



**ATHENS UNIVERSITY
OF ECONOMICS AND BUSINESS**

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

**LOGISTIC REGRESSION:
A STANDARD METHOD OF ANALYSIS
IN MEDICAL RESEARCH
-AN APPLICATION TO MEDICAL DATA**

By

Isabella D. Sourla

A THESIS

Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
2004





ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

Logistic Regression: A Standard Method of Analysis in
Medical Research – An Application to Medical Data

By

Isabella D. Sourla

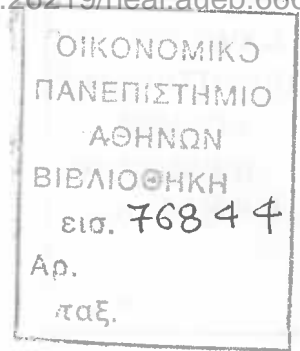


A THESIS

Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
September 2004





ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

Λογιστική Παλινδρόμηση: Βασική Μέθοδος Ανάλυσης
Στην Ιατρική Έρευνα- Εφαρμογή σε Ιατρικά Δεδομένα

Ισαβέλλα Δ. Σούρλα



ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα
Σεπτέμβριος 2004





**ATHENS UNIVERSITY
OF ECONOMICS AND BUSINESS
DEPARTMENT OF STATISTICS**

A Thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science

**LOGISTIC REGRESSION: A STANDARD METHOD
OF ANALYSIS IN MEDICAL RESEARCH
-AN APPLICATION TO MEDICAL DATA**

Isabella D. Sourla



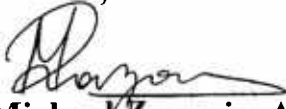
Approved by the Graduate Committee

A. Kostaki
Assistant Professor
Thesis Supervisor

E. Moustaki
Assistant Professor

St. Psarakis
Assistant Professor
Members of the Committee

Athens, November 2004


**Michael Lazanis, Associate Professor
Director of the Graduate Program**



ACKNOWLEDGEMENTS

I would like to thank my supervisor Assistant Professor Anastasia Kostaki for suggesting this problem of analysis and her guidance.

I also thank Aglaia Manousaki from Department of Dermatology and Androniki Tosca from Department of Surgical Oncology, Heraklion University Hospital, for providing data for this analysis.

Finally, I am grateful to my intimate friend Evgenia Tsompanaki for her valuable help and encouragement during the whole period of my studies at the Master Program in Statistics at the Department of Statistics of the Athens University of Economics and Business.





VITA

I was born in Larisa in 1973. In 1990 I entered the Department of Mathematics of the University of Patras and in 1994 I graduated with a degree in Mathematics with major field in Statistics and Operational research. I was accepted in the M.Sc. Program in Statistics of the Athens University of Economics and Business in October of 1999. During these studies my lovely son Alexandros was born. This thesis is the last part of the requirements of the program.





ABSTRACT

Isabella Sourla

Logistic Regression: Standard Method of Analysis in Medical Research – An Application to Medical Data

September 2004

Logistic regression is a statistical technique that can be used in binary response problems. It allows one to predict a discrete outcome, such as group membership, from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these.

Over the last years the logistic regression has become, in many fields including medical research, the standard method of analysis.

Melanoma is considered to be one of the most life threatening tumors with an alarmingly increasing incidence and therefore much attention is drawn towards early detection, recognition and excision of lesions with suspicion for melanoma. An image processing melanoma predictive system developed, taking advantage of high computer technology in nowadays, succeeding high accuracy of melanoma prediction, based on histopathological verification.

Based on this image predictive system for melanoma diagnosis in an everyday melanocytic skin lesion unit, logistic regression technique applied to provide a prognostic tool for melanoma probability.



ABSTRACT

Isabella Souris

Logistic Regression: Standard Method of Analysis in Medical Research - An Application to Medical Data

September 2004

Logistic regression is a statistical technique that can be used to solve binary response problems. It allows one to predict a discrete outcome, such as group membership, from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these.

Over the last years the logistic regression has become, in many fields including medical research, the standard method of analysis.

Melanoma is considered to be one of the most life threatening tumors with an alarmingly increasing incidence and therefore much attention is drawn towards early detection, recognition and curative of lesions with suspicion for melanoma. An image processing melanoma predictive system developed, taking advantage of high computer technology in nowadays, succeeding high accuracy of melanoma prediction, based on histopathological verification.

Based on this image predictive system for melanoma diagnosis in an everyday melanocytic skin lesion unit, logistic regression technique applied to provide a prognostic tool for melanoma probability.



ΠΕΡΙΛΗΨΗ

Ισαβέλλα Σούρλα

Λογιστική Παλινδρόμηση: Βασική Μέθοδος Ανάλυσης Στην Ιατρική Έρευνα- Εφαρμογή σε Ιατρικά Δεδομένα

Σεπτέμβριος 2004

Η Λογιστική παλινδρόμηση αποτελεί μία τεχνική της Στατιστικής η οποία χρησιμοποιείται σε προβλήματα όπου η απάντηση στο ερώτημα που απασχολεί τον ερευνητή είναι διττή. Η τεχνική αυτή επιτρέπει την πρόβλεψη ενός διακριτού αποτελέσματος, όπως η συμμετοχή ή μη σε μία ομάδα, μέσω ενός συνόλου μεταβλητών κάθε είδους, είτε αυτές είναι συνεχείς, διακριτές, δίτιμες ή μείξη των ειδών αυτών.

Τα τελευταία χρόνια η λογιστική παλινδρόμηση έχει καθιερωθεί σε πολλούς τομείς, συμπεριλαμβανομένου και του τομέα της ιατρικής έρευνας, ως η βασική μέθοδος ανάλυσης.

Το μελάνωμα του δέρματος θεωρείται ένας από τους πιο επικίνδυνους όγκους με μία ανησυχητική αύξηση εμφάνισης συμβάντων . Για το λόγο αυτό δίνεται μεγάλη προσοχή στην όσο το δυνατόν γρηγορότερη ανακάλυψη, αναγνώριση και αφαίρεση του σημείου όπου εμφανίζεται βλάβη η οποία οδηγεί σε υποψία μελανώματος. Ένα σύστημα πρόβλεψης παρουσίας μελανώματος αναπτύχθηκε μέσω της επεξεργασίας εικόνας και των μεγάλων δυνατοτήτων που παρέχει η τεχνολογία των υπολογιστών στις μέρες μας , το οποίο σύμφωνα με ιστοπαθολογική επαλήθευση, επιτυγχάνει υψηλή ακρίβεια στην πρόβλεψη του μελανώματος.

