

# The analysis of a grouped response variable using maximum likelihood estimation under constraints

by

Johannes Jurgens Hendriks

Submitted in partial fulfillment of the requirements for the degree

**Master of Science (Mathematical Statistics)**

In the Faculty of Natural and Agricultural Sciences

University of Pretoria

Pretoria

December 2016

**The analysis of a grouped response variable using maximum likelihood estimation under constraints**

by

**Johannes Jurgens Hendriks**

Mini-dissertation supervisor: Dr G Crafford - University of Pretoria

# Declaration

I, Johannes Jurgens Hendriks, hereby declare that the work submitted by me in this mini-dissertation is my own work and has not been copied in any way or submitted by my previously to any other tertiary institution.

Signature: \_\_\_\_\_

Date: \_\_\_\_\_

© University of Pretoria 2016  
All rights reserved.

# Acknowledgements

I would like to thank my Creator for giving me the ability to complete this mini dissertation. The financial support received from the National Research Foundation (NRF) and the Bureau for Statistical and Survey Methodology (STATOMET) is highly appreciated. Without it, the study would not have been possible. I would also like to thank my supervisor, friends and family for all their support. Thank you Dr Arulsivanathan Naidoo at StatsSA for providing the 10% census data of South Africa for 2011. A special word of gratitude goes out to Carla Stegmann.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Background . . . . .	1
1.2	Objectives . . . . .	2
1.3	Dissertation outline . . . . .	2
1.4	Census 2011 data . . . . .	3
<b>2</b>	<b>Fitting a log-logistic distribution</b>	<b>8</b>
2.1	Introduction . . . . .	8
2.2	Formulation . . . . .	9
2.3	The effect of the defined income categories . . . . .	17
2.4	Summary . . . . .	19
<b>3</b>	<b>Fitting other distributions</b>	<b>20</b>
3.1	The Normal Distribution . . . . .	20
3.2	The Exponential Distribution . . . . .	22
3.3	Summary . . . . .	24
<b>4</b>	<b>Single factor design</b>	<b>25</b>
4.1	Introduction . . . . .	25
4.2	Population group . . . . .	25
4.2.1	Distribution fitting . . . . .	26
4.2.2	Saturated model . . . . .	30
4.3	Gender . . . . .	33
4.4	Level of education . . . . .	34
4.4.1	Model for the medians . . . . .	36
4.5	Age . . . . .	39
4.6	Summary . . . . .	42
<b>5</b>	<b>Multifactor design</b>	<b>43</b>
5.1	Gender and population group . . . . .	43
5.1.1	Distribution fitting . . . . .	43

5.1.2	Saturated model . . . . .	46
5.1.3	Independence model . . . . .	49
5.2	Gender and level of education . . . . .	50
5.2.1	Distribution fitting with the saturated model . . . . .	51
5.2.2	Linear model . . . . .	53
5.2.3	Quadratic model . . . . .	57
5.3	Population group and level of education . . . . .	59
5.3.1	Distribution fitting with the saturated model . . . . .	60
5.3.2	Linear model . . . . .	63
5.3.3	Linear model with equal gradients . . . . .	65
5.3.4	Quadratic model . . . . .	66
5.4	Gender, population group and level of education . . . . .	67
5.4.1	Distribution fitting . . . . .	67
5.4.2	Linear model . . . . .	69
5.4.3	Linear model with population group and level of education independent . . . . .	73
5.5	Summary . . . . .	75
<b>6</b>	<b>The Logit model</b>	<b>76</b>
6.1	Introduction . . . . .	76
6.2	Level of education . . . . .	76
6.2.1	Saturated model . . . . .	76
6.2.2	Linear model . . . . .	80
6.3	Population group and level of education . . . . .	82
6.3.1	Saturated model . . . . .	82
6.3.2	The linear model . . . . .	84
6.4	Gender, population group and level of education . . . . .	86
6.4.1	Linear model . . . . .	86
6.5	Summary . . . . .	88
<b>7</b>	<b>Conclusion</b>	<b>89</b>
<b>8</b>	<b>Appendix</b>	<b>92</b>
8.1	Appendix A: Fitting a log-logistic distribution . . . . .	92
8.2	Appendix B: Single factor model . . . . .	95
8.3	Appendix C: Two factor model . . . . .	99
8.4	Appendix D: Three factor model . . . . .	105

8.5	Appendix E: Logit model with one explanatory variable . . . . .	111
8.6	Appendix F: Logit model with two explanatory variables . . . . .	112
8.7	Appendix G: Logit model with three explanatory variables . . . . .	114

# List of symbols

$\mathbf{g}$	Vector of constraints
$\mathbf{G}$	Matrix of first order partial derivatives of $g$
$n$	Sample size
$x_i$	Upper class boundary
$\mathbf{x}$	Vector of upper class boundaries
$f_i$	Frequency in class $i$
$\mathbf{f}$	Vector of frequencies
$p_0$	Vector of relative frequencies
$\pi_0$	Vector of expected relative frequencies
$\mathbf{V}_0$	Covariance matrix of $p_0$
$\mathbf{p}$	Vector of cumulative relative frequencies
$\pi$	Vector of expected cumulative relative frequencies
$\mathbf{V}$	Covariance matrix of $p$
$F(x)$	Cumulative distribution function
$\chi^2$	Pearson chi-square statistic
$W$	Wald statistic
$D$	Measure of discrepancy
$\Phi(z)$	Normal cumulative distribution function
$\alpha$	Vector of natural parameters
$\mathbf{X}'$	Indicates the transpose of matrix $\mathbf{X}$
$\mathbf{P}_X$	Projection matrix onto the vector space of $\mathbf{X}$
$\mathbf{Q}_X$	Projection matrix onto the error space of $\mathbf{X}$
$\mathbf{F}_M$	Matrix of cross tabulated frequencies
$vec(\mathbf{F}_M)$	Row-wise concatenation of matrix $\mathbf{F}_M$
$T$	Amount of cross tabulated cells
$\mathbf{n}$	Vector of frequencies of the $T$ cells



# List of Figures

1-1	Histogram of focus group . . . . .	5
2-1	Final histogram of focus group . . . . .	9
2-2	Fitted log-logistic distribution . . . . .	16
2-3	Pdf of log-logistic distribution fits for different grouping scenarios . . . . .	18
3-1	Fitted normal distribution . . . . .	22
3-2	Fitted exponential distribution . . . . .	24
4-1	Median vs level of education . . . . .	36
4-2	Medians of different levels of education with fitted models . . . . .	39
4-3	Medians of different age groups with linear model . . . . .	42
5-1	Medians under different population groups with level of education as ordinal variable . . . . .	54
5-2	Linear functions for income under different genders with level of education as ordinal variable . . . . .	57
5-3	Quadratic functions for income under different genders with level of education as ordinal variable . . . . .	59
5-4	Medians under different population groups with level of education as ordinal variable . . . . .	63
5-5	Income for different population groups with level of education as ordinal variable . . . . .	64
5-6	Income for different population groups with level of education as ordinal variable with equal gradients . . . . .	65
5-7	Income for different population groups with level of education as ordinal variable with a quadratic model . . . . .	66
5-8	Median income for Females: Ordinal trend in level of education for each population group . . . . .	70
5-9	Median income for Males: Ordinal trend in level of education for each population group . . . . .	70
5-10	Linear models for Females: Ordinal trend in level of education for each population group . . . . .	72
5-11	Linear models for Males: Ordinal trend in level of education for each population group . . . . .	72
5-12	Linear models for Females: Ordinal trend in level of education for each population group with level of education and population group independent . . . . .	74
5-13	Linear models for Males: Ordinal trend in level of education for each population group with level of education and population group independent . . . . .	74
6-1	Log-odds with level of education as ordinal variable . . . . .	79
6-2	Log-odds with level of education as ordinal variable with linear model . . . . .	81

6-3	Log-odds of different population groups with level of education as ordinal variable with linear model . . . . .	85
6-4	Log-odds for Females: Linear models for different population groups with level of education as ordinal variable . . . . .	87
6-5	Log-odds for Males: Linear models for different population groups with level of education as ordinal variable . . . . .	87

# List of Tables

1.1	Income distribution of the 10 percent census . . . . .	3
1.2	Income distribution of focus group . . . . .	4
1.3	Frequencies of focus group for different population groups . . . . .	5
1.4	Frequencies of focus group for different genders . . . . .	6
1.5	Frequencies of focus group for different levels of education . . . . .	6
1.6	Frequencies of focus group for different age groups . . . . .	6
2.1	Final income distribution of focus group . . . . .	8
2.2	Expected frequencies . . . . .	15
2.3	Goodness-of-fit statistics for log-logistic distribution . . . . .	17
2.4	Parameter estimates for different grouping scenarios . . . . .	18
3.1	Final income distribution of focus group . . . . .	20
4.1	Histograms of income under different population groups . . . . .	25
4.2	Frequencies of population group . . . . .	26
4.3	Estimated frequencies of population group . . . . .	29
4.4	Single factor for population group . . . . .	32
4.5	Cumulative relative frequencies of gender . . . . .	33
4.6	Estimated cumulative relative frequencies of gender . . . . .	33
4.7	Single factor for gender . . . . .	34
4.8	Single factor for level of education . . . . .	35
4.9	Significance tests for the estimated parameters . . . . .	38
4.10	Single factor for age groups . . . . .	40
5.1	Histograms for gender and population group . . . . .	43
5.2	Frequencies for gender and population group . . . . .	44
5.3	Cumulative relative frequencies for gender and population group . . . . .	45
5.4	Expected cumulative relative frequencies for gender and population group . . . . .	46
5.5	Fitted distributions for gender and population group . . . . .	46
5.6	Fitted distributions with saturated model for gender and population group . . . . .	48
5.7	Independence model of gender and population group . . . . .	50
5.8	Gender and level of education . . . . .	51
5.9	Fitted distributions for gender and level of education . . . . .	53

5.10	Fitted distributions for gender and level of education where level of education is ordinal	56
5.11	Population group and level of education . . . . .	60
5.12	Fitted distributions for population group and level of education . . . . .	62
5.13	Equations for linear models for different population groups . . . . .	64
5.14	Equations for linear models for different population groups with equal gradients . . . . .	65
5.15	Equations for quadratic models for different population groups with equal gradients . . . . .	67
5.16	Fitted distributions for Females: Population group vs level of education . . . . .	68
5.17	Fitted distributions for Males: Population group vs level of education . . . . .	69
5.18	Sub-matrices of the design matrix $Y$ . . . . .	71
5.19	Equations for linear models . . . . .	73
5.20	Equations for linear models: Equal gradients under identical genders . . . . .	75
6.1	Frequencies for different levels of education . . . . .	76
6.2	Indices for levels of education . . . . .	78
6.3	Frequencies for population group and level of education . . . . .	82
6.4	Indices for population group and levels of education . . . . .	83
6.5	Equations to find the estimated odds under different population groups . . . . .	85
6.6	Submatrices for saturated model . . . . .	86



# Chapter 1

## Introduction

### 1.1 Background

A continuous variable observed in grouped format is an occurrence frequently encountered in highly quantified fields. This is especially the case if official statistics are considered where income or age may be regarded in grouped format. If a continuous variable is considered as a grouped response variable, the statistical analysis might be limited to cross-tabulation methods with multivariate regression techniques being inapplicable and hence valuable information might be lost.

In Matthews & Crowther (1995) [10] a technique is developed which allows one to find the maximum likelihood (ML) estimate of the expected value  $\boldsymbol{\mu}$  of a random vector  $\mathbf{x}$ , the distribution of which belongs to the exponential family, under a vector of constraints  $\mathbf{g}(\boldsymbol{\mu}) = \mathbf{0}$  in an iterative manner. Matthews (1995) [9] also discusses this technique by considering a variety of different models. Crafford & Crowther (2009) [6] expanded on this idea by assuming a grouped response variable being multinomial distributed and estimating the expected cumulative relative frequencies such that it follows a certain distribution at the upper class boundaries of the grouped response variable. If a set of explanatory variables are cross-tabulated with the response variable to create  $T$  so-called cells, then the expected cumulative relative frequencies of each cell can be estimated simultaneously such that

1. the expected cumulative relative frequencies of each cell follow a certain distribution at the upper class boundaries of the response variable, and
2. the parameter(s) of the distribution will follow a specified model, where this model may be designed using the set of given explanatory variables.

In this mini dissertation it will be endeavoured to study the technique discussed by [6] by making use of the 10% sample of the South African Census 2011 [11]. The same data will be used to incorporate the technique with the logit model as considered by [9].

## 1.2 Objectives

The key objectives of this study can be summarised as follows:

- Consider the technique developed by [10] and [6] to fit distributions to a grouped response variable,
- Cross-tabulate explanatory variables to create  $T$  cells and fit distributions to each cell such that the median level of the distributions follow a defined model,
- Outline the elegance and simplicity of this technique by considering a range of different models for the median, and
- Consider how the iterative procedure can be used in conjunction with the logit model.

## 1.3 Dissertation outline

Section 1.4 revolves around setting up an appropriate subset of the 10% sample of the Census 2011 data that acts as a focus group and doing an exploratory analysis of the data at hand. Here, the grouped response variable INCOME is defined with a set of explanatory variables and the row frequencies are used to indicate the possible effects of the explanatory variables.

Chapter 2 defines the theory needed to fit a log-logistic distribution to a grouped response variable. Chapter 3 in turn considers how the technique is applied to fit a normal- and exponential distribution.

Chapter 4 focusses on the single factor model where log-logistic distributions are fitted to the grouped response variable that is cross-tabulated with the categories of a single explanatory variable. This is also done such that the medians of the log-logistic distributions will follow a defined model.

Following the same ideology as in Chapter 4, the multi-factor model is defined in Chapter 5 where more than one explanatory variable is cross-tabulated.

In Chapter 6 the logit model is used in conjunction with the iterative procedure.

The dissertation ends with the conclusion in Chapter 7 and the relevant code can be found in the Appendix.

## 1.4 Census 2011 data

The purpose of a census is to collect data on a country's population. Results of a vast assortment of different variables are collected that aim to provide information on a demographic, economic and social level. These variables may in turn be used by different entities such as government institutions or industry to provide solutions to pressing issues that they may be facing.

The process of collecting all of the data to compile a census is repeated every 10 years. The census data that will be used here to display the technique at hand was collected the night of 9/10 October 2011, the most recent census available. Note that only a 10% sample of the census data [11] is used since the aim of this dissertation is not necessarily to analyse the results attained from the data but more on the technique used to attain them.

The variables considered in the census were divided into three main groups, namely Person, Household and Mortality variables. Person variables can be divided up further into groups like demographics, parental survival and income, level of education and employment. The main group of variables that will be focussed on are Person variables. Specifically, the technique presented by [10] and further expanded by [5] and [9] will be used to provide insight into the grouped response variable, income. The frequency distribution of income provided by [11] is given in Table 1.1.

Code	Monthly	Frequency	Percent
1	No Income	1790847	40.53
2	R 1-R 400	785636	17.78
3	R 401 - R 800	141645	3.21
4	R801 - R 1 600	497321	11.26
5	R 1 601 - R 3 200	251223	5.69
6	R 3 201 - R 6 400	187879	4.25
7	R 6 401 - R 12 800	159575	3.61
8	R 12 801 - R 25 600	114689	2.60
9	R 25 601 - R 51 200	46532	1.05
10	R 51 201 - R 102 400	14061	0.32
11	R 102 401 - R 204 800	5191	0.12
12	R 204 801 or more	3567	0.08
99	Unspecified	339531	7.68
.	Missing values	80897	1.83
<b>Total</b>		<b>4 418 594</b>	<b>100</b>

**Table 1.1: Income distribution of the 10 percent census**

One notes that 7.68% of individuals chose not to indicate their income level and 1.83% of values are missing. Also note that 40.53% of the sample is registered as not having an income. To overcome these



and other obstacles a focus group will be defined. Different demographic variables will also be used as explanatory variables for income. Hence, they will also be considered in the process of designing the focus group. These explanatory variables are population group, gender, level of education and age. The specifications of the focus group are described next:

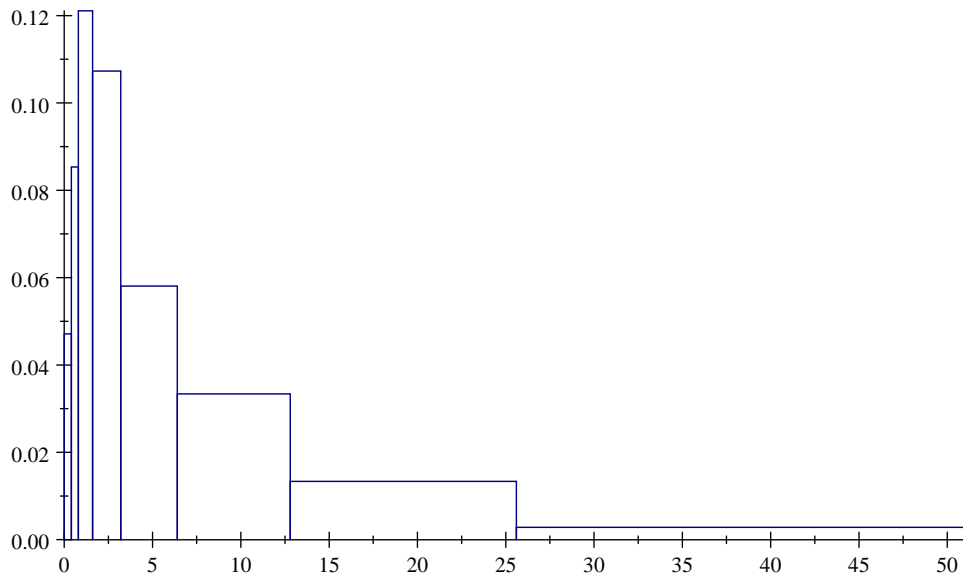
1. The person should have an income level that is greater than R1,
2. The person should have worked in the last 7 days,
3. The age of the person should be between 18 and 65,
4. The person should at least have finished secondary schooling, and
5. The population groups considered will be Black, White, Indian and Coloured.

These specifications define the focus group and reduces the sample to size  $n = 489697$  with the income categories presented in Table 1.2.

Code	Monthly	Frequency	Percent
2	R 1-R 400	9231	1.89
3	R 401 - R 800	16717	3.41
4	R801 - R 1 600	47445	9.69
5	R 1 601 - R 3 200	84065	17.17
6	R 3 201 - R 6 400	90983	18.58
7	R 6 401 - R 12 800	104642	21.37
8	R 12 801 - R 25 600	83686	17.09
9	R 25 601 - R 51 200	35316	7.21
10	R 51 201 - R 102 400	11465	2.34
11	R 102 401 - R 204 800	3560	0.73
12	R 204 801 or more	2587	0.53
<b>Total</b>		<b>489697</b>	<b>100</b>

**Table 1.2: Income distribution of focus group**

Figure 1-1 gives a graphical representation in the form of a histogram of the designed focus group.



**Figure 1-1: Histogram of focus group**

The effects of the explanatory variables can be considered if one divides income into two categories. Individuals earning more than R12 800 will be classified as falling in a high income category and individuals earning less than R12 800 will fall in a low income category. The frequencies in the high and low income categories given a certain demographic variable can now be considered. The values in brackets for all the Tables are the respective column percentages. The cross tabulated frequencies of population group are presented in Table 1.3.

Income group	Population group				Total
	Black	Coloured	Indian	White	
High	51417 (17.22)	11118 (25.19)	10027 (37.58)	64052 (53.28)	<b>136614</b> (27.90)
Low	247239 (82.78)	33011 (74.81)	16657 (62.42)	56176 (46.72)	<b>353083</b> (72.10)
<b>Total</b>	<b>298656</b>	<b>44129</b>	<b>26684</b>	<b>120228</b>	<b>489697</b>

**Table 1.3: Frequencies of focus group for different population groups**

Note that population group is recorded in the census on more levels as what is presented here. Individuals also had the option to indicate 'Asian' or 'other' as a population group. The column percentages already reveal what can be expected from the effect of population group on income. Only 17.22% of individuals from the Black population group earn more than R12 800 per month whereas 53.28% from the White population group earn more than R12 800 per month. Gender is considered next in Table 1.4.

Income group	Gender		Total
	Female	Male	
High	55914 (24.38)	80700 (30.99)	<b>136614</b> (27.90)
Low	173418 (75.62)	179665 (69.01)	<b>353083</b> (72.10)
<b>Total</b>	<b>229332</b>	<b>260365</b>	<b>489697</b>

**Table 1.4: Frequencies of focus group for different genders**

Gender was recorded as only being male or female. From the column percentages, one should note that more males earn more than R12 800 per month compared to females. The cross-tabulated frequencies for level of education are given in Table 1.5.

Income group	Education					Total
	Grade 12	Certificate	Diploma	B Degree	Post Grad	
High	39909 (13.58)	9358 (27.49)	37189 (43.30)	27938 (61.29)	22220 (73.08)	<b>136614</b> (27.90)
Low	253877 (86.42)	24685 (72.51)	48694 (56.70)	17643 (38.71)	8184 (26.92)	<b>353083</b> (72.10)
<b>Total</b>	<b>293786</b>	<b>34043</b>	<b>85883</b>	<b>45581</b>	<b>30404</b>	<b>489697</b>

**Table 1.5: Frequencies of focus group for different levels of education**

Education was recorded on a considerable amount of levels. Since the census concentrated on all individuals in South Africa, individuals were given the option to indicate from a grade 0 level up and to a masters/doctoral level. The option for no schooling or other levels of education was also presented. Only individuals with a Grade 12 level of education and up are considered in the focus group. Note that the certificate and diploma categories in Table 1.5 include higher certificates and higher diplomas as well. The post graduate category includes an honours, masters or doctoral degree. The column percentages indicate a clear ordinal trend on the categories of education. The final explanatory variable that will be considered is age group and the cross-tabulated frequencies are given in Table 1.6.

Income group	Age group							Total
	18 - 25	26 - 30	31 - 40	41 - 45	46 - 50	51 - 55	56 - 65	
High	5599 (8.12)	17162 (18.05)	45103 (27.50)	22242 (37.19)	18833 (42.62)	14287 (47.19)	13388 (48.93)	<b>136614</b> (27.10)
Low	63384 (91.88)	77905 (81.95)	118920 (72.50)	37564 (62.81)	25352 (57.38)	15986 (52.81)	13972 (51.07)	<b>353083</b> (72.10)
<b>Total</b>	<b>68983</b>	<b>95067</b>	<b>164023</b>	<b>59806</b>	<b>44185</b>	<b>30273</b>	<b>27360</b>	<b>489697</b>

**Table 1.6: Frequencies of focus group for different age groups**

Ages are recorded as the individual's current age at the time of the census. It is given in integer form but is presented in categories here. If one considers the column percentages, one should note that one becomes more likely to earn more than R12 800 per month as age increases.

Before the explanatory variables are used to provide further insight on the grouped response variable, income, one may first consider estimating the distribution of the grouped response variable. From Figure 1-1 one can see that income is skewed to the right. Hence, a log-logistic distribution is likely to provide an adequate fit to the income distribution.

# Chapter 2

## Fitting a log-logistic distribution

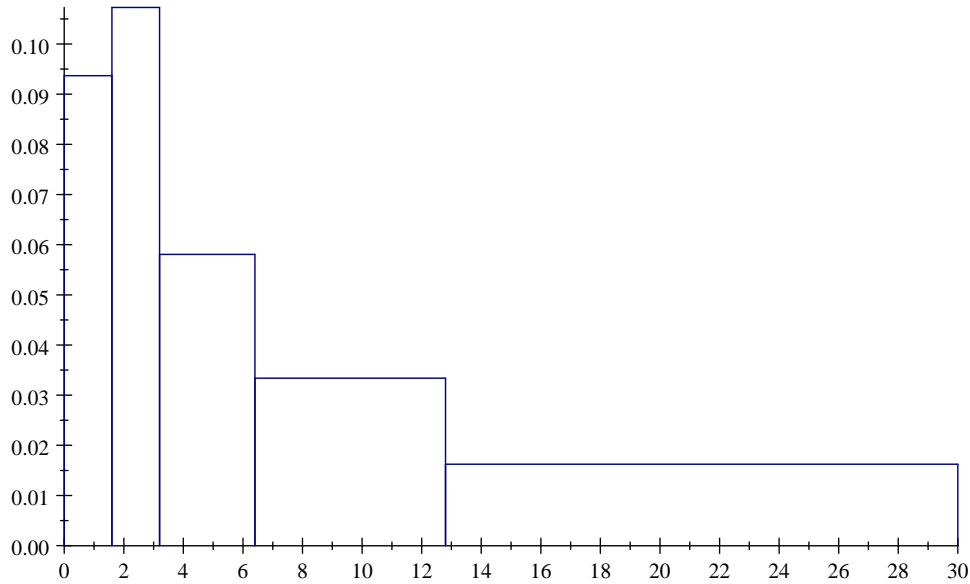
### 2.1 Introduction

The focus of Chapter 2 is to fit a log-logistic distribution to the frequency distribution of income presented in Table 2.1. Note that the income intervals presented here are joined to form only 5 intervals instead of 11. It will be shown that distribution fitting of only 5 intervals can be done in an as effective manner as if one uses 11 intervals. The observed frequencies for the 5 intervals are given in Table 2.1.

Code	Monthly	Frequency	Percent
2 + 3 + 4	R 1 - R 1600	73393	14.99
5	R 1 601 - R 3 200	84065	17.17
6	R 3 201 - R 6 400	90983	18.58
7	R6 400 - 12 800	104642	21.37
8 - 12	R 12 801 or more	136614	27.90
<b>Total</b>		<b>489697</b>	<b>100</b>

**Table 2.1: Final income distribution of focus group**

The corresponding graphical representation of the frequency distribution is given in Figure 2-1.



**Figure 2-1: Final histogram of focus group**

Since the distribution seems to be positively skewed, a log-logistic distribution will be fitted to income. The way in which the log-logistic distribution is defined is identical to the way in which the log-normal distribution is defined. Say the variable  $Y = \log(X)$  follows a logistic distribution. Then the variable  $X$  follows a log-logistic distribution.

The reason why this distribution is considered is because it gives a better fit to the lower intervals whilst not ignoring the heavy tail that the income distribution of the focus group possesses. The probability density function (pdf) of the log-logistic distribution is given by

$$f(x; \kappa, \theta) = \frac{e^{\theta} \kappa x^{\kappa-1}}{(1 + e^{\theta} x^{\kappa})^2}, \dots x > 0 \quad (2.1)$$

where  $\theta, \kappa > 0$ , with corresponding cumulative distribution function (cdf)

$$F(x; \kappa, \theta) = \frac{e^{\theta} x^{\kappa}}{1 + e^{\theta} x^{\kappa}}, \dots x > 0 \quad (2.2)$$

## 2.2 Formulation

In order to fit the log-logistic distribution to the data in Table 2.1, let  $\mathbf{f}$  denote the frequencies of the first  $(5 - 1)$  intervals

$$\mathbf{f} = \begin{pmatrix} 73393 \\ 84065 \\ 90983 \\ 104642 \end{pmatrix} \quad (2.3)$$

and let  $\mathbf{x}$  be the upper class boundaries

$$\mathbf{x} = \begin{pmatrix} 1.6 \\ 3.2 \\ 6.4 \\ 12.8 \end{pmatrix} \quad (2.4)$$

in ( $R1000$ ) units. Since the frequencies add up to a total of  $n$ , only the first  $(k - 1)$  frequencies are considered, where  $k$  is the number of intervals. This notation makes it convenient to define the vector of upper class boundaries since the last interval is open-ended. Assuming  $\mathbf{f}$  is multinomial distributed with a vector of probabilities  $\boldsymbol{\pi}_0$ , the expected value and covariance matrix of  $\mathbf{f}$  is

$$\begin{aligned} E(\mathbf{f}) &= n\boldsymbol{\pi}_0 \\ &= \mathbf{F} \end{aligned} \quad (2.5)$$

and

$$\begin{aligned} Cov(\mathbf{f}) &= n(diag[\boldsymbol{\pi}_0] - \boldsymbol{\pi}_0\boldsymbol{\pi}_0') \\ &= diag(\mathbf{F}) - \frac{1}{n}\mathbf{F}\mathbf{F}' \\ &= \mathbf{V}_F \end{aligned} \quad (2.6)$$

respectively. The vector of relative frequencies

$$\mathbf{p}_0 = \frac{1}{n}\mathbf{f} = \begin{pmatrix} 0.1499 \\ 0.1717 \\ 0.1858 \\ 0.2137 \end{pmatrix} \quad (2.7)$$

will have an expected value and covariance matrix of

$$E(\mathbf{p}_0) = \boldsymbol{\pi}_0$$

and

$$Cov(\mathbf{p}_0) = \frac{1}{n}(diag[\boldsymbol{\pi}_0] - \boldsymbol{\pi}_0\boldsymbol{\pi}_0') = \mathbf{V}_0$$

respectively. Finally, the vector  $\mathbf{p}$  is defined as the vector of cumulative relative frequencies

$$\mathbf{p} = \mathbf{C}\mathbf{p}_0 = \begin{pmatrix} 0.1499 \\ 0.3215 \\ 0.5073 \\ 0.7210 \end{pmatrix} \quad (2.8)$$

where  $\mathbf{C} : 4 \times 4$  is a lower triangular matrix of the form

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}. \quad (2.9)$$

The expected value of  $\mathbf{p}$  is given by

$$\begin{aligned} E(\mathbf{p}) &= \mathbf{C}\boldsymbol{\pi}_0 \\ &= \boldsymbol{\pi} \end{aligned} \quad (2.10)$$

with covariance matrix

$$\begin{aligned} Cov(\mathbf{p}) &= \mathbf{C}\mathbf{V}_0\mathbf{C}' \\ &= \mathbf{C}\left\{\frac{1}{n}(\text{diag}[\boldsymbol{\pi}_0] - \boldsymbol{\pi}_0\boldsymbol{\pi}_0')\right\}\mathbf{C}' \\ &= \frac{1}{n}\{\mathbf{C}\text{diag}[\mathbf{C}^{-1}\boldsymbol{\pi}]\mathbf{C}' - \boldsymbol{\pi}\boldsymbol{\pi}'\} \\ &= \mathbf{V}. \end{aligned} \quad (2.11)$$

The ML procedure developed by [10] will be used to estimate the expected cumulative relative frequencies  $\hat{\boldsymbol{\pi}}$  such that it follows a cumulative log-logistic distribution curve at the upper class boundaries  $\mathbf{x}$ . The ML estimation procedure is outlined in Proposition 1.

*Proposition 1 (ML estimation procedure)*

*Consider a random vector of cumulative relative frequencies  $\mathbf{p}$ , which may be considered as a non-singular transformation of the canonical vector of observations, having a distribution belonging to the exponential family, with*

$$E(\mathbf{p}) = \boldsymbol{\pi} \text{ and } Cov(\mathbf{p}) = \mathbf{V}.$$

*The observed  $\mathbf{p}$  is the unrestricted ML estimate of  $\boldsymbol{\pi}$  and the covariance matrix  $\mathbf{V}$  may be a function of  $\boldsymbol{\pi}$ . Let  $\mathbf{g}(\boldsymbol{\pi})$  be a continuous vector valued function of  $\boldsymbol{\pi}$ , for which the first order partial derivatives,*

$$\mathbf{G}_\pi = \frac{\partial \mathbf{g}(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}}$$

*with respect to  $\boldsymbol{\pi}$  exist. The ML estimate of  $\boldsymbol{\pi}$ , subject to the vector of constraints  $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$  is obtained iteratively from*

$$\hat{\boldsymbol{\pi}} = \mathbf{p} - (\mathbf{G}_\pi\mathbf{V})'(\mathbf{G}_p\mathbf{V}\mathbf{G}_\pi')^*\mathbf{g}(\mathbf{p}) \quad (2.12)$$



where  $\mathbf{G}_p = \left. \frac{\partial \mathbf{g}(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}} \right|_{\boldsymbol{\pi}=\mathbf{p}}$  and  $(\mathbf{G}_p \mathbf{V} \mathbf{G}'_p)^*$  is a generalized inverse of  $(\mathbf{G}_p \mathbf{V} \mathbf{G}'_p)$ .

The asymptotic covariance matrix for the ML estimate for  $\boldsymbol{\pi}$  under constraints is also provided by [10] and is given in Proposition 2.

*Proposition 2* The asymptotic covariance matrix of  $\hat{\boldsymbol{\pi}}$ , under  $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$ , is given by

$$\text{Cov}(\hat{\boldsymbol{\pi}}) = \mathbf{V} - (\mathbf{G}_\pi \mathbf{V})' (\mathbf{G}_\pi \mathbf{V} \mathbf{G}'_\pi)^* (\mathbf{G}_\pi \mathbf{V})$$

which is estimated by replacing  $\boldsymbol{\pi}$  with  $\hat{\boldsymbol{\pi}}$ .

One should note that if the constraints are not linearly independent, the generalized inverse for the matrix  $(\mathbf{G}_\pi \mathbf{V} \mathbf{G}'_\pi)$  is used.

With  $E(\mathbf{p}) = \boldsymbol{\pi}$  and  $\text{Cov}(\mathbf{p}) = \mathbf{V}$  defined, the next step discussed in Proposition 1 is to define the vector of constraints  $\mathbf{g}(\boldsymbol{\pi})$ . Since the aim is to find the ML estimate of  $\boldsymbol{\pi}$  under the constraint that the expected cumulative relative frequencies follow a cumulative log-logistic distribution curve at the upper class boundaries  $\mathbf{x}$ , a natural constraint would be that the difference between  $\mathbf{F}(\mathbf{x}; \kappa, \theta)$  and  $\boldsymbol{\pi}$  should equal zero. Hence,

$$\mathbf{g}(\boldsymbol{\pi}) = \mathbf{F}(\mathbf{x}; \kappa, \theta) - \boldsymbol{\pi} = \mathbf{0}. \quad (2.13)$$

In general, one aims to find a simple expression for the matrix of partial derivatives and the parameters of the fitted distribution. To achieve this, a vector of constraints in terms of a linear model is developed by [5] which will still imply (2.13). Assuming (2.13) holds, it is clear to see that

$$\begin{aligned} \boldsymbol{\pi} &= \mathbf{F}(\mathbf{x}; \kappa, \theta) \\ &= \frac{e^{\theta \mathbf{x}^\kappa}}{\mathbf{1} + e^{\theta \mathbf{x}^\kappa}} \end{aligned} \quad (2.14)$$

and therefore

$$\begin{aligned} \mathbf{1} - \boldsymbol{\pi} &= \mathbf{1} - \frac{e^{\theta \mathbf{x}^\kappa}}{\mathbf{1} + e^{\theta \mathbf{x}^\kappa}} \\ &= \frac{\mathbf{1}}{\mathbf{1} + e^{\theta \mathbf{x}^\kappa}} \end{aligned} \quad (2.15)$$

leading to

$$\begin{aligned} \frac{\boldsymbol{\pi}}{\mathbf{1} - \boldsymbol{\pi}} &= \frac{e^{\theta \mathbf{x}^\kappa}}{\mathbf{1} + e^{\theta \mathbf{x}^\kappa}} \times \frac{\mathbf{1} + e^{\theta \mathbf{x}^\kappa}}{\mathbf{1}} \\ &= e^{\theta \mathbf{x}^\kappa}. \end{aligned}$$

The constraint can then be redefined in terms of a linear model

$$\begin{aligned}
 \ln\left(\frac{\boldsymbol{\pi}}{\mathbf{1} - \boldsymbol{\pi}}\right) &= \boldsymbol{\kappa} \ln \mathbf{x} + \boldsymbol{\theta} \mathbf{1} \\
 &= \begin{pmatrix} \ln \mathbf{x} & \mathbf{1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\kappa} \\ \boldsymbol{\theta} \end{pmatrix} \\
 &= \mathbf{X}\boldsymbol{\alpha}
 \end{aligned}
 \tag{2.16}$$

where

$$\mathbf{X} = \begin{pmatrix} \ln \mathbf{x} & \mathbf{1} \end{pmatrix} = \begin{pmatrix} 0.4700 & 1 \\ 1.1632 & 1 \\ 1.8563 & 1 \\ 2.5495 & 1 \end{pmatrix}
 \tag{2.17}$$

and

$$\boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{\kappa} \\ \boldsymbol{\theta} \end{pmatrix}.
 \tag{2.18}$$

To see how the linear model will be applied, one can define a projection matrix onto the vector space of the design matrix  $\mathbf{X}$  as

$$\begin{aligned}
 \mathbf{P}_{\mathbf{X}} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\
 &= \begin{pmatrix} 0.7 & 0.4 & 0.1 & -0.2 \\ 0.4 & 0.3 & 0.2 & 0.1 \\ 0.1 & 0.2 & 0.3 & 0.4 \\ -0.2 & 0.1 & 0.4 & 0.7 \end{pmatrix}
 \end{aligned}$$

with the projection matrix onto the error space of  $\mathbf{X}$  being

$$\begin{aligned}
 \mathbf{Q}_{\mathbf{X}} &= \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\
 &= \begin{pmatrix} 0.3 & -0.4 & -0.1 & 0.2 \\ -0.4 & 0.7 & -0.2 & -0.1 \\ -0.1 & -0.2 & 0.7 & -0.4 \\ 0.2 & -0.1 & -0.4 & 0.3 \end{pmatrix}.
 \end{aligned}$$

If one multiplies a vector with the matrix  $\mathbf{P}_{\mathbf{X}}$ , it will project the vector onto the vector space of  $\mathbf{X}$ . If one multiplies a vector that is already in the vector space of  $\mathbf{X}$  with the matrix  $\mathbf{Q}_{\mathbf{X}}$ , the result will be a zero vector. If one assumes that  $\ln\left(\frac{\pi}{\mathbf{1}-\pi}\right)$  is in the vector space of  $\mathbf{X}$ , i.e. the vector of constraints in (2.13) holds, then the vector of constraints can be redefined using the linear model as

$$g(\boldsymbol{\pi}) = \mathbf{Q}_{\mathbf{X}} \ln\left(\frac{\boldsymbol{\pi}}{\mathbf{1}-\boldsymbol{\pi}}\right) = \mathbf{0}. \quad (2.19)$$

Note that (2.13) and (2.19) imply the same constraints. With the amount of linearly independent functions of the vector of constraints being

$$\begin{aligned} r &= \text{rank}(\mathbf{Q}_{\mathbf{X}}) \\ &= \text{rank}(\mathbf{I}) - \text{rank}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\ &= (5 - 1) - \text{rank}(\mathbf{x}) \\ &= 4 - 2 \\ &= 2 \end{aligned} \quad (2.20)$$

one can see that the amount of intervals used for income may not be less than 4. The matrix of partial derivatives of  $g(\boldsymbol{\pi})$  is given by

$$\begin{aligned} \mathbf{G}_{\boldsymbol{\pi}} &= \frac{\partial \mathbf{g}(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}} \\ &= \frac{\partial}{\partial \boldsymbol{\pi}} \left\{ \mathbf{Q}_x \ln\left(\frac{\boldsymbol{\pi}}{\mathbf{1}-\boldsymbol{\pi}}\right) \right\} \\ &= \mathbf{Q}_x \mathbf{D}_{\boldsymbol{\pi}} \end{aligned} \quad (2.21)$$

where

$$\begin{aligned} \mathbf{D}_{\boldsymbol{\pi}} &= \frac{\partial}{\partial \boldsymbol{\pi}} \left\{ \ln\left(\frac{\boldsymbol{\pi}}{\mathbf{1}-\boldsymbol{\pi}}\right) \right\} \\ &= \frac{\partial}{\partial \boldsymbol{\pi}} \{ \ln(\boldsymbol{\pi}) - \ln(\mathbf{1}-\boldsymbol{\pi}) \} \\ &= \{ \text{diag}[\boldsymbol{\pi}] \}^{-1} + \{ \text{diag}[\mathbf{1}-\boldsymbol{\pi}] \}^{-1}. \end{aligned} \quad (2.22)$$

The ML estimation procedure can now be employed by making use of a double iterative procedure described next.

