = 1	-.738	1.297	.324	1	.569	.478	.038	6.074
	HID = 2	-.231	1.164	.039	1	.843	.794	.081	7.780
	HID = 3	0 ^b			0				
	PROV = 1	-2.485	1.020	5.931	1	.015	.083	.011	.616
	PROV = 2	-1.474	1.197	1.515	1	.218	.229	.022	2.394
	PROV = 3	0 ^b			0				

INC = 3	Intercept	-1.724	2.006	.738	1	.390			
	OCC = 1	-.789	1.533	.265	1	.607	.454	.023	9.171
	OCC = 2	2.763	1.398	3.906	1	.048	15.850	1.023	245.517
	OCC = 3	2.403	1.427	2.836	1	.092	11.062	.674	181.444
	OCC = 4	0 ^b			0				
	GEN = 1	-.770	.992	.603	1	.437	.463	.066	3.235
	GEN = 2	0 ^b			0				
	AGE = 1	-2.281	1.714	1.770	1	.183	.102	.004	2.943
	AGE = 2	-2.363	1.339	3.116	1	.078	.094	.007	1.298
	AGE = 3	0 ^b			0				
	MTS = 1	1.414	1.419	.994	1	.319	4.114	.255	66.388
	MTS = 2	1.693	1.548	1.196	1	.274	5.436	.261	112.987
	MTS = 3	0 ^b			0				
	HID = 1	1.644	1.746	.887	1	.346	5.174	.169	158.393
	HID = 2	1.157	1.349	.735	1	.391	3.179	.226	44.754
	HID = 3	0 ^b			0				
	PROV = 1	-.307	1.333	.053	1	.818	.736	.054	10.027
	PROV = 2	-1.400	1.595	.770	1	.380	.247	.011	5.623
	PROV = 3	0 ^b			0				

a. The reference category is: 4.00.

b. This parameter is set to zero because it is redundant.

Source: Alumni survey data set; author's

4.4 Cross-tabulation empirical results

4.4.1 NIDS data set empirical results

Cross-tabulation was used to determine whether or not statistically significant differences occur between the various income categories with regards to the highest degree obtained. The results of the cross-tabulation are presented in Table 4-9. To determine whether there is a statistically significant difference between the income groups (INC) and the highest degree obtained (HID), the Pearson Chi-Square significance value will be considered. A Pearson Chi-square sig. value smaller than 0.05 would indicate that there is a significant difference between the income groups and the highest degree obtained (Pallant, 2011:219).

Table 4-9 shows the cross-tabulation estimates between the income groups and the highest degree obtained. The Pearson Chi-Square test results (sig. value of 0.000) indicated that, given the sample, the level of education (highest degree obtained) has a statistically significant effect on the category of income. From Table 4-9 it becomes evident that the level of education has some effect on income, since those with a bachelor's degree represent 47.5% of the individuals within the first income category (R1 – R13 000), while representing 42.6% of individuals within the second income

category (R13 001 – R22 000) and only 22.9% of individuals within the third income category (> R22 000).

Table 4-9: Cross-tabulation results for the NIDS data set (INC*HID)

		Income (Before tax and general deductions)			Total	
		R1 – R13 000	R13 001 – R22 000	> R22 000		
Highest degree obtained	Bachelor's degree	Count	29	23	8	60
		% within HID	48.3%	38.3%	13.3%	100.0%
		% within INC	47.5%	42.6%	22.9%	40.0%
		% of Total	19.3%	15.3%	5.3%	40.0%
	Bachelor's degree & Diploma	Count	17	18	7	42
		% within HID	40.5%	42.9%	16.7%	100.0%
		% within INC	27.9%	33.3%	20.0%	28.0%
		% of Total	11.3%	12.0%	4.7%	28.0%
	Honours degree	Count	13	9	7	29
		% within HID	44.8%	31.0%	24.1%	100.0%
		% within INC	21.3%	16.7%	20.0%	19.3%
		% of Total	8.7%	6.0%	4.7%	19.3%
	Master's degree & Ph.D.	Count	2	4	13	19
		% within HID	10.5%	21.1%	68.4%	100.0%
		% within INC	3.3%	7.4%	37.1%	12.7%
		% of Total	1.3%	2.7%	8.7%	12.7%
Total	Count	61	54	35	150	
	% within HID	40.7%	36.0%	23.3%	100.0%	
	% within INC	100.0%	100.0%	100.0%	100.0%	
	% of Total	40.7%	36.0%	23.3%	100.0%	