Η τεχνική της λογιστικής παλινδρόμησης εφαρμόζεται σε στοιχεία παραγόμενα από το σύστημα αυτό, της επεξεργασίας της εικόνας για την πρόβλεψη ύπαρξης μελανώματος του δέρματος, συντελώντας στην προσπάθεια δημιουργίας ενός εργαλείου πρόγνωσης της πιθανότητας παρουσίας μελανώματος.





TABLE OF CONTENTS

Chapter 1	Introduction	1
Chapter 2	Logistic Regression Model	3
2.1	Introduction	3
2.2	Why Logistic Regression?.....	4
2.3	Simple Logistic Regression.....	15
2.3.1	Fitting a Logistic Response Function.....	15
2.3.2	Simple Logistic Regression Model	18
2.3.3	Likelihood Function.....	18
2.4	Multiple Logistic Regression	22
2.4.1	Multiple Logistic Regression Model.....	22
2.4.2	Likelihood Function.....	22





LIST OF TABLES

Table 2.1	Age and Coronary Heart Disease Status (CHD) of 100 Subjects .	4
Table 2.2	Summary Table of Age Group By Coronary Heart Disease.....	6
Table 2.3	Age (Midpoint) and Coronary Heart Disease Status (CHD).....	7
Table 2.4	Logits of Coronary Heart Disease for Age Midpoint	15
Table 2.5	Results of fitting the Logistic Regression Model to the Data in Table 2.3	21
Table 2.6	Interpretation of a diagnostic test based on test results and disease status.....	28
Table 2.7	Classification table for data of Table 2.3	30
Table 3.1	Descriptives for the five selected covariates	46
Table 3.2	Analysis of Maximum Likelihood Estimates for Multivariate Logistic with independent variables “RM”, “LAC_GREY”, “CV_SHARP”, “MEAN_RED”, “RANGE BLUE”	47
Table 3.3	Analysis of Maximum Likelihood Estimates for Multivariate Logistic with independent variables “RM”, “LAC_GREY”, “MEAN_RED” and “RANGE BLUE”	48
Table 3.4	Analysis of Maximum Likelihood Estimates for Multivariate Logistic with independent variables “RM”, “LAC_GREY”, “CV_SHARP”, and “RANGE BLUE”	48
Table 3.5	Analysis of Maximum Likelihood Estimates for Multivariate Logistic with independent variables “RM”, “LAC_GREY”, “RANGE BLUE”	48
Table 3.6	Model Fit Statistics for models with five & three covariates	49
Table 3.7	Classification Table for model with three covariates for different cut-off points.....	50
Table 3.8	Classification Table for model with three covariates for cut-off point of 0.5.....	52
Table 3.9	Classification Table for jackknife validation.....	57
Table 3.10	Statistics of predictive probability for model described by equation (3.1).....	58
Table A. 1	The Studied Geometric, Colour, Sharpness and Colour Texture Variables (n=43).....	63
Table A. 2	Analysis of Maximum Likelihood and Odd Ratio Estimates from Univariate Analysis.....	64
Table A. 3	Spearman’s Correlation Coefficient for Parameters of fractal geometry for lesion color texture	65
Table A. 4	Spearman’s Correlation Coefficient for Parameters of Geometry	66
Table A. 5	Spearman’s Correlation Coefficient Spearman’s Correlation Coefficient for Parameters of color including skewness from the Gaussian curve of normal distribution for the 4 color intensities.....	71
Table A. 6	Spearman’s Correlation Coefficient for estimates of sharpness of the lesion border from the surrounding skin	71
Table A. 7	Association of Predicted Probabilities and Observed Responses for model with five variables –Rm, Lac_Grey, Cv_Sharp, Mean_Red, Range Blue- and three variables - Rm, Lac_Grey, Range Blue.....	72



Table A. 8 Group Statistics of predictive probability for model described by equation (3.1) 72

Table A. 9 Results of independent sample t-test for predictive probability for model described by equation (3.1)..... 73

Table A. 1 Association of Predicted Probabilities and Observed Responses for model with five variables - km, lac_grey, cv_sharp, Mean_Red, Range_Blue - and three variables - km, lac_grey, Range_Blue 73

Table A. 2 Specimen's Correlation Coefficient for Parameters of Gaussian curve of normal distribution for the 4 color intensities 71

Table A. 3 Specimen's Correlation Coefficient for Parameters of Geometry for lesion color texture 65

Table A. 4 Specimen's Correlation Coefficient for Parameters of Geometry 66

Table A. 5 Specimen's Correlation Coefficient for Parameters of Geometry for lesion color texture 65

Table A. 6 Specimen's Correlation Coefficient for Parameters of Geometry for lesion color texture 65

Table A. 7 Specimen's Correlation Coefficient for Parameters of Geometry for lesion color texture 65

Table A. 8 Specimen's Correlation Coefficient for Parameters of Geometry for lesion color texture 65

Table A. 9 Specimen's Correlation Coefficient for Parameters of Geometry for lesion color texture 65

Table A. 10 Statistics of predictive probability for model described by equation (3.1) 58

Table 3.9 Classification Table for jackknife validation 57

Table 3.8 Classification Table for model with three covariates for cut-off point at 0.5 52

Table 3.7 Classification Table for model with three covariates for different cut-off points 50

Table 3.6 Model Fit Statistics for models with five & three covariates 49

Table 3.5 Analysis of Maximum Likelihood Estimates for Multivariate Logistic with independent variables "KM", "LAC_GREY", "RANGE_BLUE" 48

Table 3.4 Analysis of Maximum Likelihood Estimates for Multivariate Logistic with independent variables "KM", "LAC_GREY", "CV_SHARP", and "RANGE_BLUE" 48

Table 3.3 Analysis of Maximum Likelihood Estimates for Multivariate Logistic with independent variables "KM", "LAC_GREY", "MEAN_RED", and "RANGE_BLUE" 48

Table 3.2 Analysis of Maximum Likelihood Estimates for Multivariate Logistic with independent variables "KM", "LAC_GREY", "CV_SHARP", "MEAN_RED", and "RANGE_BLUE" 47

Table 3.1 Descriptors for the five selected covariates 46

Table 3.0 Classification table for data of Table 3.1 46

Table 2.6 Interpretation of a diagnostic test based on test results and disease status 38