### The double iterative procedure

1. Set  $\mathbf{p}$  equal to the unrestricted ML estimate for  $\boldsymbol{\pi}$ , i.e. the observed cumulative relative frequencies.
2. Do until convergence over  $\boldsymbol{\pi}$ 
  - a. Set  $\boldsymbol{\pi}$  equal to  $\mathbf{p}$  and calculate  $\mathbf{V}$  and  $\mathbf{G}_{\boldsymbol{\pi}}$
  - b. Set  $\mathbf{p}$  equal to the unrestricted ML estimate for  $\boldsymbol{\pi}$ .

- c. Do until convergence over  $\mathbf{p}$ 
  - i. Calculate  $\mathbf{g}(\mathbf{p})$  and  $\mathbf{G}_p$ .
  - ii. Calculate  $\mathbf{p}$  using (2.12).

This will result in the ML estimate

$$\hat{\boldsymbol{\pi}} = \begin{pmatrix} 0.1548869 \\ 0.3068302 \\ 0.5166974 \\ 0.7208354 \end{pmatrix}$$

under the constraints set out in (2.19). Since  $\boldsymbol{\pi} = \mathbf{C}\boldsymbol{\pi}_0$  where  $\boldsymbol{\pi}_0 = \frac{1}{n}\mathbf{f}$ , the ML estimate for  $\boldsymbol{\pi}$  can be transformed back into expected frequencies by applying the transformation

$$\hat{\mathbf{f}} = n\mathbf{C}^{-1}\hat{\boldsymbol{\pi}}. \tag{2.23}$$

The expected frequencies are then given in Table 2.2.

Code	Monthly	Expected Frequency	Percent
2 + 3 + 4	R 1 - R 1600	75847.642	15.49
5	R 1 601 - R 3 200	74406.19	15.19
6	R 3 201 - R 6 400	102771.34	20.99
7	R6 400 - 12 800	99965.781	20.41
8 - 12	R 12 801 or more	136706.05	27.92
<b>Total</b>		<b>489697</b>	<b>100</b>

**Table 2.2: Expected frequencies**

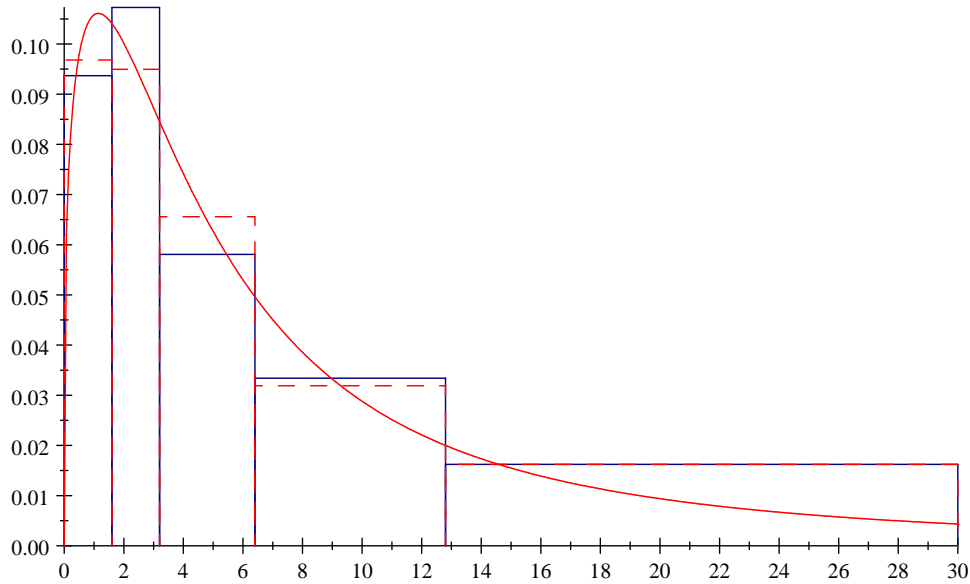
The ML estimates of the parameters of the log-logistic distribution  $\hat{\boldsymbol{\alpha}}$  can now be determined by

$$\begin{aligned} \hat{\boldsymbol{\alpha}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \ln \left( \frac{\hat{\boldsymbol{\pi}}}{\mathbf{1} - \hat{\boldsymbol{\pi}}} \right) \\ &= \begin{pmatrix} \hat{\kappa} \\ \hat{\theta} \end{pmatrix} \\ &= \begin{pmatrix} 1.2722 \\ -2.2947 \end{pmatrix} \end{aligned} \tag{2.24}$$

with estimated log-logistic pdf given by

$$\hat{f}(x; \hat{\kappa}, \hat{\theta}) = \frac{e^{-2.2947} (1.2722) x^{1.2722-1}}{(1 + e^{-2.2947} x^{1.2722})^2}. \tag{2.25}$$

The distribution can now be presented with the observed and expected frequencies in Figure 2-2. The blue solid histogram represents the observed frequencies whereas the red dashed histogram represents the expected frequencies.



**Figure 2-2: Fitted log-logistic distribution**

Using the multivariate delta theorem in conjunction with Proposition 2, one can also determine the covariance matrix of  $\hat{\boldsymbol{\alpha}}$

$$\begin{aligned}
 Cov(\hat{\boldsymbol{\alpha}}) &= \left( \frac{\partial \hat{\boldsymbol{\alpha}}}{\partial \boldsymbol{\pi}} \right) Cov(\hat{\boldsymbol{\pi}}) \left( \frac{\partial \hat{\boldsymbol{\alpha}}}{\partial \boldsymbol{\pi}} \right)' \\
 &= \left\{ (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{D}_{\boldsymbol{\pi}} \right\} Cov(\hat{\boldsymbol{\pi}}) \left\{ (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{D}_{\boldsymbol{\pi}} \right\}' \\
 &= \begin{pmatrix} 4.22 \times 10^{-6} & 0 \\ 0 & 1.9 \times 10^{-6} \end{pmatrix}
 \end{aligned}$$

and since the parameters are approximately normally distributed, one can set up confidence intervals for the true population values or do hypothesis testing.

From Figure 2-2 one can see that the estimated frequencies are reasonably similar to the observed frequencies. A convenient way to test if  $\mathbf{p}$  deviates significantly from the ML estimate  $\hat{\boldsymbol{\pi}}$ , attained under the vector of constraints  $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$ , is to formulate the null hypothesis

$$\mathbf{H}_0 : \mathbf{g}(\boldsymbol{\pi}) = \mathbf{0} \quad (2.26)$$

where this hypothesis can be tested by using some goodness-of-fit statistic like the Pearson  $\chi^2$ -statistic

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - \hat{f}_i)^2}{\hat{f}_i} \quad (2.27)$$

with  $f_i$  and  $\hat{f}_i$  being the observed and expected frequency in interval  $i$  and  $k$  being the amount of intervals for the response variable. The transformation from  $\hat{\boldsymbol{\pi}}$  to  $\hat{\mathbf{f}}$  is achieved by using (2.23).

The Pearson  $\chi^2$ -statistic will follow a  $\chi^2(r)$  distribution where  $r = 2$ , the number of linear indepen-

dent functions of the vector of constraints  $\mathbf{g}(\boldsymbol{\pi})$  that was calculated in (2.20).

By setting  $\mathbf{p}$  and  $\boldsymbol{\pi}$  equal to the unrestricted ML estimate of  $\boldsymbol{\pi}$ , the Wald statistic can be calculated using

$$W = \mathbf{g}(\mathbf{p})'(\mathbf{G}_p \mathbf{V} \mathbf{G}'_p)^{-1} \mathbf{g}(\mathbf{p}) \quad (2.28)$$

where  $W$  also follows a  $\chi^2(r)$  distribution under the null hypothesis.

Finally, one can consider the measure of discrepancy

$$D = \frac{W}{n} \quad (2.29)$$

which will give a more conservative result for large sample sizes. A rule of thumb is that the null hypothesis can not be rejected if the measure of discrepancy is less than 0.05.

The results attained for these goodness of fit statistics are given in Table 2.3.

Goodness-of-fit statistics	Value	p-value	Discrepancy
Pearson	2904.2515	0	N/A
Wald	2941.99	0	0.0060078

**Table 2.3: Goodness-of-fit statistics for log-logistic distribution**

The measure of discrepancy in Table 2.3 is 0.006. Since it is less than 0.05 it indicates that the estimates for  $\boldsymbol{\pi}$  does not differ significantly from the observed cumulative relative frequencies.

An important aspect to note is that this process can easily be applied to cases where  $k$  is not equal to 5. Also note that a different formulation of the vector of constraints can also be used as long as it is still equivalent to (2.13).

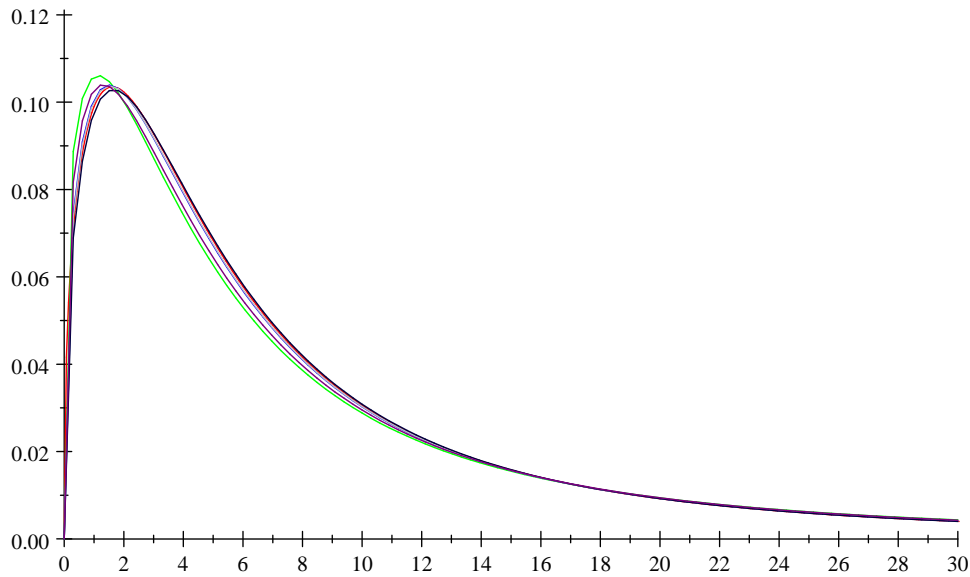
## 2.3 The effect of the defined income categories

Different scenarios on the categorization of the response variable will be considered and are given in Table 2.4. Although completely different groupings of the response variable are used, if the estimated parameters and distribution curves in Figure 2-3 is considered, it is evident that the fitted log-logistic distributions are very close to each other.

	Light red	Navy	Light green	Purple	Blue	Light grey
$\mathbf{x}$	$\begin{pmatrix} 0.4 \\ 0.8 \\ 1.6 \\ 3.2 \\ 6.4 \\ 12.8 \\ 25.6 \\ 51.2 \\ 102.4 \\ 204.8 \end{pmatrix}$	$\begin{pmatrix} 1.6 \\ 6.4 \\ 51.2 \end{pmatrix}$	$\begin{pmatrix} 1.6 \\ 3.2 \\ 6.4 \\ 12.8 \end{pmatrix}$	$\begin{pmatrix} 0.8 \\ 1.6 \\ 3.2 \\ 6.4 \\ 12.8 \end{pmatrix}$	$\begin{pmatrix} 0.8 \\ 1.6 \\ 3.2 \\ 6.4 \\ 12.8 \\ 25.6 \end{pmatrix}$	$\begin{pmatrix} 0.4 \\ 0.8 \\ 1.6 \\ 3.2 \\ 6.4 \\ 12.8 \\ 25.6 \end{pmatrix}$
$\begin{pmatrix} \hat{\kappa} \\ \hat{\theta} \end{pmatrix}$	$\begin{pmatrix} 1.3878 \\ -2.4710 \end{pmatrix}$	$\begin{pmatrix} 1.3984 \\ -2.5044 \end{pmatrix}$	$\begin{pmatrix} 1.2722 \\ -2.2947 \end{pmatrix}$	$\begin{pmatrix} 1.3097 \\ -2.3657 \end{pmatrix}$	$\begin{pmatrix} 1.3638 \\ -2.4402 \end{pmatrix}$	$\begin{pmatrix} 1.3686 \\ -2.4491 \end{pmatrix}$

**Table 2.4: Parameter estimates for different grouping scenarios**

Applying the ML estimation procedure to each of the scenarios, one will find the ML estimates in Table 2.4 and the respective log-logistic distributions plotted in Figure 2-3.



**Figure 2-3: Pdf of log-logistic distribution fits for different grouping scenarios**

From Figure 2-3 one can see that the grouping does not have a significant effect on the estimated distribution if the grouping is done in an intelligent way. For simplicity, the vector of upper class boundaries used in Chapter 2 and corresponding with the light green log-logistic distribution will be used throughout this mini-dissertation.

## 2.4 Summary

In this chapter Proposition 1 developed by [10] was introduced. It was shown how this can be used to fit a log-logistic distribution to the grouped response variable, income. Specifically, the ML estimate of the cumulative relative frequencies  $\hat{\pi}$  are attained under the vector of constraints that imply that  $\boldsymbol{\pi}$  should equal the cdf of a log-logistic distribution  $\mathbf{F}(\mathbf{x};\kappa, \theta)$ . This constraint is redefined in terms of a linear model. From this, the ML estimates of the parameters for the log-logistic distribution are attained with the corresponding goodness-of-fit statistics.



# Chapter 3

## Fitting other distributions

Very often the grouped response variable may have a different underlying distribution. It will now be illustrated how to fit a normal and exponential distribution to the data presented in Table 2.1.

Code	Monthly	Frequency	Percent
2 + 3 + 4	R 1 - R 1600	73393	14.99
5	R 1 601 - R 3 200	84065	17.17
6	R 3 201 - R 6 400	90983	18.58
7	R6 400 - 12 800	104642	21.37
8 - 12	R 12 801 or more	136614	27.90
<b>Total</b>		<b>489697</b>	<b>100</b>

**Table 3.1: Final income distribution of focus group**

### 3.1 The Normal Distribution

Although the normal distribution is not known to fit income distributions well, it is used here to elucidate the application of the iterative process. The pdf of the normal distribution is given by

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right\}, \dots -\infty < x < \infty \quad (3.1)$$

where  $-\infty < \mu < \infty$  and  $\sigma > 0$ . Defining a standardized variable  $z = \frac{x - \mu}{\sigma}$ , the standard normal distribution is attained with pdf

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} z^2 \right\}. \quad (3.2)$$

The cdf will be denoted as  $\Phi(z)$ .

As stated before, to fit a distribution to the grouped response variable, it is required that the expected cumulative relative frequencies  $\pi$  equal the cdf  $F(\mathbf{x})$ , at the upper class boundaries  $\mathbf{x}$ . Hence, the vector of constraints will be

$$\mathbf{g}(\boldsymbol{\pi}) = \Phi\left(\frac{\mathbf{x} - \mu\mathbf{1}}{\sigma}\right) - \boldsymbol{\pi} = \mathbf{0}. \quad (3.3)$$

The aim is to define  $\mathbf{g}(\boldsymbol{\pi})$  in the simplest way as to still imply (3.3). This was done by [5] and the results are shown here with the relevant output attained from the ML iterative procedure.

The standardized upper class boundaries can be expressed in terms of a linear model

$$\begin{aligned} \mathbf{z} &= \left(\frac{\mathbf{x} - \mu\mathbf{1}}{\sigma}\right) \\ &= \mathbf{X}\boldsymbol{\alpha} \end{aligned} \quad (3.4)$$

where

$$\mathbf{X} = \begin{pmatrix} \mathbf{x} & -\mathbf{1} \end{pmatrix} \quad (3.5)$$

and

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma} \\ \frac{\mu}{\sigma} \end{pmatrix}. \quad (3.6)$$

Under the constraint

$$\mathbf{g}(\boldsymbol{\pi}) = \Phi(\mathbf{z}) - \boldsymbol{\pi} = \mathbf{0}$$

it is observed that

$$\begin{aligned} \Phi^{-1}(\boldsymbol{\pi}) &= \mathbf{z} \\ &= \mathbf{X}\boldsymbol{\alpha}. \end{aligned}$$

Hence, if one defines  $\mathbf{Q}_\mathbf{X} = \mathbf{I} - \mathbf{P}_\mathbf{X}$  to be the projection matrix onto the error space of  $\mathbf{X}$ , a new set of constraints can be defined as

$$\mathbf{g}(\boldsymbol{\pi}) = \mathbf{Q}_\mathbf{X}\Phi^{-1}(\boldsymbol{\pi}) = \mathbf{0} \quad (3.7)$$

with matrix of partial derivatives

$$\begin{aligned} \mathbf{G}_\pi &= \frac{\partial \mathbf{g}(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}} \\ &= \mathbf{Q}_\mathbf{X}\mathbf{D}_\pi \end{aligned} \quad (3.8)$$

where  $\mathbf{D}_\pi = (\text{diag}[\phi(\Phi^{-1}(\boldsymbol{\pi}))])^{-1}$ .

Utilizing the ML estimation procedure, the ML estimate for  $\boldsymbol{\pi}$  under the constraints in 3.7 can be

attained. The natural parameter estimates are then given by

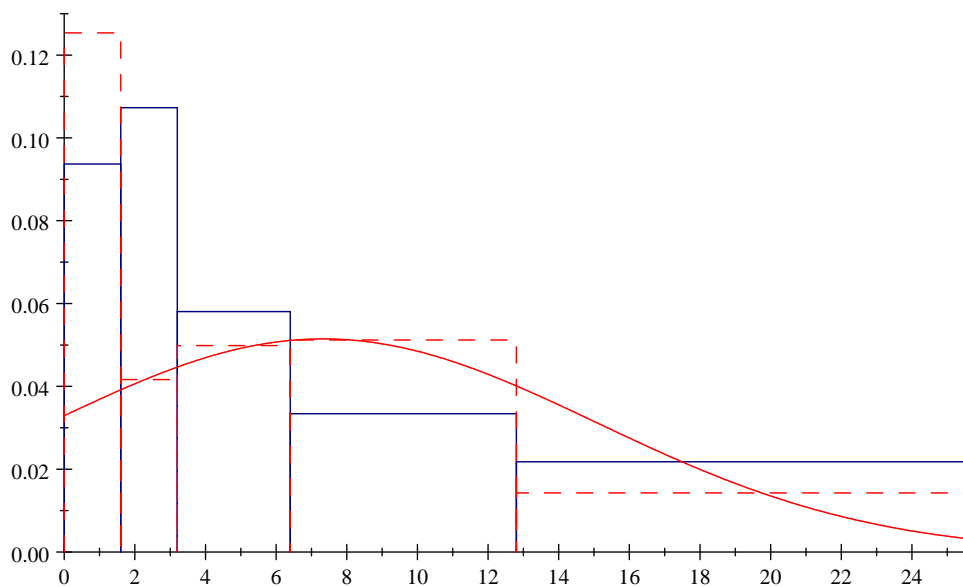
$$\begin{aligned}\hat{\alpha} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Phi^{-1}(\hat{\pi}) \\ &= \begin{pmatrix} 0.1365 \\ 1.0580 \end{pmatrix}.\end{aligned}$$

One can then use (3.6) to attain the estimated mean and standard deviation for the fitted normal distribution:

$$\hat{\mu} = 7.7529$$

$$\hat{\sigma} = 7.3279$$

The variances for these estimates can be found using the multivariate delta theorem and Proposition 2. The fitted distribution is then given in Figure 3-1.



**Figure 3-1: Fitted normal distribution**

It is clear that the normal distribution does not fit the data well. The measure of discrepancy of 0.1283 is significantly larger than 0.05. This further confirms that the normal distribution does not give an adequate fit.

## 3.2 The Exponential Distribution

The normal distribution fails to take into account that the distribution is positively skewed. A study was done by [7] where they attempted to fit an exponential distribution to individual income for the USA using census data from 1996. Although not using the same method used here, they found that the exponential distribution indicated an adequate fit. Hence, in this section the exponential distribution

is fitted to the grouped response variable presented earlier using the method described in Proposition 1. The theory discussed here is fully derived by [5]. The reader is referred there for further detail.

The pdf and cdf of the exponential distribution with mean parameter  $\mu$  is given by

$$f(x; \mu) = \frac{1}{\mu} e^{-\frac{x}{\mu}}, \dots x > 0 \quad (3.9)$$

where  $\mu > 0$ , and

$$F(x; \mu) = 1 - e^{-\frac{x}{\mu}}. \quad (3.10)$$

The vector of constraints that will be used is

$$\begin{aligned} \mathbf{g}(\boldsymbol{\pi}) &= \mathbf{F}(\mathbf{x}; \mu) - \boldsymbol{\pi} \\ &= \{\mathbf{1} - \exp(-\theta \mathbf{x})\} - \boldsymbol{\pi} = \mathbf{0} \end{aligned} \quad (3.11)$$

from which the linear model

$$\begin{aligned} \{\mathbf{1} - \exp(-\theta \mathbf{x})\} - \boldsymbol{\pi} &= \mathbf{0} \\ \exp(-\theta \mathbf{x}) &= \mathbf{1} - \boldsymbol{\pi} \\ \ln(\mathbf{1} - \boldsymbol{\pi}) &= -\theta \mathbf{x} \end{aligned} \quad (3.12)$$

is developed. This implies that, under the constraint,  $\ln(\mathbf{1} - \boldsymbol{\pi})$  is a scalar multiple of the vector of upper class boundaries  $\mathbf{x}$ . Hence,  $\ln(\mathbf{1} - \boldsymbol{\pi})$  must be in the vector space generated by  $\mathbf{x}$ . If one defines the projection matrix onto the error space of  $\mathbf{x}$  as  $\mathbf{Q}_x = \mathbf{I} - \mathbf{P}_x$  then the vector of constraints can be expressed by

$$\mathbf{g}(\boldsymbol{\pi}) = \mathbf{Q}_x \ln(\mathbf{1} - \boldsymbol{\pi}) \quad (3.13)$$

with matrix of partial derivatives

$$\begin{aligned} \mathbf{G}_\pi &= \frac{\partial \mathbf{g}(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}} \\ &= \frac{\partial}{\partial \boldsymbol{\pi}} \{\mathbf{Q}_x \ln(\mathbf{1} - \boldsymbol{\pi})\} \\ &= \mathbf{Q}_x \mathbf{D}_\pi \end{aligned} \quad (3.14)$$

where  $\mathbf{D}_\pi = -(\text{diag}[\mathbf{1} - \boldsymbol{\pi}])^{-1}$ . By applying the iterative procedure, the ML estimate for  $\boldsymbol{\pi}$  under the constraints is obtained, implying the ML estimator for  $\theta$  is given by

$$\begin{aligned} \hat{\theta} &= -\frac{\mathbf{x}' \ln(\mathbf{1} - \hat{\boldsymbol{\pi}})}{\mathbf{x}' \mathbf{x}} \\ &= 0.1038 \end{aligned} \quad (3.15)$$

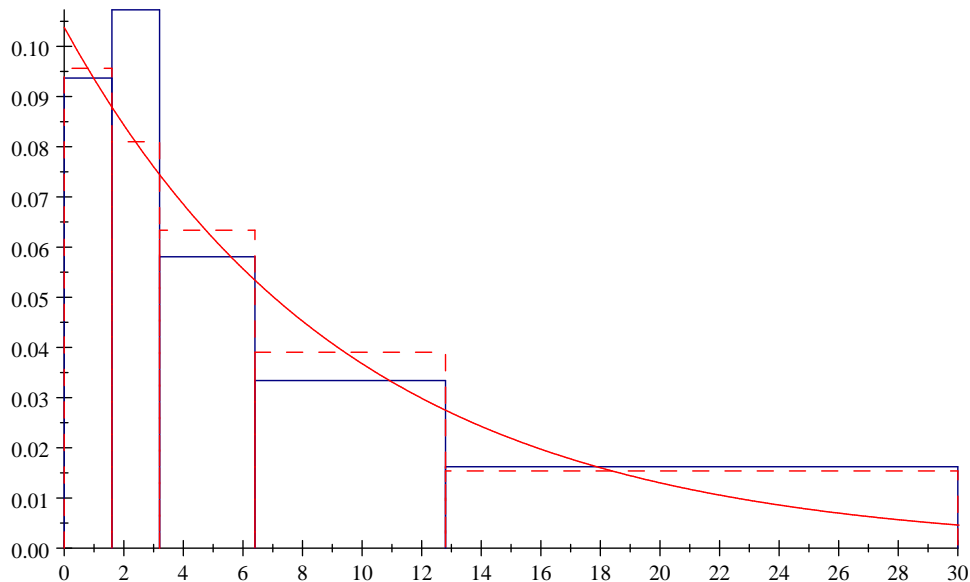
with the corresponding ML estimator for  $\mu$  being

$$\begin{aligned}\hat{\mu} &= \frac{1}{\hat{\theta}} \\ &= 9.6354.\end{aligned}\tag{3.16}$$

The variance of the parameter is attained using the multivariate delta theorem with Proposition 2. Note the multivariate delta theorem should first be applied to attain the variance of  $\hat{\theta}$ . After doing so, the theorem is applied again to attain the variance of  $\hat{\mu}$ .

$$\text{Var}(\hat{\mu}) \cong 0.0003\tag{3.17}$$

The fitted distribution is given in Figure 3-2.



**Figure 3-2: Fitted exponential distribution**

The measure of discrepancy found was 0.01765. This indicates that the exponential distribution does fit the income distribution adequately. Upon further inspection one should note though that since the exponential distribution is a strictly decreasing function of income, it will not take into account the spike in frequencies observed in the second income category. This spike is in fact taken into account when the log-logistic distribution is considered in Chapter 2.

### 3.3 Summary

It should be noted with what ease the method used can be expanded to accommodate different types of distributions. In Chapter 3 the aim was to show how to apply the technique to estimate the expected cumulative relative frequencies such that they follow different distributions at the upper class boundaries  $\mathbf{x}$  of the grouped response variable.

# Chapter 4

## Single factor design

### 4.1 Introduction

To study the effect of an explanatory variable on the grouped response variable income, the intervals of income will be cross-tabulated with the categories of the explanatory variable being considered. From this, different log-logistic distributions will be fitted simultaneously to each of the categories, i.e. cells, of the explanatory variable. Since the log-logistic distribution is positively skewed, the medians of the different log-logistic distributions will be studied to see what the effect of the explanatory variable is on the median income level. The explanatory variables that will be considered are

- population group,
- gender,
- level of education, and
- age.

### 4.2 Population group

The histograms attained from cross-tabulating population group with the frequency distribution of income are given in Table 4.1.

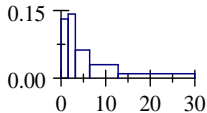
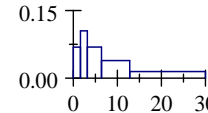
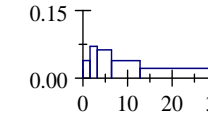
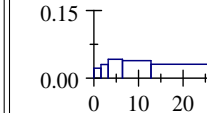
Black	Coloured	Indian	White
			
$n = 298656$	$n = 44129$	$n = 26684$	$n = 120228$

Table 4.1: Histograms of income under different population groups

### 4.2.1 Distribution fitting

The ML estimates of the expected cumulative relative frequencies  $\pi_1, \pi_2, \pi_3$  and  $\pi_4$  of each of the  $T = 4$  cells need to be found simultaneously under the constraint that  $\pi_1, \pi_2, \pi_3$  and  $\pi_4$  follow log-logistic distributions at the upper class boundaries  $\mathbf{x}$ . For simplicity, only a constant vector of upper class boundaries  $\mathbf{x}$  will be considered but this methodology can be extended so that the vector of upper class boundaries is different for each cell.

Table 4.2 presents the cross-tabulated frequencies of the intervals of income and the categories of population group that will be used in the estimation procedure.

		Income				
Population group	Cell	1 – 1600	1601 – 3200	3201 – 6400	6400 – 12800	12801 – $\infty$
Black	1	62620	67767	59686	57136	51417
Coloured	2	4847	7403	9738	11023	11118
Indian	3	1668	3009	5365	6615	10027
White	4	4258	5856	16194	29868	64052

**Table 4.2: Frequencies of population group**

Defining

$$\mathbf{F}_M = \begin{pmatrix} 62620 & 67767 & 59686 & 57136 \\ 4847 & 7403 & 9738 & 11023 \\ 1668 & 3009 & 5365 & 6615 \\ 4258 & 5856 & 16194 & 29868 \end{pmatrix}$$

as a  $T \times (k - 1)$  matrix containing the first  $(k - 1) = 4$  frequencies of each of the  $T = 4$  cells, the vector of frequencies

$$\mathbf{f} = \text{vec}(\mathbf{F}_M) = \begin{pmatrix} 62620 \\ 67797 \\ 59686 \\ 57136 \\ 4847 \\ 7403 \\ 9738 \\ 11023 \\ \vdots \\ 4258 \\ 5856 \\ 16194 \\ 29868 \end{pmatrix} \quad (4.1)$$

can be defined by concatenating  $\mathbf{F}_M$  row wise and is assumed to follow a product multinomial distribution with a block diagonal covariance matrix. This will ensure that the sub-sample sizes for each population group will be kept fixed during the estimation procedure. With the vector of row totals being defined as

$$\mathbf{n} = \begin{pmatrix} 298656 & 44129 & 26684 & 120228 \end{pmatrix}' \quad (4.2)$$

the relative and cumulative relative frequencies

$$\mathbf{p}_0 = \begin{pmatrix} \mathbf{p}_{01} \\ \mathbf{p}_{02} \\ \mathbf{p}_{03} \\ \mathbf{p}_{04} \end{pmatrix} = \begin{pmatrix} \frac{1}{n_1} \mathbf{f}_1 \\ \frac{1}{n_2} \mathbf{f}_2 \\ \frac{1}{n_3} \mathbf{f}_3 \\ \frac{1}{n_4} \mathbf{f}_4 \end{pmatrix} = ((diag(\mathbf{n}))^{-1} \otimes \mathbf{I}_4) \mathbf{f}$$

and

$$\mathbf{p} = \begin{pmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \mathbf{p}_3 \\ \mathbf{p}_4 \end{pmatrix} = \begin{pmatrix} \mathbf{Cp}_{01} \\ \mathbf{Cp}_{02} \\ \mathbf{Cp}_{03} \\ \mathbf{Cp}_{04} \end{pmatrix} = (\mathbf{I}_4 \otimes \mathbf{C}) \mathbf{p}_0 \quad (4.3)$$

can be defined, where the matrix  $\mathbf{C}$  is as defined in (2.9).

The expected values of  $\mathbf{p}_0$  and  $\mathbf{p}$  are defined as

$$E(\mathbf{p}_0) = \begin{pmatrix} \pi_{01} \\ \pi_{02} \\ \pi_{03} \\ \pi_{04} \end{pmatrix} = \boldsymbol{\pi}_0$$

and

$$E(\mathbf{p}) = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{pmatrix} = \boldsymbol{\pi} \quad (4.4)$$

respectively. The covariance of  $\mathbf{p}_0$  is

$$Cov(\mathbf{p}_0) = \begin{pmatrix} \mathbf{V}_{01} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{02} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{V}_{03} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{V}_{04} \end{pmatrix} = \mathbf{V}_0$$

where



$$Cov(\mathbf{p}_{0t}) = \frac{1}{n_t} (diag(\boldsymbol{\pi}_{0t}) - \boldsymbol{\pi}_{0t}\boldsymbol{\pi}'_{0t}) = \mathbf{V}_{0t}$$

is on its diagonals. This is also expressed in terms of Kronecker products in [5] with

$$\mathbf{V}_0 = \{diag[\mathbf{n}]^{-1} \otimes \mathbf{I}_{k-1}\} \{diag[\boldsymbol{\pi}_0] - diag[\boldsymbol{\pi}_0] (\mathbf{I}_T \otimes (\mathbf{1}_{k-1}\mathbf{1}'_{k-1})) diag[\boldsymbol{\pi}_0]\}.$$

If  $k = 5$  and  $T = 4$ , the covariance matrix of  $\mathbf{p}$  follows where

$$\mathbf{V} = (\mathbf{I}_T \otimes \mathbf{C}) \mathbf{V}_0 (\mathbf{I}_T \otimes \mathbf{C})'. \quad (4.5)$$

The block structure of the covariance matrix assumes independence between the different cells. If this assumption is not made, the vector of row totals  $\mathbf{n}$  will not stay constant and one may find expected cumulative relative frequencies being greater than 1. Log-logistic distributions is now fitted to each of the  $T = 4$  cells simultaneously such that

$$\begin{pmatrix} F_1(\mathbf{x}, \boldsymbol{\alpha}_1) \\ F_2(\mathbf{x}, \boldsymbol{\alpha}_2) \\ F_3(\mathbf{x}, \boldsymbol{\alpha}_3) \\ F_4(\mathbf{x}, \boldsymbol{\alpha}_4) \end{pmatrix} = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{pmatrix}$$

where  $\boldsymbol{\alpha}_t = \begin{pmatrix} \kappa_t \\ \theta_t \end{pmatrix}$  for  $t = 1, 2, 3, 4$ , and

$$\boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \\ \boldsymbol{\alpha}_3 \\ \boldsymbol{\alpha}_4 \end{pmatrix}$$

with vector of constraints

$$g(\boldsymbol{\pi}) = \begin{pmatrix} F_1(\mathbf{x}, \boldsymbol{\alpha}_1) \\ F_2(\mathbf{x}, \boldsymbol{\alpha}_2) \\ F_3(\mathbf{x}, \boldsymbol{\alpha}_3) \\ F_4(\mathbf{x}, \boldsymbol{\alpha}_4) \end{pmatrix} - \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{pmatrix} = \mathbf{0}. \quad (4.6)$$

Using

$$\ln\left(\frac{\pi_t}{\mathbf{1} - \pi_t}\right) = \kappa_t \ln \mathbf{x} + \theta_t \mathbf{1} = \mathbf{X}\boldsymbol{\alpha}_t$$

from (2.16) with  $\mathbf{X} = \begin{pmatrix} \ln \mathbf{x} & \mathbf{1} \end{pmatrix}$ , a linear model

$$\ln \left( \frac{\boldsymbol{\pi}}{\mathbf{1} - \boldsymbol{\pi}} \right) = \begin{pmatrix} \mathbf{X}\boldsymbol{\alpha}_1 \\ \mathbf{X}\boldsymbol{\alpha}_2 \\ \mathbf{X}\boldsymbol{\alpha}_3 \\ \mathbf{X}\boldsymbol{\alpha}_4 \end{pmatrix} = (\mathbf{I}_4 \otimes \mathbf{X}) \boldsymbol{\alpha}$$

can be defined. The vector of constraints  $g(\boldsymbol{\pi}) = \mathbf{0}$  in terms of the linear model follows in a similar manner as before where

$$\mathbf{g}_{\log}(\boldsymbol{\pi}) = \begin{pmatrix} \mathbf{Q}_X \ln \left( \frac{\pi_1}{1 - \pi_1} \right) \\ \mathbf{Q}_X \ln \left( \frac{\pi_2}{1 - \pi_2} \right) \\ \mathbf{Q}_X \ln \left( \frac{\pi_3}{1 - \pi_3} \right) \\ \mathbf{Q}_X \ln \left( \frac{\pi_4}{1 - \pi_4} \right) \end{pmatrix} = (\mathbf{I}_T \otimes \mathbf{Q}_X) \ln \left( \frac{\boldsymbol{\pi}}{\mathbf{1} - \boldsymbol{\pi}} \right) \quad (4.7)$$

and  $\mathbf{Q}_X$  is the projection matrix onto the error space of  $\mathbf{X}$ . The matrix of partial derivatives is then given by

$$\mathbf{G}_{\log}(\boldsymbol{\pi}) = (\mathbf{I}_T \otimes \mathbf{Q}_X) \mathbf{D}_{\boldsymbol{\pi}} \quad (4.8)$$

where  $\mathbf{D}_{\boldsymbol{\pi}} = \{\text{diag}[\boldsymbol{\pi}]\}^{-1} + \{\text{diag}[\mathbf{1} - \boldsymbol{\pi}]\}^{-1}$ . The ML estimate for  $\boldsymbol{\pi}$  under the constraints can now be determined using the iterative procedure which will produce the estimated frequencies given in Table 4.3.