South African labour and development research unit; author's

Furthermore, those individuals with a master's degree or Ph.D. only represent 3.3% of the individuals within the first income category, 7.4% in the second income category and 37.1% within the third income category. This indicates that the lowest level of education is more concentrated within the lower income categories, while the highest level of education is more concentrated within the higher income categories. The results obtained from Table 4-9 are in line with the relevant literature, where higher levels of education are associated with higher levels of income (Bhorat, 2000:3; Rospabe, 2001:21; Keswell & Poswell, 2004:849). When considering those individuals with a bachelor's degree & Diploma, it is evident that the majority of those individuals are found within the second income category while, oddly, the majority of those individuals with an honours degree can be found in the first income category. The reason as to why those individuals with an honours degree are more likely to be found within the first

income category remains a subject yet to be debated. Since the majority of these individuals listed their occupation as professional (79%), it is difficult to determine whether their occupational category may be the reason for the association with lower levels of income. It should be noted that those individuals with an honours degree (AWH < 40 hours; 69%) work fewer hours per week on average than those individuals with a master's degree or Ph.D. (AWH < 40 hours; 37%).

It should also be noted that 91.7% of the cells have expected frequencies of 5 or more, this comes to say that the Chi-Square assumption was not violated; where the assumption states that at least 80% of the cells should have a minimum expected cell frequency of 5 or greater (Pallant, 2011:219). The minimum expected count is 4.43, indicating that all the expected cell frequencies are greater than 4.43.

4.4.2 Alumni data set empirical results

Table 4-10 provides the cross-tabulation results for the Alumni data set to determine whether a statistically significant difference exists between the income groups (INC) and the highest degree obtained (HID). Firstly, it should be noted that the income categories had to be reduced from four categories to two, the reason for this is because 58.3% of the cells have expected frequencies of 5 or more, this means that the Chi-Square assumption was violated when considering four income categories. With the reduction in income categories from four to two, 83.3% of the cells have expected frequencies of 5 or more, and as a result the Chi-Square assumption was not violated in the case of a cross-tabulation with only two income categories. The minimum expected count is 4.77, indicating that all the expected cell frequencies are greater than 4.77.

Secondly, the Pearson Chi-square sig. value indicates that there is not a significant difference between the income groups and the highest degree obtained, reporting a sig. value of 0.299. Although the significance of the cross-tabulation between INC and HID, for this sample is seen as undistinguished, there is still some value within these results which is worthy of interpretation. From Table 4-10 it is evident that lower levels of education are more likely to be associated with the first income category, while higher levels of education are more likely to be associated with the second income category. The table thus shows that 66.7% of those individuals with a bachelor's degree are likely to earn within the first income category, while 59.3% of those with an honours and 41.7% of those with a master's degree or Ph.D. are likely to earn within the first income

category. It is thus evident from Table 4-10 that higher levels of education are, to a greater extent, associated with higher income categories. This is in line with the relevant literature stating that higher levels of education are associated with higher levels of income (Keswell & Poswell, 2004:849; Solidarity Research Institute, 2012:3; Statistics South Africa, 2012:xvi).