Table 2.5 Results of fitting the Logistic Regression Model to the Data 31

Table 2.4 Logit of Coronary Heart Disease for Age Midpoint 13

Table 2.3 Age (Midpoint) and Coronary Heart Disease Status (CHD) 7



LIST OF FIGURES

Figure 2.1 Scatterplot of Coronary Heart Disease by Age.....	5
Figure 2.2 Mosaic Plot of Coronary Heart Disease By Age Group	6
Figure 2.3 Fitting a linear model to describe the relationship between Coronary Heart Disease and Age (Midpoint).....	9
Figure 2.4 Plot of the Mean & Predicted Mean (O.L.S.) of Coronary Heart Disease By Age (Midpoint)	10
Figure 2.5 Plot of the proportion p versus standard error of p	11
Figure 2.6 Increasing and Decreasing Logistic Plots	14
Figure 2.7 Plot of Logit(p) versus Age Midpoint	16
Figure 2.8 Relationship between <i>Sensitivity - Specificity</i> and <i>Cut-off Point</i>	29
Figure 2.9 Comparing ROC curves.....	32
Figure 3.1 ROC curves for models with three and five covariates	50
Figure 3.2 ROC curve for the selected model with three covariates	51
Figure 3.3 Plots of Pearson and Deviance Residuals versus predicted probabilities or case number for the model described by equation (3.1).....	54
Figure 3.4 Plots of Hat Diagonal versus predicted probabilities or case number for the model described by equation (3.1).....	55
Figure 3.5 Plots of DIFCHISQ and DIFDEV versus case number for the model described by equation (3.1)	56



Chapter 1 Introduction

Regression methods have become an integral component of any data analysis concerned with describing the relationship between a response variable and one or more explanatory variables. It is often the case that the outcome variable is discrete, taking two or more possible values. This is the point that distinguishes a logistic regression model from the linear regression model; the outcome variable in logistic regression is usually binary or dichotomous. This difference between logistic and linear regression is reflected both in the choice of a parametric model and in the assumptions. Once this difference is accounted for, the methods employed in an analysis using logistic regression follow the same general principles used in linear regression. Thus, the techniques used in linear regression analysis motivate the approach to logistic regression.

Over the last decades the logistic regression model has become, in many fields, the standard method of analysis.

In this thesis logistic regression analysis was used in order to identify potential prognostic parameters based on an image processing analysis for melanoma diagnosis in an everyday melanocytic skin lesion unit, establishing a statistical model to predict the risk for skin melanomas and no-melanomas.

Melanoma is considered to be one of the most life threatening tumors with an alarmingly increasing incidence (Armstrong and Kricger, 1994). Early diagnosed primary cutaneous melanoma has an excellent overall prognosis whereas the prognosis of advanced melanoma is poor (Balch, 1992), (Morton, et.al., 1993). Therefore much attention is drawn towards early detection, recognition and excision of lesions with suspicion for melanoma.

With technology prevailing in most sciences, the idea of computer assisted melanoma diagnosis was brought about in the early 80s. Since then many efforts have brought promising results to this goal (Green et. al., 1994), (Sober and Burstein, 1994), (Andreassi et. al., 1999). Still, there is no method claiming to offer definite accuracy in melanoma diagnosis from other



Melanocytic Skin Lesions (MSL). Most computer-based research methods use either clinical view (naked eye) images, calculating various algorithms with very good results or epiluminescence microscopy with equally good results.

In an attempt to aid in early melanoma diagnosis a simple to use and low- cost image-processing program was developed, evaluating naked- eye, good quality digital images for the analysis and mapping of malignant and benign MSL. A wide array of algorithms evaluated trying to reduce inherent algorithm limitations, and an image processing melanoma predictive system developed succeeding high accuracy of melanoma prediction, based on histopathological verification.



Chapter 2 Logistic Regression Model

2.1 Introduction

Logistic regression is a statistical technique that can be used in binary response problems. It allows one to predict a discrete outcome, such as group membership, from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these.

Generally, the dependent or response variable is dichotomous, that is, the dependent variable can take the value 1 with a probability of success p , or the value 0 with probability of failure $1 - p$.

Applications of logistic regression have also been extended to cases where the dependent variable is of more than two cases, known as multinomial or polytomous regression (Tabachnick and Fidell, 1996). Discriminant analysis is also a statistical tool used to predict group membership with only two groups. However, discriminant analysis can only be used with continuous independent variables. Thus, in instances where the independent variables are categorical one, or a mix of continuous and categorical ones, logistic regression is preferred.

It is important to understand that the goal of an analysis using this method is the same as that of any model-building statistical technique. To find the best fitting and most parsimonious, yet biologically reasonable model to describe the relationship between an outcome (dependent or response variable) and a set of independent (predictor or explanatory) variables. These independent variables are often called covariates.

Logistic regression is a predictive analysis, like linear regression, with predictors could be continuous or dichotomous. However ordinary least squares regression (OLS) is not appropriate if the outcome is dichotomous. Whereas the OLS regression uses normal probability theory, logistic regression uses binomial probability theory. This makes things a bit more complicated mathematically.



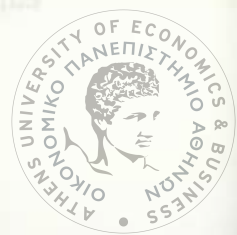
2.2 Why Logistic Regression?

As our motivation for use of *logistic regression* we will consider an example study that looks at the relationship between age and coronary heart disease. Table 2.1 lists age in years (AGE), and presence or absence of evidence of significant coronary heart disease (CHD) for 100 subjects selected to participate in a study. The table also contains an identifier variable (ID) and an age group variable (AGRP). The outcome variable is CHD, which is coded with a value of zero to indicate CHD is absent, or 1 to indicate that it is present in the individual.