	Income				
Population group	1 – 1600	1601 – 3200	3201 – 6400	6400 – 12800	12801 – ∞
Black	64692.736	59958.617	69431.103	53154.83	51418.714
Coloured	4989.2283	6757.3987	10668.246	10506.867	11207.26
Indian	1811.689	2775.4955	5331.8327	6831.3258	9933.6569
White	3662.3323	6770.5372	16405.348	29067.027	64322.756

**Table 4.3: Estimated frequencies of population group**

The fitted log-logistic distributions for each cell with the associated parameters are given in Table 4.4. Equations (4.9) and (4.10) provide simple expressions to find  $\hat{\kappa}_t$  and  $\hat{\theta}_t$  in separate vectors.

$$\begin{aligned} \hat{\boldsymbol{\kappa}} &= \begin{pmatrix} \hat{\kappa}_1 \\ \hat{\kappa}_2 \\ \hat{\kappa}_3 \\ \hat{\kappa}_4 \end{pmatrix} = \left[ \left( \mathbf{I}_T \otimes \left[ (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \right]_1 \right) \right] \ln \left( \frac{\hat{\boldsymbol{\pi}}}{\mathbf{1} - \hat{\boldsymbol{\pi}}} \right) \\ &= \begin{pmatrix} 1.3734 \\ 1.5088 \\ 1.5110 \\ 1.5966 \end{pmatrix} \end{aligned} \quad (4.9)$$

and

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \hat{\theta}_3 \\ \hat{\theta}_4 \end{pmatrix} = \left( \mathbf{I}_T \otimes [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']_2 \right) \ln \left( \frac{\hat{\boldsymbol{\pi}}}{\mathbf{1} - \hat{\boldsymbol{\pi}}} \right) \\ &= \begin{pmatrix} -1.9310 \\ -2.7690 \\ -3.3210 \\ -4.2108 \end{pmatrix}\end{aligned}\quad (4.10)$$

where  $[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']_1$  and  $[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']_2$  refer to the 1<sup>st</sup> and 2<sup>nd</sup> row of  $[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']$ , respectively. The covariance matrix of  $\boldsymbol{\alpha}$  is given by

$$\begin{aligned}Cov(\hat{\boldsymbol{\alpha}}) &= \left( \frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{\pi}} \right) Cov(\hat{\boldsymbol{\pi}}) \left( \frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{\pi}} \right)' \\ &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \mathbf{D}_{\boldsymbol{\pi}} Cov(\hat{\boldsymbol{\pi}}) \mathbf{D}'_{\boldsymbol{\pi}} [\mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}]'\end{aligned}$$

#### 4.2.2 Saturated model

The effect of population group on the median income of a fitted log-logistic distribution is investigated using the median income

$$\nu = \exp \left( -\frac{\boldsymbol{\theta}}{\boldsymbol{\kappa}} \right) \quad (4.11)$$

of each cell as a representative measure due to the log-logistic distribution being positively skew. The covariance matrix of  $\hat{\boldsymbol{\nu}}$  is derived in [5] and is given by

$$Cov(\hat{\boldsymbol{\nu}}) = \left( \frac{\partial \boldsymbol{\nu}}{\partial \boldsymbol{\alpha}} \right) Cov(\hat{\boldsymbol{\alpha}}) \left( \frac{\partial \boldsymbol{\nu}}{\partial \boldsymbol{\alpha}} \right)'$$

where the matrix  $\mathbf{A} = \left( \frac{\partial \boldsymbol{\nu}}{\partial \boldsymbol{\alpha}} \right)$  can be expressed as

$$\mathbf{A} = \left( \text{diag}[\mathbf{a}_{\boldsymbol{\kappa}}] \otimes \begin{pmatrix} 1 & 0 \end{pmatrix} \right) + \left( \text{diag}[\mathbf{a}_{\boldsymbol{\theta}}] \otimes \begin{pmatrix} 0 & 1 \end{pmatrix} \right) \quad (4.12)$$

with

$$\mathbf{a}_{\boldsymbol{\kappa}} = \frac{\boldsymbol{\theta}}{\boldsymbol{\kappa}^2} \exp \left( -\frac{\boldsymbol{\theta}}{\boldsymbol{\kappa}} \right) \quad (4.13)$$

and

$$\mathbf{a}_\theta = -\frac{\mathbf{1}}{\kappa} \exp\left(-\frac{\boldsymbol{\theta}}{\kappa}\right). \quad (4.14)$$

To investigate the effect of population group, the medians are reparameterised as

$$\nu_j = \tau_0 + \tau_j^P$$

for  $j = 1, 2, 3, 4$  for the different population groups Black, Coloured, Indian and White respectively. The saturated model can be written in matrix notation

$$\begin{aligned} \boldsymbol{\nu} &= \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 \end{pmatrix} \begin{pmatrix} \tau_0 \\ \tau_1^P \\ \tau_2^P \\ \tau_3^P \end{pmatrix} \\ &= \mathbf{Z}\boldsymbol{\tau} \end{aligned}$$

where  $\tau_4^P = -\sum_{j=1}^3 \tau_j^P$ . To accommodate the last category, the formulation

$$\boldsymbol{\tau}^* = \mathbf{S}\boldsymbol{\tau} \quad (4.15)$$

is used where

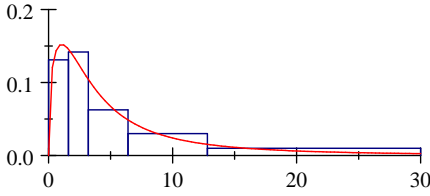
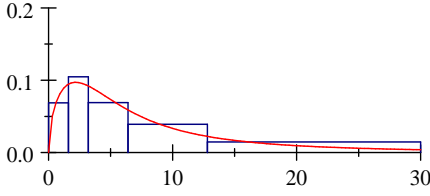
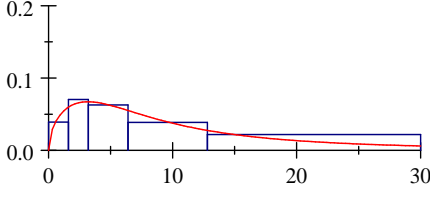
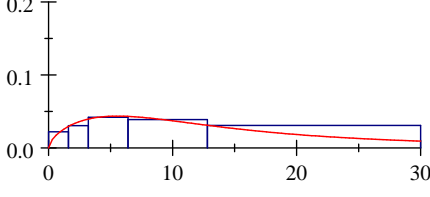
$$\boldsymbol{\tau}^* = \begin{pmatrix} \tau_0 \\ \tau_1^P \\ \tau_2^P \\ \tau_3^P \\ \tau_4^P \end{pmatrix} \text{ and } \mathbf{S} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & -1 & -1 & -1 \end{pmatrix}.$$

The covariance matrix of  $\hat{\boldsymbol{\tau}}^*$  is then

$$\begin{aligned} \text{Cov}(\hat{\boldsymbol{\tau}}^*) &= \left(\frac{\partial \boldsymbol{\tau}^*}{\partial \boldsymbol{\tau}}\right) \text{Cov}(\hat{\boldsymbol{\tau}}) \left(\frac{\partial \boldsymbol{\tau}^*}{\partial \boldsymbol{\tau}}\right)' \\ &= \left(\frac{\partial \boldsymbol{\tau}^*}{\partial \boldsymbol{\tau}}\right) \left(\frac{\partial \boldsymbol{\tau}}{\partial \boldsymbol{\nu}}\right) \text{Cov}(\hat{\boldsymbol{\nu}}) \left(\frac{\partial \boldsymbol{\tau}}{\partial \boldsymbol{\nu}}\right)' \left(\frac{\partial \boldsymbol{\tau}^*}{\partial \boldsymbol{\tau}}\right)' \\ &= \mathbf{S}(\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}' \text{Cov}(\hat{\boldsymbol{\nu}}) \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{S}'. \end{aligned}$$

The distributions of the different population groups with their respective estimated parameters are given in Table 4.4.



Pop grp	Income	$\hat{\nu}$ ( $\hat{\sigma}_{\hat{\nu}}$ )	$\begin{pmatrix} \hat{\kappa} \\ \hat{\theta} \end{pmatrix}$	$\hat{\tau}_j$ ( $\hat{\sigma}_{\hat{\tau}_j}$ )
<b>Black</b>		4.080 (0.010)	$\begin{pmatrix} 1.373 \\ -1.93 \end{pmatrix}$	-4.27 (0.024)
<b>Coloured</b>		6.267 (0.035)	$\begin{pmatrix} 1.509 \\ -2.77 \end{pmatrix}$	-2.08 (0.034)
<b>Indian</b>		9.058 (0.067)	$\begin{pmatrix} 1.511 \\ -3.33 \end{pmatrix}$	0.713 (0.053)
<b>White</b>		13.98 (0.052)	$\begin{pmatrix} 1.597 \\ -4.21 \end{pmatrix}$	5.630 (0.043)
$\hat{\tau}_0$ ( $\hat{\sigma}_{\hat{\tau}_0}$ )				8.345 (0.023)

**Table 4.4: Single factor for population group**

From Table 4.4 it can be seen that the log-logistic distributions fit the data extremely well. This is confirmed with a measure of discrepancy of 0.0068. Although the last income interval for the White population group does not seem to be presented efficiently by the distribution, one should note that it is in fact an open-ended interval and is cut off at an upper bound of 30. What is also apparent are the differences in the distributions for the different population groups. The Black population group seems to have the most skew distribution whereas the distribution for the White population group seems to have a heavier right tail. The deviations from the overall median,  $\hat{\tau}_j^P$ , for each population group should also be studied. With the overall median income being  $\hat{\tau}_0 = R8345$  over the 4 population groups, it is interesting to note that the median income for the Black population group is  $R4270$  lower than the average median whereas the median income for the White population group is  $R5630$  higher than the average median.

### 4.3 Gender

The effect of gender on income will be investigated by fitting a single factor model with  $T = 2$  cells. The cross-tabulated frequencies can be given in the same manner as in Table 4.2 from which the vector of cumulative relative frequencies  $\mathbf{p}$  can be attained. The cumulative relative frequencies are given in Table 4.5.

		Income			
Gender	Cell	1 – 1600	1601 – 3200	3201 – 6400	6400 – 12800
Female	1	0.1696	0.3411	0.5231	0.7562
Male	2	0.1325	0.3043	0.4934	0.6901

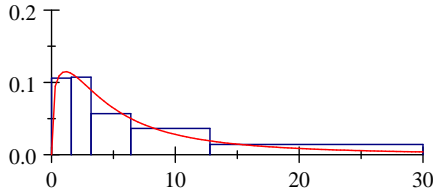
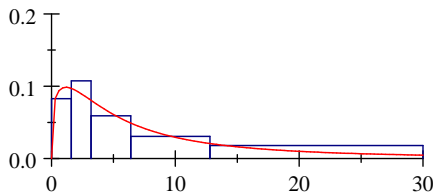
**Table 4.5: Cumulative relative frequencies of gender**

The vector  $\mathbf{p}$  has an expected value and variance as given in equations (4.4) and (4.5) respectively. The objective is the same as in the previous section: attain the ML estimate for  $\boldsymbol{\pi}$  such that  $\pi_1$  and  $\pi_2$  equal cumulative log-logistic distributions at the upper class boundaries  $\mathbf{x}$ . The resultant  $\hat{\boldsymbol{\pi}}$  under the vector of constraints  $\mathbf{g}_{\log}(\boldsymbol{\pi})$  as given in (4.7) is given in Table 4.6.

		Income			
Gender	Cell	1 – 1600	1601 – 3200	3201 – 6400	6400 – 12800
Female	1	0.1667	0.3292	0.5462	0.7470
Male	2	0.1445	0.2874	0.4906	0.6970

**Table 4.6: Estimated cumulative relative frequencies of gender**

The ML estimate for  $\hat{\boldsymbol{\pi}}$  is used to calculate the ML estimates of the parameters of the  $T = 2$  log-logistic distributions using equations (4.9) and (4.10). The parameter estimates are in turn used to construct the pdfs and calculate the medians for the distributions for each cell. These results are summarized in Table 4.7.

Gender	Income	$\hat{\nu}$ ( $\hat{\sigma}_{\hat{\nu}}$ )	$\begin{pmatrix} \hat{\kappa} \\ \hat{\theta} \end{pmatrix}$	$\hat{\tau}_i$ ( $\hat{\sigma}_{\hat{\tau}_i}$ )
Female		5.546 (0.016)	$\begin{pmatrix} 1.294 \\ -2.22 \end{pmatrix}$	-0.524 (0.012)
Male		6.594 (0.018)	$\begin{pmatrix} 1.256 \\ -2.37 \end{pmatrix}$	0.524 (0.012)
$\hat{\tau}_0$ ( $\hat{\sigma}_{\hat{\tau}_0}$ )				6.070 (0.012)

**Table 4.7: Single factor for gender**

The estimated distributions seem to fit the data well. This is supported with a measure of discrepancy of 0.0096. If the income distributions of the two different genders are inspected visually, they do not seem that different. The deviations from the average median do however indicate a considerable difference. With the average median income being R6070, the median income of females is R524 lower than the average median income, i.e. R1048 lower than the median income for males.

## 4.4 Level of education

Level of education was observed to have an ordinal relationship with income in Chapter 1. After the distributions of income for the different categories of level of education are fitted, this ordinal relationship will be studied further in a model for the medians.

The process to estimate the income distributions under different levels of education is the same as in the previous two sections. Once again one starts off with the frequencies of the different income intervals cross-tabulated with the 5 different categories of level of education. The frequencies of the 5 different cells are transformed to form the vector of cumulative relative frequencies  $\mathbf{p}$  with expected value and covariance matrix represented in equation (4.4) and (4.5), respectively. The ML estimate of the expected value  $\boldsymbol{\pi}$  under the constraint  $\mathbf{g}_{\log}(\boldsymbol{\pi})$  as given in (4.7) is attained and is used to find the 5 different sets of log-logistic parameters using equations (4.9) and (4.10). The parameter estimates are in turn used to construct the pdfs and medians for each cell. The result of this is given in Table 4.8.



Ed grp	Income	$\hat{\nu}$ ( $\hat{\sigma}_{\hat{\nu}}$ )	$\begin{pmatrix} \hat{\kappa} \\ \hat{\theta} \end{pmatrix}$	$\hat{\tau}_k$ ( $\hat{\sigma}_{\hat{\tau}_k}$ )
Gr 12		3.717 (0.008)	$\begin{pmatrix} 1.494 \\ -1.96 \end{pmatrix}$	-9.19 (0.069)
Certificate		6.858 (0.045)	$\begin{pmatrix} 1.490 \\ -2.87 \end{pmatrix}$	-6.05 (0.077)
Diploma		11.17 (0.043)	$\begin{pmatrix} 1.672 \\ -4.03 \end{pmatrix}$	-1.74 (0.076)
B Degree		17.30 (0.118)	$\begin{pmatrix} 1.597 \\ -4.55 \end{pmatrix}$	4.386 (0.114)
Post Grad		25.50 (0.317)	$\begin{pmatrix} 1.471 \\ -4.76 \end{pmatrix}$	12.60 (0.255)
$\hat{\tau}_0$ ( $\hat{\sigma}_{\hat{\tau}_0}$ )				12.91 (0.069)

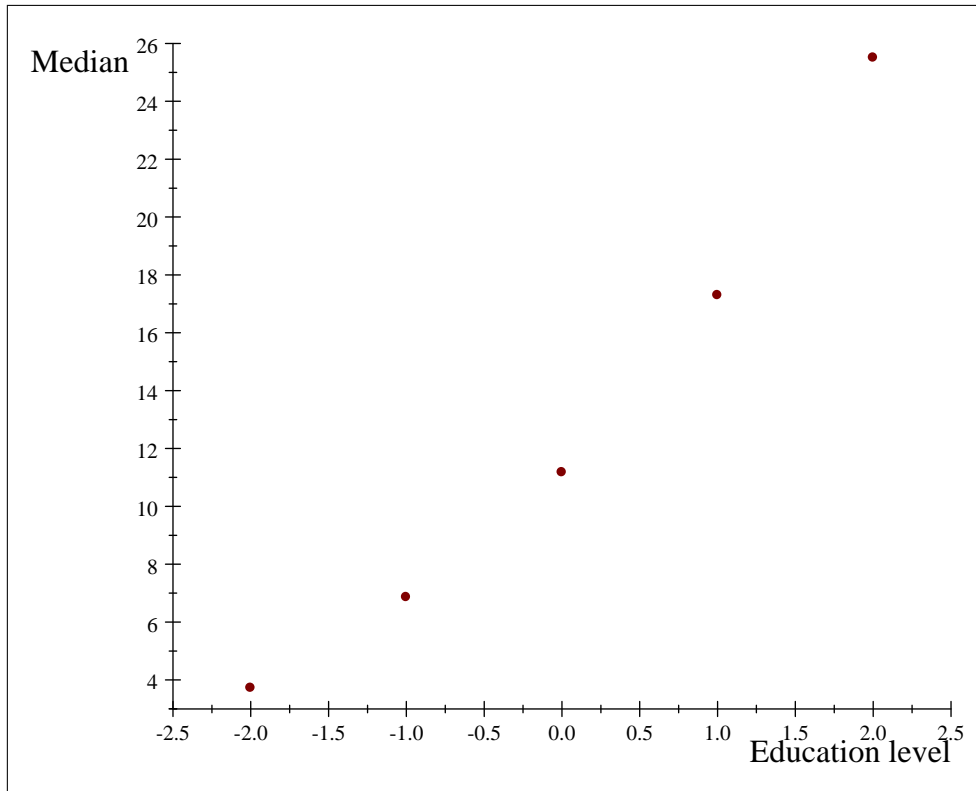
**Table 4.8: Single factor for level of education**

The attained measure of discrepancy is 0.0072, indicating the estimated distributions fit the data well. The effect of having a higher level of education can clearly be seen in the estimated distributions and in the estimates for  $\hat{\tau}_k$ . The distributions seem to be less skewed to the right as the level of education increases and the deviations from the average median indicate a clear ordinal trend between the median income levels and level of education.

#### 4.4.1 Model for the medians



To further inspect the effect of level of education on income, the medians can be plotted to see if a trend is observed. Education level is coded as  $\{-2, -1, 0, 1, 2\}$  for Grade 12, Diploma, Certificate, Bachelors degree an Post graduate degree, respectively.



**Figure 4-1: Median vs level of education**

If Figure 4-1 is considered, it seems like a linear or quadratic model can be estimated to fit this trend in median income. This monotone trend in income over the categories of education can be modelled by

$$\begin{aligned} \boldsymbol{\nu} &= \mathbf{Y}\boldsymbol{\gamma} \\ &= \begin{pmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} \gamma_0 \\ \gamma_1 \end{pmatrix} \end{aligned} \tag{4.16}$$

where a linear increase in income over the categories of education is assumed. The vector of constraints to be incorporated in the ML estimation procedure is

$$g_{\text{lin}}(\boldsymbol{\pi}) = \mathbf{Q}_Y \boldsymbol{\nu} = \mathbf{0} \tag{4.17}$$

with the matrix of partial derivatives

$$\begin{aligned}
\mathbf{G}_{\text{lin}}(\boldsymbol{\pi}) &= \frac{\partial \mathbf{Q}_Y \boldsymbol{\nu}}{\partial \boldsymbol{\pi}} \\
&= \mathbf{Q}_Y \frac{\partial \boldsymbol{\nu}}{\partial \boldsymbol{\alpha}} \frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{\pi}} \\
&= \mathbf{Q}_Y \cdot \mathbf{A} \cdot \left( \mathbf{I}_T \otimes (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \right) \mathbf{D}_{\boldsymbol{\pi}}
\end{aligned} \tag{4.18}$$

where  $\mathbf{A}$  is as defined before in (4.12). The two sets of constraints  $\mathbf{g}_{\text{log}}(\boldsymbol{\pi})$  and  $\mathbf{g}_{\text{lin}}(\boldsymbol{\pi})$  will be imposed simultaneously leading to

$$\mathbf{g}(\boldsymbol{\pi}) = \begin{pmatrix} \mathbf{g}_{\text{log}}(\boldsymbol{\pi}) \\ \mathbf{g}_{\text{lin}}(\boldsymbol{\pi}) \end{pmatrix} = \mathbf{0} \tag{4.19}$$

with matrix of partial derivatives

$$\mathbf{G}(\boldsymbol{\pi}) = \begin{pmatrix} \mathbf{G}_{\text{log}}(\boldsymbol{\pi}) \\ \mathbf{G}_{\text{lin}}(\boldsymbol{\pi}) \end{pmatrix} \tag{4.20}$$

and if the iterative procedure is applied using the vector of constraints (4.19), the ML estimate for  $\boldsymbol{\pi}$  is attained such that:

1. the frequencies of the five cells follow separate log-logistic distributions at the upper class boundaries  $\mathbf{x}$ , and
2. the median income levels of the five log-logistic distributions are in the vector space of  $\mathbf{Y}$ .

Once the ML estimation procedure is applied, the ML estimates for the parameters of (4.16) is given by

$$\begin{aligned}
\hat{\boldsymbol{\gamma}} &= (\mathbf{Y}'\mathbf{Y})^{-1} \mathbf{Y}'\hat{\boldsymbol{\nu}} \\
&= \begin{pmatrix} 11.7780 \\ 4.0429 \end{pmatrix}.
\end{aligned}$$

To see if the estimated parameters of the model defined by (4.16) are significant, the variances for  $\hat{\boldsymbol{\gamma}}$  can be attained using the multivariate delta theorem

$$\begin{aligned} Cov(\hat{\gamma}) &= \left( \frac{\partial \gamma}{\partial \nu} \right) Cov(\hat{\nu}) \left( \frac{\partial \gamma}{\partial \nu} \right)' \\ &= \begin{pmatrix} 0.0011 & 0.0006 \\ 0.00006 & 0.00003 \end{pmatrix} \end{aligned}$$

which can be used to calculate t-statistics and the associated p-values given in Table 4.9.

Parameter	Value	t-value	p-value
$\hat{\gamma}_0$	11.7780	355.4144	< 0.000001
$\hat{\gamma}_1$	4.0429	237.7978	< 0.000001

**Table 4.9: Significance tests for the estimated parameters**

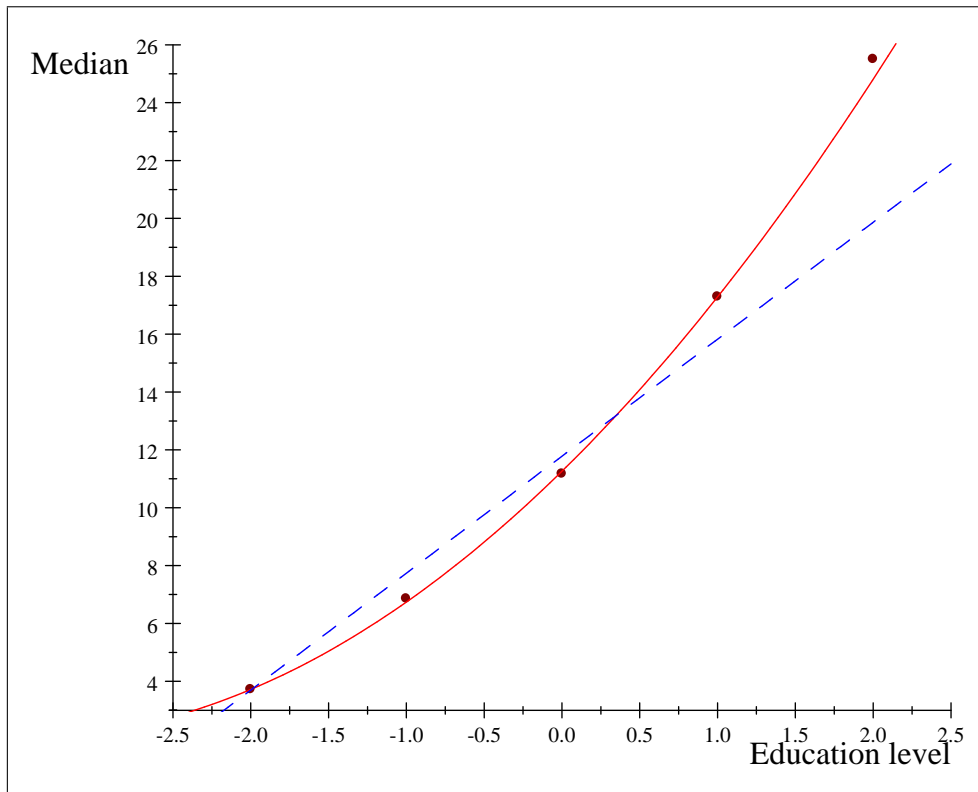
The test statistics indicate that the parameters are all significant. The extension to the quadratic model is done by redefining the design matrix to include the quadratic effect in the design matrix  $\mathbf{Y}$ .

$$\begin{aligned} \nu &= \mathbf{Y}\gamma \\ &= \begin{pmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{pmatrix} \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{pmatrix}. \end{aligned} \tag{4.21}$$

Following the same procedure as with the linear model, the ML estimate of  $\gamma$  is given by

$$\begin{aligned} \hat{\gamma} &= (\mathbf{Y}'\mathbf{Y})^{-1} \mathbf{Y}'\hat{\nu} \\ &= \begin{pmatrix} 11.2610 \\ 5.2735 \\ 0.7511 \end{pmatrix}. \end{aligned}$$

The linear and quadratic models are drawn in Figure 4-2 with the observed medians.



**Figure 4-2: Medians of different levels of education with fitted models**

The level of discrepancy for the linear and quadratic model are found to be 0.0085 and 0.0072, respectively. When the linear and quadratic models are compared, the quadratic model is preferred.

## 4.5 Age

The final explanatory variable that is considered is age. The vector of constraints  $\mathbf{g}_{\log}(\boldsymbol{\pi})$  as in (4.7) is first used to fit log-logistic distributions to the expected cumulative relative frequencies of the seven cells simultaneously. The results of doing so is given in Table 4.10.



agegrp	Income	$\hat{\nu}$ ( $\hat{\sigma}_{\hat{\nu}}$ )	$\begin{pmatrix} \hat{\kappa} \\ \hat{\theta} \end{pmatrix}$	$\hat{\tau}_1$ ( $\hat{\sigma}_{\hat{\tau}_1}$ )
18 - 25		2.970 (0.012)	$\begin{pmatrix} 1.628 \\ -1.77 \end{pmatrix}$	-5.20 (0.027)
26 - 30		4.453 (0.018)	$\begin{pmatrix} 1.425 \\ -2.13 \end{pmatrix}$	-3.72 (0.029)
31 - 40		6.021 (0.020)	$\begin{pmatrix} 1.290 \\ -2.32 \end{pmatrix}$	-2.15 (0.030)
41 - 45		8.628 (0.050)	$\begin{pmatrix} 1.283 \\ -2.76 \end{pmatrix}$	0.455 (0.049)
46 - 50		10.44 (0.070)	$\begin{pmatrix} 1.338 \\ -3.14 \end{pmatrix}$	2.272 (0.064)
51 - 55		12.04 (0.097)	$\begin{pmatrix} 1.381 \\ -3.44 \end{pmatrix}$	3.872 (0.086)
56 - 65		12.65 (0.109)	$\begin{pmatrix} 1.381 \\ -3.50 \end{pmatrix}$	4.476 (0.095)
$\hat{\tau}_0$ ( $\hat{\sigma}_{\hat{\tau}_0}$ )				8.173 (0.025)

Table 4.10: Single factor for age groups

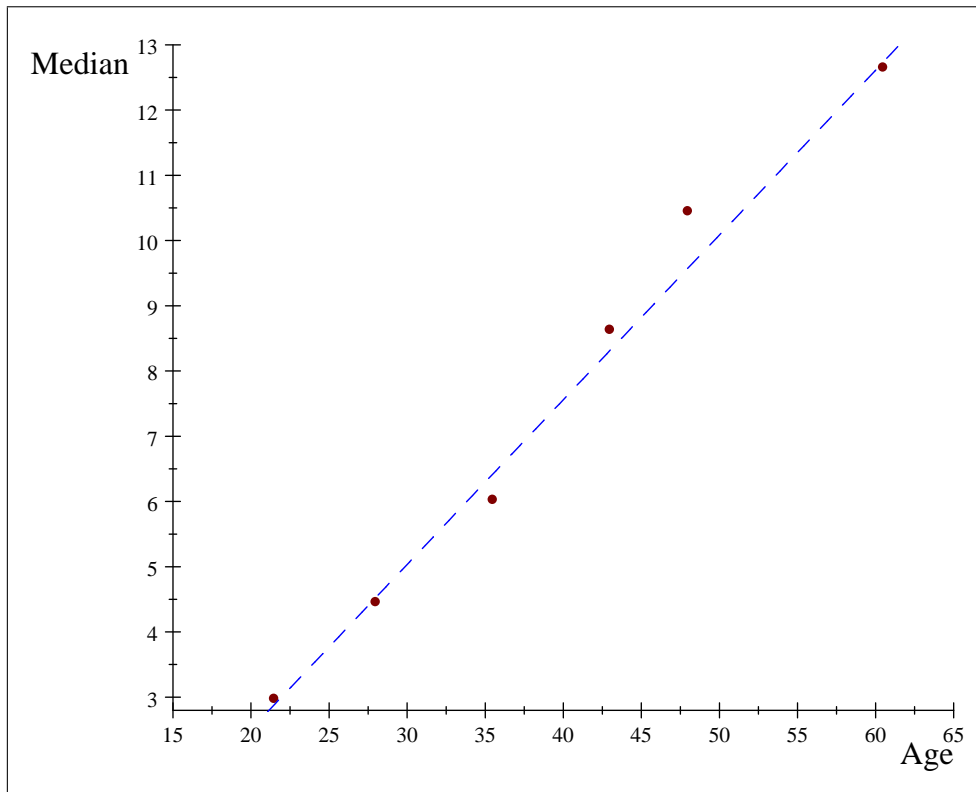
The measure of discrepancy attained is 0.0063. As age increases, one will note that the distribution seems to be less skewed to the right, i.e. income increases. The same trend can be seen if the deviations from the average median are considered an linear model can be used to capture this ordinal relationship between age and the median income level. The linear model that will be used is

$$\begin{aligned} \boldsymbol{\nu} &= \mathbf{Y}\boldsymbol{\gamma} \\ &= \begin{pmatrix} 1 & 21.5 \\ 1 & 28 \\ 1 & 35.5 \\ 1 & 43 \\ 1 & 48 \\ 1 & 53 \\ 1 & 60.5 \end{pmatrix} \begin{pmatrix} \gamma_0 \\ \gamma_1 \end{pmatrix} \end{aligned} \quad (4.22)$$

with the vector of constraints and matrix of partial derivatives being identical in form to equations (4.19) and (4.20) respectively, with the design matrix for the linear model now being given by (4.22). Applying the iterative procedure, the ML estimate for  $\boldsymbol{\pi}$  is attained with the ML estimate for  $\boldsymbol{\gamma}$  being

$$\begin{aligned} \hat{\boldsymbol{\gamma}} &= (\mathbf{Y}'\mathbf{Y})^{-1} \mathbf{Y}'\hat{\boldsymbol{\nu}} \\ &= \begin{pmatrix} -2.5452 \\ 0.2526 \end{pmatrix}. \end{aligned}$$

These estimates are finally used to plot the linear model given in Figure 4-3.



**Figure 4-3: Medians of different age groups with linear model**

The measure of discrepancy is 0.0078. One can see that the linear model does well in predicting the median income levels.

## 4.6 Summary

This chapter concentrated on using different explanatory variables to better understand how the income acts under different conditions. Specifically, the aim was to estimate different log-logistic distributions for each cell, i.e. category, of an explanatory variable. The medians were used as representative measures for each cell and it was shown how relevant constraints can be placed on the median levels as well.

# Chapter 5

## Multifactor design

The aim of this chapter is to expand on the technique explained in Chapter 4 by cross tabulating more than one explanatory variable with the frequency distribution of income. Log-logistic distributions will be estimated simultaneously for each cross-tabulated cell of the multifactor design such that the medians of the cells also follow a specified model.

### 5.1 Gender and population group

To present the expansion to a two-factor model with the technique at hand, the two variables gender and population group will be cross-tabulated and used as explanatory variables for the grouped response variable, income. The resultant histograms are given in Table 5.1.

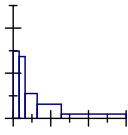
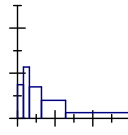
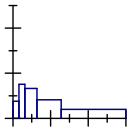
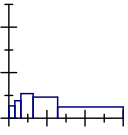
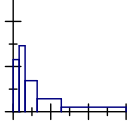
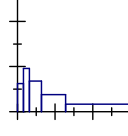
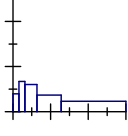
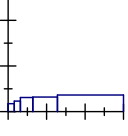
Gender	Population group			
	Black	Coloured	Indian	White
Female	 $n = 138851$	 $n = 22411$	 $n = 10944$	 $n = 57126$
Male	 $n = 159805$	 $n = 21718$	 $n = 15740$	 $n = 63102$

Table 5.1: Histograms for gender and population group

#### 5.1.1 Distribution fitting

Distributions will first be fitted to each of the  $T = 8$  cells of the two-factor design before models for the medians are considered by estimating the expected cumulative relative frequencies of each



cell simultaneously such that they will follow log-logistic distributions at the upper class boundaries  $\mathbf{x} = \left( 1.6 \ 3.2 \ 6.4 \ 12.8 \right)'$ . Table 5.2 presents the frequency distribution of income for each cell when gender and population group is cross-tabulated.

Cell	Gender	Popgrp	Income (Upper Class Limits)					Total $n_t$
			R1600	R3200	R6400	R12800	R12800+	
1	F	Black	33064	30420	24498	27878	22991	138851
2	F	Coloured	2678	4075	5007	5739	4912	22411
3	F	Indian	669	1325	2322	2878	3750	10944
4	F	White	2485	3514	9911	16955	24261	57126
5	M	Black	29556	37377	35188	29258	28426	159805
6	M	Coloured	2169	3328	4731	5284	6206	21718
7	M	Indian	999	1684	3043	3737	6277	15740
8	M	White	1774	2342	6283	12913	39791	63102

**Table 5.2: Frequencies for gender and population group**

The matrix

$$\mathbf{F}_M = \begin{pmatrix} 33064 & 30420 & 24498 & 27878 \\ 2678 & 4075 & 5007 & 5739 \\ \vdots & \vdots & \vdots & \vdots \\ 1774 & 2342 & 6283 & 12913 \end{pmatrix} = \begin{pmatrix} \mathbf{f}'_1 \\ \mathbf{f}'_2 \\ \vdots \\ \mathbf{f}'_8 \end{pmatrix} : 8 \times 4$$

contains the first  $(k - 1)$  frequencies of each cell and is transformed into the concatenated vector of frequencies

$$vec(\mathbf{F}_M) = \mathbf{f} = \begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \\ \mathbf{f}_8 \end{pmatrix} : 32 \times 1$$

which is distributed as product multinomial with fixed subtotals

$$\mathbf{n} = \begin{pmatrix} n_1 \\ n_2 \\ \vdots \\ n_8 \end{pmatrix} = \begin{pmatrix} 138851 \\ 22411 \\ \vdots \\ 63102 \end{pmatrix}$$

with a block diagonal covariance matrix. The observed frequencies are transformed into cumulative relative frequencies using  $\mathbf{p} = (\mathbf{I}_8 \otimes \mathbf{C}) \mathbf{p}_0$  and are presented in Table 5.3.

Cell	Gender	Popgrp	(Upper Class Limits) ( $\mathbf{x}$ )			
			1.6	3.2	6.4	12.8
1	F	Black	0.238	0.457	0.634	0.834
2	F	Coloured	0.119	0.301	0.525	0.781
3	F	Indian	0.061	0.182	0.394	0.657
4	F	White	0.044	0.105	0.279	0.575
5	M	Black	0.185	0.419	0.639	0.822
6	M	Coloured	0.10	0.253	0.471	0.714
7	M	Indian	0.063	0.170	0.364	0.601
8	M	White	0.028	0.065	0.165	0.369

**Table 5.3: Cumulative relative frequencies for gender and population group**

The cumulative relative frequencies has an expected value and covariance matrix given by

$$E(\mathbf{p}) = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_8 \end{pmatrix} = \boldsymbol{\pi}$$

and

$$Cov(\mathbf{p}) = \mathbf{V}$$

as in (4.4) and (4.5), respectively. The vector of constraints that is to be used should imply that the expected cumulative relative frequencies of each cell follow a log-logistic distribution at the upper class boundaries  $\mathbf{x}$ . By following the same rationale as in Chapter 4, the vector of constraints

$$\begin{pmatrix} F_1(\mathbf{x}, \theta_1, \kappa_1) \\ F_2(\mathbf{x}, \theta_2, \kappa_2) \\ \vdots \\ F_8(\mathbf{x}, \theta_8, \kappa_8) \end{pmatrix} - \begin{pmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_8 \end{pmatrix} = \mathbf{0}$$

can be expressed in terms of a linear model by using the transformation

$$\begin{aligned} \ln\left(\frac{\pi_t}{\mathbf{1} - \pi_t}\right) &= \kappa_t \ln \mathbf{x} + \theta_t \mathbf{1} \\ &= \begin{pmatrix} \ln \mathbf{x} & \mathbf{1} \end{pmatrix} \begin{pmatrix} \kappa_t \\ \theta_t \end{pmatrix} \\ &= \mathbf{X}\boldsymbol{\alpha}_t \end{aligned} \tag{5.1}$$

for  $t = 1, 2, \dots, 8$ . The vector of constraints  $\mathbf{g}_{\log}(\boldsymbol{\pi}) = \mathbf{0}$  with corresponding matrix of partial derivatives  $\mathbf{G}_{\log}(\boldsymbol{\pi})$  which will be used in the iterative procedure are given by

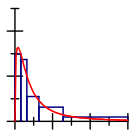
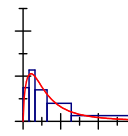
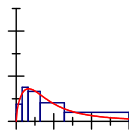
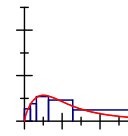
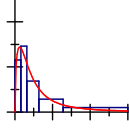
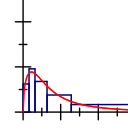
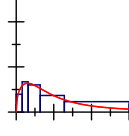
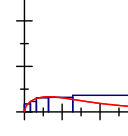
$$\mathbf{g}_{\log}(\boldsymbol{\pi}) = (\mathbf{I}_8 \otimes \mathbf{Q}_{\mathbf{X}}) \ln\left(\frac{\boldsymbol{\pi}}{1-\boldsymbol{\pi}}\right) \quad \text{and} \quad \mathbf{G}_{\log}(\boldsymbol{\pi}) = (\mathbf{I}_8 \otimes \mathbf{Q}_{\mathbf{X}}) \mathbf{D}_{\boldsymbol{\pi}}$$

where  $\mathbf{Q}_{\mathbf{X}}$  is the projection matrix onto the error space of  $\mathbf{X}$ . The attained ML estimate  $\hat{\boldsymbol{\pi}}$  under  $\mathbf{g}_{\log}(\boldsymbol{\pi}) = \mathbf{0}$  is given in Table 5.4.