Table 4-10: Cross-tabulation results for the Alumni data set (INC*HID)

			Income (Before tax and general deductions)		Total
			R1 - R15 000	> R15 000	
Highest degree obtained	Bachelor's degree	Count	26	13	39
		% within HID	66.7%	33.3%	100.0%
		% within INC	55.3%	41.9%	50.0%
		% of Total	33.3%	16.7%	50.0%
	Honours degree	Count	16	11	27
		% within HID	59.3%	40.7%	100.0%
		% within INC	34.0%	35.5%	34.6%
		% of Total	20.5%	14.1%	34.6%
	Master's degree and Ph.D.	Count	5	7	12
		% within HID	41.7%	58.3%	100.0%
		% within INC	10.6%	22.6%	15.4%
		% of Total	6.4%	9.0%	15.4%
Total	Count		47	31	78
	% within HID		60.3%	39.7%	100.0%
	% within INC		100.0%	100.0%	100.0%
	% of Total		60.3%	39.7%	100.0%

Source: Alumni survey data set; author's

4.5 Conclusion

This chapter considered the multinomial logistic regression model and cross-tabulation empirical results for both the NIDS and Alumni data sets. The variables, age, population group, occupation and average weekly work hours were eliminated from the Multinomial logistic regression model for the NIDS data set, since these variables either resulted in missing cases being reported or that these variables suffered from multicollinearity, or both. The likelihood ratio chi-square statistic was significant at the 5% level for this model and as a result it was concluded that there is a presence of a relationship between income and the independent variables, for this sample. Furthermore, the classification accuracy rate of the model was found to be larger than the probability accuracy criterion, indicating that the model is feasible, yet it was also determined that

the model tends to provide an overestimate of the predicted values. The parameter estimates of the model indicated that males are more likely to earn a wage within higher income categories than their female counterparts, while married individuals are also more likely to earn a wage within higher income categories than individuals from other marital status categories. The model also indicated that lower levels of education are associated with lower income categories, while higher levels of education are associated with higher income categories. The finding that the level of tertiary education plays a significant role within this sample of individuals is in line with the relevant literature and also allows an assumption to be made that the rate of return to education would be higher as the level of education increases, when only considering the earnings component and not the cost component of the rate of return to education.

The likelihood ratio chi-square statistic for the multinomial logistic regression model for the Alumni data set was found to be significant, indicating that there is a relationship between income and the independent variables. The classification accuracy rate of the model was found to be larger than the calculated probability accuracy criterion, indicating that the model is feasible, yet renders an overestimate of the predicted values. The parameter estimates obtained from the model indicated that younger individuals are more likely to earn within lower income categories than older individuals, who are more likely to earn within higher income categories. It was also found that the Gauteng province was the only significant category of province of the primary employment variable, and indicated that those individuals who are located in the Gauteng province are more likely to earn within higher income categories than those individuals from other provinces. Both these findings are supported by the literature findings which were discussed in chapter 2. Furthermore, the empirical results also indicated that those individuals who listed their occupations under general management, operations, accounting, analyst, logistics, or marketing, were more likely to earn within the third income category than individuals who listed their occupations as other, with reference to the fourth income category. This indicates that these occupations were less likely to earn within the fourth income category, compared to other occupations.

This chapter also reported the cross-tabulation results obtained for both data sets, where it was used to determine whether or not statistically significant differences occur between the various income categories with regards to the highest degree obtained.

The Pearson Chi-Square test results indicated that, given the NIDS data sample, the level of education has a statistically significant effect on the category of income. The cross-tabulation estimates indicated that higher levels of education are associated with higher income categories, while lower levels of education are associated with lower income categories. This result was found to be true for all cases apart from those with an honours degree, which were associated with similar income categories as those with a bachelor's degree.

The cross-tabulation estimates for the Alumni data set were found to be insignificant, since the Pearson Chi-square sig. value reported a sig. value higher than 0.05. It should be noted that for both the NIDS and Alumni data set, the Chi-Square assumption was not violated. Although the Alumni data set cross-tabulation estimates are insignificant, it showed similar results to those of the NIDS data set. The reason as to why the level of education proved to be mostly insignificant for the Alumni data set may be as a result of the data set itself, since only individuals who have graduated less than four years prior to the year in which this study was conducted were considered in this sample. It can therefore be assumed that the level of education does not play a significant role within the first few years after graduation, for this sample of individuals. Although the NIDS data set came to prove that the level of education does play a significant role when considering all majors, years within which individuals have graduated, and universities from which individuals have graduated. The rate of return to education is also assumed to increase as the level of education increases within this sample, only considering the earnings component and not the cost component of the rate for return to education.