ID	AGRP	AGE	CHD	ID	AGRP	AGE	CHD	ID	AGRP	AGE	CHD
1	1	20	0	35	3	38	0	68	6	51	0
2	1	23	0	36	3	39	0	69	6	52	0
3	1	24	0	37	3	39	1	70	6	52	1
4	1	25	0	38	4	40	0	71	6	53	1
5	1	25	1	39	4	40	1	72	6	53	1
6	1	26	0	40	4	41	0	73	6	54	1
7	1	26	0	41	4	41	0	74	7	55	0
8	1	28	0	42	4	42	0	75	7	55	1
9	1	28	0	43	4	42	0	76	7	55	1
10	1	29	0	44	4	42	0	77	7	56	1
11	2	30	0	45	4	42	1	78	7	56	1
12	2	30	0	46	4	43	0	79	7	56	1
13	2	30	0	47	4	43	0	80	7	57	0
14	2	30	0	48	4	43	1	81	7	57	0
15	2	30	0	49	4	44	0	82	7	57	1
16	2	30	1	50	4	44	0	83	7	57	1
17	2	32	0	51	4	44	1	84	7	57	1
18	2	32	0	52	4	44	1	85	7	57	1
19	2	33	0	53	5	45	0	86	7	58	0
20	2	33	0	54	5	45	1	87	7	58	1
21	2	34	0	55	5	46	0	88	7	58	1
22	2	34	0	56	5	46	1	89	7	59	1
23	2	34	1	57	5	47	0	90	7	59	1
24	2	34	0	58	5	47	0	91	8	60	0
25	2	34	0	59	5	47	1	92	8	60	1
26	3	35	0	60	5	48	0	93	8	61	1
27	3	35	0	61	5	48	1	94	8	62	1
28	3	36	0	62	5	48	1	95	8	62	1
29	3	36	1	63	5	49	0	96	8	63	1
30	3	36	0	64	5	49	0	97	8	64	0
31	3	37	0	65	5	49	1	98	8	64	1
32	3	37	1	66	6	50	0	99	8	65	1
33	3	37	0	67	6	50	1	100	8	69	1
34	3	38	0								

Table 2.1 Age and Coronary Heart Disease Status (CHD) of 100 Subjects

(Source: Hosmer and Lemeshow, 1989)



It is of interest to explore the relationship between age and the presence or absence of Coronary Heart Disease (CHD) in this study population. According to Chambers et.al. (1983), "there is no statistical tool that is as powerful as a well-chosen graph". To have a visual impression of the nature and strength of any relationship between the outcome and the independent variable, we use a scatterplot of the outcome versus the independent variable.

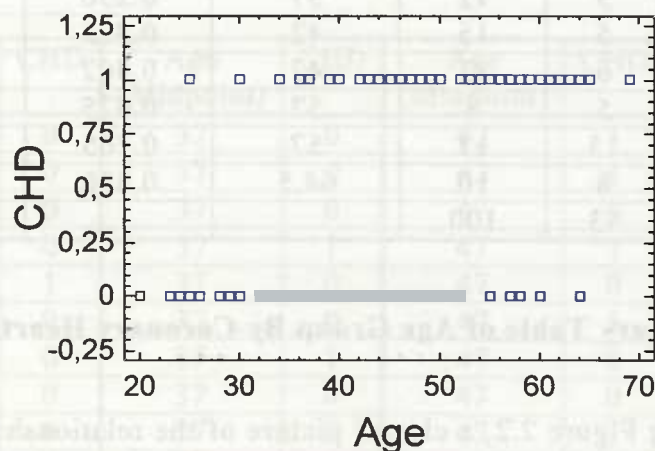


Figure 2.1 Scatterplot of Coronary Heart Disease by Age

In this scatterplot all points fall on one of two parallel lines representing the absence of CHD ($y = 0$) and the presence of CHD ($y = 1$). There is some tendency for the individuals with no evidence of CHD to be younger than those with evidence of CHD. While this plot does depict the dichotomous nature of the outcome variable quite clearly, it does not provide a clear picture of the nature of the relationship between CHD and age.

Another problem as mentioned by Hosmer et.al. (Hosmer and Lemeshow, 1989) with Figure 2.1 is that the variability in CHD at all ages is large. This makes it difficult to describe the functional relationship between age and CHD. One common method of removing some variation while still maintaining the structure of the relationship between the outcome and the independent variable is to create intervals for the independent variable and compute the mean of the outcome variable within each group. We carry out this strategy by categorizing the variable age, forming the age group variable,

named AGRP. Table 1.2 contains, for each age group, the frequency of occurrence of each outcome as well as the mean of the group (level), the mean (proportion with CHD present) and the standard deviation (std. deviation) for each group.

Age Group	Coronary Heart Disease			Midpoint Age	Mean (Proportion)	Std Deviation
	Absent	Present	n			
20 – 29	9	1	10	24.5	0.100	0.316
30 – 34	13	2	15	32	0.133	0.352
35 – 39	9	3	12	37	0.250	0.452
40-44	10	5	15	42	0.333	0.488
45 – 49	7	6	13	47	0.462	0.519
50-54	3	5	8	52	0.625	0.518
55 – 59	4	13	17	57	0.765	0.437
60 – 69	2	8	10	64.5	0.800	0.422
Total	57	43	100			

Table 2.2 Summary Table of Age Group By Coronary Heart Disease

By examining Figure 2.2, a clearer picture of the relationship begins to emerge. It appears that as age increases, the proportion of individuals with evidence of CHD increases.

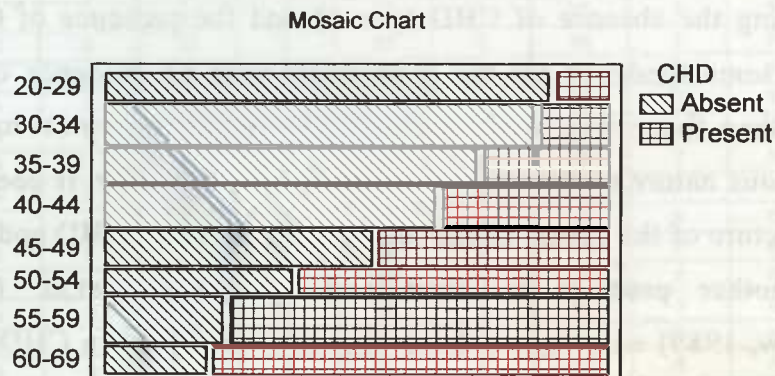


Figure 2.2 Mosaic Plot of Coronary Heart Disease By Age Group

Figure 2.2 displays mosaic plot of CHD by Age Group with rectangles' area proportional to the cell counts of individuals with and without CHD for



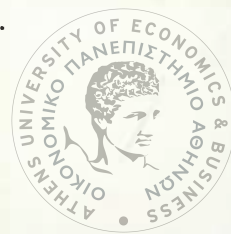
the midpoint of each age interval. In addition, the width of the bars is proportional to the percentage distribution of Age Group. While this provides considerable insight into the relationship between Coronary Heart Disease and Age in this study, a functional form for this relationship needs to be described.

If Age is the midpoint of the age-range of a group of subjects, that is Age=24,5 was coded when age fell between 20-29, Age=32 when age fell between 30-34, and so on, whereas CHD was coded as present =1 and absent=0, we have Table 2.3.