Cell	Gender	Popgrp	(Upper Class Limits) (x)			
			1.6	3.2	6.4	12.8
1	F	Black	0.240	0.438	0.659	0.827
2	F	Coloured	0.121	0.287	0.541	0.775
3	F	Indian	0.067	0.177	0.392	0.660
4	F	White	0.035	0.108	0.286	0.570
5	M	Black	0.197	0.399	0.642	0.829
6	M	Coloured	0.104	0.245	0.474	0.715
7	M	Indian	0.069	0.168	0.358	0.605
8	M	White	0.024	0.067	0.169	0.368

**Table 5.4: Expected cumulative relative frequencies for gender and population group**

Using these estimates, the parameters  $\boldsymbol{\kappa}$  and  $\boldsymbol{\theta}$  of the separate log-logistic distributions can be estimated using (4.9) and (4.10), respectively, whereafter the vector of estimated medians  $\hat{\boldsymbol{\nu}}$  can be estimated using (4.11). The results of doing so are given in Table 5.5.

Gender	Population group			
	Black	Coloured	Indian	White
F				
	$\hat{\nu} = 3.869$	$\hat{\nu} = 5.754$	$\hat{\nu} = 8.427$	$\hat{\nu} = 10.863$
M				
	$\hat{\nu} = 4.256$	$\hat{\nu} = 6.863$	$\hat{\nu} = 9.559$	$\hat{\nu} = 18.296$

**Table 5.5: Fitted distributions for gender and population group**

A measure of discrepancy of 0.0091 is observed. The distributions over different genders for identical population groups seem to differ but is most evident in the White population group. The distributions are also vastly different across population groups if gender is kept constant. To further inspect the effects of the explanatory variables, a saturated model for the medians will be considered next.

### 5.1.2 Saturated model

The medians can be reparameterized by

$$\nu_{ij} = \tau_0 + \tau_i^G + \tau_j^P + \tau_{ij}^{GP}$$

for  $i = 1, 2$  for the genders Female and Male, respectively, and  $j = 1, 2, 3, 4$  for the population groups Black, Coloured, Indian and White, respectively. The reparameterization is given in matrix notation by

$$\boldsymbol{\nu} = \mathbf{Z}\boldsymbol{\tau}$$

where

$$\begin{aligned} \mathbf{Z} &= (\mathbf{1}_8 : \mathbf{Z}_G : \mathbf{Z}_P : \mathbf{Z}_{GP}) \\ &= \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 & 1 & 0 & 0 & -1 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 \end{pmatrix} \end{aligned}$$

and  $\boldsymbol{\tau} = \begin{pmatrix} \tau_0 \\ \boldsymbol{\tau}^P \\ \boldsymbol{\tau}^G \\ \boldsymbol{\tau}^{GP} \end{pmatrix}$ . The submatrices of  $\mathbf{Z}$  are defined by making use of design matrices  $\mathbf{D}_P$  and  $\mathbf{D}_G$ , where

$$\mathbf{D}_P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \end{pmatrix} \text{ and } \mathbf{D}_G = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

from which

$$\begin{aligned} \mathbf{Z}_P &= \mathbf{D}_P \otimes \mathbf{1}_2 \\ \mathbf{Z}_G &= \mathbf{1}_4 \otimes \mathbf{D}_G \\ \mathbf{Z}_{PG} &= \mathbf{Z}_P \odot \mathbf{Z}_G \end{aligned}$$

are constructed, where  $\otimes$  and  $\odot$  are Kronecker and direct products, respectively.

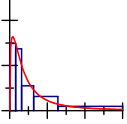
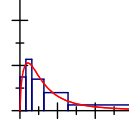
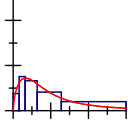
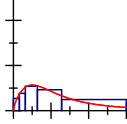
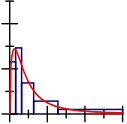
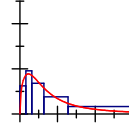
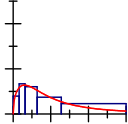
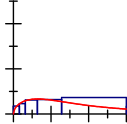
The submatrices can be defined by using the following expressions in SAS:

$$\mathbf{Z}_P = \text{DESIGNF}(\text{CUSUM}(\text{J}(4, 1, 1)))@ \text{J}(2, 1, 1);$$

$$\mathbf{Z}_G = \text{J}(4, 1, 1)@ \text{DESIGNF}(\text{CUSUM}(\text{J}(2, 1, 1)));$$

$$\mathbf{Z}_{PG} = \text{HDIR}(\mathbf{Z}_P, \mathbf{Z}_G);^1$$

Table 5.6 presents the results for the reparameterization of the median.

Gender	Population group				$\hat{\tau}^G$
	Black	Coloured	Indian	White	
Female					-1.258
	$\hat{\nu} = 3.869$ $\hat{\tau}^{GP} = 1.064$	$\hat{\nu} = 5.754$ $\hat{\tau}^{GP} = 0.703$	$\hat{\nu} = 8.427$ $\hat{\tau}^{GP} = 0.692$	$\hat{\nu} = 10.863$ $\hat{\tau}^{GP} = -2.459$	
Male					1.258
	$\hat{\nu} = 4.256$ $\hat{\tau}^{GP} = -1.064$	$\hat{\nu} = 6.863$ $\hat{\tau}^{GP} = -0.703$	$\hat{\nu} = 9.559$ $\hat{\tau}^{GP} = -0.692$	$\hat{\nu} = 18.296$ $\hat{\tau}^{GP} = 2.459$	
$\hat{\tau}^P$	-4.423	-2.177	0.507	6.093	$\hat{\tau}_0 = 8.486$

**Table 5.6: Fitted distributions with saturated model for gender and population group**

Instead of studying the distributions, one can now consider the deviations from the overall median  $\tau_0$  caused by the explanatory variables. Since the model is simply a reparameterization, the original values for the medians can be attained by using the overall median, the corresponding marginal and interaction effects. As an example, one can consider calculating the median income for a White Male with

$$\begin{aligned} \nu_{24} &= \tau_0 + \tau_2^G + \tau_4^P + \tau_{24}^{GP} \\ &= 8.486 + 1.258 + 6.093 + 2.459 \\ &= 18.296. \end{aligned}$$

The marginal deviations for gender and population group indicate that both explanatory variables play a substantial role. The marginal deviations caused by the different genders indicate that a Male is expected to earn  $R1258$  more than the overall median, hence  $R2516$  more than a Female. The effect of population group is also evident with the White population group earning  $R6093$  more than the overall median and the Black population group earning  $R4423$  less than the overall median. The marginal

<sup>1</sup>The  $\text{CUSUM}(\langle \mathbf{A} \rangle)$  function calculates the cumulative sum vector/matrix of a vector/matrix. The  $\text{J}(\langle n \text{ row} \rangle, \langle n \text{ col} \rangle, \langle \text{element} \rangle)$  operator creates a vector/matrix of dimension  $\langle n \text{ row} \rangle \times \langle n \text{ col} \rangle$  with each element equaling  $\langle \text{element} \rangle$ . The  $@$  operator calculates a Kronecker product of two matrices. The  $\text{HDIR}(\langle \mathbf{A} \rangle, \langle \mathbf{B} \rangle)$  calculates the direct product of the two matrices  $\langle \mathbf{A} \rangle$  and  $\langle \mathbf{B} \rangle$ .

deviation caused by the interaction of the explanatory variables seem to only be prominent in the Black and White population groups, with White Males doing relatively better than White Females. From the interaction effect it is clear that apart from the marginal effects of gender and population group, an additional  $R2459$  is caused by the interaction effect of a White Male. The reverse situation is true for the Black population group where Black Females are doing relatively better. If one wished to estimate  $\boldsymbol{\pi}$  under the added constraint that the explanatory variables are independent, it could be incorporated in the vector of constraints  $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$  that is used in the iterative procedure.

### 5.1.3 Independence model

If one forces gender and population group to be independent, the model for the medians would be

$$\nu_{ij} = \tau_0 + \tau_i^G + \tau_j^P$$

for  $i = 1, 2$  and  $j = 1, 2, 3, 4$  with the model given in matrix notation by

$$\boldsymbol{\nu} = \mathbf{Z}\boldsymbol{\tau}$$

where

$$\begin{aligned} \mathbf{Z} &= (\mathbf{1}_8 : \mathbf{Z}_G : \mathbf{Z}_P) \\ &= \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 \\ 1 & -1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 & -1 \end{pmatrix} \end{aligned}$$

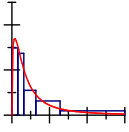
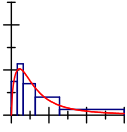
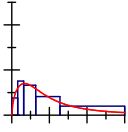
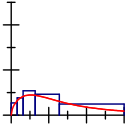
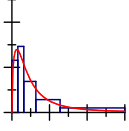
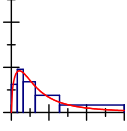
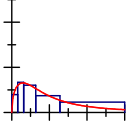
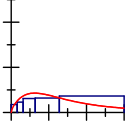
Defining  $\mathbf{Q}_Z$  as the projection matrix onto the error space of  $\mathbf{Z}$ , the vector of constraints  $\mathbf{g}_{\text{mod}}(\boldsymbol{\pi}) = \mathbf{Q}_Z\boldsymbol{\nu} = \mathbf{0}$  and corresponding matrix of partial derivatives  $\mathbf{G}_{\text{mod}}(\boldsymbol{\pi})$  will be used in conjunction with  $\mathbf{g}_{\text{log}}(\boldsymbol{\pi}) = \mathbf{0}$  to form the vector of constraints  $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$  and matrix of partial derivatives  $\mathbf{G}(\boldsymbol{\pi})$  given by

$$\mathbf{g}(\boldsymbol{\pi}) = \begin{pmatrix} \mathbf{g}_{\text{log}}(\boldsymbol{\pi}) \\ \mathbf{g}_{\text{mod}}(\boldsymbol{\pi}) \end{pmatrix} \quad \text{and} \quad \mathbf{G}(\boldsymbol{\pi}) = \begin{pmatrix} \mathbf{G}_{\text{log}}(\boldsymbol{\pi}) \\ \mathbf{G}_{\text{mod}}(\boldsymbol{\pi}) \end{pmatrix}.$$

Note that the desired effect could also be achieved by setting  $\mathbf{g}_{\text{mod}}(\boldsymbol{\pi}) = \mathbf{Z}_{GP}\boldsymbol{\nu} = \mathbf{0}$  since this vector of constraints explicitly sets the interaction effects equal to zero. Using  $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$  as the vector of constraints in the iterative procedure will lead to the ML estimate of  $\boldsymbol{\pi}$  such that:

1. the elements of  $\hat{\pi}_1, \hat{\pi}_2, \dots, \hat{\pi}_8$  follow cumulative log-logistic curves at the upper class boundaries  $\mathbf{x}$ , and
2. the ML estimate  $\hat{\nu}$  is a linear combination of columns of  $\mathbf{Z}$ .

The resultant distributions and corresponding estimates are then given in Table 5.7.

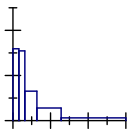
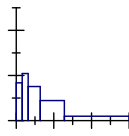
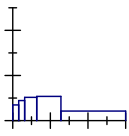
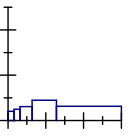
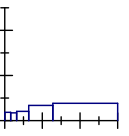
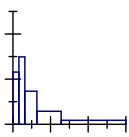
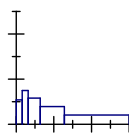
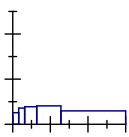
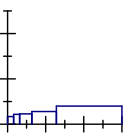
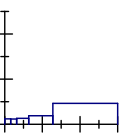
Gender	Population group				$\hat{\tau}^G$
	Black	Coloured	Indian	White	
F					-0.332
	$\hat{\nu} = 3.728$ $\hat{\tau}^{GP} = 0.000$	$\hat{\nu} = 5.935$ $\hat{\tau}^{GP} = 0.000$	$\hat{\nu} = 8.662$ $\hat{\tau}^{GP} = 0.000$	$\hat{\nu} = 13.605$ $\hat{\tau}^{GP} = 0.000$	
M					0.332
	$\hat{\nu} = 4.392$ $\hat{\tau}^{GP} = 0.000$	$\hat{\nu} = 6.599$ $\hat{\tau}^{GP} = 0.000$	$\hat{\nu} = 9.326$ $\hat{\tau}^{GP} = 0.000$	$\hat{\nu} = 14.269$ $\hat{\tau}^{GP} = 0.000$	
$\hat{\tau}^P$	-4.255	-2.047	0.679	5.622	$\hat{\tau}_0 = 8.314$

**Table 5.7: Independence model of gender and population group**

The attained measure of discrepancy is 0.0144, indicating that the independence model for the medians is sufficient to explain the effects of the explanatory variables. The medians will now proportionately reflect the marginal effects under the independence model. The same inferences regarding the explanatory variables that was made earlier are seen here if the marginal deviations from the overall median are considered. Since the medians are all linear combinations of the columns of the design matrix  $\mathbf{Z}$ , the deviations caused by the interactions between the explanatory variables are all equal to 0.

## 5.2 Gender and level of education

Replicating the ideology of the previous section, the two-factor model will now be used to study the effects of gender and level of education on the grouped response variable, income. The resultant histograms are given in Table 5.8 when the two explanatory variables are cross-tabulated with the frequency distribution of income.

Gender	Level of education				
	Grade 12	Certif	Diploma	B Degree	Post Grad
<b>Female</b>	 $n = 130013$	 $n = 15147$	 $n = 46347$	 $n = 23261$	 $n = 14564$
<b>Male</b>	 $n = 163773$	 $n = 18896$	 $n = 39536$	 $n = 22320$	 $n = 15840$

**Table 5.8: Gender and level of education**

One should note that the last income interval is defined as open-ended but is restricted to 30 for display purposes.

### 5.2.1 Distribution fitting with the saturated model

To fit log-logistic distributions to each of the  $T = 10$  cells, one needs to transform the vector of observed frequencies  $\mathbf{f}$  into the vector of cumulative relative frequencies  $\mathbf{p}$ , as was done in Section 5.1.1, and estimate the vector of expected cumulative relative frequencies  $\boldsymbol{\pi}$  such that each cell follows a log-logistic distribution at the upper class boundaries  $\mathbf{x}$ . When this is done,  $\hat{\boldsymbol{\pi}}$  is used to estimate  $\boldsymbol{\kappa}$  and  $\boldsymbol{\theta}$  using (4.9) and (4.10), respectively, whereafter the vector of estimated medians  $\hat{\boldsymbol{\nu}}$  can be estimated using (4.11). Note that the estimated medians can be reparameterized to find the effects of the explanatory variables with

$$\nu_{ik} = \tau_0 + \tau_i^G + \tau_k^E + \tau_{ik}^{GE}$$

where  $i = 1, 2$  for Female and Male respectively, and  $k = 1, 2, 3, 4, 5$  for the different levels of education Grade 12, Diploma, Certificate, Bachelor's degree and Postgraduate degree, respectively. The reparameterization is given in matrix notation by

$$\boldsymbol{\nu} = \mathbf{Z}\boldsymbol{\tau}$$



where

$$\mathbf{Z} = (\mathbf{1}_{10} : \mathbf{Z}_G : \mathbf{Z}_E : \mathbf{Z}_{GE})$$

$$= \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

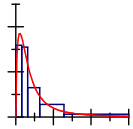
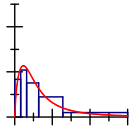
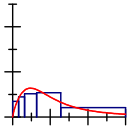
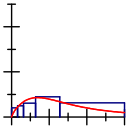
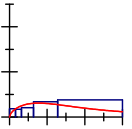
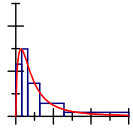
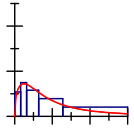
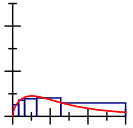
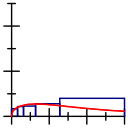
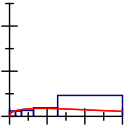
and  $\boldsymbol{\tau} = \begin{pmatrix} \tau_0 \\ \boldsymbol{\tau}^G \\ \boldsymbol{\tau}^E \\ \boldsymbol{\tau}^{GE} \end{pmatrix}$ . The submatrices of  $\mathbf{Z}$  are defined by making use of design matrices  $\mathbf{D}_G$  and  $\mathbf{D}_E$ , where  $\mathbf{D}_G$  is as defined in Section 5.1.2 and

$$\mathbf{D}_E = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & -1 & -1 & -1 \end{pmatrix}$$

from which

$$\begin{aligned} \mathbf{Z}_G &= \mathbf{D}_G \otimes \mathbf{1}_5 \\ \mathbf{Z}_E &= \mathbf{1}_2 \otimes \mathbf{D}_E \\ \mathbf{Z}_{GE} &= \mathbf{Z}_G \odot \mathbf{Z}_E \end{aligned}$$

are constructed. The distributions and relevant parameters are given in Table 5.9.

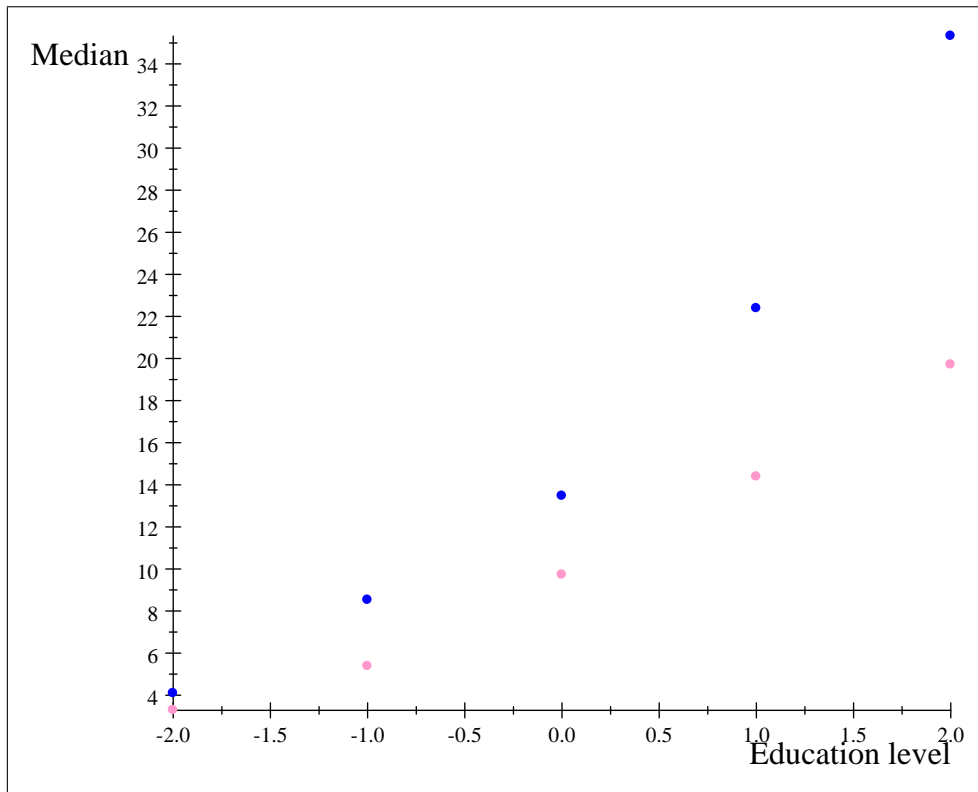
G	Level of education					$\hat{\tau}^G$
	Grade 12	Certif	Diploma	B Degree	Post Grad	
F	 $\hat{\nu} = 3.287$ $\hat{\tau}^{GE} = 2.729$	 $\hat{\nu} = 5.379$ $\hat{\tau}^{GE} = 1.558$	 $\hat{\nu} = 9.723$ $\hat{\tau}^{GE} = 1.258$	 $\hat{\nu} = 14.385$ $\hat{\tau}^{GE} = -0.869$	 $\hat{\nu} = 19.710$ $\hat{\tau}^{GE} = -4.677$	-3.132
M	 $\hat{\nu} = 4.092$ $\hat{\tau}^{GE} = -2.729$	 $\hat{\nu} = 8.526$ $\hat{\tau}^{GE} = -1.558$	 $\hat{\nu} = 13.470$ $\hat{\tau}^{GE} = -1.258$	 $\hat{\nu} = 22.388$ $\hat{\tau}^{GE} = 0.869$	 $\hat{\nu} = 35.328$ $\hat{\tau}^{GE} = 4.677$	3.132
$\hat{\tau}^E$	-9.939	-6.676	-2.032	4.758	13.890	$\hat{\tau}_0 = 13.629$

**Table 5.9: Fitted distributions for gender and level of education**

With the measure of discrepancy being calculated as 0.0093, one can conclude that the estimated frequencies do not differ significantly from the observed frequencies. The marginal deviations from the overall median caused by gender seem to indicate that gender still has a considerable effect in the median income level, even if level of education is taken into account. The marginal deviations from the overall median caused by level of education shows that it has a prominent effect on income. An individual with a Post graduate level of education earns R13890 more than the overall median whereas a person with only a Grade 12 level of education earns R9939 less than the overall median. The marginal deviations also indicate that an ordinal trend exists between income and level of education. This ordinal trend is also observed in the interaction effects where the interaction between education and gender follows a decreasing trend for Females but increasing for Males. How level of education impacts the median income under different genders can be studied further by fitting different linear models to the median income levels for different genders.

## 5.2.2 Linear model

Figure 5-1 presents the median income levels against level of education for different genders where the pink and blue dots represent the median levels for Female and Male, respectively. Note that level of education is coded as in Section 4.4.1.



**Figure 5-1: Medians under different population groups with level of education as ordinal variable**

The ordinal relationship between level of education and income can be captured by the model

$$\nu = \mathbf{Y}\gamma$$

where the design matrix

$$\begin{aligned}
 \mathbf{Y} &= (\mathbf{1}_{10} : \mathbf{Z}_G : \mathbf{Y}_E : \mathbf{Y}_{GE}) \\
 &= \begin{pmatrix} 1 & 1 & -2 & -2 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 \\ 1 & -1 & -2 & 2 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 0 & 0 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 2 & -2 \end{pmatrix}
 \end{aligned}$$

with corresponding vector of parameters  $\boldsymbol{\gamma} = \begin{pmatrix} \mu \\ \tau^G \\ \beta^E \\ \eta^{GE} \end{pmatrix}$  is used. The vectors  $\mathbf{1}_{10}$  and  $\mathbf{Y}_E = \mathbf{1}_2 \otimes \begin{pmatrix} -2 \\ -1 \\ 0 \\ 1 \\ 2 \end{pmatrix}$

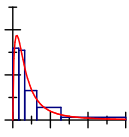
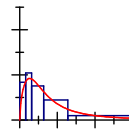
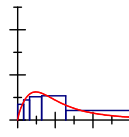
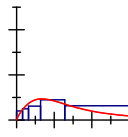
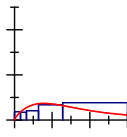
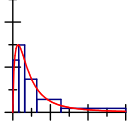
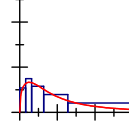
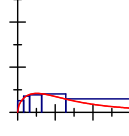
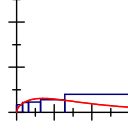
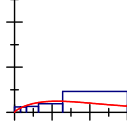
coincide with the overall intercept and overall gradient caused by level of education, respectively. The matrices  $\mathbf{Z}_G = \mathbf{D}_G \otimes \mathbf{1}_5$  and  $\mathbf{Y}_{GE} = \mathbf{Z}_G \odot \mathbf{Y}_E$  coincide with the deviations caused by different genders from the overall intercept and the overall gradient, respectively. If  $\mathbf{Q}_Y$  is defined as the projection matrix onto the error space of  $\mathbf{Y}$ , the vector of constraints  $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$  with matrix of partial derivatives  $\mathbf{G}(\boldsymbol{\pi})$  follow as before as

$$\mathbf{g}(\boldsymbol{\pi}) = \begin{pmatrix} \mathbf{g}_{\log}(\boldsymbol{\pi}) \\ \mathbf{g}_{\text{lin}}(\boldsymbol{\pi}) \end{pmatrix} \quad \text{and} \quad \mathbf{G}(\boldsymbol{\pi}) = \begin{pmatrix} \mathbf{G}_{\log}(\boldsymbol{\pi}) \\ \mathbf{G}_{\text{lin}}(\boldsymbol{\pi}) \end{pmatrix}$$

as in (4.19) and (4.20), respectively where  $\mathbf{g}_{\text{lin}}(\boldsymbol{\pi}) = \mathbf{Q}_Y \boldsymbol{\nu} = \mathbf{0}$ . The ML estimate for  $\boldsymbol{\pi}$  under  $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$  is estimated using the iterative procedure and finally the ML estimates of the parameters  $\boldsymbol{\gamma}$  can be attained using

$$\begin{aligned} \hat{\boldsymbol{\gamma}} &= (\mathbf{Y}'\mathbf{Y})^{-1} \mathbf{Y}'\hat{\boldsymbol{\nu}} \\ &= \begin{pmatrix} 12.2165 \\ -2.2435 \\ 4.2770 \\ -0.9150 \end{pmatrix}. \end{aligned}$$

One should note how these parameters can be interpreted. The value  $\hat{\mu} = 12.2165$  corresponds to the estimated overall median  $\hat{\tau}_0$ , the marginal effect of gender on the overall median for Females is given by  $\hat{\tau}^G = -2.2435$ , the overall gradient of the lines  $\hat{\beta}^E = 4.2770$  corresponds with what increment the effect of education  $\hat{\tau}^E$  will change by and the deviation from the overall gradient caused by gender  $\hat{\eta}^{GE} = -0.9150$  corresponds with what increment the interaction effects  $\hat{\tau}^{GE}$  will change by for Females. The estimated distributions are given in Table 5.10.

G	Level of education					$\hat{\tau}^G$
	Grade 12	Certif	Diploma	B Degree	Post Grad	
F						-2.243
	$\hat{\nu} = 3.249$ $\hat{\tau}^{GE} = 1.830$	$\hat{\nu} = 6.611$ $\hat{\tau}^{GE} = 0.915$	$\hat{\nu} = 9.973$ $\hat{\tau}^{GE} = 0.000$	$\hat{\nu} = 13.335$ $\hat{\tau}^{GE} = -0.915$	$\hat{\nu} = 16.697$ $\hat{\tau}^{GE} = -1.830$	
M						2.243
	$\hat{\nu} = 4.076$ $\hat{\tau}^{GE} = -1.830$	$\hat{\nu} = 9.268$ $\hat{\tau}^{GE} = -0.915$	$\hat{\nu} = 14.460$ $\hat{\tau}^{GE} = 0.000$	$\hat{\nu} = 19.652$ $\hat{\tau}^{GE} = 0.915$	$\hat{\nu} = 24.844$ $\hat{\tau}^{GE} = 1.830$	
$\hat{\tau}^E$	-8.554	-4.277	0.000	4.277	8.554	$\hat{\tau}_0 = 12.217$

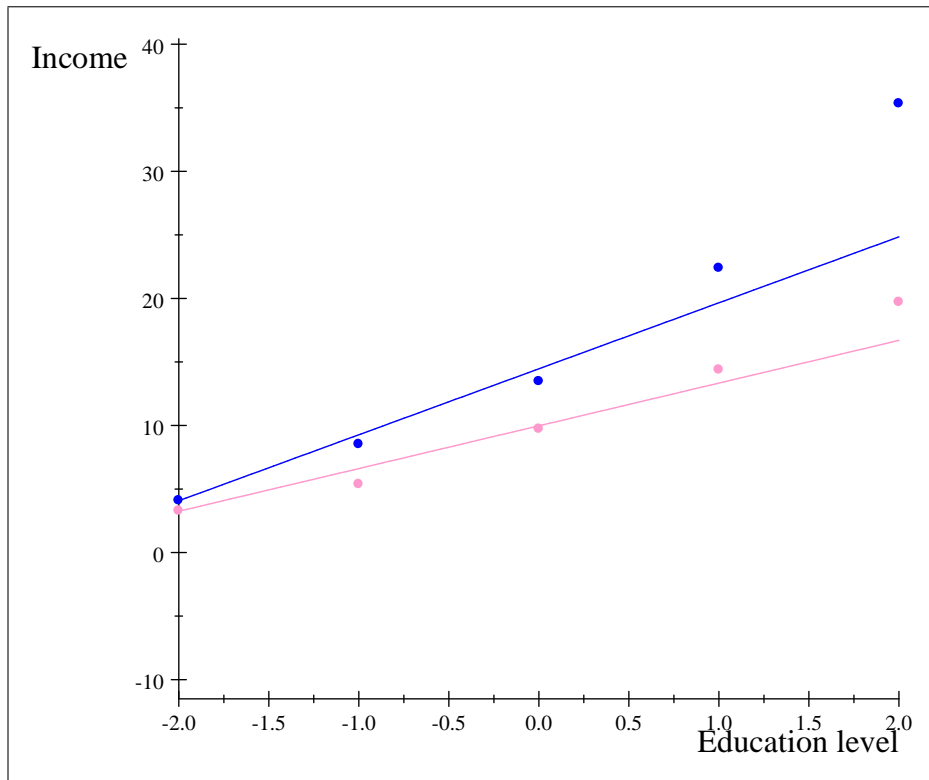
**Table 5.10: Fitted distributions for gender and level of education where level of education is ordinal**

With the measure of discrepancy being 0.0112, one can conclude that the model fits sufficiently. Note that the respective medians under different genders now increase with a linear trend. The marginal effects of level of education and the interaction effects between gender and level of education can be observed as being symmetrical about Diploma. The linear functions

$$\hat{\nu}_F = 9.973 + 3.362x_E$$

$$\hat{\nu}_M = 14.46 + 5.192x_E$$

can also be drawn and are given in Figure 5-2 where the pink and blue lines represent Females and Males, respectively.



**Figure 5-2: Linear functions for income under different genders with level of education as ordinal variable**

From these linear models one can see that Males attain the highest income whereas Females attain the lowest income over all levels of education. The effect of having a higher level of education is also more prominent with Males. Although the linear model does give an adequate measure of discrepancy, the highest value for income that is observed at a post graduate level of education for Males is not properly described. To rectify this observation, a quadratic model will be fitted to the medians.

### 5.2.3 Quadratic model

The quadratic model

$$\nu = Y\gamma$$

can be estimated where the design matrix is given by

$$\mathbf{Y} = (\mathbf{1} : \mathbf{Y}_G : \mathbf{Y}_E : \mathbf{Y}_E^2 : \mathbf{Y}_{GE} : \mathbf{Y}_{GE}^2)$$

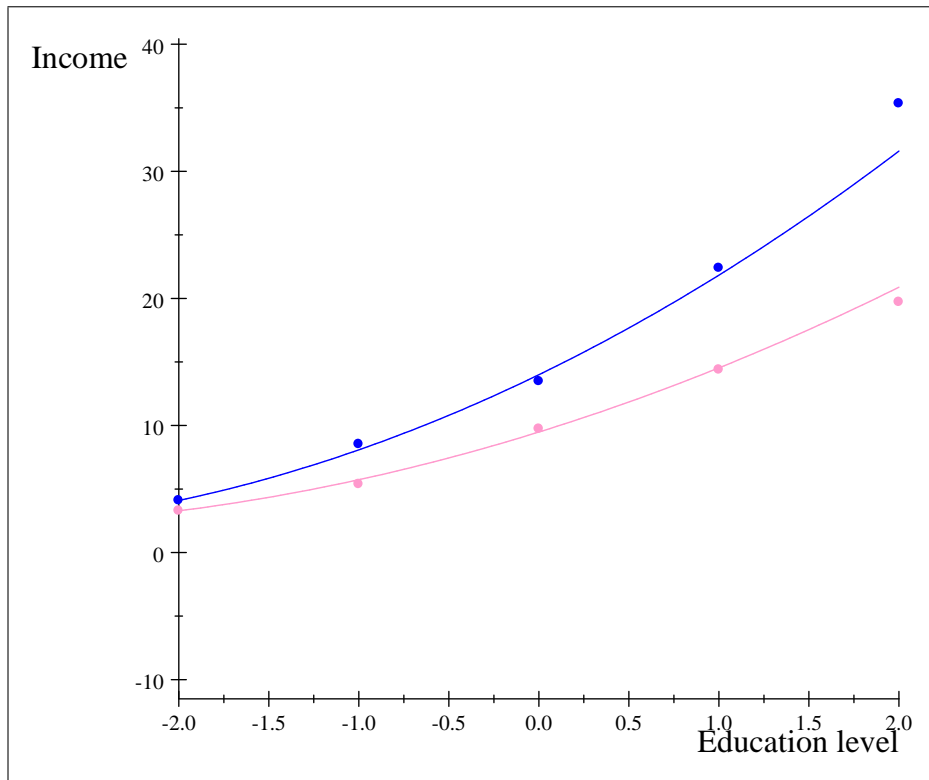
$$= \begin{pmatrix} 1 & 1 & -2 & 4 & -2 & 4 \\ 1 & 1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 4 & 2 & 4 \\ 1 & -1 & -2 & 4 & 2 & -4 \\ 1 & -1 & -1 & 1 & 1 & -1 \\ 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & 2 & 4 & -2 & -4 \end{pmatrix}$$

with a corresponding vector of parameters  $\boldsymbol{\gamma}$ . This will represent a quadratic model where the intercepts, linear coefficients and quadratic coefficients vary for different genders. Using the same process as before, the ML estimate of  $\boldsymbol{\pi}$  can be determined under the vector of constraints  $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$  to find the ML estimate of  $\boldsymbol{\gamma}$  from which the functions

$$\hat{v}_F = 9.4837 + 4.4x_E + 0.6492x_E^2$$

$$\hat{v}_M = 13.9957 + 6.8742x_E + 0.9622x_E^2$$

are determined. These functions are illustrated graphically in Figure 5-3.



**Figure 5-3: Quadratic functions for income under different genders with level of education as ordinal variable**

A measure of discrepancy of 0.0096 was attained indicating the quadratic model does give a good fit. This can also be seen in Figure 5-3 where the income level for a Male with a post graduate level of education is more accurately described.

### 5.3 Population group and level of education

The two-factor model will now be used to study the effects of population group and level of education on the grouped response variable, income. The resultant histograms are given in Table 5.11 when the two explanatory variables are cross-tabulated with the frequency distribution of income.





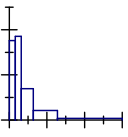
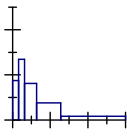
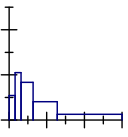
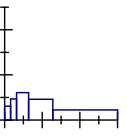
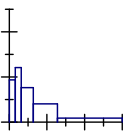
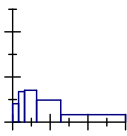
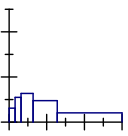
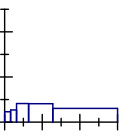
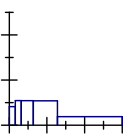
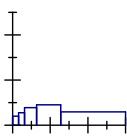
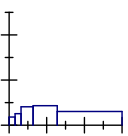
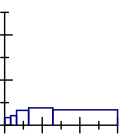
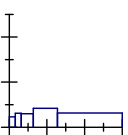
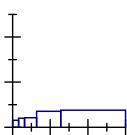
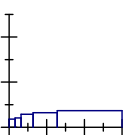
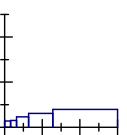
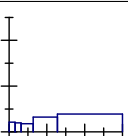
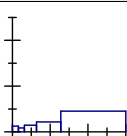
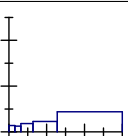
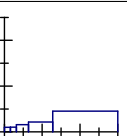
Education	Population group			
	Black	Coloured	Indian	White
Grade 12	 $n = 195887$	 $n = 31335$	 $n = 15150$	 $n = 51414$
Certif	 $n = 20495$	 $n = 2708$	 $n = 1452$	 $n = 9388$
Diploma	 $n = 48176$	 $n = 6174$	 $n = 4144$	 $n = 27389$
B Degree	 $n = 21444$	 $n = 2468$	 $n = 3650$	 $n = 18019$
Post Grad	 $n = 12654$	 $n = 1444$	 $n = 2288$	 $n = 14018$

Table 5.11: Population group and level of education

### 5.3.1 Distribution fitting with the saturated model

The iterative procedure is used such that the expected relative frequencies  $\hat{\pi}_1, \hat{\pi}_2, \dots, \hat{\pi}_{20}$  for each of the  $T = 20$  cells follow log-logistic distributions at the upper class boundaries  $\mathbf{x}$ . The estimated medians are reparameterized to find the effects of the explanatory variables with

$$\nu_{jk} = \tau_0 + \tau_j^P + \tau_k^E + \tau_{jk}^{PE}$$

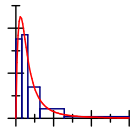
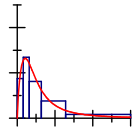
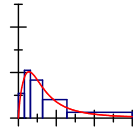
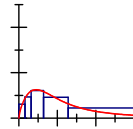
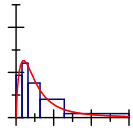
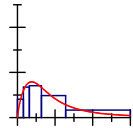
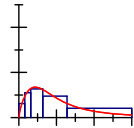
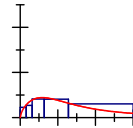
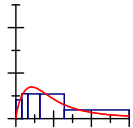
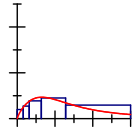
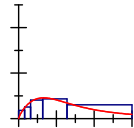
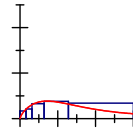
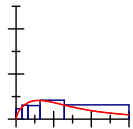
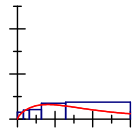
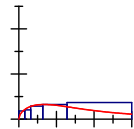
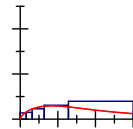
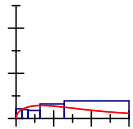
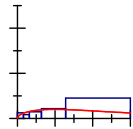
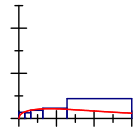
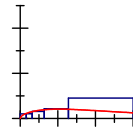
where  $j = 1, 2, 3, 4$  for the different population groups Black, Coloured, Indian and White, respectively, and  $k = 1, 2, 3, 4, 5$  for the different levels of education Grade 12, Diploma, Certificate, Bachelor's degree and Postgraduate degree, respectively. The reparameterization is given in matrix notation by

$$\boldsymbol{\nu} = \mathbf{Z}\boldsymbol{\tau}$$

where  $\mathbf{Z} = (\mathbf{1}_{20} : \mathbf{Z}_P : \mathbf{Z}_E : \mathbf{Z}_{PE})$  and  $\boldsymbol{\tau} = \begin{pmatrix} \tau_0 \\ \boldsymbol{\tau}^P \\ \boldsymbol{\tau}^E \\ \boldsymbol{\tau}^{PE} \end{pmatrix}$ . The submatrices of  $\mathbf{Z}$  are defined by making use of design matrices  $\mathbf{D}_P$  and  $\mathbf{D}_E$  as defined in Section 5.1.2 and Section 5.2.1, respectively from which

$$\begin{aligned} \mathbf{Z}_P &= \mathbf{D}_P \otimes \mathbf{1}_5 \\ \mathbf{Z}_E &= \mathbf{1}_4 \otimes \mathbf{D}_E \\ \mathbf{Z}_{PE} &= \mathbf{Z}_P \odot \mathbf{Z}_E \end{aligned}$$

are constructed. The fitted distributions and corresponding parameters are given in Table 5.12.