Age (Midpoint)	CHD	Age (Midpoint)	CHD	Age (Midpoint)	CHD	Age (Midpoint)	CHD
24,5	0	37	0	42	1	57	1
24,5	0	37	0	42	1	57	1
24,5	0	37	0	47	0	57	1
24,5	0	37	1	47	1	57	1
24,5	1	37	0	47	0	57	0
24,5	0	37	0	47	1	57	0
24,5	0	37	1	47	0	57	1
24,5	0	37	0	47	0	57	1
24,5	0	37	0	47	1	57	1
24,5	0	37	0	47	0	57	1
32	0	37	0	47	1	57	0
32	0	37	1	47	1	57	1
32	0	42	0	47	0	57	1
32	0	42	1	47	0	57	1
32	0	42	0	47	1	57	1
32	1	42	0	52	0	64,5	0
32	0	42	0	52	1	64,5	1
32	0	42	0	52	0	64,5	1
32	0	42	0	52	0	64,5	1
32	0	42	1	52	1	64,5	1
32	0	42	0	52	1	64,5	1
32	0	42	0	52	1	64,5	0
32	1	42	1	52	1	64,5	1
32	0	42	0	57	0	64,5	1
32	0	42	0	57	1	64,5	1

Table 2.3 Age (Midpoint) and Coronary Heart Disease Status (CHD)

Before proceeding we have to recall that when the response variable is binary, taking on the values 1 and 0 with probabilities p and $1-p$, respectively, Y is a Bernoulli random variable with parameter $E\{Y\} = p$.



Therefore, proportion and probability of 1 are the same in such cases and the mean of a binary distribution could be denoted as p , the proportion of 1's.

We will use $p_i = p(x_i)$ and $1 - p_i = 1 - p(x_i)$ to represent the probability that $Y=1$ and $Y=0$, respectively. These probabilities are written in the following form:

$$p_i = p(x_i) = P(Y = 1 | X_1, X_2, \dots, X_n)$$

$$1 - p_i = 1 - p(x_i) = P(Y = 0 | X_1, X_2, \dots, X_n)$$

We should note that the x_i in the expression, $p_i = p(x_i)$, is a vector representing the set of the independent predictor variables, X_1, X_2, \dots, X_n .

Let assume that the form of the relationship between CHD, notated Y , and Age, notated X , is linear. In other words let assume that the predicted value of the outcome variable CHD, given the value of the independent variable Age, may be expressed as an equation linear in X (or some transformation of X or Y), such as

$$\hat{y} = b_0 + b_1x$$

If we tried to draw a straight (best fitting) line through the points, fitting a linear model to describe the relationship between CHD and Age (Midpoint), the equation of the fitted model, shown as a solid line at next graph, is:

$$\text{CHD} = -0,49936 + 0,0209315 \text{ Age} \quad (2.1)$$



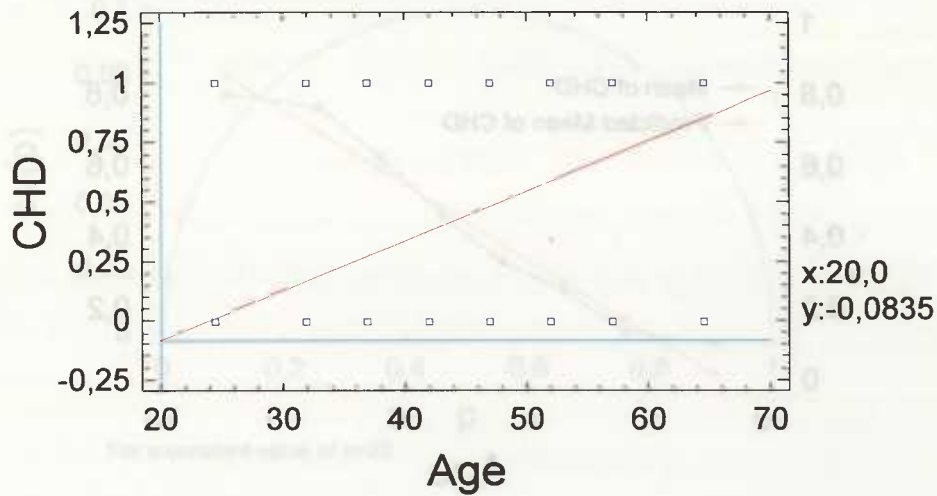
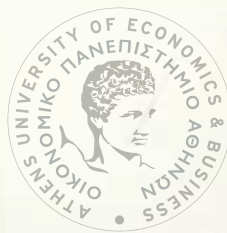


Figure 2.3 Fitting a linear model to describe the relationship between Coronary Heart Disease and Age (Midpoint).

Thus, note that the predicted proportion of CHD for age equals 20 is approximately -0.1, a negative proportion. Note also that the linear trend predicts that at some age above 70 the proportion of individuals with CHD will exceed 1.0. Thus, the model can produce nonsense results; proportions less than zero or greater than one. So fitting a straight line to proportion does not make sense.

Generally, assuming linear relationship between the mean response and the predictor variable implies insensibly that, it is possible for $E\{Y_i|X_i\}$ to take on any value as values of X range between $-\infty$ and $+\infty$.

One solution would be to convert or transform these numbers into probabilities. We might compute the average of the y values at each point on the x axis. The y values can only be 0 or 1, so an average of them will be between 0 and 1. This average is the same as the probability p_i of having a value of 1 on the y variable, given a certain value of x. So, we could then plot the probabilities of y, the column labeled "Mean -Proportion" in Table 2.2, at each value of x. Figure 2.4 illustrates proportion of CHD as well as predicted proportion derived by fitting the O.L.S. regression line, equation (2.1), versus age midpoint.



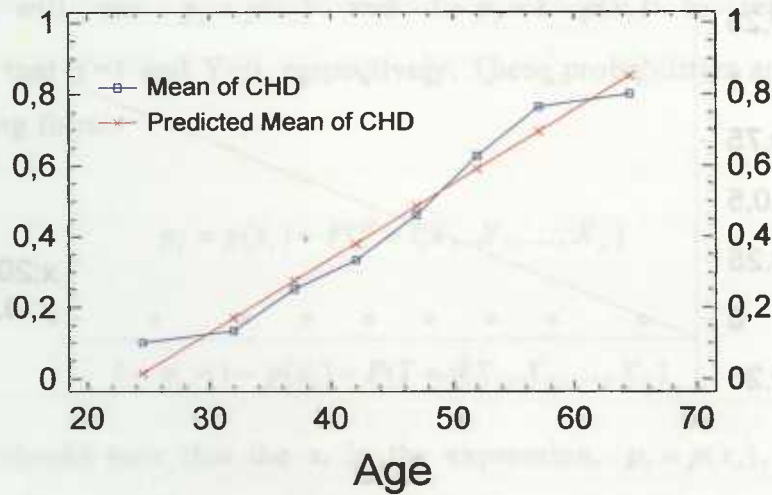


Figure 2.4 Plot of the Mean & Predicted Mean (O.L.S.) of Coronary Heart Disease By Age (Midpoint)

Examining the above graph we observe a deviation between the two lines. As at any linear model the i th observation is made up of the mean-predicted value plus residual noise, $y_i = \bar{y} + \varepsilon_i$. The quantity ε_i is called the error and expresses an observation's deviation from the conditional mean.