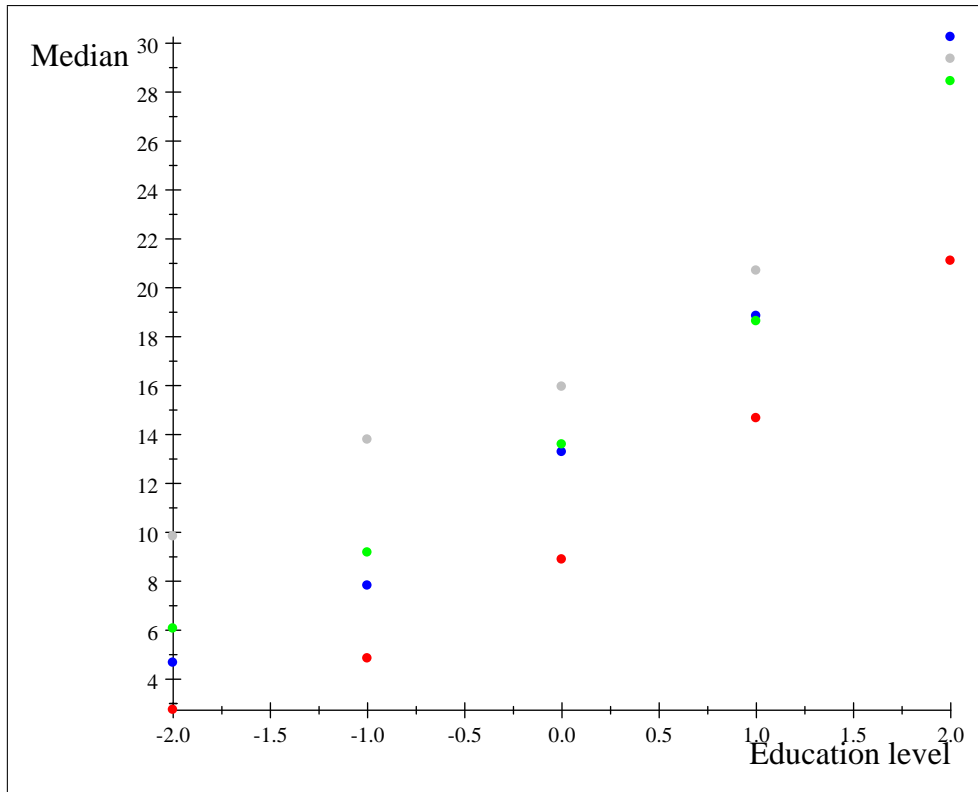
Education	Population group				$\hat{\tau}^E$
	Black	Coloured	Indian	White	
Grade 12	 $\hat{\nu} = 2.732$ $\hat{\tau}^{PE} = 1.095$	 $\hat{\nu} = 4.665$ $\hat{\tau}^{PE} = -1.498$	 $\hat{\nu} = 6.061$ $\hat{\tau}^{PE} = -0.310$	 $\hat{\nu} = 9.830$ $\hat{\tau}^{PE} = 0.713$	-8.805
Certif	 $\hat{\nu} = 4.837$ $\hat{\tau}^{PE} = 0.122$	 $\hat{\nu} = 7.815$ $\hat{\tau}^{PE} = -1.427$	 $\hat{\nu} = 9.168$ $\hat{\tau}^{PE} = -0.282$	 $\hat{\nu} = 13.784$ $\hat{\tau}^{PE} = 1.588$	-5.726
Diploma	 $\hat{\nu} = 8.881$ $\hat{\tau}^{PE} = 0.144$	 $\hat{\nu} = 13.274$ $\hat{\tau}^{PE} = 0.010$	 $\hat{\nu} = 13.589$ $\hat{\tau}^{PE} = 0.117$	 $\hat{\nu} = 15.947$ $\hat{\tau}^{PE} = -0.271$	-1.705
B Degree	 $\hat{\nu} = 14.661$ $\hat{\tau}^{PE} = 0.640$	 $\hat{\nu} = 18.841$ $\hat{\tau}^{PE} = 0.294$	 $\hat{\nu} = 18.627$ $\hat{\tau}^{PE} = -0.129$	 $\hat{\nu} = 20.697$ $\hat{\tau}^{PE} = -0.805$	3.579
Post Grad	 $\hat{\nu} = 21.098$ $\hat{\tau}^{PE} = -2.000$	 $\hat{\nu} = 30.246$ $\hat{\tau}^{PE} = 2.621$	 $\hat{\nu} = 28.437$ $\hat{\tau}^{PE} = 0.604$	 $\hat{\nu} = 29.355$ $\hat{\tau}^{PE} = -1.225$	12.657
$\hat{\tau}^P$	-4.185	0.341	0.549	3.295	$\hat{\tau}_0 = 14.627$

**Table 5.12: Fitted distributions for population group and level of education**

With the measure of discrepancy being calculated as 0.0083, one can conclude that the estimated frequencies gives an adequate estimation of the observed frequencies. The marginal deviations from the overall median caused by population group also still seem to indicate that population group has an effect even if level of education is taken into account. The marginal effects for level of education indicate that an ordinal trend exists between income and level of education. This can be studied further by fitting linear models to the different median income levels.

### 5.3.2 Linear model

Figure 5-4 presents the median income levels against level of education for different population groups where the red, blue, green and grey dots represent the Black, Coloured, Indian and White population groups, respectively. Note that level of education is coded as in Section 4.4.1.



**Figure 5-4: Medians under different population groups with level of education as ordinal variable**

The ordinal relationship between level of education and income can be captured by the model

$$\nu = \mathbf{Y}\gamma$$

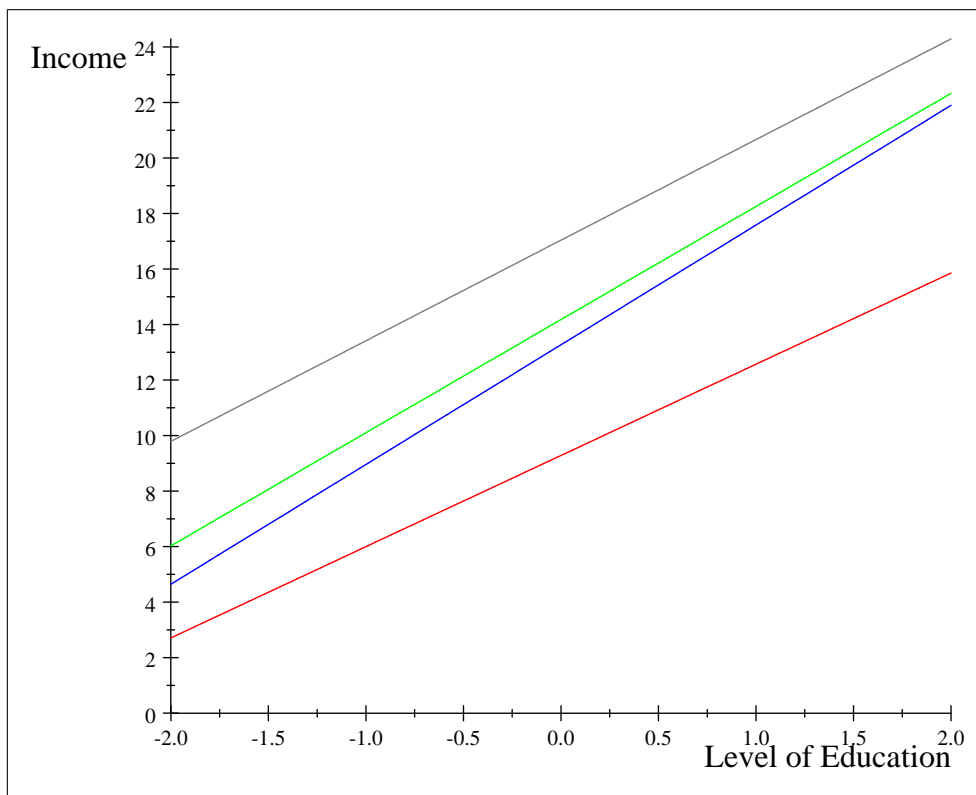
where the design matrix is  $\mathbf{Y} = (\mathbf{1}_{20} : \mathbf{Z}_p : \mathbf{Y}_E : \mathbf{Y}_{PE})$  with corresponding vector of parameters  $\gamma = \begin{pmatrix} \mu \\ \tau^P \\ \beta^E \\ \eta_1^{PE} \\ \eta_2^{PE} \\ \eta_3^{PE} \end{pmatrix}$ .

The vectors  $\mathbf{1}_{20}$  and  $\mathbf{Y}_E = \mathbf{1}_4 \otimes \begin{pmatrix} -2 \\ -1 \\ 0 \\ 1 \\ 2 \end{pmatrix}$  coincide with the overall intercept and overall gradient caused

by level of education respectively. The matrices  $\mathbf{Z}_p = \mathbf{D}_P \otimes \mathbf{1}_5$  and  $\mathbf{Y}_{PE} = \mathbf{Z}_p \odot \mathbf{Y}_E$  coincide with the deviations caused by different population groups from the overall intercept and the overall gradient respectively. If  $\mathbf{Q}_Y$  is defined as the projection matrix onto the error space of  $\mathbf{Y}$ , the vector of constraints  $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$  with matrix of partial derivatives  $\mathbf{G}(\boldsymbol{\pi})$  follow as before as

$$\mathbf{g}(\boldsymbol{\pi}) = \begin{pmatrix} \mathbf{g}_{\log}(\boldsymbol{\pi}) \\ \mathbf{g}_{\text{lin}}(\boldsymbol{\pi}) \end{pmatrix} \quad \text{and} \quad \mathbf{G}(\boldsymbol{\pi}) = \begin{pmatrix} \mathbf{G}_{\log}(\boldsymbol{\pi}) \\ \mathbf{G}_{\text{lin}}(\boldsymbol{\pi}) \end{pmatrix}$$

as in (4.19) and (4.20), respectively where  $\mathbf{g}_{\text{lin}}(\boldsymbol{\pi}) = \mathbf{Q}_Y \boldsymbol{\nu} = \mathbf{0}$ . The ML estimate for  $\boldsymbol{\pi}$  under  $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$  is estimated using the iterative procedure and finally the ML estimates of the parameters  $\boldsymbol{\gamma}$  can be attained. The linear functions can now be drawn and are given in Figure 5-5 where the red, blue, green and grey lines represent the Black, Coloured, Indian and White population groups, respectively.



**Figure 5-5: Income for different population groups with level of education as ordinal variable**

The measure of discrepancy attained for this model is 0.0109. The equations for the linear models are given in Table 5.13.

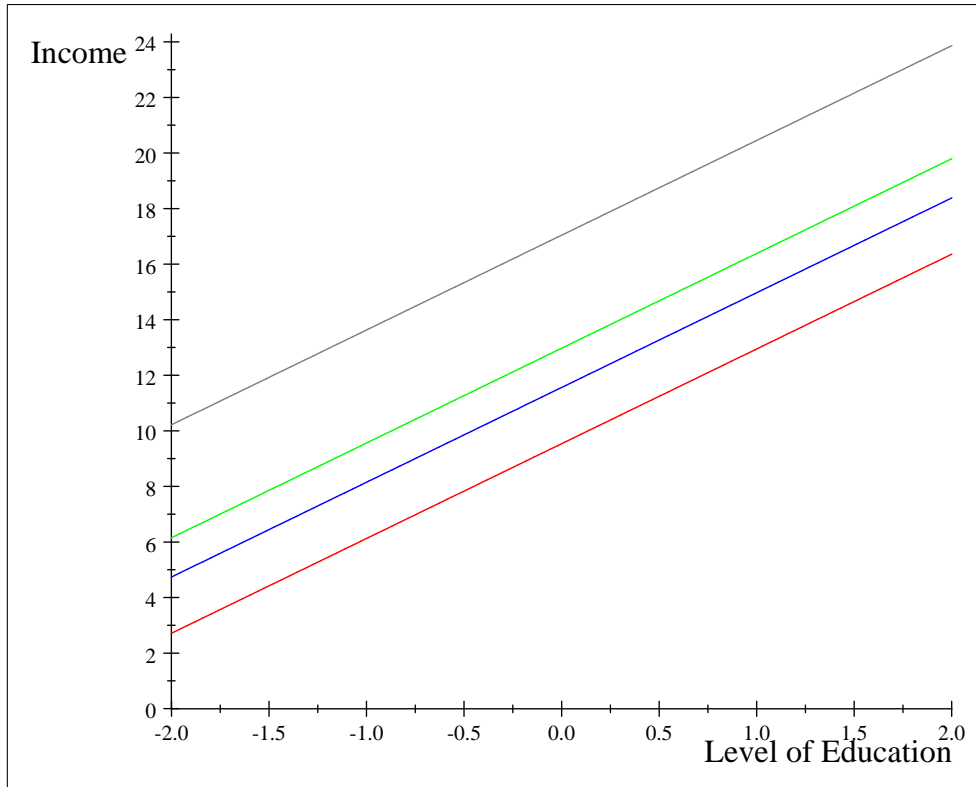
		Income
Population group	Black	$9.2836 + 3.2862x_k$
	Coloured	$13.2708 + 4.3128x_k$
	Indian	$14.1786 + 4.0758x_k$
	White	$17.0402 + 3.6243x_k$

**Table 5.13: Equations for linear models for different population groups**

From the linear models one can see that the White population group attains the highest income whereas the Black population group attains the lowest income over all levels of education. The gradients over different population groups do seem to be comparable but the intercepts are clearly different. This can be incorporated in the estimation procedure by forcing the gradients to be equal.

### 5.3.3 Linear model with equal gradients

The interaction effect between population group and level of education can be discarded by changing the design matrix  $\mathbf{Y}$  to  $\mathbf{Y} = (\mathbf{1}_{20}; \mathbf{Z}_p; \mathbf{Y}_E)$ . The resultant linear functions are given in Figure 5-6.



**Figure 5-6: Income for different population groups with level of education as ordinal variable with equal gradients**

The measure of discrepancy attained for this model is 0.01139 which indicates that the trends for each population group can be taken as equal. The linear models are given in Table 5.14.

		Income
Popgrp	Black	$9.5351 + 3.4111x_k$
	Coloured	$11.5597 + 3.4111x_k$
	Indian	$12.9723 + 3.4111x_k$
	White	$17.0402 + 3.4111x_k$

**Table 5.14: Equations for linear models for different population groups with equal gradients**

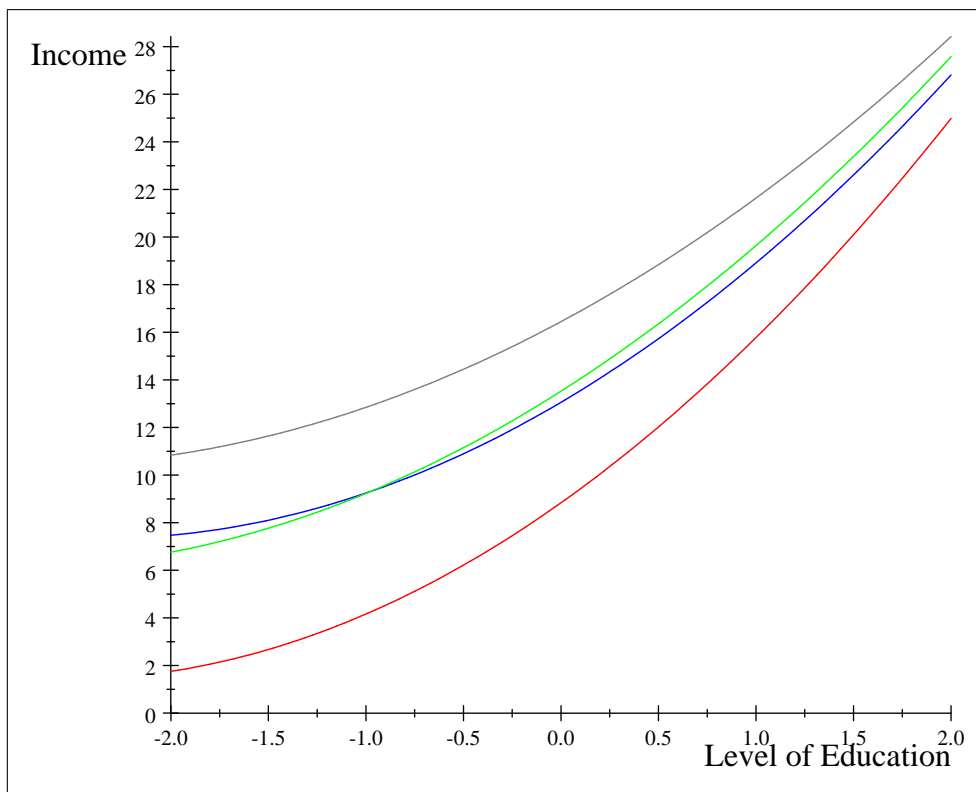
As observed previously, the income values are not properly described with linear models. To this end, a higher order model will also be considered.

### 5.3.4 Quadratic model

The extension to a higher order model is simple. To illustrate this, a model of the form

$$\nu = \mathbf{Y}\gamma$$

can be estimated where the design matrix  $\mathbf{Y} = (\mathbf{1} : \mathbf{Z}_P : \mathbf{Y}_E : \mathbf{Y}_E^2 : \mathbf{Y}_{PE} : \mathbf{Y}_{PE}^2)$  is used with a corresponding vector of parameters  $\gamma$ . This will represent a quadratic model where the intercepts, linear coefficients and quadratic coefficients vary for different population groups. Using the same process as before, the ML estimate of  $\pi$  can be determined under the vector of constraints  $\mathbf{g}(\pi) = \mathbf{0}$  to finally find the ML estimate of  $\gamma$ . A measure of discrepancy of 0.0084 is attained. Note that this is the smallest measure of discrepancy attained between the three models. The result of this model is given in Figure 5-7.



**Figure 5-7: Income for different population groups with level of education as ordinal variable with a quadratic model**

The quadratic functions are given in Table 5.15.

		<b>Income</b>
<b>Population group</b>	<b>Black</b>	$8.8361 + 5.8082x_k + 1.1330x_k^2$
	<b>Coloured</b>	$13.0582 + 4.8347x_k + 1.0211x_k^2$
	<b>Indian</b>	$13.5242 + 5.2053x_k + 0.9114x_k^2$
	<b>White</b>	$16.4462 + 4.3965x_k + 0.7958x_k^2$

**Table 5.15: Equations for quadratic models for different population groups with equal gradients**

If a quadratic model is fitted to the median income levels, one will note that the White population group still earns the highest income whereas the Black population group earns the lowest income over all levels of education. The quadratic model does provide more information regarding the Coloured and Indian population groups. If the level of education is higher than a certificate level, the Indian population group earns a higher income when compared to the Coloured population group.

## 5.4 Gender, population group and level of education

The same technique described in Chapter 4 and in Sections 5.1, 5.2 and 5.3 can be used to fit a three factor model to income. If gender, population group and level of education are considered as explanatory variables, a table with  $T = 2 \times 4 \times 5$  cells can be created where the different explanatory variables are cross-tabulated.

### 5.4.1 Distribution fitting

The aim is to estimate different log-logistic distributions at the upper class boundaries  $\mathbf{x}$  for each of the  $T = 40$  cells simultaneously from which the medians of the fitted log-logistic distributions can be analyzed to see the effects of the explanatory variables. The results of doing so are summarized in Tables 5.16 and 5.17.





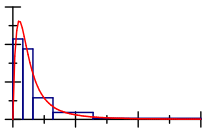
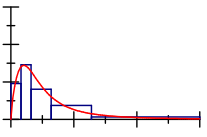
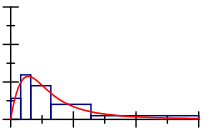
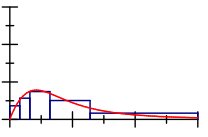
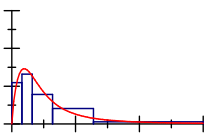
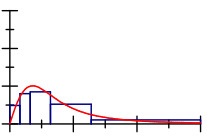
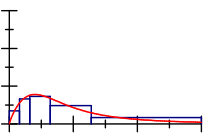
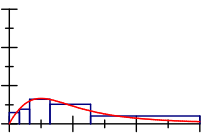
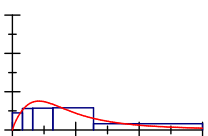
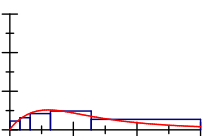
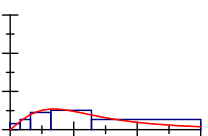
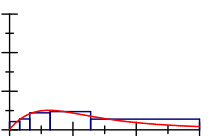
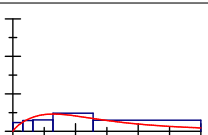
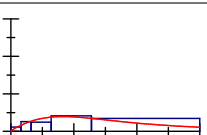
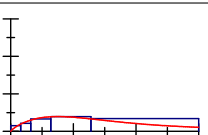
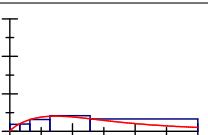
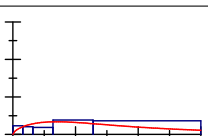
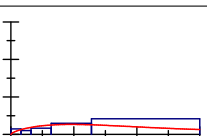
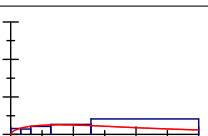
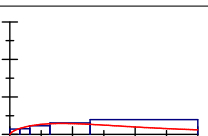
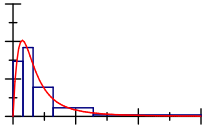
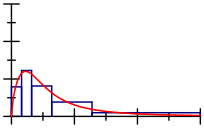
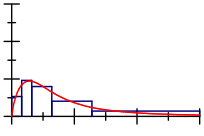
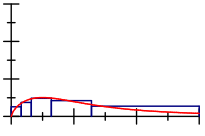
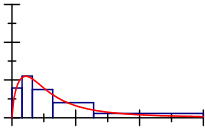
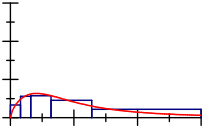
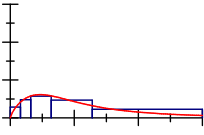
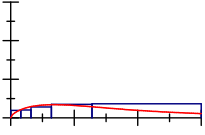
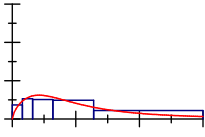
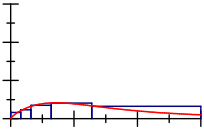
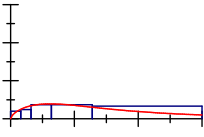
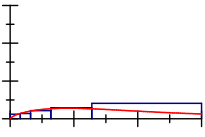
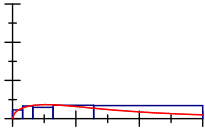
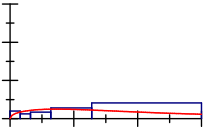
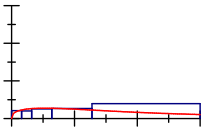
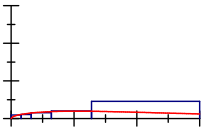
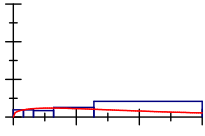
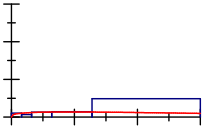
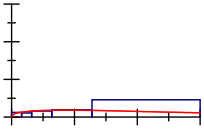
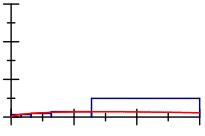
Female Education	Population group			
	Black	Coloured	Indian	White
Grade 12	 $\hat{\nu} = 2.328$	 $\hat{\nu} = 4.287$	 $\hat{\nu} = 5.467$	 $\hat{\nu} = 8.055$
Certificate	 $\hat{\nu} = 4.217$	 $\hat{\nu} = 6.367$	 $\hat{\nu} = 7.983$	 $\hat{\nu} = 9.489$
Diploma	 $\hat{\nu} = 8.271$	 $\hat{\nu} = 12.11$	 $\hat{\nu} = 11.97$	 $\hat{\nu} = 12.36$
B Degree	 $\hat{\nu} = 13.45$	 $\hat{\nu} = 15.95$	 $\hat{\nu} = 15.67$	 $\hat{\nu} = 15.25$
Post Grad	 $\hat{\nu} = 18.02$	 $\hat{\nu} = 22.85$	 $\hat{\nu} = 23.83$	 $\hat{\nu} = 20.76$

Table 5.16: Fitted distributions for Females: Population group vs level of education



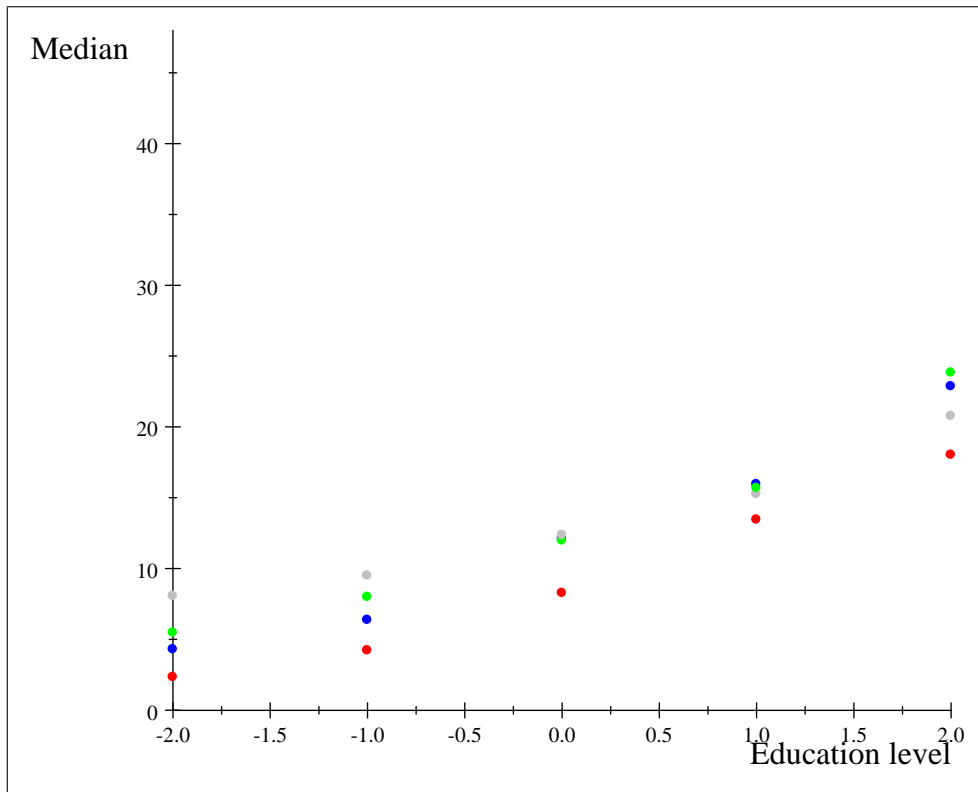
Male Education	Population group			
	Black	Coloured	Indian	White
Grade 12	 $\hat{\nu} = 3.064$	 $\hat{\nu} = 5.090$	 $\hat{\nu} = 6.485$	 $\hat{\nu} = 12.11$
Certificate	 $\hat{\nu} = 5.537$	 $\hat{\nu} = 9.625$	 $\hat{\nu} = 10.02$	 $\hat{\nu} = 17.81$
Diploma	 $\hat{\nu} = 9.822$	 $\hat{\nu} = 14.89$	 $\hat{\nu} = 15.50$	 $\hat{\nu} = 21.90$
B Degree	 $\hat{\nu} = 16.60$	 $\hat{\nu} = 24.49$	 $\hat{\nu} = 22.73$	 $\hat{\nu} = 31.01$
Post Grad	 $\hat{\nu} = 26.42$	 $\hat{\nu} = 47.17$	 $\hat{\nu} = 33.81$	 $\hat{\nu} = 42.63$

**Table 5.17: Fitted distributions for Males: Population group vs level of education**

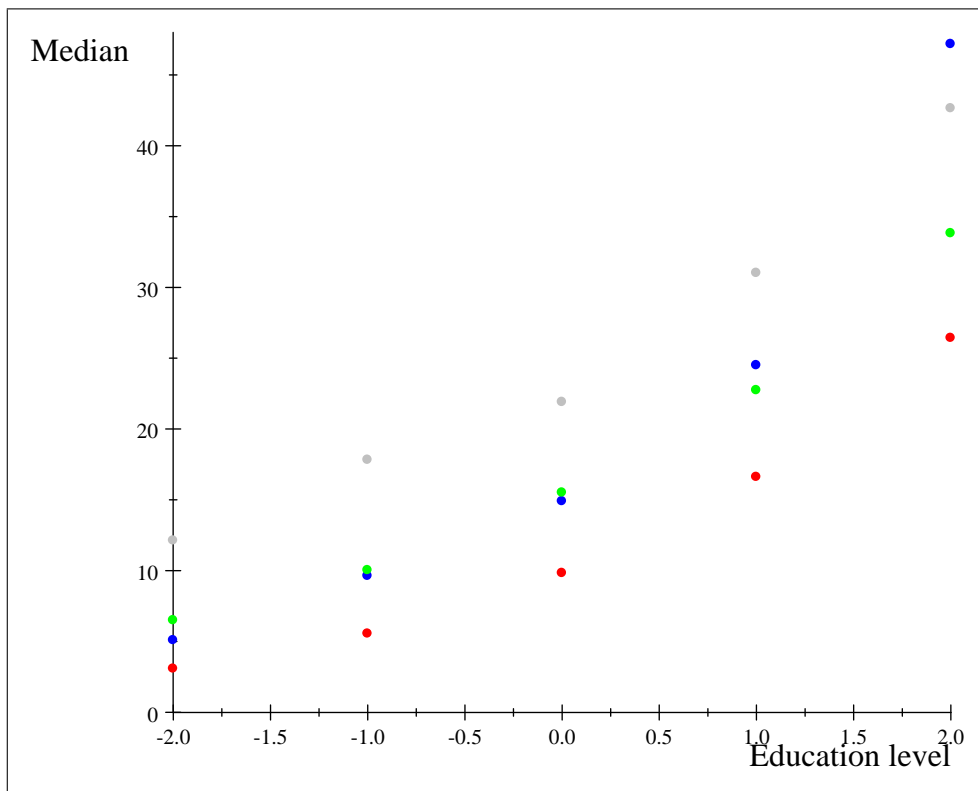
A measure of discrepancy of 0.0097 was attained. It follows that the estimated frequencies of each cell do not differ significantly from the observed frequencies. Although Table 5.16 and 5.17 present the estimated pdfs and medians, it may be difficult to make inferences regarding the effects of the explanatory variables on income. The medians can be reparameterized in a similar manner as before to find the deviations from the overall median income caused by the explanatory variables but instead of doing so, one can consider fitting different models to the medians.

### 5.4.2 Linear model

The medians of each fitted distribution will be plotted against the ordinal variable, level of education, with the red, blue, green and grey dots representing the Black, Coloured, Indian and White population groups respectively, whereafter a model can be suggested.



**Figure 5-8: Median income for Females: Ordinal trend in level of education for each population group**



**Figure 5-9: Median income for Males: Ordinal trend in level of education for each population group**

By simply considering the medians in Figure 5-8 and 5-9 one should note that level of education

still plays an ordinal role with regards to the median income but the effect of level of education between the different genders is considerably different. To delve further into this observation, one can fit linear models to the median income levels for the different combinations of gender and population group using level of education as the ordinal variable. The linear model will be of the form

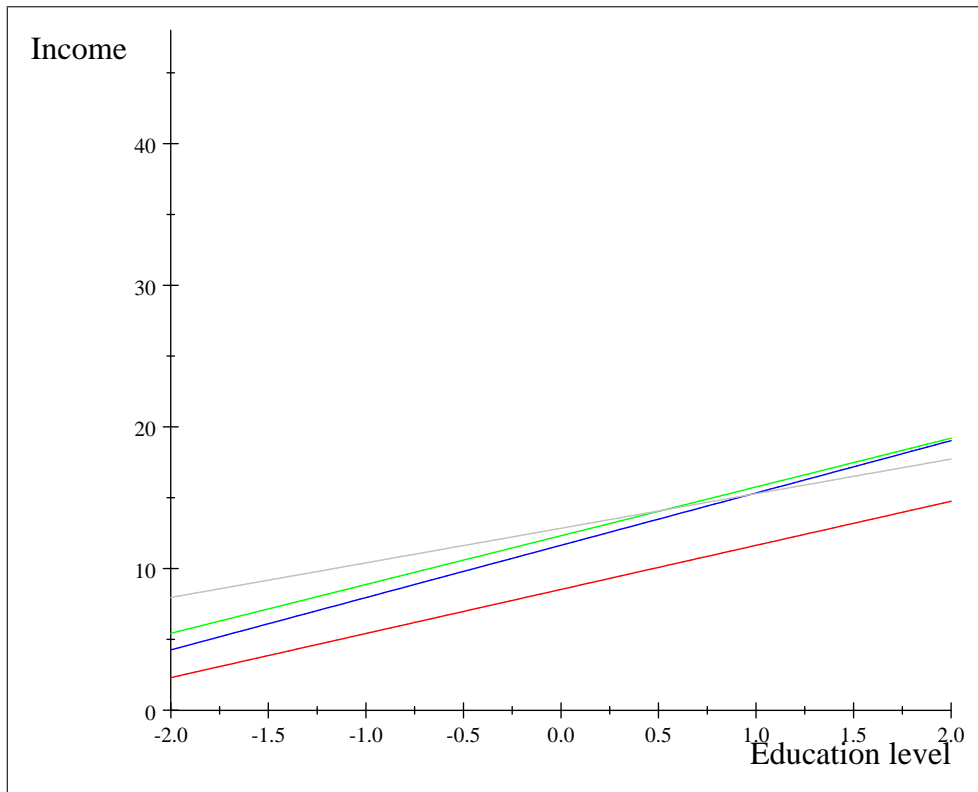
$$\nu = \mathbf{Y}\gamma$$

where the design matrix  $\mathbf{Y} = (\mathbf{1}_{40} : \mathbf{Z}_G : \mathbf{Z}_P : \mathbf{Y}_E : \mathbf{Z}_{GP} : \mathbf{Y}_{GE} : \mathbf{Y}_{PE} : \mathbf{Y}_{GPE})$  with the submatrices of  $\mathbf{Y}$  in given Table 5.18.

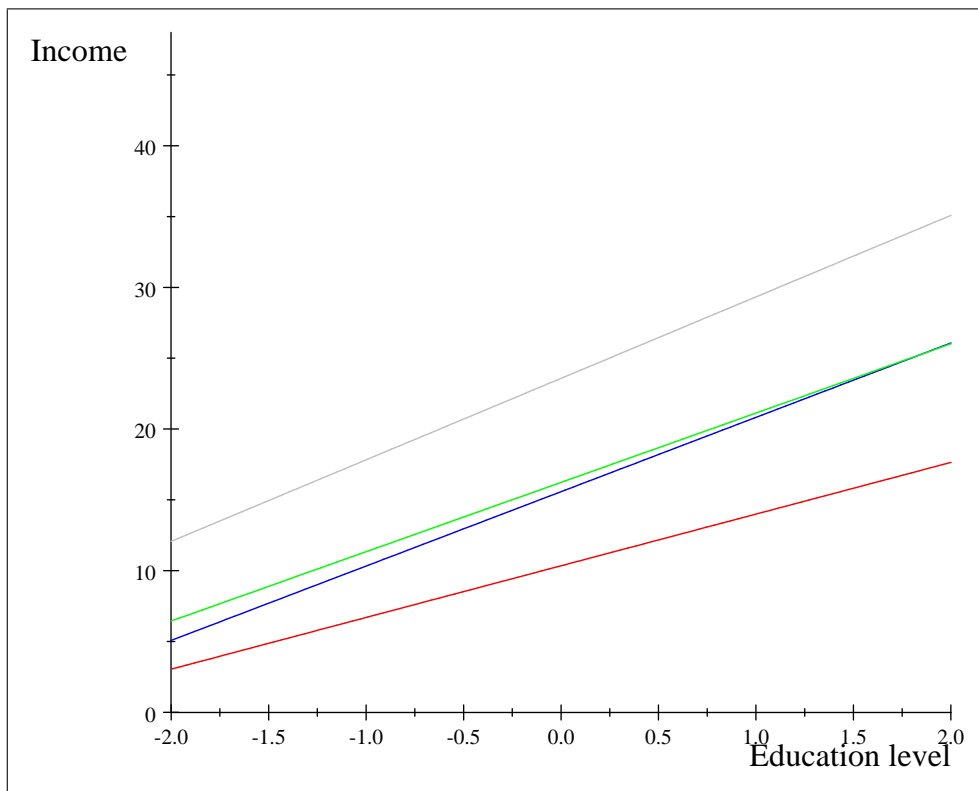
Main Effects	First-order interactions	Second-order interactions
$\mathbf{Z}_G = \mathbf{D}_G \otimes \mathbf{1}_4 \otimes \mathbf{1}_5$	$\mathbf{Z}_{GP} = \mathbf{Z}_G \odot \mathbf{Z}_P$	$\mathbf{Y}_{GPE} = \mathbf{Z}_{GP} \odot \mathbf{Y}_E$
$\mathbf{Z}_P = \mathbf{1}_2 \otimes \mathbf{D}_P \otimes \mathbf{1}_5$	$\mathbf{Y}_{PE} = \mathbf{Z}_P \odot \mathbf{Y}_E$	
$\mathbf{Y}_E = \mathbf{1}_2 \otimes \mathbf{1}_4 \otimes \begin{pmatrix} -2 \\ -1 \\ 0 \\ 1 \\ 2 \end{pmatrix}$	$\mathbf{Y}_{GE} = \mathbf{Z}_G \odot \mathbf{Y}_E$	

**Table 5.18: Sub-matrices of the design matrix  $\mathbf{Y}$**

In the design matrix, the vectors  $\mathbf{1}_{40}$  and  $\mathbf{Y}_E$  correspond with the overall intercept and gradient caused by level of education respectively. The matrices  $\mathbf{Z}_G$ ,  $\mathbf{Z}_P$  and  $\mathbf{Z}_{GP}$  correspond with the deviations from the overall intercept caused by gender, population group and the interaction of gender and population group respectively. The matrices  $\mathbf{Y}_{GE}$ ,  $\mathbf{Y}_{PE}$  and  $\mathbf{Y}_{PGE}$  correspond with the deviations from the overall gradient caused by gender, population group and the interaction of gender and population group, respectively. The ML estimate for  $\pi$  is then found such that for each of the  $T = 40$  cells, the expected cumulative relative frequencies will equal a cumulative log-logistic distribution at the upper class boundaries  $\mathbf{x}$  and that each of the  $T = 40$  medians of the log-logistic distributions are in the vector space of  $\mathbf{Y}$ . The resultant linear functions are given in Figures 5-10 and 5-11.



**Figure 5-10: Linear models for Females: Ordinal trend in level of education for each population group**



**Figure 5-11: Linear models for Males: Ordinal trend in level of education for each population group**

A measure of discrepancy of 0.0127 was attained for this model. The equations are summarised in

Table 5.19.

		Gender	
		Female	Male
Popgrp	Black	$8.5276 + 3.1116x$	$10.3472 + 3.6498x$
	Coloured	$11.6444 + 3.6923x$	$15.5754 + 5.2498x$
	Indian	$12.3183 + 3.4430x$	$16.2363 + 4.8938x$
	White	$12.8450 + 2.4428x$	$23.5782 + 5.7535x$

**Table 5.19: Equations for linear models**

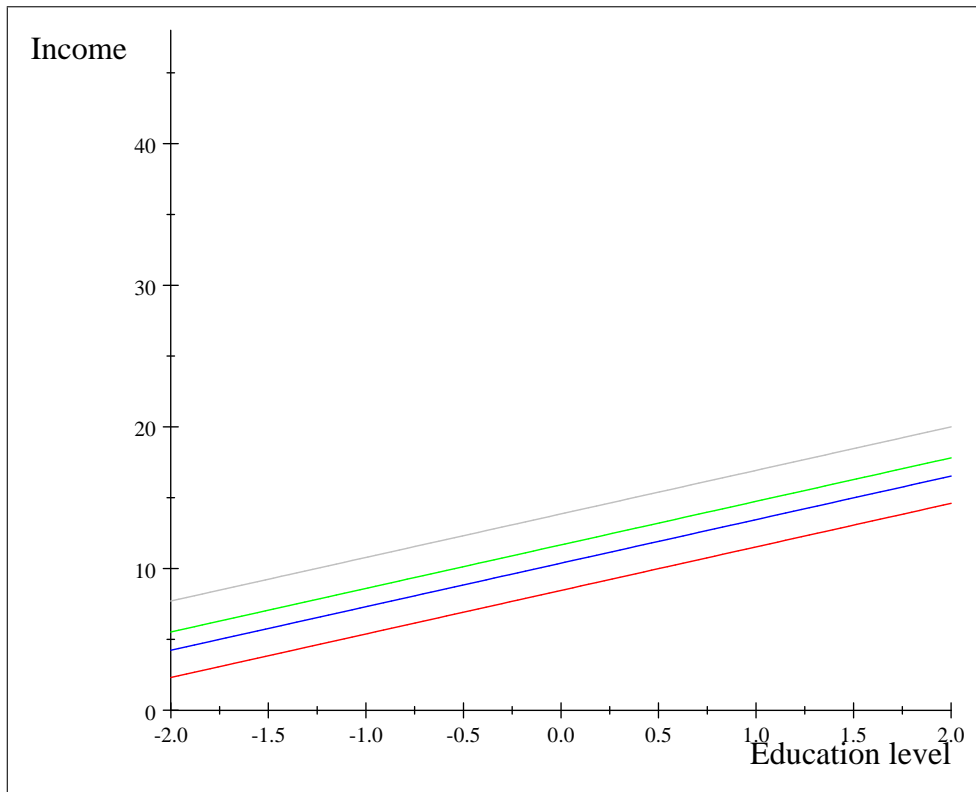
The effect of level of education is clearly more prominent for Males, regardless of population group. For Males and Females, one can note that the gradients over the different population groups seem to be comparable. The gradients over different genders for the same population group, however, do seem to be different. This is mainly seen in the White population group.

### 5.4.3 Linear model with population group and level of education independent

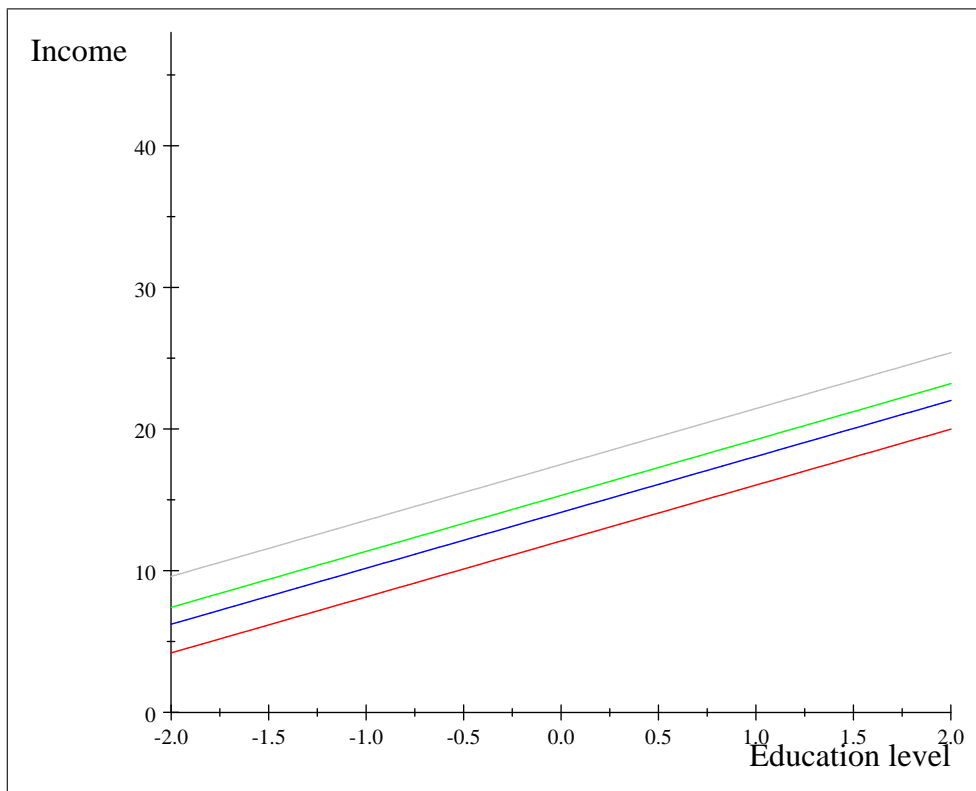
The iterative procedure will be used to estimate linear models where the gradients for different population groups are constant under the same gender. The model of the form

$$\nu = \mathbf{Y}\gamma$$

is used with design matrix  $\mathbf{Y} = (\mathbf{1}_{40} : \mathbf{Z}_G : \mathbf{Z}_P : \mathbf{Y}_E : \mathbf{Z}_{GP} : \mathbf{Y}_{GE})$ . Since this design matrix does not have an interaction effect between population group and level of education, parallel lines are implied if the genders are the same. The ML estimate for  $\pi$  is then found such that for each of the  $T = 40$  cells, the expected cumulative relative frequencies will equal a cumulative log-logistic distribution at the upper class boundaries  $\mathbf{x}$  and that each of the  $T = 40$  medians of the log-logistic distributions are in the vector space of  $\mathbf{Y}$ . The resultant linear functions are given in Figures 5-12 and 5-13.



**Figure 5-12: Linear models for Females: Ordinal trend in level of education for each population group with level of education and population group independent**



**Figure 5-13: Linear models for Males: Ordinal trend in level of education for each population group with level of education and population group independent**

From the latter figures and Table 5.20, the gradients of the linear functions for different population

groups with identical genders are forced to be equal.

		Gender	
		Female	Male
Popgrp	Black	$8.4554 + 3.0735x$	$12.0942 + 3.9488x$
	Coloured	$10.3805 + 3.0735x$	$14.1183 + 3.9488x$
	Indian	$11.67122 + 3.0735x$	$15.31 + 3.9488x$
	White	$13.8575 + 3.0735x$	$17.4963 + 3.9488x$

**Table 5.20: Equations for linear models: Equal gradients under identical genders**

Under this model, a measure of discrepancy of 0.0135 was attained indicating that the model gives a adequate fit.

## 5.5 Summary

Chapter 5 focussed on the extension of the single-factor model to the multifactor model. Specifically, two-factor models were used where gender and population group, gender and level of education and population group and level of education were used as explanatory variables for the grouped response variable, income. After this was done, a three-factor model was used where gender, population group and level of education acted as the explanatory variables. Different linear models were applied to the medians of the defined cells of the two- and three-factor models to further analyze the effect of the explanatory variables on income.



# Chapter 6

## The Logit model

### 6.1 Introduction

The logit model is a simple technique that can also be used to model a grouped response variable. In this chapter, the grouped response variable is coded as a binary response variable and the logit model will be used in conjunction with the iterative procedure if one wishes to find the ML estimates of the frequencies under a certain vector of constraints.

### 6.2 Level of education

To see how level of education effects income, the saturated model will first be used with level of education being an explanatory variable whereafter the iterative procedure will be used to capture the ordinal relationship between level of education and income. Using the same ideology as in Chapter 1, the grouped response variable can be divided up into low- and high income categories from where the table of frequencies for level of education as an explanatory variable can be set up as in Table 6.1. An individual is coded as being in the low income group if income is less than  $R12800$  and in the high income group if income is greater than or equal to  $R12800$ .

	Level of education					
Income group	Grade 12	Certificate	Diploma	B Degree	Post Grad	Total
High (H)	39909	9358	37189	27938	22220	<b>136614</b>
Low (L)	253877	24685	48694	17643	8184	<b>353083</b>
<b>Total</b>	<b>293786</b>	<b>34043</b>	<b>85883</b>	<b>45581</b>	<b>30404</b>	$n = 489697$

Table 6.1: Frequencies for different levels of education

#### 6.2.1 Saturated model

The frequencies in Table 6.1 can be used to estimate simple probabilities. For example, one can estimate the probability of an individual falling in a high income category, given the individual has a Bachelors degree, with  $p_4 = \frac{27938}{45581} = 0.6129$ . The odds of the event occurring can also be calculated as

$odds_4 = \frac{p_4}{1-p_4} = 1.5835$ . If one defines  $f_{H,k}$  and  $f_{L,k}$  as the frequencies in the  $(H, k)^{th}$  and  $(L, k)^{th}$  cells, respectively, then the observed frequencies can be used to find  $odds_4$  using using

$$\begin{aligned} odds_4 &= \frac{\frac{f_{H,4}}{n_4}}{\frac{f_{L,4}}{n_4}} \\ &= \frac{f_{H,4}}{f_{L,4}}. \end{aligned}$$

By taking the log of the odds, the logit model takes form with

$$\ln\left(\frac{f_{H,k}}{f_{L,k}}\right) = \mu + \lambda_k^E$$

for  $k = 1, 2, 3, 4, 5$  for level of education and  $\lambda_5^E = -\sum_{k=1}^4 \lambda_k^E$ . By simply removing the log component, the indices can also be inspected with

$$\begin{aligned} \frac{f_{H,k}}{f_{L,k}} &= e^\mu e^{\lambda_k^E} \\ &= I^\mu I_k^E. \end{aligned}$$

$I^\mu$  is the geometric average of the odds over all the categories of level of education whereas  $I_k^E$  is the index indicating the effect of a certain level of education on the respective odds. If  $0 < I_k^E < 1$ , it reduces the odds of falling in the high income category whereas if  $I_k^E > 1$ , the odds increases. Note that since  $\lambda_5^E = -\sum_{k=1}^4 \lambda_k^E$ , one has that  $I_5^E = \prod_{k=1}^4 (I_k^E)^{-1}$ .

If one were to define  $\mathbf{f}'_H$  and  $\mathbf{f}'_L$  as the row vectors containing the frequencies corresponding with the high- and low income categories, respectively, then the vector of frequencies  $\mathbf{f} = \begin{pmatrix} \mathbf{f}_H \\ \mathbf{f}_L \end{pmatrix}$  can be used to write the logit model in matrix notation. Since

$$\begin{aligned} \ln\left(\frac{\mathbf{f}_H}{\mathbf{f}_L}\right) &= \ln(\mathbf{f}_H) - \ln(\mathbf{f}_L) \\ &= \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 \end{pmatrix} \ln(\mathbf{f}) \\ &= \mathbf{A} \ln(\mathbf{f}) \end{aligned}$$

it follows that

$$\begin{aligned} \mathbf{A} \ln(\mathbf{f}) &= \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 & -1 \end{pmatrix} \begin{pmatrix} \mu \\ \lambda_1^E \\ \lambda_2^E \\ \lambda_3^E \\ \lambda_4^E \end{pmatrix} \\ &= \mathbf{Z}\boldsymbol{\lambda}. \end{aligned}$$

Note that the vector  $\mathbf{f}$  is multinomial distributed with  $E(\mathbf{f}) = \mathbf{F}$  and  $Cov(\mathbf{f}) = \mathbf{D}_{\mathbf{F}} - \frac{1}{n}\mathbf{F}\mathbf{F}' = \mathbf{V}_{\mathbf{F}}$  as in (2.5) and (2.6) with  $\mathbf{D}_{\mathbf{F}} = \text{diag}(\mathbf{F})$ . The vector of parameters  $\boldsymbol{\lambda}$  for the saturated model can be determined by

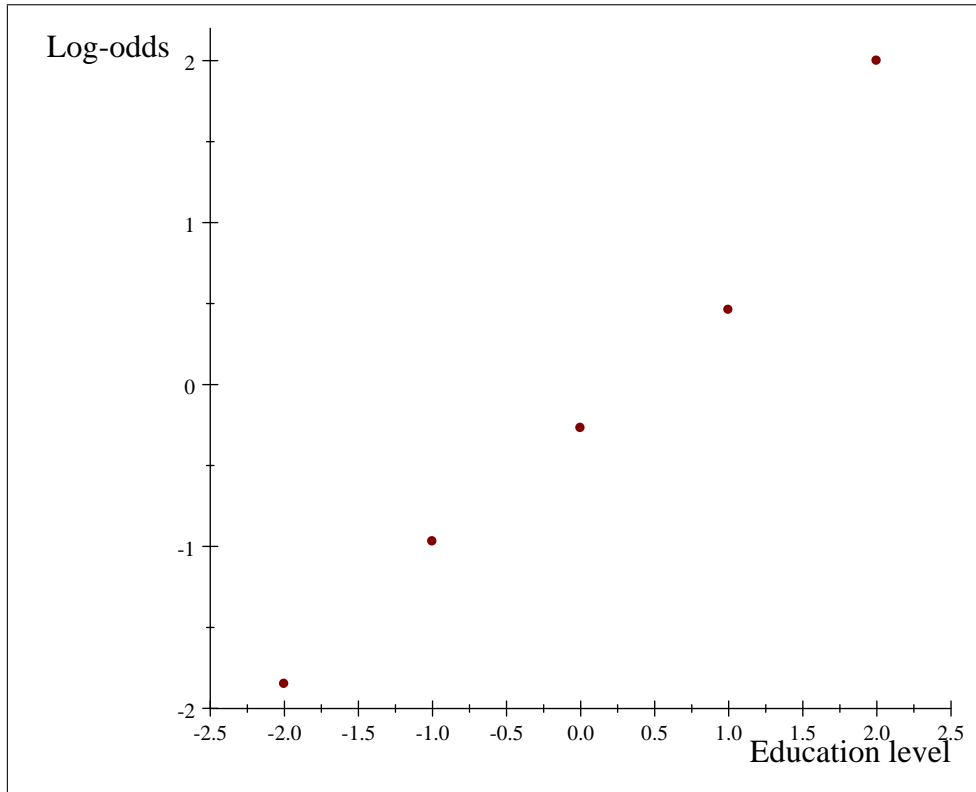
$$\begin{aligned} \boldsymbol{\lambda} &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{A} \ln(\mathbf{f}) \\ &= \begin{pmatrix} -0.3263 \\ -1.5240 \\ -0.6437 \\ 0.0567 \\ 0.7859 \end{pmatrix} \end{aligned}$$

from which the indices can be found as given in Table 6.2.

Indices		Level of education					
		Grade 12	Certificate	Diploma	B Degree	Post Grad	Overall
Income group	High:Low index	0.2178	0.5253	1.0584	2.1944	3.7625	0.7216

**Table 6.2: Indices for levels of education**

Note that the logit model is simply a reparameterization of the log-odds. Hence, using either the indices or the estimated parameters, one is able to find the exact odds or log-odds. The indices do however indicate that an exponential trend might exist in the odds of falling in the high income category and hence a linear trend might exist in the log-odds. Figure 6-1 shows a plot of the log-odds against level of education with education being coded as in Section 4.4.1.



**Figure 6-1: Log-odds with level of education as ordinal variable**

Since  $\mathbf{f}$  is multinomial distributed and thus belongs to the exponential family, Proposition 1 can be redefined in terms of frequencies to find the expected value  $E(\mathbf{f}) = \mathbf{F}$  under a vector of constraints  $\mathbf{g}(\mathbf{F}) = \mathbf{0}$ . It follows that the linear relationship between the log-odds and level of education can be captured using an appropriate vector of constraints. Proposition 3 outlines the iterative procedure needed in this scenario.

*Proposition 3 (ML estimation procedure)*

*Consider a random vector of frequencies  $\mathbf{f}$  with distribution belonging to the exponential family, with*

$$E(\mathbf{f}) = \mathbf{F} \text{ and } Cov(\mathbf{f}) = \mathbf{V}_{\mathbf{F}}.$$

*The observed  $\mathbf{f}$  is the unrestricted ML estimate of  $\mathbf{F}$  and the covariance matrix  $\mathbf{V}_{\mathbf{F}}$  may be a function of  $\mathbf{F}$ . Let  $\mathbf{g}(\mathbf{F})$  be a continuous vector valued function of  $\mathbf{F}$ , for which the first order partial derivatives,*

$$\mathbf{G}_{\mathbf{F}} = \frac{\partial \mathbf{g}(\mathbf{F})}{\partial \mathbf{F}}$$

*with respect to  $\mathbf{F}$  exist. The ML estimate of  $\mathbf{F}$ , subject to the vector of constraints  $\mathbf{g}(\mathbf{F}) = \mathbf{0}$  is obtained iteratively from*

$$\hat{\mathbf{F}} = \mathbf{f} - (\mathbf{G}_{\mathbf{F}} \mathbf{V}_{\mathbf{F}})' (\mathbf{G}_{\mathbf{F}} \mathbf{V}_{\mathbf{F}} \mathbf{G}_{\mathbf{F}}')^* \mathbf{g}(\mathbf{f}) \tag{6.1}$$

where  $\mathbf{G}_{\mathbf{f}} = \left. \frac{\partial \mathbf{g}(\mathbf{F})}{\partial \mathbf{F}} \right|_{\mathbf{F}=\mathbf{f}}$  and  $(\mathbf{G}_{\mathbf{f}} \mathbf{V}_{\mathbf{F}} \mathbf{G}_{\mathbf{f}}')^*$  is a generalized inverse of  $(\mathbf{G}_{\mathbf{f}} \mathbf{V}_{\mathbf{F}} \mathbf{G}_{\mathbf{f}}')$ .

## 6.2.2 Linear model

Using the iterative procedure outlined in Proposition 3, the linear trend in the log-odds can be captured with an appropriate vector of constraints. By attaining the ML estimate  $\hat{\mathbf{F}}$  under the relevant vector of constraints, the ML estimates for the linear model can be estimated. If the linear model is of the form

$$\begin{aligned} \mathbf{A} \ln(\mathbf{F}) &= \begin{pmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} \mu \\ \beta \end{pmatrix} \\ &= \mathbf{Y}\boldsymbol{\gamma} \end{aligned}$$

then  $\mathbf{F}$  needs to be estimated such that  $\mathbf{A} \ln(\mathbf{F})$  is in the vector space of  $\mathbf{Y}$ . Defining  $\mathbf{Q}_Y$  as the projection matrix onto the error space of  $\mathbf{Y}$ , the vector of constraints

$$\mathbf{g}(\mathbf{F}) = \mathbf{Q}_Y \mathbf{A} \ln(\mathbf{F}) = \mathbf{0}$$

with matrix of partial derivatives

$$\begin{aligned} \mathbf{G}_F &= \frac{\partial g(\mathbf{F})}{\partial \mathbf{F}} \\ &= \mathbf{Q}_Y \mathbf{A} \mathbf{D}_F^{-1} \end{aligned}$$

can be used in the iterative procedure. Before the procedure is applied, (6.1) will first be simplified to see if a double iterative procedure is necessary. Hence,

$$\begin{aligned} \mathbf{G}_F \mathbf{V}_F &= \mathbf{Q}_Y \mathbf{A} \mathbf{D}_F^{-1} (\mathbf{D}_F - \frac{1}{n} \mathbf{F} \mathbf{F}') \\ &= \mathbf{Q}_Y \mathbf{A} (\mathbf{I} - \mathbf{1} \mathbf{F} \frac{1}{n}) \\ &= \mathbf{Q}_Y \mathbf{A} - \mathbf{Q}_Y \underbrace{\mathbf{A} \mathbf{1} \mathbf{F}}_0 \frac{1}{n} \\ &= \mathbf{Q}_Y \mathbf{A} - \mathbf{0} \\ &= \mathbf{Q}_Y \mathbf{A} \end{aligned}$$

with the result holding in general if  $\mathbf{A} = (\mathbf{I} : -\mathbf{I})$ . Also,

$$\begin{aligned} \mathbf{G}_f \mathbf{V}_F \mathbf{G}_F' &= \mathbf{G}_f (\mathbf{G}_F \mathbf{V}_F)' \\ &= \mathbf{Q}_Y \mathbf{A} \mathbf{D}_f^{-1} \mathbf{A}' \mathbf{Q}_Y. \end{aligned}$$

If these expressions are substituted into (6.1), the equation simplifies to

$$\hat{\mathbf{F}} = \mathbf{f} - (\mathbf{Q}_Y \mathbf{A})' (\mathbf{Q}_Y \mathbf{A} \mathbf{D}_f^{-1} \mathbf{A}' \mathbf{Q}_Y)^{-1} \mathbf{Q}_Y \mathbf{A} \ln(\mathbf{f}) \quad (6.2)$$

which is only a function of  $\mathbf{f}$ . It follows that only convergence over  $\mathbf{f}$  is required. The procedure is outlined in the following steps:

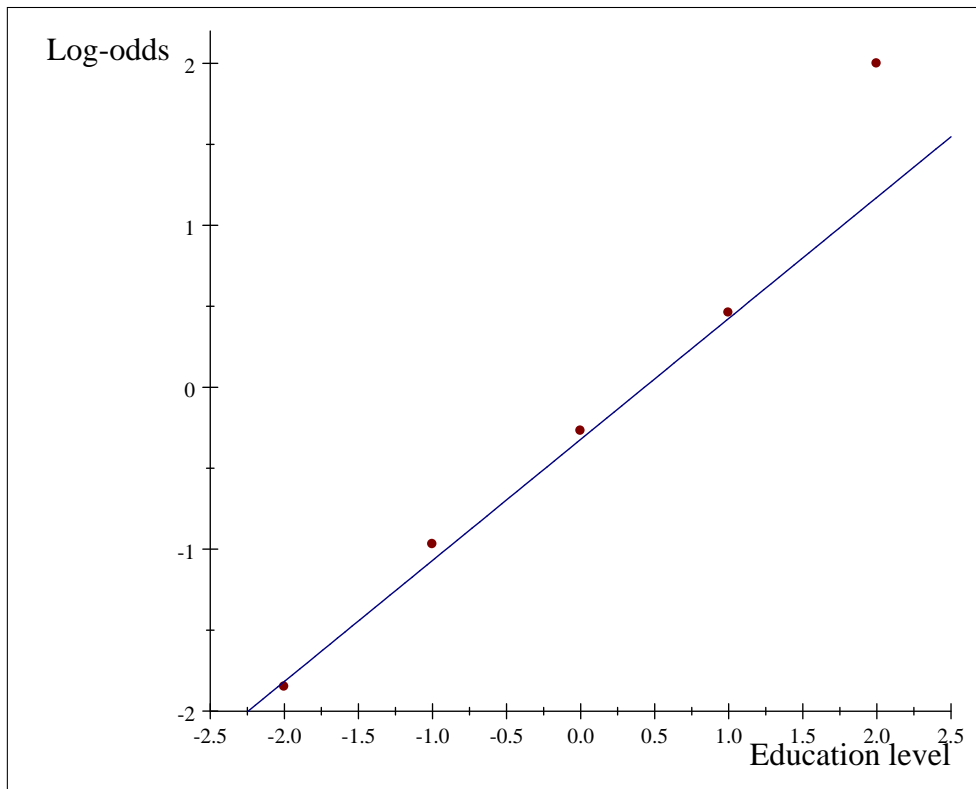
**The iterative procedure**

1. Set  $\mathbf{f}$  equal to the unrestricted ML estimate for  $\mathbf{F}$
2. Set the matrices  $\mathbf{A}$  and  $\mathbf{Q}_Y$
3. Do until convergence over  $\mathbf{f}$ 
  - i. Calculate  $\mathbf{g}(\mathbf{f})$  and  $\mathbf{D}_f$
  - ii. Calculate  $\mathbf{f}$  using (6.2).

An appropriate stopping criteria would be to test if  $\mathbf{g}(\mathbf{f})$  is less than an appropriate error. When this is completed, the ML estimate  $\hat{\mathbf{F}}$  under  $\mathbf{g}(\mathbf{F}) = \mathbf{0}$  is attained and the ML estimates for the parameters of the linear function are estimated with

$$\begin{aligned} \hat{\boldsymbol{\gamma}} &= (\mathbf{Y}'\mathbf{Y})^{-1} \mathbf{Y}'\mathbf{C} \ln(\hat{\mathbf{F}}) \\ &= \begin{pmatrix} -0.3223 \\ 0.7476 \end{pmatrix}. \end{aligned}$$

Using the parameters, a linear function can be drawn which estimates the value of the log-odds at a certain level of education.



**Figure 6-2: Log-odds with level of education as ordinal variable with linear model**

A measure of discrepancy of 0.0007297 was attained. This indicates that the linear model fits the log-odds significantly. The odds can then be attained with the linear model

$$\begin{aligned} \ln\left(\frac{F_{H,k}}{F_{L,k}}\right) &= -0.3223 + 0.7476x_k \\ \left(\frac{F_{H,k}}{F_{L,k}}\right) &= \exp(-0.3223 + 0.7476x_k) \\ &= 0.7245(2.1119)^{x_k} \end{aligned}$$

where  $x_k = -2, -1, 0, 1, 2$ . This shows that as level of education increases, there is an exponential increase in the odds of landing in the high income category. In fact, the odds will more than double as one's level of education increases from one level to the next.

### 6.3 Population group and level of education

Using the logit model, one can also consider using more than one explanatory variable. The first combination of explanatory variables that will be considered is population group and level of education. An extract of the frequencies is given in Table 6.3.

Pop Group	Education	High income	Low income
Black	Grade 12	12349	183538
	Certificate	3128	17367
	Diploma	15756	32420
	B Degree	11719	9725
	Post Grad	8465	4189
⋮	⋮	⋮	⋮
White	Grade 12	19798	31616
	Certificate	4939	4449
	Diploma	16119	11270
	B Degree	12291	5728
	Post Grad	10905	3113

**Table 6.3: Frequencies for population group and level of education**

#### 6.3.1 Saturated model

With the logit model being defined as

$$\ln\left(\frac{f_{H,j,k}}{f_{L,j,k}}\right) = \mu + \lambda_j^P + \lambda_k^E + \lambda_{jk}^{PE} \tag{6.3}$$

with  $j = 1, 2, 3, 4$  for population group and  $k = 1, 2, 3, 4, 5$  for level of education, one can concatenate the high- and low income frequencies to form  $\mathbf{f} = \begin{pmatrix} \mathbf{f}_H \\ \mathbf{f}_L \end{pmatrix}$  to express the logit model in matrix notation

$$\mathbf{A} \ln(\mathbf{f}) = \mathbf{Z}\boldsymbol{\lambda} \quad (6.4)$$

with  $\mathbf{A} = (\mathbf{I}_{20} : -\mathbf{I}_{20})$  and  $\mathbf{Z} = (\mathbf{1} : \mathbf{Z}_P : \mathbf{Z}_E : \mathbf{Z}_{PE})$ . Note that the submatrices of  $\mathbf{Z}$  are identical to the submatrices defined in Section 5.3.1 for the saturated model for the medians.

The corresponding vector of parameters is given by  $\boldsymbol{\lambda} = \begin{pmatrix} \mu \\ \lambda^P \\ \lambda^E \\ \lambda^{PE} \end{pmatrix}$ . From this,  $\boldsymbol{\lambda}$  and finally the indices can be determined.

Indices		Level of Education					Marginal
		Grade 12	Certificate	Diploma	B Degree	Post Grad	
Population Group	Black	0.6401	0.7856	1.0297	1.4141	1.3657	0.5003
	Coloured	0.7915	0.8789	1.1038	1.0910	1.1937	0.9991
	Indian	1.1577	1.0452	1.0146	0.8994	0.9056	1.1441
	White	1.7049	1.3857	0.8672	0.7207	0.6774	1.7486
	Marginal	0.2457	0.5358	1.1031	1.9915	3.4578	$I_\mu = 0.8552$

**Table 6.4: Indices for population group and levels of education**

Using these indices, one can calculate the odds of an individual of a certain population group and a certain level of education to be in the high income category. To show the simplicity of the calculation, consider the odds of an individual from the Black population group with a post graduate level of education to land in the high income category:

$$\begin{aligned} \frac{f_{H,1,5}}{f_{L,1,5}} &= I_\mu \times I_1^P \times I_5^E \times I_{1,5}^{PE} \\ &= 0.8552 \times 0.5003 \times 3.4578 \times 1.3657 \\ &= 2.0205. \end{aligned}$$

One can also consider the odds of an individual from the White population group with a post graduate level of education to land in the high income category:

$$\begin{aligned} \frac{f_{H,4,5}}{f_{L,4,5}} &= I_\mu \times I_4^P \times I_5^E \times I_{4,5}^{PE} \\ &= 0.8552 \times 1.7486 \times 3.4578 \times 0.6774 \\ &= 3.5027. \end{aligned}$$

From these two odds, one can see that even at a post graduate level, population group still has a substantial effect on income. The marginal indices also indicate that population group has a significant



effect on the odds of falling in the high income category. The marginal indices for level of education indicate an exponential trend in the odds which may indicate a linear trend in the log-odds.

### 6.3.2 The linear model

The iterative procedure can now be used to further study the effect of level of education on the log-odds of falling in a high income category for the different population groups. The linear model that will be used is

$$\mathbf{A} \ln(\mathbf{F}) = \mathbf{Y}\boldsymbol{\gamma} \quad (6.5)$$

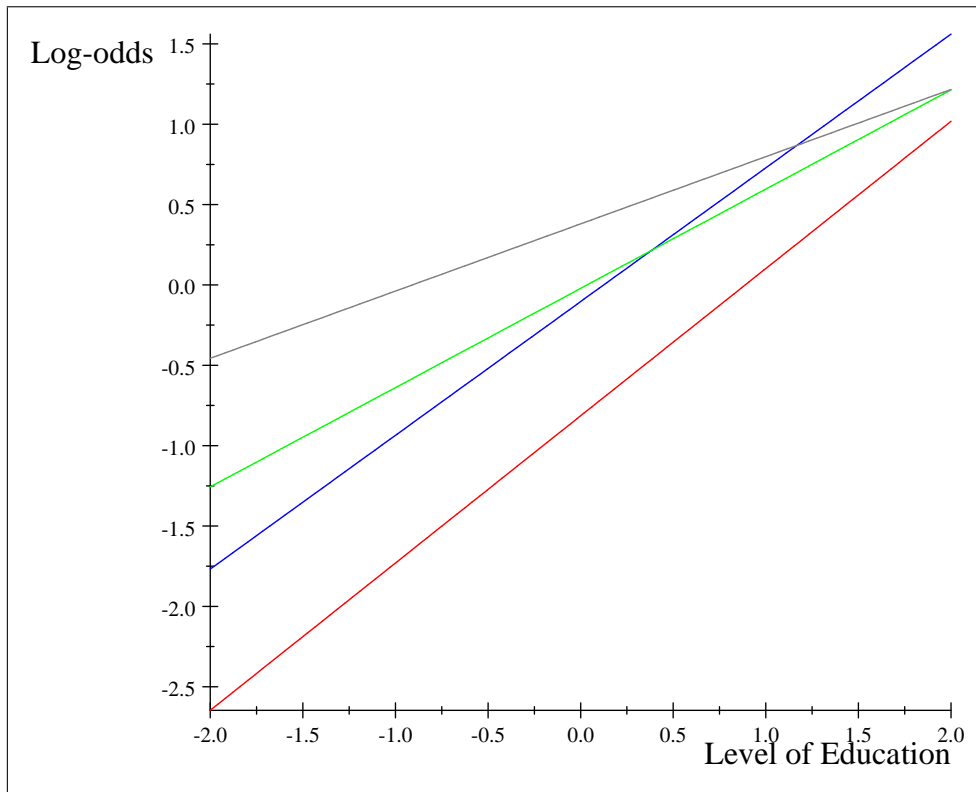
with  $\mathbf{Y} = (\mathbf{1}_{20} : \mathbf{Z}_p : \mathbf{Y}_E : \mathbf{Y}_{PE})$ . The submatrices of  $\mathbf{Y}$  are identical to the submatrices defined in Section 5.3.2 where a linear model was estimated for the median income levels. Defining  $\mathbf{Q}_Y$  as the projection matrix onto the error space of  $\mathbf{Y}$ , the ML estimate of  $\mathbf{F}$  can then be attained such that  $\mathbf{A} \ln(\mathbf{F})$  is in the vector space of  $\mathbf{Y}$ . The vector of constraints is then

$$\mathbf{g}(\mathbf{F}) = \mathbf{Q}_Y \mathbf{A} \ln(\mathbf{F}) = \mathbf{0}$$

with matrix of partial derivatives

$$\begin{aligned} \mathbf{G}_F &= \frac{\partial \mathbf{g}(\mathbf{F})}{\partial \mathbf{F}} \\ &= \mathbf{Q}_Y \mathbf{A} \mathbf{D}_F^{-1}. \end{aligned}$$

Applying the iterative procedure, the ML estimate of  $\mathbf{F}$  under  $\mathbf{g}(\mathbf{F}) = \mathbf{0}$  with a measure of discrepancy of 0.0012 is attained. From  $\hat{\mathbf{F}}$ , the parameter estimates can be attained and the linear functions in Figure 6-3 can be drawn. The red, blue, green and grey line represents the Black, Coloured, Indian and White population groups respectively.



**Figure 6-3: Log-odds of different population groups with level of education as ordinal variable with linear model**

By taking the anti-logarithm as before, the functions can be converted to find the estimated odds and are given in Table 6.5.

		Odds
<b>Population group</b>	<b>Black</b>	$0.4431(2.1993)^{x_k}$
	<b>Coloured</b>	$0.9016(2.2981)^{x_k}$
	<b>Indian</b>	$0.9790(1.8543)^{x_k}$
	<b>White</b>	$1.4623(1.5192)^{x_k}$

**Table 6.5: Equations to find the estimated odds under different population groups**

One will note that the White population group starts off with a higher log-odds but is overtaken by the Coloured population group as soon as level of education is higher than a bachelors degree. The Indian population group is also overtaken by the Coloured population group when level of education is higher than certificate level. The same phenomenon is observed in Table 6.5 where the effect of level of education is the greatest for the Coloured population group. The Black population group seems to have the lowest log-odds for all levels of education. Although the effect of level of education is relatively high in Table 6.5 for the Black population group, the initial value of the estimated odds is just

$$\begin{aligned}
 odds_1 &= 0.4431(2.1993)^{-2} \\
 &= 0.0916.
 \end{aligned}$$

## 6.4 Gender, population group and level of education

Finally, three of the explanatory variables can be combined to fit the logit model. The logit model can be written as

$$\mathbf{A} \ln(\mathbf{f}) = \mathbf{Z}\boldsymbol{\lambda} \quad (6.6)$$

with  $\mathbf{A} = (\mathbf{I}_{40} : -\mathbf{I}_{40})$  and  $\mathbf{Z} = (\mathbf{1} : \mathbf{Z}_G : \mathbf{Z}_P : \mathbf{Z}_E : \mathbf{Z}_{GP} : \mathbf{Z}_{PE} : \mathbf{Z}_{GE} : \mathbf{Z}_{GPE})$ . The design submatrices are set up in a similar manner as in the previous sections using Kronecker and direct products as given in Table 6.6.

Direct Effects	First order interactions	Second order interactions
$\mathbf{Z}_G = \mathbf{D}_G \otimes \mathbf{1}_2 \otimes \mathbf{1}_5$	$\mathbf{Z}_{GP} = \mathbf{Z}_G \odot \mathbf{Z}_P$	$\mathbf{Z}_{GPE} = \mathbf{Z}_{GP} \odot \mathbf{Z}_E$
$\mathbf{Z}_P = \mathbf{1}_2 \otimes \mathbf{D}_P \otimes \mathbf{1}_5$	$\mathbf{Z}_{PE} = \mathbf{Z}_P \odot \mathbf{Z}_E$	
$\mathbf{Z}_E = \mathbf{1}_2 \otimes \mathbf{1}_4 \otimes \mathbf{D}_E$	$\mathbf{Z}_{GE} = \mathbf{Z}_G \odot \mathbf{Z}_E$	

Table 6.6: Submatrices for saturated model

The indices can be attained using the saturated model and interpreted to see the effect of the explanatory variables. Instead of doing so, a linear model will be fitted to the log-odds.