The most common assumption is that ε_i follow a normal distribution with mean zero and some variance σ^2 that is constant across levels of the independent variable/variables. But as could be seen at Figure 2.5 this is not the case if outcome values are dichotomous, zero or one. The standard error of the proportion p , $se(p) = \sqrt{\frac{p(1-p)}{n}}$, depends on p .



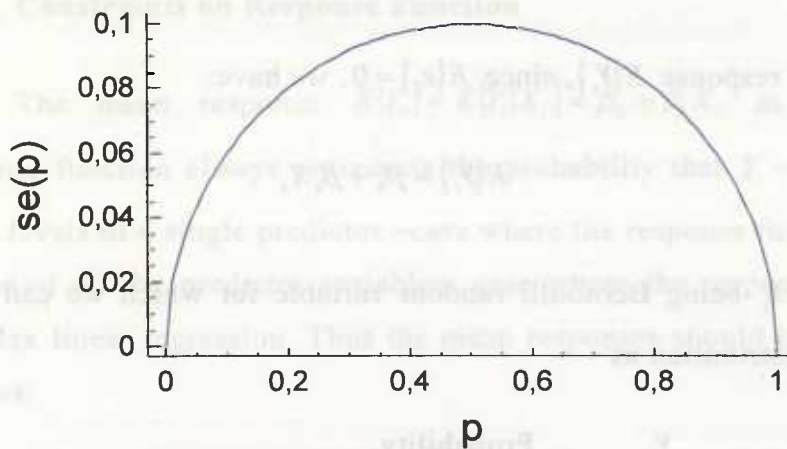


Figure 2.5 Plot of the proportion p versus standard error of p

So, if we use regression to analyze a binary outcome Y_i with one covariate X_i , $i = 1, \dots, n$ we encounter these problems:

- **Non-normal Error Terms.**

For a binary 0, 1 response variable, each error term $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$ can take on only two values:

$$\text{When } Y_i = 1: \varepsilon_i = 1 - \beta_0 - \beta_1 X_i$$

$$\text{When } Y_i = 0: \varepsilon_i = -\beta_0 - \beta_1 X_i$$

Clearly, normal error regression model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, which assumes that the ε_i are normally distributed, is not appropriate.

- **Non-Constant Error Variance**

Another problem with the error terms ε_i is that they do not have equal variances when the response variable is an indicator variable.

For the linear model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, with $Y_i = 0, 1$ we have:

$$\sigma^2(Y_i) = E\{(Y_i - E\{Y_i\})^2\} = (E\{Y_i\})(1 - E\{Y_i\}) \quad (2.2)$$

For the expected response $E\{Y_i\}$, since $E\{\varepsilon_i\} = 0$, we have:

$$E\{Y_i\} = \beta_0 + \beta_1 X_i \quad (2.3)$$

Consider Y_i being Bernoulli random variable for which we can state the probability distribution as

Y_i	Probability
1	$P(Y_i = 1) = p_i$
0	$P(Y_i = 0) = 1 - p_i$

By the definition of expected value of a random variable we obtain:

$$E\{Y_i\} = 1(p_i) + 0(1 - p_i) = p_i \quad (2.4)$$

Equating (2.3) and (2.4) we thus find:

$$E\{Y_i\} = \beta_0 + \beta_1 X_i = p_i.$$

Based on this result, we obtain from (2.2):

$$\begin{aligned} \sigma^2\{Y_i\} &= (1 - p_i)^2 p_i + (0 - p_i)^2 (1 - p_i) \\ &= p_i(1 - p_i) \end{aligned}$$

The variance of ε_i is the same as that of Y_i , since $\varepsilon_i = Y_i - p_i$ and p_i is a constant.

So,
$$\sigma^2\{\varepsilon_i\} = p_i(1 - p_i) = (\beta_0 + \beta_1 X_i)(1 - \beta_0 + \beta_1 X_i)$$

In other words $\sigma^2\{\varepsilon_i\}$ depends on X_i , which means that the error variances will differ at different levels of X and ordinary least squares will no longer be optimal.



- **Constraints on Response Function**

The mean response $E\{Y_i\} = E\{Y_i|X_i\} = \beta_0 + \beta_1 X_i$, as given by the response function always represents the probability that $Y_i = 1$ either for the given levels of a single predictor –case where the response function is a linear one or of all the predictor variables- case where the response function is a complex linear regression. Thus the mean responses should be constrained as follows:

$$0 \leq E\{Y\} = p \leq 1$$

The plot at Figure 2.4 shows that this mean approaches zero and 1 "gradually". The change in the $E(Y_i|X_i)$ per-unit change in X_i becomes progressively smaller as the conditional mean gets closer to zero or 1. The curve is said to be S-shaped and it is approximately linear except at the ends. This response function belongs to the "family" of sigmoidal response functions, which have asymptotes at 0 and 1 and thus automatically meet the constraints on $E\{Y\}$. Many response functions do not automatically possess this constraint. A linear response function may fall outside the constraint limits within the range of the predictor variable in the scope of the model.