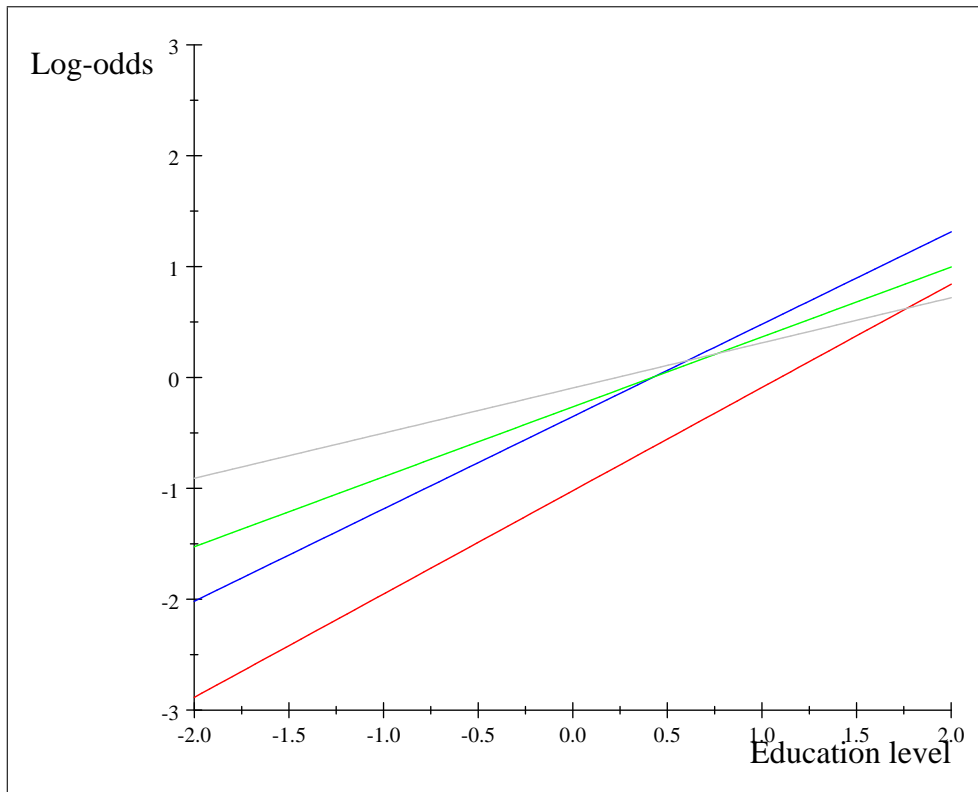
### 6.4.1 Linear model

A model of the form

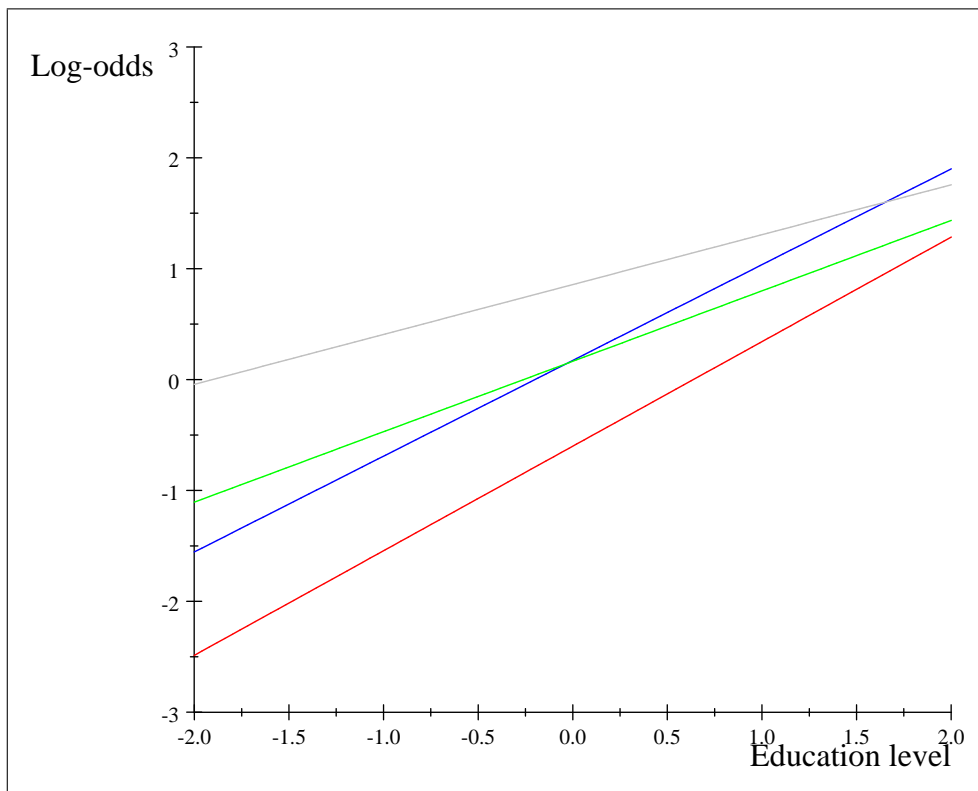
$$\mathbf{A} \ln(\mathbf{F}) = \mathbf{Y}\boldsymbol{\gamma} \quad (6.7)$$

can be estimated to fit a linear model to the log-odds for different genders and population groups with level of education acting as an ordinal variable. In this case, one will have the design matrix  $\mathbf{Y} = (\mathbf{1}_{40} : \mathbf{Z}_G : \mathbf{Z}_P : \mathbf{Y}_E : \mathbf{Z}_{GP} : \mathbf{Y}_{GE} : \mathbf{Y}_{PE} : \mathbf{Y}_{GPE})$  with the submatrices of  $\mathbf{Y}$  being identical to the submatrices defined in Section 5.4.2.

Using the vector of constraints  $\mathbf{g}(\mathbf{F}) = \mathbf{Q}_Y \mathbf{A} \ln(\mathbf{F}) = \mathbf{0}$  with matrix of partial derivatives  $\mathbf{G}_F = \mathbf{Q}_Y \mathbf{A} \mathbf{D}_F^{-1}$ , the ML estimate for  $\mathbf{F}$  can be found from which the ML for  $\boldsymbol{\gamma}$  is attained. The resultant linear functions are given in Figures 6-4 and 6-5 where the red, blue, green and grey line represents the Black, Coloured, Indian and White population groups respectively.



**Figure 6-4: Log-odds for Females: Linear models for different population groups with level of education as ordinal variable**



**Figure 6-5: Log-odds for Males: Linear models for different population groups with level of education as ordinal variable**

For all levels of education, males seem to have a higher log-odds of falling in the high income category.

One should note that the gradients between the genders with population group being constant do seem similar but the intercepts are vastly different. This indicates that even with the same level of education, males are expected to have a higher log-odds of falling in the high income category and the effect of education on the log-odds is relatively constant over different genders.

## 6.5 Summary

Chapter 6 focussed on using the logit model to extrapolate more information from the given data. This was done by transforming the log-odds of falling in the high income category into indices to see what the effects of the explanatory variables are. Level of education was then used as an ordinal variable to fit linear models to the log-odds. Different explanatory variables were also used to gain further insights by fitting separate linear models for different categories of the explanatory variables.

One should note that the results attained in Chapter 6 are vastly different to the results attained in Chapter 5. A clear example of this is when one compares Figure 6-3 to Figure 5-5. Both analyse the effect of education on income for different population groups but the inferences that can be made from the two figures differ. This is caused by the fact that in Chapter 5 the continuous nature of the grouped response variable is taken into account by first estimating the distributions whereas in Chapter 6 the grouped response variable is simply coded as a binary response variable an essential information may be lost.

# Chapter 7

## Conclusion

Data with a continuous underlying nature that is recorded in grouped format is often encountered in data analysis. Although it is a simple and popular method used for data collection, it may limit the researcher from using techniques such as multivariate regression to analyse the effect of a set of explanatory variables on a grouped response variable. One may be tempted to ignore the underlying continuous nature of a grouped response variable but in doing so important information is lost.

This mini-dissertation concentrates on a technique developed by Matthews & Crowther (1995) [10] and Crafford & Crowther (2009) [6] that allows one to find the Maximum Likelihood (ML) estimate of the expected value  $\boldsymbol{\pi}$  of a vector  $\mathbf{p}$  belonging to the exponential family, under a vector of constraints  $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$ . This technique is used in conjunction with the 10% sample of the Census 2011 to analyse the effect of a set of explanatory variables on the grouped response variable, income.

Chapter 1 initiates the study with an exploratory analysis of the data with Chapter 2 following by fitting a log-logistic distribution to the frequency distribution of income. If  $\mathbf{p}$  is the vector of cumulative relative frequencies of income, then the latter is achieved by defining the vector of constraints  $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{g}_{\log}(\boldsymbol{\pi}) = \mathbf{0}$  in such a way that the expected value  $\boldsymbol{\pi}$  of  $\mathbf{p}$  under the vector of constraints  $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$  will follow a cumulative log-logistic curve at the grouped response variable's upper class boundaries  $\mathbf{x}$ . When the ML estimate of  $\boldsymbol{\pi}$  is attained, the parameters of the log-logistic distribution can be found. Chapter 3 repeats the procedure by fitting Normal and Exponential distributions to the frequency distribution of income.

Chapter 4 and Chapter 5 introduces explanatory variables by firstly cross-tabulating the explanatory variables with the frequency distribution of the grouped response variable, leading to  $T$  so-called cells in a multifactor design. By defining  $\mathbf{p}$  as the concatenated vector of cumulative relative frequencies of the  $T$  cells, the vector of constraints defined by  $\mathbf{g}(\boldsymbol{\pi}) = \begin{pmatrix} \mathbf{g}_{\log}(\boldsymbol{\pi}) \\ \mathbf{g}_{\text{mod}}(\boldsymbol{\pi}) \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}$  is used where  $\mathbf{g}_{\log}(\boldsymbol{\pi}) = \mathbf{0}$  forces the estimates of  $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \dots, \boldsymbol{\pi}_T$  to follow cumulative log-logistic curves at the grouped response variable's upper class boundaries  $\mathbf{x}$ , and  $\mathbf{g}_{\text{mod}}(\boldsymbol{\pi}) = \mathbf{0}$  acts as a model for the medians to examine the effect of the explanatory variables. This will lead to the ML estimate of  $\boldsymbol{\pi}$  under the vector of constraints  $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$  such that

- $\hat{\boldsymbol{\pi}}_1, \hat{\boldsymbol{\pi}}_2, \dots, \hat{\boldsymbol{\pi}}_T$  follow cumulative log-logistic curves at the upper class boundaries  $\mathbf{x}$ , and
- The  $T$  medians of the fitted distributions will adhere to a specified model.

In Chapter 6 the logit model is combined with the iterative procedure where the grouped response variable is transformed into a binary response variable and the expected odds of falling in the high income category is modeled under a defined set of constraints.

Future considerations might include developing an algorithm where the vector of upper class boundaries is allowed to vary for each cell. Different types of models for the parameters of the fitted distributions should also be studied and confidence bands to the estimated linear models will also add to the visual representation of the models.

In conclusion, the technique developed by Matthews & Crowther (1995) [10] and Crafford (2009) [6] offers a different methodology to the analysis of grouped data where a grouped response variable is considered and provides the fundamentals from which new research areas can stem off of.

# Bibliography

- [1] Aitchison, J. & Silvey, S. D. Maximum likelihood estimation of parameters subject to restraints, *The Annals of Mathematical Statistics*, 1958, **29**, 813-828
- [2] Benabou, R. Unequal Societies: Income Distribution and the Social Contract, *The American Economic Review*, 2000, **90**, 96-129
- [3] Cirera, X. & Masset, E. Income distribution trends and future food demand, *Philosophical Transactions of the Royal Society B*, 2010, **365**, 2821-2834
- [4] Cowell, F. A. & Flachaire, E. Income distribution and inequality measurement: The problem of extreme values, *Journal of Econometrics*, 2007, **141**, 1044-1072
- [5] Crafford, G. Statistical analysis of grouped data. Ph.D. Thesis, University of Pretoria, 2006
- [6] Crafford, G. & Crowther, N.A.S. Linear models for grouped data, *South African Statistics Journal*, 2009, **43**, 151-176
- [7] Dragulescu, A. & Yakovenko, V. M. Evidence for the exponential distribution of income in the USA, *Physics of Condensed Matter*, 2001, **20**, 585-589
- [8] Haber, M. & Brown, M. Maximum Likelihood Methods for Log-Linear Models When Expected Frequencies are Subject to Linear Constraints, *Journal of the American Statistical Association*, 1986, **81**, 477-482
- [9] Matthews, G. B. Maximum likelihood estimation when modelling in terms of constraints. Ph.D. Thesis, University of Pretoria, 1995
- [10] Matthews, G. & Crowther, N.A.S. A maximum likelihood estimation procedure when modelling in terms of constraints, 1995, *South African Statistics Journal*, **29**, 29-51
- [11] Statistics South Africa. 2012. Census 2011 Ten Percent Sample Statistical Release P0301.4.





# Chapter 8

## Appendix

### 8.1 Appendix A: Fitting a log-logistic distribution

```
libname census11 "C:\Users\Jurgens\Desktop\1. Dissertation New\Dr Crafford 19 July 2016\  
DATA";
```

```
options dquote;
```

```
proc freq data=Census11.data_analyse;  
tables income / out=f_edit;  
*where 12<=p17<=20;  
run;
```

```
proc iml worksize= 60;  
*****;  
* Exponential ='E' *;  
* Normal      ='N' *;  
* Log-logistic='L' *;  
*****;
```

```
filename sw "C:\Jurgens\tex\pgm1.tex";  
file sw;  
put '%TCIDATA{LaTeXparent=0,0,master.tex}';
```

```
%macro mac(distr,d);  
print "The &distr distribution";  
distr="&d";  
use f_edit; read all var{count} into f;  
use f_edit; read all var{income} into xu;
```

```
n=f[+];  
k=nrow(f); k1=k-1;  
x=xu[1:k1]; print f xu x;  
C=J(k1,1,1)@cusum(J(1,k1,1))<=J(1,k1,1)@cusum(J(k1,1,1));  
CI=inv(C);  
v1=J(k1,1,1);  
p=C*f[1:k1]/n;
```

```
start X;  
if distr='E' then XD=-x;  
if distr='N' then XD=x||(-v1);  
if distr='L' then XD=log(x)||v1;  
finish;
```

```
start h;  
if distr='E' then h=log(v1-p);  
if distr='N' then h=probit(p);
```



```
if distr='L' then h=log(p/(v1-p));
finish;

start D(Dp,p) global(distr,v1);
if distr='E' then Dp=inv(diag(p-v1));
if distr='N' then Dp=inv(diag(pdf('normal',probit(p))));
if distr='L' then Dp=inv(diag(p))+inv(diag(v1-p));
finish;

start beta;
if distr='E' then beta=1/alpha;
if distr='N' then do;
beta[1]=alpha[2]/alpha[1];
beta[2]=1/alpha[1];
end;
if distr='L' then beta=alpha;
finish;

start B;
if distr='E' then B=-1/(alpha**2);
if distr='N' then do;
B[1,1]=-alpha[2]/((alpha[1])**2);
B[1,2]=1/(alpha[1]);
B[2,1]=-1/((alpha[1])**2);
end;
if distr='L' then B=I(nrow(alpha));
finish;

start wald;
Wald=g'*ginv(Gp*V*Gp')*g;
GpV=Gp*V;
df=trace(GpV*ginv(GpV'*GpV)*GpV');
finish;

start mu;
if distr='E' then mu=beta;
if distr='N' then mu=beta[1];
if distr='L' then mu=exp(-beta[2]/beta[1])*gamma(1+1/beta[1])*gamma(1-1/beta[1]);
finish;

start sigma;
if distr='E' then sigma=beta;
if distr='N' then sigma=beta[2];
*if distr='L' then sigma=sqrt(exp(-2*beta[2]/beta[1])*
(gamma(1+2/beta[1])*gamma(1-2/beta[1]) - (gamma(1+1/beta[1])*gamma(1-1/beta[1]))**2));
if distr='L' then sigma=.;
finish;

run X;
Q=I(k1)-XD*inv(XD'*XD)*XD';

i=0; p0=p; diff1=1;
do while (diff1 > 1e-9);
i=i+1; pi=p; p=p0;
run D(Dpi,pi);
Gpi=Q*Dpi;
V=(C*diag(CI*pi)*C' - pi*pi')/n;
j=0; diff=1;
do while (diff > 1e-9);
j=j+1; pv=p;
run D(Dp,p);
run h;
Gp=Q*Dp;
g=Q*h;
print i j pi p g;
```



```

p=p-(Gpi*V)'*ginv(Gp*V*Gpi')*g;
diff=sqrt((p-pv)'*(p-pv));
if i=1 & j=1 then run wald;
end;
diff1=sqrt((p-pi)'*(p-pi));
end;
Cov_pi=V-(Gpi*V)'*ginv(Gpi*V*Gpi')*(Gpi*V);

alpha=inv(XD'*XD)*XD'*h;
Cov_alpha=(inv(XD'*XD)*XD'*Dpi)*Cov_pi*(inv(XD'*XD)*XD'*Dpi)';
SE_alpha=sqrt(diag(Cov_alpha)*J(nrow(alpha),1,1));
print alpha Cov_alpha SE_alpha;

beta=J(nrow(alpha),1,0); run beta;
B=J(nrow(alpha),nrow(alpha),0); run B;
Cov_beta=B*Cov_alpha*B';
SE_beta=sqrt(diag(Cov_beta)*J(nrow(beta),1,1));
print beta Cov_beta SE_beta;

run mu; run sigma;
print mu sigma;

e=(CI*pi*n)/(n-(CI*pi*n)[+]);
Pearson=((f-e)##2)/e[+];
P_pvalue=1-probchi(Pearson,df);
W_pvalue=1-probchi(Wald,df);
discr=wald/n;
print Pearson P_pvalue Wald W_pvalue df discr;

print f e;
p_0_hat = e*(1/n);
print p_0_hat pi;
toets=Q*h; print toets;

class=0//x//30;
width=class[2:k+1]-class[1:k];
fx=e/n/width;
put "The &distr distribution";
put ' ';
put 'Pearson: ' Pearson ' p-value: ' P_pvalue ' Wald: ' Wald ' p-value: ' W_pvalue
' df: ' df ' Discr ' discr;
put '\FRAME{dtbpF}{10cm}{6cm}{0pt}{-}{Plot}';
put '\{special{language "Scientific Word";type "MAPLEPLOT"}';
put 'width 15cm;height 8cm;depth 0pt;';
put 'display "USEDEF";plot_snapshots TRUE;mustRecompute FALSE;lastEngine "MuPAD";';
put 'xviewmin "0";xviewmax "'(max(class))"' ;yviewmin "0";yviewmax "'(max(fx))"' ;viewset
"XY";';
put 'plotticks 1;num-x-ticks 7;num-y-ticks 6;';
put 'plottype 4;labeloverrides 3;numpoints 100;plotstyle "patch";';
put 'axesstyle "normal";xis \TEXUX{x};yis \TEXUX{y};var1name \TEXUX{$x$};var2name
\TEXUX{$y$};';
put 'function \TEXUX{$\left(' ;
put '\MATRIX{2,'(k*4)'}{c}\VR{,,c,,}{,c,,}{,,,,}\HR{,,,,,,,,,,,,,,,,,,,,,,,,,,,,,}';
* \MATRIX 20=k*4 en \HR 20 kommas;
do i=1 to k;
put '\CELL{'(class[i])'}\CELL{0}';
put '\CELL{'(class[i])'}\CELL{'(fx[i])'}';
put '\CELL{'(class[i+1])'}\CELL{'(fx[i])'}';
put '\CELL{'(class[i+1])'}\CELL{0}';
end;

put '\right) $);';
put 'linecolor "blue";linestyle 1;pointstyle "point";linethickness 1;lineAttributes
"Solid";';
put 'curveColor "[flat::RGB:0x00800000]";curveStyle "Line";discont FALSE;';

```



```
if distr='E' then put "function\TEXUX{\$E(x,"(mu)")\$}";
else put "function\TEXUX{\$&d.(x,"(beta[2])","(beta[1])")\$}";
/*put 'function\TEXUX{\$l(x,'(beta[1]'),'(beta[2])')\$}'; */
/*put 'function\TEXUX{\$p(x,'(beta[1]'),'(beta[2])')\$}'; */
put 'linecolor "red";linestyle 1;pointstyle "point";linethickness 1;lineAttributes
"Solid";';
put 'varirange "0,50";rangeset"X";num-x-gridlines 700;';
put 'curveColor "[flat::RGB:0x000000ff]";curveStyle "Line";discont False;';
put '}}';
%mend mac;
%mac (Exponential,E);
*%mac (Normal,N);
*%mac (Log-logistic,L);
closefile sw;

quit;
```

## 8.2 Appendix B: Single factor model

```
libname census11 "C:\Users\Jurgens\Desktop\1. Dissertation New\Dr Crafford 19 July 2016
\DATA";
```

```
options pageno=1 nocenter pagesize=500 linesize=200;
```

```
%macro mac(response,factor1);
proc freq data=Census11.Data_analyse;
tables &factor1/out=factor1;
tables &response/out=response;
tables &factor1*&response/out=f list;
run;
```

```
proc transpose data=f out=freq prefix=c;
by &factor1;
var count;
run;
```

```
proc iml workspace=200 symsize=2000;
use freq; read all var _num_ into freq;
use response; read all var{income} into response;
use factor1; read all var _char_ into factor1;

use f;read all var _char_ into names;names=rowcat(names);
effects="Intercept"//factor1;
class="R1600"//"R3200"//"R6400"//"R12800"//"R30000";
n=freq[+];
nfac1=nrow(factor1);
nt=nrow(freq);
print response;
k=nrow(response); k1=k-1; response = response[1:k1]//30;
x=response[1:k1];
nn=freq[,+];
print nn;
f=colvec(freq[,1:k1]); f=f<>0.0001;
C=J(k1,1,1)@cusum(J(1,k1,1))<=J(1,k1,1)@cusum(J(k1,1,1));
CI=inv(C);
v1=J(k1,1,1);
po=inv(diag(nn)@I(k1))*f;
p=(I(nt)@C)*po;
```

```
*** Start: Design matrix for log-logistic distributions ***;
XD=log(x)||v1;
XXX=inv(XD'*XD)*XD'; XXX1=XXX[1,]; XXX2=XXX[2,];
Px=XD*inv(XD'*XD)*XD';
```



```
Qx=I(k1)-Px;
*** Finish: Design matrix for log-logistic distributions ***;

*** Start: Single-factor model ***;
Y1=designf(cusum(J(nfac1,1,1))); *<=== Factor1: dummy;

*Y1={-2,-1,0,1,2};*<=== Factor1: ordinal;
*Y1={21.5, 28, 35.5, 43, 48, 53, 60.5};*<=== Factor1: linear;

*Y1=Y1||Y1#2;

YD=J(nt,1,1)||Y1; *<=== Only main effects;
Py=YD*inv(YD'*YD)*YD';
Qy=I(nt)-Py;
*** Finish: Single-factor model ***;

start GGG(p,g,GG) global(nt,v1,Qx,XXX,XXX1,XXX2,Qy,h,D,kappa,theta,nu,A,Y12);
h=log(p/((J(nt,1,1)*v1)-p));
D=inv(diag(p))+inv(diag((J(nt,1,1)*v1)-p));

glog=(I(nt)*Qx)*h;
GGlog=(I(nt)*Qx)*D;

kappa=(I(nt)*XXX1)*h;
theta=(I(nt)*XXX2)*h;
nu=exp(-theta/kappa);
A1=nu*(theta/(kappa*kappa));
A2=nu*(-1/kappa);
A=diag(A1)*I(1) + diag(A2)*I(1);

greg=Qy*nu;
GGreg=Qy*A*(I(nt)*XXX)*D;

g=glog; *<=== Model 1;
GG=GGlog; *<=== Model 1;
g=glog//greg; *<=== Model 2-4;
GG=GGlog//GGreg; *<=== Model 2-4;
finish;
print p;
i=0; p0=p; diff1=1;
do while (diff1 > 1e-9);
i=i+1; pi=p; p=p0;
pio=(I(nt)*CI)*pi;
Vo=inv(diag(nn)*I(k1))*(diag(pio)- (diag(pio))*(I(nt)*(v1*v1'))*(diag(pio))');
V=(I(nt)*C)*Vo*(I(nt)*C)';
run GGG(pi,gpi,GGpi);
j=0; diff=1;
do while (diff > 1e-9);
j=j+1; pv=p;
run GGG(p,gp,GGp);
*print i j p pi gp;
p=p-(GGpi*V)'*ginv(GGp*V*GGpi')*gp;
diff=sqrt((p-pv)'*(p-pv));
if i=1 & j=1 then do;
Wald=gp'*ginv(GGp*V*GGp')*gp;
GpV=GGp*V;
df=trace(GpV*ginv(GpV'*GpV)*GpV');
end;
end;
diff1=sqrt((p-pi)'*(p-pi));
end;

print pi;
print XXX XXX1 XXX2;
Cov_pi=V-(GGpi*V)'*ginv(GGpi*V*GGpi')*(GGpi*V);
```



```
Cov_alpha=((I(nt)@XXX)*D)*Cov_pi*((I(nt)@XXX)*D)';

*mu=exp(-theta/kappa)#gamma(J(nt,1,1)+1/kappa)#gamma(J(nt,1,1)-1/kappa);
*sigma=sqrt(exp(-2*theta/kappa)
#(gamma(J(nt,1,1)+2/kappa)#gamma(J(nt,1,1)-2/kappa)
-(gamma(J(nt,1,1)+1/kappa)#gamma(J(nt,1,1)-1/kappa))##2));
*mum=shape(mu,nfac1);
*sigmam=shape(sigma,nfac1);

Cov_nu=A*Cov_alpha*A';
SE_nu=sqrt(diag(Cov_nu)*J(nrow(nu),1,1));

YYY=inv(YD'*YD)*YD';
gamma=YYY*nu;
Cov_gamma=YYY*Cov_nu*YYY';
SE_gamma=sqrt(diag(Cov_gamma)*J(nrow(gamma),1,1));
t_gamma = gamma / SE_gamma;
/*
D1=designf(cusum(J(nfac2,1,1)));
DDD=block(1,1,D1,D1);
delta=DDD*gamma;
Cov_delta=DDD*Cov_gamma*DDD';
SE_delta=sqrt(diag(Cov_delta)*J(nrow(delta),1,1));
*print delta SE_delta;
*/
Z1=designf(cusum(J(nfac1,1,1)));
LD=J(nt,1,1)||Z1;
LLL=inv(LD'*LD)*LD';
lambda=LLL*nu;
lambda=choose(abs(lambda)<1e-9,0,lambda);
Cov_lambda=LLL*Cov_nu*LLL';
Cov_lambda=choose(abs(Cov_lambda)<1e-9,0,Cov_lambda);

S1=designf(cusum(J(nfac1,1,1)));
S=block(1,S1);
print S;
tau=S*lambda;
Cov_tau=S*Cov_lambda*S';
SE_tau=sqrt(diag(Cov_tau)*J(nrow(tau),1,1));
*print tau[rowname=effects] SE_tau;

count=cusum(1//nfac1);
tau0=tau[1:1]; SE_tau0=SE_tau[1:1];
tau1=tau[count[1]+1:count[2]]; SE_tau1=SE_tau[count[1]+1:count[2]];

piom=(shape(pio,nt));
exp1=piom#(repeat(nn,1,k1));
exp2=nn-exp1[,+];
exp=exp1||exp2;

Pearson=((freq-exp)##2)/exp[+];
P_pvalue=1-probchi(Pearson,df);
W_pvalue=1-probchi(Wald,df);
discr=wald/n;

print "Grouped response variable: &Response";
print "Single factor model: &Factor1";
print "Number of observations: " n ;
print "Number of cells: " nt;
print "Goodness of fit:" Pearson P_pvalue Wald W_pvalue df discr ;
print ' ';
print "Observed frequencies:", freq[rowname=factor1 colname=class] nn;
print "Expected frequencies:", exp[rowname=factor1 colname=class] nn;
print ' ';
print "Fitted log-logistic distributions and medians:";
```



```
print nn[rowname=factor1] nu kappa theta ;
print ' ';
print "The single-factor model:";
print YD gamma cov_gamma SE_gamma t_gamma;
print ' ';

print "Effects for the single-factor model: Intercept";
print tau0;
print ' ';
print "Effects for the single-factor model: Main effects";
print "&Factor1", tau1[rowname=factor1];

*** Start: Graph ***;
*response[k]=20;
xl=0//response[1:k1];
xu=response;
width=xu-xl;

start graph;
put '\FRAME{dtbpF}{4cm}{2cm}{0.5pt}{Plot}';
put '{\special{language "Scientific Word";type "MAPLEPLOT"}';
put 'width 6cm;height 3cm;depth 0.5pt;';
put 'display "USEDEF";plot_snapshots True;mustRecompute FALSE;lastEngine "MuPAD";';
put 'xmin "0";xmax "30";xviewmin "0";xviewmax "30";yviewmin "0";yviewmax ".15";';
put 'viewset"XY";rangeset"X";plottype 4;labeloverrides 3;numpoints 100;plotstyle
"patch";';
put 'plotticks 1;num-x-ticks 7;num-y-ticks 6;';
put 'axesstyle "normal";xis \TEXUX{x};yis \TEXUX{y};var1name \TEXUX{$x$};var2name
\TEXUX{$y$};';
put 'function \TEXUX{$\left(';
put '\MATRIX{2,20}{c}{\VR{,,c,,}{,c,,}{,,,,}\HR{,,,,,,,,,,,,,,,,,,,,,}}';
do ii=1 to k;
put '\CELL{'(xl[ii])'}\CELL{0}';
put '\CELL{'(xl[ii])'}\CELL{'(pt[ii])'}';
put '\CELL{'(xu[ii])'}\CELL{'(pt[ii])'}';
put '\CELL{'(xu[ii])'}\CELL{0}';
end;
put '\right) $}';;
put 'linecolor "blue";linestyle 1;pointstyle "point";linethickness 1;lineAttributes
"Solid";';
put 'curveColor "[flat::RGB:0x00800000]";curveStyle "Line";';
put 'discont FALSE;';
/*
put 'function\TEXUX{$l(x,'(kappa[i,])','(theta[i,])')$}';
*/
put 'function\TEXUX{$(x,\frac{e^{'(theta[i,])'}}\times '(kappa[i,])'\times x^{
{'(kappa[i,])'-1}}{\left( 1+e^{'(theta[i,])'}x^{'(kappa[i,])'}\right) ^{2}})}$}';;

put 'linecolor "blue";linestyle 1;pointstyle "point";linethickness 1;lineAttributes
"Solid";';
put 'discont FALSE;';
put 'varirange "0,30";num-x-gridlines 100;';
put 'curveColor "[flat::RGB:0x000000ff]";curveStyle "Line";';
put '}}';;
finish;

filename table "C:\Jurgens\tex\factor1.tex";
file table;
put '%TCIDATA{LaTeXparent=0,0,master.tex}';
put '\renewcommand{\arraystretch}{0.8}';
put '\begin{center}';
put '\begin{tabular}{|c|c|c|c|c|c|}';
put '\hline\hline';
put "\textbf{&Factor1} & \textbf{&Response} & \textbf{$n$}";
put ' & $ \begin{array}{c} \widehat{\nu} \\ \backslash \backslash (\widehat{\sigma}_{\widehat{\nu}}) \end{array}
```



```

\end{array} $' ;
put ' & $ \left( \begin{array}{c} \widehat{\kappa} \\ \widehat{\theta} \end{array} \right) $' ;
put ' & $ \begin{array}{c} \widehat{\tau}_1 \\ (\widehat{\sigma}_{\widehat{\tau}_1}) \end{array} \end{array} $ \\' ;
put '\hline\hline';
do i=1 to nfac1;
pt=(freq[i,]/width)/mn[i];
pt=choose(pt<1e-6,0,pt);
put '\textbf{'(factor1[i]) '}' &';
run graph;
put '& ' (nn[i]);
put ' & $ \begin{array}{c} (\nu[i])5.3 \\ (' (SE_\nu[i])5.3 ') \end{array} $' ;
*put ' & $ \begin{array}{c} (\kappa[i,])5.3 \\ (' (\theta[i,])5.3 ') \end{array} $' ;
put ' & $ \left( \begin{array}{c} (\kappa[i,])5.3 \\ (' (\theta[i,])5.3 ') \end{array} \right) $' ;
put ' & $ \begin{array}{c} (\tau_1[i])5.3 \\ (' (SE_\tau_1[i])5.3 ') \end{array} $ \\' ;
put '\hline';
end;
put ' $ \begin{array}{c} \widehat{\tau}_0 \\ (\widehat{\sigma}_{\widehat{\tau}_0}) \end{array} \end{array} $';
put '&& n '&&& $ \begin{array}{c} (\tau_0)5.3 \\ (' (SE_\tau_0)5.3 ') \end{array} $ \\' ;
put '\hline\hline';
put '\end{tabular}';
put '\end{center}';
closefile table;
*** Finish: Graph ***;
%mend mac;
%mac(Income,Race);
*%mac(Income,Gender);
*%mac(Income,educationgrp);
*%mac(Income,agegrp);

```

## 8.3 Appendix C: Two factor model

```

libname census11 "C:\Users\Jurgens\Desktop\1. Dissertation New\Dr Crafford 19 July 2016
\DATA";

```

```

%macro mac(response,factor1,factor2);
proc freq data=Census11.Data_analyse;
tables &factor1/out=factor1;
tables &factor2/out=factor2;
tables &response/out=response;
tables &factor1*&factor2*&response/out=f list;
run;

proc transpose data=f out=freq prefix=c;
by &factor1 &factor2;
var count;
run;

proc iml worksize=200 symsize=2000;
use freq; read all var _num_ into freq;
use response; read all var {income} into response;
use factor1; read all var _char_ into factor1;
use factor2;read all var _char_ into factor2;
print factor1;
print factor2;
use f;read all var _char_ into names;names=rowcat(names);
effects="Intercept"//factor1//factor2;
class="R1600"//"R3200"//"R6400"//"R12800"//"R30000";

```





```
n=freq[+];
nfac1=nrow(factor1);
nfac2=nrow(factor2);
nt=nrow(freq);
print response;
k=nrow(response); k1=k-1; response = response[1:k1]/30;
x=response[1:k1];
nn=freq[,+];
nmm=shape(nn,nfac1);
print nmm;
f=colvec(freq[,1:k1]); f=f<>0.0001;
C=J(k1,1,1)@cusum(J(1,k1,1))<=J(1,k1,1)@cusum(J(k1,1,1));
CI=inv(C);
v1=J(k1,1,1);
po=inv(diag(nn)@I(k1))*f;
p=(I(nt)@C)*po;

*** Start: Design matrix for log-logistic distributions ***;
XD=log(x)||v1;
XXX=inv(XD'*XD)*XD'; XXX1=XXX[1,]; XXX2=XXX[2,];
Px=XD*inv(XD'*XD)*XD';
Qx=I(k1)-Px;
*** Finish: Design matrix for log-logistic distributions ***;

*** Start: Single-factor model ***;
/*
Dep1=designf(J(nfac1,1,1)@cusum(J(nfac2,1,1)))[1:nfac2,];
Dep2=Dep1;
if nfac1>2 then do;
do jj=1 to nfac1-2;
Dep2=block(Dep2,Dep1);
end;
end;
Dep=Dep2//(-repeat(Dep1,1,nfac1-1));
*/
Y1=designf(cusum(J(nfac1,1,1))@J(nfac2,1,1));*<=== dummy1;
*Y2=designf(J(nfac1,1,1)@cusum(J(nfac2,1,1))); *<=== dummy2;
*Y1={24.5,34.5,44.5,54.5}@J(nfac2,1,1); *<=== Factor1: linear;
*Y2=J(nfac1,1,1)@{24.5,34.5,44.5,54.5}; *<=== Factor2: linear;
*Y1={-2,-1,0,1,2}@J(nfac2,1,1);*<=== Factor1: ordinal;
Y2=J(nfac1,1,1)@{-2,-1,0,1,2};*<=== Factor2: ordinal;

*Y2=Y2||Y2##2;

Y12 = hdir(Y1,Y2);

*Y1 = Y1||Y2;*<=== independent;
Y1 = Y1||Y2||Y12; *<=== dependent;

YD=J(nt,1,1)||Y1;
Py=YD*inv(YD'*YD)*YD';
Qy=I(nt)-Py;
*** Finish: two-factor model ***;
print YD;

start GGG(p,g,GG) global(nt,v1,Qx,XXX,XXX1,XXX2,Qy,h,D,kappa,theta,nu,A,Y12);
h=log(p/((J(nt,1,1)@v1)-p));
D=inv(diag(p))+inv(diag((J(nt,1,1)@v1)-p));

glog=(I(nt)@Qx)*h;
GGlog=(I(nt)@Qx)*D;

kappa=(I(nt)@XXX1)*h;
theta=(I(nt)@XXX2)*h;
```



```

nu=exp(-theta/kappa);
A1=nu#(theta/(kappa#kappa));
A2=nu#(-1/kappa);
A=diag(A1)@{1 0} + diag(A2)@{0 1};

greg=Qy*nu;
GGreg=Qy*A*(I(nt)@XXX)*D;

g=glog;          *<=== Model 1;
GG=GGlog;       *<=== Model 1;
g=glog//greg;   *<=== Model 2-4;
GG=GGlog//GGreg; *<=== Model 2-4;
finish;

i=0; p0=p; diff1=1;
p0m=shape(p0,nfac1*nfac2);
print p0m[format=5.3];
do while (diff1 > 1e-9);
i=i+1; pi=p; p=p0;
    pio=(I(nt)@CI)*pi;
Vo=inv(diag(nn)@I(k1))*(diag(pio)- (diag(pio))*(I(nt)@(v1*v1'))*(diag(pio))');
V=(I(nt)@C)*Vo*(I(nt)@C)';
run GGG(pi,gpi,GGpi);
j=0; diff=1;
do while (diff > 1e-9);
j=j+1; pv=p;
run GGG(p,gp,GGp);
*print i j p pi gp;
p=p-(GGpi*V)'*ginv(GGp*V*GGpi')*gp;
diff=sqrt((p-pv)'*(p-pv));
if i=1 & j=1 then do;
Wald=gp'*ginv(GGp*V*GGp')*gp;
GpV=GGp*V;
df=trace(GpV*ginv(GpV'*GpV)*GpV');
end;
end;
diff1=sqrt((p-pi)'*(p-pi));
end;

thetam=shape(theta,nfac1);
kappam=shape(kappa,nfac1);
pm=shape(p,nfac1*nfac2);
print pm[format=5.3];

Cov_pi=V-(GGpi*V)'*ginv(GGpi*V*GGpi')*(GGpi*V);
Cov_alpha=((I(nt)@XXX)*D)*Cov_pi*((I(nt)@XXX)*D)';

*mu=exp(-theta/kappa)#gamma(J(nt,1,1)+1/kappa)#gamma(J(nt,1,1)-1/kappa);
*sigma=sqrt(exp(-2*theta/kappa)
#(gamma(J(nt,1,1)+2/kappa)#gamma(J(nt,1,1)-2/kappa)
-(gamma(J(nt,1,1)+1/kappa)#gamma(J(nt,1,1)-1/kappa))##2));
*mum=shape(mu,nfac1);
*sigmam=shape(sigma,nfac1);

Cov_nu=A*Cov_alpha*A';
SE_nu=sqrt(diag(Cov_nu)*J(nrow(nu),1,1));
num=shape(nu,nfac1); SE_num=shape(SE_nu,nfac1);

YYY=inv(YD'*YD)*YD';
gamma=YYY*nu;
Cov_gamma=YYY*Cov_nu*YYY';
SE_gamma=sqrt(diag(Cov_gamma)*J(nrow(gamma),1,1));
t_gamma = gamma / SE_gamma;
/*