The response functions plotted in Figure 2.6 are called logistic response functions and the relationship between p and X is of the form:

$$p_i = P(Y_i = 1) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} \quad (2.5)$$

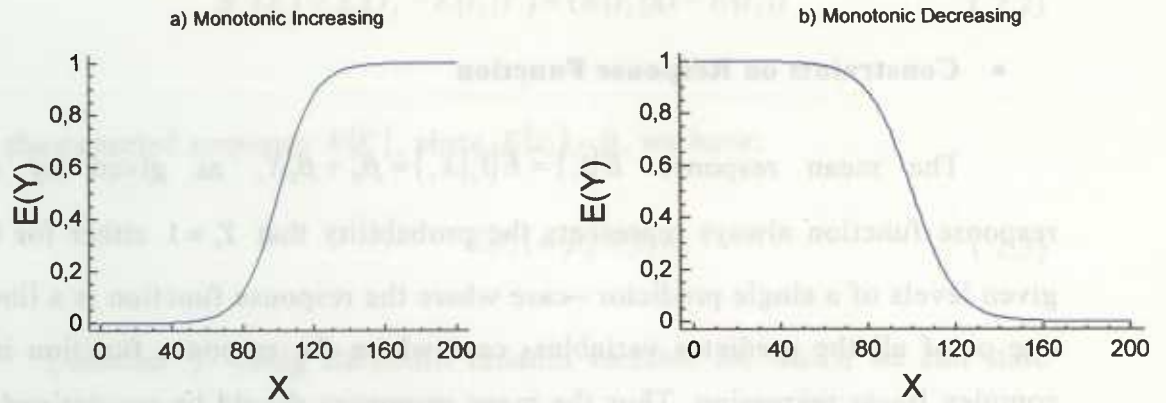


Figure 2.6 Increasing and Decreasing Logistic Plots

Since $E\{Y_i\} = p_i$, an equivalent form of (2.5) is given by:

$$E\{Y_i\} = [1 + \exp(-\beta_0 - \beta_1 X_i)]^{-1} \quad (2.6)$$

It is apparent that the variability is highest when p is close to 0.5 and decreases, as p gets closer to zero or one. Different choices of β give curves that are steeper (larger $|\beta_1|$) or flatter (smaller $|\beta_1|$), or shaped as an inverted S (negative β_1), case of monotonic decreasing logistic response function.

Among the three previous mentioned problems the problem of unequal error variances could be handled by the use of weighted least squares. In addition with large sample sizes the method of least squares provides estimators that are asymptotically normal under quite general conditions, even if the distribution of the error terms is far from normal. The difficulty created by the need of the model to give expected values between 0 and 1 is the most serious. As we would see below this problem can be solved by transformation.

2.3 Simple Logistic Regression

2.3.1 Fitting a Logistic Response Function

Let denote $p = E(Y|X)$ to represent the conditional mean of Y given X since the mean response is a probability when the response variable is a 0, 1 indicator variable.

Consider a transformed p' , the logit transformation of p , to be as follows:

$$p' = \left[\ln \frac{p}{1-p} \right] \quad (2.7)$$

where p is the observed proportion of responses in the i th group of observations.

The ratio $\frac{p}{1-p} = \frac{\text{probability of presence of characteristic}}{\text{probability of absence of characteristic}}$ is called the odds.

So, in our case the logits are as follows:

Age (Midpoint)	Observed	
	p	logit
24,5	0,100	-2,1972
32	0,133	-1,8718
37	0,250	-1,0986
42	0,333	-0,6931
47	0,462	-0,1542
52	0,625	0,5108
57	0,765	1,1787
64,5	0,800	1,3863

Table 2.4 Logits of Coronary Heart Disease for Age Midpoint

Plotting this new p' against age (midpoint) we see the following:

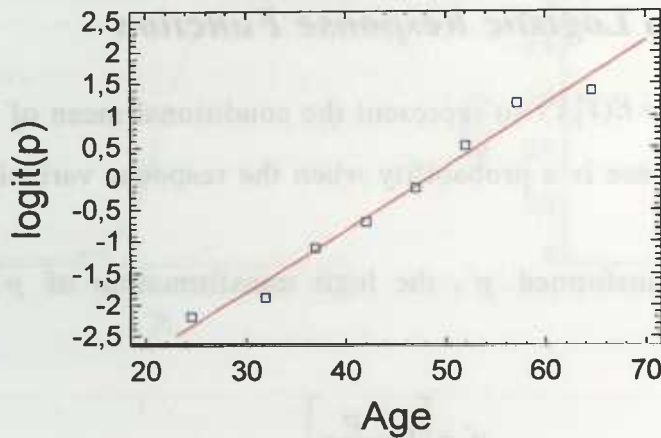


Figure 2.7 Plot of Logit(p) versus Age Midpoint

Based on the logit transformation and (2.3) we obtain the linear logistic model described by

$$p' = \text{logit}(p) = \beta_0 + \beta_1 X \tag{ 2.8}$$

which referred as the logit response function and p' is called the logit mean response. In our example this is the straight line in Figure 2.7 described by the equation:

$$p' = \text{logit}(p) = -5,310 + 0,111 \text{ Age (Midpoint)}$$

This equation will be discussed and explained further in details at next sections.

The logistic model shares a common feature with a more general class of linear models, that a function $g = g(\mu)$ of the mean of the response variable is assumed to be linearly related to the explanatory variables. Since the mean μ implicitly depends on the stochastic behavior of the response, and the explanatory variables are assumed to be fixed, the function g provides the link between the random (stochastic) component and the systematic (deterministic) component of the response variable Y . For this reason, Nelder and Wedderburn (1972) refer to $g(\mu)$ as a link function. One advantage of the logit function over other link functions is that differences on the logistic scale



are interpretable regardless of whether the data are sampled prospectively or retrospectively (McCullagh and Nelder, 1989).

All the intuition developed in regression is useful in logistic regression. Logistic regression uses regression models (the right hand side of the equation) to model a binary response (the left hand side of the equation).

The importance of this transformation is that p' has many of the desirable properties of a linear regression model. The logit, p' is linear in its parameters, may be continuous, and may range from $-\infty$ to $+\infty$., depending on the range of X.

The second important note concerns the conditional distribution of the outcome variable. In this situation we may express the value of the outcome variable given X as $Y = E\{Y|X\} + \varepsilon = p + \varepsilon$. Here the quantity ε may assume one of two possible values. If $Y=1$ then $\varepsilon = 1 - p$ with probability p , and if $Y=0$ then $\varepsilon = -p$ with probability $1 - p$. Thus, ε has a distribution with mean zero and variance equal to $p(1 - p)$. That is, the conditional distribution of the outcome variable follows a binomial distribution with probability given by the conditional mean, p .

Cox (1970) and Cox & Snell (1989) have proposed many distribution functions for use in the analysis of a dichotomous outcome variable. There are two primary reasons for choosing the logistic distribution. These are:

from a mathematical point of view, it is an extremely flexible and easily used function. A curvilinear response function of almost the same shape as logistic function is obtained by transforming p by means of the cumulative normal distribution, the probit transformation. But the probit regression model is less flexible than the logistic regression model since it cannot be readily extended to more than one predictor variable. Additionally, the complementary log-log transformation of the probability p given by $\ln[-\ln(1 - p)]$, another curvilinear response function of almost the same shape as the logistic, unlike the logistic and probit transformation (Aldrich et.al., 1984) is not symmetric about $p = 0,5$.

it lends itself to a biologically meaningful interpretation.