```



```
D1=designf(cusum(J(nfac2,1,1)));
DDD=block(1,1,D1,D1);
delta=DDD*gamma;
Cov_delta=DDD*Cov_gamma*DDD';
SE_delta=sqrt(diag(Cov_delta)*J(nrow(delta),1,1));
*print delta SE_delta;
*/
Z1=designf(cusum(J(nfac1,1,1))@J(nfac2,1,1))||designf(J(nfac1,1,1)@cusum(J(nfac2,1,1)))
; *<=== independent;
Dep = hdir(designf(cusum(J(nfac1,1,1))@J(nfac2,1,1)),designf(J(nfac1,1,1)
@cusum(J(nfac2,1,1)))));
Z1=Z1||Dep;*<=== dependent;
LD=J(nt,1,1)||Z1;
LLL=inv(LD'*LD)*LD';
lambda=LLL*nu;
lambda=choose(abs(lambda)<1e-9,0,lambda);
Cov_lambda=LLL*Cov_nu*LLL';
Cov_lambda=choose(abs(Cov_lambda)<1e-9,0,Cov_lambda);

*independent;
S=(1//J(nfac1+nfac2,1,0))||(J(1,nfac1-1,0)//designf(cusum(J(nfac1,1,1)))
//J(nfac2,nfac1-1,0))||(J(nfac1+1,nfac2-1,0)//designf(cusum(J(nfac2,1,1)))));
*dependent;
S=block(S,Dep);

tau=S*lambda; *<=== dependent;
*tau=tau//J(nfac1*nfac2,1,0);*<=== independent;
print tau;
Cov_tau=S*Cov_lambda*S';
SE_tau=sqrt(vecdiag(Cov_tau));
*print tau[rowname=effects] SE_tau;

count=cusum(1//nfac1//nfac2//(nfac1*nfac2));
tau0=tau[1:1]; SE_tau0=SE_tau[1:1];
tau1=tau[count[1]+1:count[2]]; SE_tau1=SE_tau[count[1]+1:count[2]];
tau2=tau[count[2]+1:count[3]]; SE_tau2=SE_tau[count[2]+1:count[3]];
tau12=tau[count[3]+1:count[4]]; SE_tau12=SE_tau[count[3]+1:count[4]];
tau12m=shape(tau12,nfac1); SE_tau12m=shape(SE_tau12,nfac1);
piom=(shape(pio,nt));
exp1=piom#(repeat(nn,1,k1));
exp2=nn-exp1[,+];
exp=exp1||exp2;

Pearson=((freq-exp)##2)/exp[+];
P_pvalue=1-probchi(Pearson,df);
W_pvalue=1-probchi(Wald,df);
discr=wald/n;
name1 = colvec(repeat(factor1,1,nfac2));
name2 = repeat(factor2,nfac1,1);
print "Grouped response variable: &Response";
print "Two factor model: &Factor1 &Factor2";
print "Number of observations: " n ;
print "Number of cells: " nt;
print "Goodness of fit:" Pearson P_pvalue Wald W_pvalue df discr ;
print ' ';
print "Observed frequencies:", name1 name2 freq[colname=class] nn;
print "Expected frequencies:", name1 name2 exp nn;
print ' ';
print "Fitted log-logistic distributions and medians:";
print name1 name2 nn nu kappa theta ;
print ' ';
print "The two factor model:";
print YD gamma SE_gamma t_gamma;
print ' ';
```



```

print "Effects for the two factor model: Intercept";
print tau0;
print ' ';
print "Effects for the two factor model: Main effects";
print tau1[rowname=(factor1)];
print tau2[rowname=(factor2)];
print "Effects for the two factor model: Interaction effects";
print name1 name2 tau12;

*** Start: Graph ***;
*response[k]=20;
xl=0//response[1:k1];
xu=response;
width=xu-xl;

start graph;
put '\FRAME{dtbpF}{2.1cm}{2.1cm}{0pt}{}{}{Plot}';
put '\{special{language "Scientific Word";type "MAPLEPLOT"}';
if fig=1 then put 'width 2cm;height 2cm;depth Opt;';
if fig=2 then put 'width 2cm;height 2cm;depth Opt;';
if fig=2 then put 'width 2cm;height 2cm;depth Opt;';
if fig=1 then put 'width 2cm;height 2cm;depth Opt;';
put 'display "USEDEF";plot_snapshots TRUE;mustRecompute FALSE;lastEngine "MuPAD";';
put 'xmin "0";xmax "30";xviewmin "0";xviewmax "30";yviewmin "0";yviewmax "0.25";';
put 'viewset"XY";rangeset"X";plottype 4;labeloverrides 3;numpoints 100;plotstyle
"patch";';
put 'plotticks 1;plottickdisable TRUE;num-x-ticks 7;num-y-ticks 4;';
put 'axesstyle "normal";xis \TEXUX{x};yis \TEXUX{y};var1name \TEXUX{$x$};var2name
\TEXUX{$y$};';
put 'function \TEXUX{\left(';
put '\MATRIX{2,20}{c}\VR{,,c,,}{,c,,}{,,,,}\HR{,,,,,,,,,,,,,,,,,,,,}';
do ii=1 to k;
put '\CELL{'(xl[ii])'}\CELL{0}';
put '\CELL{'(xl[ii])'}\CELL{'(pt[ii])'}';
put '\CELL{'(xu[ii])'}\CELL{'(pt[ii])'}';
put '\CELL{'(xu[ii])'}\CELL{0}';
end;
put '\right) $}';

put 'linecolor "maroon";linestyle 1;pointstyle "point";linethickness 1;lineAttributes
"Solid";';
put 'curveColor "[flat::RGB:0x00800000]";curveStyle "Line";';
put 'discont FALSE;';
/*
put 'function\TEXUX{$l(x,'(kappam[i,j])','(thetam[i,j])')$}';
*/

put 'function \TEXUX{$(x,\frac{e^{-(thetam[i,j])}}\times '(kappam[i,j])'\times x^
{'(kappam[i,j])'-1})}\left( 1+e^{(thetam[i,j])}x^{(kappam[i,j])}\right) ^{-2}}$};';
put 'linecolor "blue";linestyle 1;pointstyle "point";linethickness 1;lineAttributes
"Solid";';
put 'discont FALSE;';
put 'var1range "0,30";num-x-gridlines 100;';
put 'curveColor "[flat::RGB:0x000000ff]";curveStyle "Line";';

put '}}';

finish;

filename table "C:\Jurgens\tex\factor2.tex";
file table;
put '%TCIDATA{LaTeXparent=0,0,master.tex}';

```



```

put '\renewcommand{\arraystretch}{0.8}';
put ' ';
/*
fig=1;
put "\begin{tabular}{|c|}(rowcat(J(1,nfac2,'c')))"|}|"; print (rowcat(J(1,nfac2,'c')));
put "\hline \hline";
put "& \multicolumn{(nfac2)}{|c|}{\textbf{&factor2}} \ \ ";
put "\textbf{&factor1}";
do j=1 to nfac2;
put "& \textbf{" (factor2[j,1]) " } ";
end;
put " \ \ ";
put "\hline \hline ";
do i=1 to nfac1;
put "\textbf{" (factor1[i,1]) " }";
do j=1 to nfac2;
pt=(freq[((i-1)*nfac2 + j),]'/nnm[i,j])/width;
put "&"; run graph;
end;
put " \ \ ";

do j=1 to nfac2;
put "&";
put '\begin{tabular}{cc}';
put '$\kappa=' (kappam[i,j])5.3 '$ & $\widehat{\nu}=' (num[i,j])5.3 '$ \ \ ';
*put '$\theta=' (thetam[i,j])5.3 '$ & $\widehat{s}_{\widehat{\nu}}=' (SE_num[i,j])5.3 '$';
put '$\theta=' (thetam[i,j])5.3 '$ & $\widehat{s}_{\widehat{\nu}}=' (nnm[i,j])5.3 '$';
put '\end{tabular}';
end;
put " \ \ ";
put "\hline";
end;
put "\hline";
put " \end{tabular}";
put ' ';
put '\renewcommand{\arraystretch}{1}';

put '\renewcommand{\arraystretch}{0.8}';
put ' ';
*/

fig=2;
put "\begin{tabular}{|c|}(rowcat(J(1,nfac2,'c')))"|}|";
put "\hline \hline";
put "& \multicolumn{(nfac2)}{|c|}{\textbf{&factor2}} & $\widehat{\tau}^G$ \ \ ";
put "\textbf{&factor1}";
do j=1 to nfac2;
put "& \textbf{" (factor2[j,1]) " } ";
end;
put " & $\widehat{\sigma}_{\widehat{\tau}^G}$ ";
put " \ \ \hline \hline ";
do i=1 to nfac1;
put "\textbf{" (factor1[i,1]) " }";
do j=1 to nfac2;
pt=(freq[((i-1)*nfac2 + j),]'/nnm[i,j])/width;
put "&"; run graph;
end;
put "& \ \ ";
do j=1 to nfac2;
put "&";
put '\begin{tabular}{c}';
*put '$\widehat{\kappa}=' (kappam[i,j])6.3 '$ & $\widehat{\theta}=' (thetam[i,j])
6.3 '$ \ \ ';
put '$\widehat{\nu}=' (num[i,j])6.3 '$ \ \ $\widehat{\tau}^{\{GE\}}=' (tau12m[i,j])
6.3 '$';

```



```

*put '$\widehat{\tau }^{AB}=' (tau12m[i,j])6.3 '$';
*put '$\widehat{\tau }^{AB}=' (tau12m[i,j])6.3 '$';
*put '$n=' (nm[i,j])6.3 '$';
*put '$\nu =' (num[i,j])6.3 '$';

*put '\\ $\widehat{\sigma }_{\widehat{\tau }^{AB}}=' (SE_tau12m[i,j])6.3 '$';
put '\end{tabular}';
end;
put "& ";
put '$\begin{array}{c}';
put (tau1[i])6.3 ;
*put '\\ (SE_tau1[i])6.3 ;
put '\end{array}$ \\' ;
put "\hline";
end;
put '$\begin{array}{c}';
put '\widehat{\tau }^E ';
put '\\ \widehat{\sigma }_{\widehat{\tau }^E} ';
put '\end{array}$ ';
do j=1 to nfac2;
put "& ";
put '$\begin{array}{c}';
put (tau2[j])6.3 ;
*put '\\ (SE_tau2[j])6.3 ;
put '\end{array}$ ';
end;
put "& " ;
put '$\begin{array}{c}';
put (tau0)6.3 ;
put '\\ (SE_tau0)6.3 ;
put '\end{array}$ ';
put " \\";
put "\hline\hline";
put "\end{tabular}";
put ' ';
put '\renewcommand{\arraystretch}{1}';

closefile table;
*** Finish: Graph ***;

%mend mac;
*%mac(Income,Gender,Race);
*%mac(Income,educationgrp,Race);
%mac(Income,Gender,educationgrp);

```

## 8.4 Appendix D: Three factor model

```

libname census11 "C:\Users\Jurgens\Desktop\1. Dissertation New\Dr Crafford 19 July 2016
\DATA";

```

```

%macro mac(response,factor1,factor2,factor3);
proc freq data=Census11.Data_analyse;
tables &factor1/out=factor1;
tables &factor2/out=factor2;
tables &factor3/out=factor3;
tables &response/out=response;
tables &factor1*&factor2*&factor3*&response/out=f list;
run;

proc transpose data=f out=freq prefix=c;
by &factor1 &factor2 &factor3;
var count;

```



```
run;

proc iml worksize=200 symsize=2000;
use freq; read all var _num_ into freq;
use response; read all var{income} into response;
use factor1; read all var _char_ into factor1;
use factor2;read all var _char_ into factor2;
use factor3;read all var _char_ into factor3;
print factor1;
print factor2;
print factor3;
print freq;
use f;read all var _char_ into names;names=rowcat(names);
effects="Intercept"//factor1//factor2//factor3;
class="R1600"//"R3200"//"R6400"//"R12800"//"R30000";
n=freq[+];
nfac1=nrow(factor1);
nfac2=nrow(factor2);
nfac3=nrow(factor3);
nt=nrow(freq);
k=nrow(response); response[k]=30; print response; k1=k-1;
x=response[1:k1];
nn=freq[,+];
nmm=shape(nn,nfac1);
f=colvec(freq[,1:k1]); f=f<>0.0001;
C=J(k1,1,1)@cusum(J(1,k1,1))<=J(1,k1,1)@cusum(J(k1,1,1));
CI=inv(C);
v1=J(k1,1,1);
po=inv(diag(nn)@I(k1))*f;
p=(I(nt)@C)*po;

*** Start: Design matrix for log-logistic distributions ***;
XD=log(x)||v1;
XXX=inv(XD'*XD)*XD'; XXX1=XXX[1,]; XXX2=XXX[2,];
Px=XD*inv(XD'*XD)*XD';
Qx=I(k1)-Px;
*** Finish: Design matrix for log-logistic distributions ***;

*** Start: Single-factor model ***;
/*
Dep1=designf(J(nfac1,1,1)@cusum(J(nfac2,1,1)))[1:nfac2,];
Dep2=Dep1;
if nfac1>2 then do;
do jj=1 to nfac1-2;
Dep2=block(Dep2,Dep1);
end;
end;
Dep=Dep2//(-repeat(Dep1,1,nfac1-1));
*/
Y1=designf(cusum(J(nfac1,1,1))@J(nfac2,1,1)@J(nfac3,1,1));*<=== dummy1;
Y2=designf(J(nfac1,1,1)@cusum(J(nfac2,1,1))@J(nfac3,1,1)); *<=== dummy2;
*Y3=designf(J(nfac1,1,1)@J(nfac2,1,1)@cusum(J(nfac3,1,1))); *<=== dummy3;

Y3=J(nfac1,1,1)@J(nfac2,1,1)@{-2,-1,0,1,2};*<=== Factor3: ordinal;
*Y3=Y3||Y3##2;

Y12=hdir(Y1,Y2);
Y13=hdir(Y1,Y3);
Y23=hdir(Y2,Y3)*0;
Y123=hdir(hdir(Y1,Y2),Y3)*0;

*YD=J(nt,1,1)||Y1||Y2||Y3; *<=== Only main effects;
*YD=J(nt,1,1)||Y1||Y2||Y3||Y12||Y13||Y23; *<=== Main effects with first order
interactions;
*YD=J(nt,1,1)||Y1||Y2||Y12||Y3; *<=== Main effects with first order interactions for
```



```

factor 1,2 and ordinal factor 3;
*YD=J(nt,1,1)||Y1||Y2||Y12||Y3||Y13||Y23||Y123;

YD=J(nt,1,1)||Y1||Y2||Y3||Y12||Y13||Y23||Y123;*<=== Main effects with first and second
order interactions;
YD=J(nt,1,1)||Y1||Y2||Y12||Y3||Y13||Y23||Y123;*<=== Main effects with first and second
order interactions for linear model;
Py=YD*inv(YD'*YD)*YD';
Qy=I(nt)-Py;
*** Finish: Three-factor model ***;
print YD;

start GGG(p,g,GG) global(nt,v1,Qx,XXX,XXX1,XXX2,Qy,h,D,kappa,theta,nu,A,Y12);
h=log(p/((J(nt,1,1)*v1)-p));
D=inv(diag(p))+inv(diag((J(nt,1,1)*v1)-p));

glog=(I(nt)*Qx)*h;
GGlog=(I(nt)*Qx)*D;

kappa=(I(nt)*XXX1)*h;
theta=(I(nt)*XXX2)*h;
nu=exp(-theta/kappa);
A1=nu*(theta/(kappa*kappa));
A2=nu*(-1/kappa);
A=diag(A1)*I(1) + diag(A2)*I(1);

greg=Qy*nu;
GGreg=Qy*A*(I(nt)*XXX)*D;

g=glog;          *<=== Model 1;
GG=GGlog;       *<=== Model 1;
g=glog/greg;    *<=== Model 2-4;
GG=GGlog/GGreg; *<=== Model 2-4;
finish;

i=0; p0=p; diff1=1;
do while (diff1 > 1e-9);
i=i+1; pi=p; p=p0;
    pio=(I(nt)*CI)*pi;
Vo=inv(diag(nn)*I(k1))*(diag(pio)- (diag(pio))*I(nt)*v1)*I(nt)*diag(pio);
V=(I(nt)*C)*Vo*(I(nt)*C)';
run GGG(pi,gpi,GGpi);
j=0; diff=1;
do while (diff > 1e-9);
j=j+1; pv=p;
run GGG(p,gp,GGp);
print i j p pi gp;
p=p-(GGpi*V)'*ginv(GGp*V*GGpi)'*gp;
diff=sqrt((p-pv)'*(p-pv));
if i=1 & j=1 then do;
Wald=gp'*ginv(GGp*V*GGp)'*gp;
GpV=GGp*V;
df=trace(GpV*ginv(GpV'*GpV)*GpV');
end;
end;
diff1=sqrt((p-pi)'*(p-pi));
end;

Cov_pi=V-(GGpi*V)'*ginv(GGpi*V*GGpi)'*(GGpi*V);
Cov_alpha=((I(nt)*XXX)*D)*Cov_pi*((I(nt)*XXX)*D)';

*mu=exp(-theta/kappa)*gamma(J(nt,1,1)+1/kappa)*gamma(J(nt,1,1)-1/kappa);
*sigma=sqrt(exp(-2*theta/kappa)
#(gamma(J(nt,1,1)+2/kappa)*gamma(J(nt,1,1)-2/kappa)
-(gamma(J(nt,1,1)+1/kappa)*gamma(J(nt,1,1)-1/kappa))##2));

```





```
*mum=shape(mu,nfac1);
*sigmam=shape(sigma,nfac1);

Cov_nu=A*Cov_alpha*A';
SE_nu=sqrt(diag(Cov_nu)*J(nrow(nu),1,1));
num=shape(nu,nfac1); SE_num=shape(SE_nu,nfac1);

YYY=inv(YD'*YD)*YD';
gamma=YYY*nu;
Cov_gamma=YYY*Cov_nu*YYY';
SE_gamma=sqrt(diag(Cov_gamma)*J(nrow(gamma),1,1));
t_gamma = gamma / SE_gamma;
Z1=designf(cusum(J(nfac1,1,1))@J(nfac2,1,1)@J(nfac3,1,1));
Z2=designf(J(nfac1,1,1)@cusum(J(nfac2,1,1))@J(nfac3,1,1));
Z3=designf(J(nfac1,1,1)@J(nfac2,1,1)@cusum(J(nfac3,1,1)));
Z12=hdir(Z1,Z2);
Z13=hdir(Z1,Z3);
Z23=hdir(Z2,Z3);
Z123=hdir(hdir(Z1,Z2),Z3);
LD=J(nt,1,1)||Z1||Z2||Z3||Z12||Z13||Z23||Z123;;
LLL=inv(LD'*LD)*LD';
lambda=LLL*nu;
lambda=choose(abs(lambda)<1e-9,0,lambda);
Cov_lambda=LLL*Cov_nu*LLL';
Cov_lambda=choose(abs(Cov_lambda)<1e-9,0,Cov_lambda);

S1=designf(cusum(J(nfac1,1,1)));
S2=designf(cusum(J(nfac2,1,1)));
S3=designf(cusum(J(nfac3,1,1)));
S12=S1@S2;
S13=S1@S3;
S23=S2@S3;
S123=S1@S2@S3;
S=block(1,S1,S2,S3,S12,S13,S23,S123);
tau=S*lambda;
Cov_tau=S*Cov_lambda*S';
SE_tau=sqrt(diag(Cov_tau)*J(nrow(tau),1,1));
*print tau[rowname=effects] SE_tau;

count=cusum(1//nfac1//nfac2//nfac3//(nfac1*nfac2)//(nfac1*nfac3)//(nfac2*nfac3)
//(nfac1*nfac2*nfac3));
tau0=tau[1:1]; SE_tau0=SE_tau[1:1];
tau1=tau[count[1]+1:count[2]]; SE_tau1=SE_tau[count[1]+1:count[2]];
tau2=tau[count[2]+1:count[3]]; SE_tau2=SE_tau[count[2]+1:count[3]];
tau3=tau[count[3]+1:count[4]]; SE_tau3=SE_tau[count[3]+1:count[4]];
tau12=tau[count[4]+1:count[5]]; SE_tau12=SE_tau[count[4]+1:count[5]];
tau13=tau[count[5]+1:count[6]]; SE_tau13=SE_tau[count[5]+1:count[6]];
tau23=tau[count[6]+1:count[7]]; SE_tau23=SE_tau[count[6]+1:count[7]];
tau123=tau[count[7]+1:count[8]]; SE_tau123=SE_tau[count[7]+1:count[8]];

tau12m=shape(tau12,nfac1); SE_tau12m=shape(SE_tau12,nfac1);
tau13m=shape(tau13,nfac1); SE_tau13m=shape(SE_tau13,nfac1);
tau23m=shape(tau23,nfac2); SE_tau23m=shape(SE_tau23,nfac2);

tau123t=shape(tau123,nfac1)';
nut=shape(nu,nfac1)';
kappat=shape(kappa,nfac1)';
thetat=shape(theta,nfac1)';

piom=(shape(pio,nt));
exp1=piom#(repeat(nn,1,k1));
exp2=nn-exp1[,+];
exp=exp1||exp2;

Pearson=((freq-exp)##2)/exp[+];
```



```
P_pvalue=1-probchi(Pearson,df);
W_pvalue=1-probchi(Wald,df);
discr=wald/n;

name1 = colvec(repeat(factor1,1,nfac2*nfac3));
name2 = repeat(colvec(repeat(factor2,1,nfac3)),nfac1,1);
name3 = repeat(factor3,nfac1*nfac2,1);
print name1 name2 name3;

print "Grouped response variable: &Response";
print "Three factor model: &Factor1 &Factor2 &Factor3";
print "Number of observations: " n ;
print "Number of cells: " nt;
print "Goodness of fit:" Pearson P_pvalue Wald W_pvalue df discr ;
print ' ';
print "Observed frequencies:", name1 name2 name3 freq[colname=class] nn;
print "Expected frequencies:", name1 name2 name3 exp nn;
print ' ';
print "Fitted log-logistic distributions and medians:";
print name1 name2 name3 nn nu kappa theta ;
print ' ';
print "The three factor model:";
print YD gamma SE_gamma t_gamma;
print ' ';

print "Effects for the three factor model: Intercept";
print tau0;
print ' ';
print "Effects for the three factor model: Main effects";
print tau1[rowname=(factor1)];
print tau2[rowname=(factor2)];
print tau3[rowname=(factor3)];
print "Effects for the three factor model: First order interaction effects";
print name1 name2 tau12;
print name1 name3 tau13;
print name2 name3 tau23;
print "Effects for the three factor model: Second order interaction effects";
print name1 name2 name3 tau123;

*** Start: Graph ***;
*response[k]=20;
xl=0//response[1:k1];
xu=response;
width=xu-xl;
start graph;
put '\FRAME{dtbpF}{3cm}{2cm}{1pt}{-}{-}{Plot}';
put '{\special{language "Scientific Word";type "MAPLEPLOT"}';
put 'width 3.0cm;height 2.0cm;depth 1pt;';
put 'display "USEDEF";plot_snapshots TRUE;mustRecompute FALSE;lastEngine "MuPAD";';
put 'xmin "0";xmax "30";xviewmin "0";xviewmax "30";yviewmin "0";yviewmax "0.3";';
put 'viewset"XY";rangeset"X";plottype 4;labeloverrides 3;numpoints 100;plotstyle
"patch";';
put 'plotticks 1;plottickdisable TRUE;num-x-ticks 7;num-y-ticks 6;';
put 'axesstyle "normal";xis \TEXUX{x};yis \TEXUX{y};var1name \TEXUX{$x$};var2name
\TEXUX{$y$};';
put 'function \TEXUX{$\left('';
put '\MATRIX{2,20}{c}\VR{,,c,,}{,c,,}{,,,,}\HR{,,,,,,,,,,,,,,,,,,,,,}';
do ii=1 to k;
put '\CELL{'(xl[ii])'}\CELL{0}';
put '\CELL{'(xl[ii])'}\CELL{'(pt[ii])'}';
put '\CELL{'(xu[ii])'}\CELL{'(pt[ii])'}';
put '\CELL{'(xu[ii])'}\CELL{0}';
end;
put '\right) $);';
put 'linecolor "maroon";linestyle 1;pointstyle "point";linethickness 1;lineAttributes
```



```

"Solid";';
put 'curveColor "[flat::RGB:0x00800000]";curveStyle "Line";';
put 'discont FALSE;';
put 'function \TEXUX{\$(x,\frac{e^{(\theta[te,l])}})\times '(kappa[te,l])'\times x^
{(\kappa[te,l])'-1}}{\left( 1+e^{(\theta[te,l])}'x^{(\kappa[te,l])}'\right) ^{2}})}$};';
put 'linecolor "black";linestyle 1;pointstyle "point";linethickness 1;lineAttributes
"Solid";';
put 'discont FALSE;';
put 'varlrange "0,30.00";num-x-gridlines 100;';
put 'curveColor "[flat::RGB:0x000000ff]";curveStyle "Line";';
put '}}';
finish;

filename table "C:\Jurgens\tex\factor3.tex";
file table;
put '%TCIDATA{LaTeXparent=0,0,master.tex}';
*put '\renewcommand{\arraystretch}{0.8}';
tel=0;
do l=1 to nfac1;
put "\newpage";
*put "{\textbf{\large &Factor1: " (Factor1[l]) " (&Factor2*&Factor3)}}";

put "\begin{tabular}{|c|}"(rowcat(J(1,nfac3,'c')))"|";
put "\hline \hline";
put "& \multicolumn{" nfac3 "}{|c|}{\textbf{\&factor3}} \ \ ";
put "\textbf{\&factor2}";
do j=1 to nfac3;
put "& \textbf{" (factor3[j,1]) " } ";
end;
put " \ ";
put "\hline \hline ";
do i=1 to nfac2;
put "\textbf{" (factor2[i,1]) " }";
do j=1 to nfac3;
tel=tel+1;
pt=(freq[te,l]'/nn[te,l])/width;
put "&"; run graph;
end;
put " \ ";
tel=tel-nfac3;
do j=1 to nfac3;
tel=tel+1;
put "&";
put '\begin{tabular}{cc}';
put '$\widehat{\nu}=' (nu[te,l])5.3 '$';
/*
put '$\kappa=' (kappa[te,l])5.3 '$ & $\widehat{\nu}=' (nu[te,l])5.3 '$ \ \ ';
put '$\theta=' (theta[te,l])5.3 '$ & $\widehat{s}_{\widehat{\nu}}=' (SE_nu[te,l])
5.3 '$';
*/
put '\end{tabular}';
end;
put " \ ";
put "\hline";
end;
put "\hline";
put " \end{tabular}";
put ' ';
end;
closefile table;
*** Finish: Graph ***;

create data_gamma from gamma;
append from gamma;

```



```
D1=designf(cusum(J(nfac1,1,1))@J(nfac2,1,1))*<=== dummy1;
D2=designf(J(nfac1,1,1)@cusum(J(nfac2,1,1))); *<=== dummy2;
D3=hdir(D1,D2); *<=== dummy3;
```

```
D=J(nfac1*nfac2,1,1)|D1|D2|D3;
D = block(D,D);
print D;
```

```
gamma_test = D*gamma;
print gamma_test;
%mend mac;
*%mac(Income,Gender);
*%mac(Income,Race);
*%mac(Income,educationgrp);
%mac(Income,Gender,race,educationgrp);
quit;
```

## 8.5 Appendix E: Logit model with one explanatory variable

```
libname census11 "C:\Users\Jurgens\Desktop\1. Dissertation New\Dr Crafford 19 July 2016
\DATA";
proc freq data=census11.data_focus;
tables incomegrp*educationgrp/out=f list;;
run;
```

```
proc iml;
```

```
age = 0;
education = 1;
linear = 1;
quadratic = 0;
```

```
use f;
read all into f1;
f = f1[,1];
n = f[+];
k = nrow(f)/2;
print k;
C = I(k) || -I(k);
print C;
logitf = C*log(f);
print logitf;
A1 = designf(do(1,k,1)');
A = J(k,1,1) || (A1);
print A;
lambda=inv(A'*A)*A'*logitf;
lambda_l = -lambda[2:nrow(lambda)][+];
lambda = lambda//lambda_l;
print lambda;
indices = exp(lambda);
print indices;
log_odds = log(J(k,1,indices[1])#indices[2:k+1]);
print log_odds;
D_f = diag(f);
D_1_f = inv(D_f);
print D_f D_1_f;
cov_lambda = inv(A'*A)*A'*C*D_1_f*(D_f-1/n*f*f')*D_1_f*C'*A*inv(A'*A);
print cov_lambda;
AH = A1;
if education = 1 then do;
if linear = 1 then do;
XD = J(k,1,1) || do(-2,2,1)';
```



```
end;
if quadratic = 1 then do;
XD = J(k,1,1) || do(-2,2,1)' || do(-2,2,1)'##2;
end;
AH = I(k) - XD*inv(XD'*XD)*XD';
end;

if age = 1 then do;
midpoints = {21.5, 28, 35.5, 43, 48, 53, 60.5};
if linear = 1 then do;
XD = J(k,1,1) || midpoints;
end;
if quadratic = 1 then do;
XD = J(k,1,1) || midpoints || midpoints##2;
end;
AH = I(k) - XD*inv(XD'*XD)*XD';
end;

print AH;
gf=AH'*logitf;
print gf;
Wald=gf'*ginv(AH'*C*diag(1/f)*C'*AH)*gf;
Prob=1-probchi(wald,ncol(AH));
MoD = Wald/n;
print Wald prob MoD;

do i=1 to 10;
f=f-C'*AH*ginv(AH'*C*diag(1/f)*C'*AH)*AH'*C*log(f);
ft=f';
print i ft[format=8.4];
end;
logitfh = C*log(f);
lambdah = inv(A'*A)*A'*logitfh;
indh = exp(lambdah);
print lambdah indh;

if linear = 1 | quadratic = 1 then do;
betah = inv(XD'*XD)*XD'*logitfh;
print betah;
end;

quit;
```

## 8.6 Appendix F: Logit model with two explanatory variables

```
libname census11 "C:\Users\Jurgens\Desktop\1. Dissertation New\Dr Crafford 19 July 2016
\DATA";
run;

proc freq data=census11.data_focus;
tables incomegrp*race*educationgrp/out=f list;
tables race/out=factor1;
tables educationgrp/out=factor2;
run;

proc iml;

age = 0;
education = 1;
linear = 1;
quadratic = 0;
independent = 1;
```



```
use f;
read all into f1;
f = f1[,1];
n = f[+];
use factor1; read all var _char_ into factor1;
use factor2; read all var _char_ into factor2;
print factor1 factor2;

nfac1 = nrow(factor1);
nfac2 = nrow(factor2);

k = nrow(f)/2;
print k;
C = I(k) || -I(k);
logitf = C*log(f);

A1 = designf(cusum(J(nfac1,1,1))@J(nfac2,1,1));
A2 = designf(J(nfac1,1,1)@cusum(J(nfac2,1,1)));
A12 = hdir(A1,A2);
A = J(k,1,1) || A1||A2||A12;
print A;
lambda=inv(A'*A)*A'*logitf;
indices = exp(lambda);
print lambda indices;

AH = A12;

if education = 1 then do;
steps = {-2,-1,0,1,2};
if linear = 1 then do;
A2 = J(nfac1,1,1)@steps;
end;
if quadratic = 1 then do;
A2 = J(nfac1,1,1)@steps||J(nfac1,1,1)@steps##2;
end;
A12 = hdir(A1,A2);
if independent = 1 then do;
XD = J(k,1,1)||A1||A2;
end;
if independent = 0 then do;
XD = J(k,1,1)||A1||A2||A12;
end;
PX = XD*inv(XD'*XD)*XD';
QX = I(k) - PX;
AH = QX;
end;

if age = 1 then do;
steps = {21.5, 28, 35.5, 43, 48, 53, 60.5};
if linear = 1 then do;
A2 = J(nfac1,1,1)@steps;
end;
if quadratic = 1 then do;
A2 = J(nfac1,1,1)@steps||J(nfac1,1,1)@steps##2;
end;
A12 = hdir(A1,A2);
if independent = 1 then do;
XD = J(k,1,1)||A1||A2;
end;
if independent = 0 then do;
XD = J(k,1,1)||A1||A2||A12;
end;
PX = XD*inv(XD'*XD)*XD';
```



```
QX = I(k) - PX;
AH = QX;
end;
print XD;
print AH;
gf=AH'*logitf;
print gf;
Wald=gf'*ginv(AH'*C*diag(1/f)*C'*AH)*gf;
Prob=1-probchi(wald,ncol(AH));
MoD = Wald/n;
print Wald prob MoD;

do i=1 to 10;
f=f-C'*AH*ginv(AH'*C*diag(1/f)*C'*AH)*AH'*C*log(f);
ft=f';
print i ft[format=8.4];
end;
logitfh = C*log(f);
lambdah = inv(A'*A)*A'*logitfh;
indh = exp(lambdah);
print lambdah indh;

if linear = 1 | quadratic = 1 then do;
betah = inv(XD'*XD)*XD'*logitfh;
print betah;
end;

quit;
```

## 8.7 Appendix G: Logit model with three explanatory variables

```
libname census11 "C:\Users\Jurgens\Desktop\1. Dissertation New\Dr Crafford 19 July 2016\DATA";
```

```
run;
```

```
proc freq data=census11.data_focus;
tables incomegrp*gender*race*educationgrp/out=f list;
tables gender/out=factor1;
tables race/out=factor2;
tables educationgrp/out=factor3;
run;
```

```
proc iml;
```

```
error = 10##(-6);
print error;
```

```
education = 1;
independent = 0;
```

```
use f;
read all into f1;
f = f1[,1];
n = f[+];
use factor1; read all var _char_ into factor1;
use factor2; read all var _char_ into factor2;
use factor3; read all var _char_ into factor3;
print factor1 factor2 factor3;
```



```
nfac1 = nrow(factor1);
nfac2 = nrow(factor2);
nfac3 = nrow(factor3);

k = nrow(f)/2;
print k;
C = I(k) || -I(k);
logitf = C*log(f);

steps = {-2,-1,0,1,2};

A1 = designf(cusum(J(nfac1,1,1)))@J(nfac2,1,1)@J(nfac3,1,1);
A2 = J(nfac1,1,1)@designf(cusum(J(nfac2,1,1)))@J(nfac3,1,1);
A3 = J(nfac1,1,1)@J(nfac2,1,1)@designf(cusum(J(nfac3,1,1)));

A12 = hdir(A1,A2);
A13 = hdir(A1,A3);
A23 = hdir(A2,A3);

A123 = hdir(hdir(A1,A2),A3);

A = J(k,1,1) || A1 || A2 || A3 || A12 || A13 || A23 || A123;
print A;
lambda=inv(A'*A)*A'*logitf;
indices = exp(lambda);
print lambda indices;

AH = A12 || A13 || A23 || A123;

if education = 1 then do;
A3 = J(nfac1,1,1)@J(nfac2,1,1)@steps;
A13 = hdir(A1,A3);
A23 = hdir(A2,A3);
A123 = hdir(hdir(A1,A2),A3);
if independent = 1 then do;
XD = J(k,1,1) || A1 || A2 || A3;
end;
if independent = 0 then do;
XD = J(k,1,1) || A1 || A2 || A12 || A3 || A13 || A23 || A123;
end;
PX = XD*inv(XD'*XD)*XD';
QX = I(k) - PX;
AH = QX;
end;

print AH;
gf=AH'*logitf;
print gf;
Wald=gf'*ginv(AH'*C*diag(1/f)*C'*AH)*gf;
Prob=1-probchi(wald,ncol(AH));
MoD = Wald/n;
print Wald prob MoD;

do i=1 to 10;
f=f-C'*AH*ginv(AH'*C*diag(1/f)*C'*AH)*AH'*C*log(f);
if i > 1 then do;
test = abs(ft - f')[+];
print i test;
if test < error then i = 999;
end;
ft=f';
end;
print test;
```





```
logitfh = C*log(f);
lambdah = inv(A'*A)*A'*logitfh;
indh = exp(lambdah);
print lambdah indh;

beta = inv(XD'*XD)*XD'*logitfh;
print beta;
D1=designf(cusum(J(nfac1,1,1))@J(nfac2,1,1));*<=== dummy1;
D2=designf(J(nfac1,1,1)@cusum(J(nfac2,1,1))); *<=== dummy2;
D3=hdir(D1,D2); *<=== dummy3;

D=J(nfac1*nfac2,1,1)||D1||D2||D3;
D = block(D,D);
print D;

beta_test = D*beta;
print beta_test;

quit;
```