In summary, we have seen that in a regression analysis when the outcome variable is dichotomous:

(1) The conditional mean of the regression equation must be formulated to be bounded between zero and 1. We have stated that the logistic regression model, p given in equation (2.8), satisfies this constraint.

(2) The binomial, not the normal, distribution describes the distribution of the errors and will be the statistical distribution upon which the analysis is based.

(3) The principles that guide an analysis using linear regression will also guide us in logistic regression.

2.3.2 *Simple Logistic Regression Model*

We state the simple logistic regression model in the form:

$$p' = \text{logit}(p) = \beta_0 + \beta_1 X$$

Equivalently,

$$p = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

With this functional form:

- if we let $\beta_0 + \beta_1 X = 0$, then $p = 0,5$ implying a 50/50 chance for the event to occur
- as $\beta_0 + \beta_1 X$ gets really big (approaches infinity), p approaches 1 , indicating a high likelihood for the event to occur.
- as $\beta_0 + \beta_1 X$ gets really small (negative infinity), p approaches 0.

2.3.3 *Likelihood Function*

A loss function is a measure of fit between a mathematical model of data and the actual data. The parameters of the model are chosen to minimize the badness-of-fit or to maximize the goodness-of-fit of the model to the data. With least squares, the sum of squared errors is minimized. With some



models, like the logistic curve, there is no mathematical solution that will produce least squares estimates of the parameters. For many of these models, the loss function chosen is called maximum likelihood. A likelihood is a conditional probability ($P(Y|X)$), the probability of Y given X). Specifically it is the probability that the observed values of the dependent may be predicted from the observed values of the independents. We can pick the parameters of the model (a and b of the logistic curve) at random or by trial-and-error and then compute the likelihood of the data given those parameters. We will choose as our parameters, those that result in the greatest likelihood computed. The estimates are called maximum likelihood because the parameters are chosen to maximize the likelihood (conditional probability of the data given parameter estimates) of the sample data. The techniques actually employed to find the maximum likelihood estimates fall under the general label numerical analysis. There are several methods of numerical analysis, but they all follow a similar series of steps. First, the computer starts with an initial arbitrary guess of what the logit coefficients should be. Then it computes the likelihood of the data given these parameter estimates. Then it determines the direction and size change in the logit coefficients, improves the parameter estimates slightly and recalculates the likelihood of the data. The process is repeated until convergence is reached, that is when the parameter estimates do not change much (usually a change .01 or .001 is small enough to tell the computer to stop). Sometimes the computer stops after a certain number of tries or iterations. However this usually indicates a problem in estimation.

The logistic regression model indirectly models the response variable based on probabilities associated with the values of Y. For a set of observations in the data (x_i, y_i) the contribution to the likelihood function is $p(x_i)^{y_i} (1 - p(x_i))^{1 - y_i}$, where $y_i = 1$ and $1 - p(x_i)$, where $y_i = 0$. The following equation results for the contribution (call it $\zeta(x_i)$) to the likelihood function for the observation, (x_i, y_i) :

$$\zeta(x_i) = p(x_i)^{y_i} (1 - p(x_i))^{1 - y_i}, \quad y_i = 0, 1; i = 1, \dots, n \quad (2.9)$$



This equation accounts for only one set of observations. Since the observations are assumed to be independent, the likelihood function is obtained as the product of the terms given in (2.9) as

$$l(\beta_0, \beta_1) = \prod_{i=1}^n \zeta(x_i) = \prod_{i=1}^n p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}$$

where $l(\beta_0, \beta_1)$ is the likelihood function of the parameters β_0, β_1 .

It will be easier to find the maximum likelihood estimates (Silverstone,1957) by working with the logarithm of the likelihood function, having the log likelihood defined as:

$$\begin{aligned} L(\beta_0, \beta_1) &= \ln l(\beta_0, \beta_1) = \ln \prod_{i=1}^n p(x_i)^{y_i} (1 - p(x_i))^{1-y_i} \\ &= \sum_{i=1}^n [y_i \ln[p(x_i)] + (1 - y_i) \ln[1 - p(x_i)]] \end{aligned} \tag{2.10}$$

We employ the techniques of calculus to determine the value of β_0, β_1 that maximizes $L(\beta_0, \beta_1)$. This is done by differentiating (2.10) with respect to β_0, β_1 and setting the resulting derivatives equal to zero. These equations are called likelihood equations and there are of the following form:

$$\sum_{i=1}^n [y_i - p(x_i)] = 0, \text{ for the intercept } \beta_0$$

and

$$\sum_{i=1}^n x_i [y_i - p(x_i)] = 0 \text{ for the coefficient } \beta_1$$

These expressions are nonlinear in β_0 and β_1 and thus require special methods for solution. The solution, the maximum likelihood estimate of



β_0, β_1 , denoted $\hat{\beta}_0, \hat{\beta}_1$ ¹ could be solved for by using iterative methods available into logistic regression software (McCullagh and Nelder, 1983).

As we have discussed in section 2.3.1 use of logistic regression with continuous variable Age (midpoint) as the independent variable produces, the results of the Table 2.5.

Variable	B	S.E.	Wald	df	Sig.	Exp(B)
AGE	0,111	0,024	21,254	1	,000	1,117
Constant	-5,310	1,134	21,935	1	,000	,005

-2log-likelihood Intercept Only=136.664

-2log-likelihood Intercept and Covariates = 107,354

Table 2.5 Results of fitting the Logistic Regression Model to the Data in Table 2.3

Logistic regression fits an intercept and a slope. The maximum likelihood estimates of β_0 and β_1 are thus seen to be $\hat{\beta}_0 = -5.310$ and $\hat{\beta}_1 = 0.111$.

The fitted values are given by the equation:

$$p = \frac{1}{1 + e^{-\text{logit}(p)}} = \frac{1}{1 + e^{-5.310 + 0.111\text{Age}}}$$

and the estimated logit , is given by the equation

$$p' = \text{logit}(p) = -5,310 + 0,111 \text{ Age (Midpoint)}$$

The -2log-likelihood given in the above table is the value of equation (2.10) multiplied by minus twice, a result usually printed out by logistic regression software.

¹The symbol ^ will denote the maximum likelihood estimate



2.4 Multiple Logistic Regression

2.4.1 Multiple Logistic Regression Model

Generalising the logistic model to the case of more than one independent variable we have the so-called multivariate logistic regression model

$$p' = \text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_j X_j = \beta_0 + \sum_{j=1}^n \beta_j X_j$$

Equivalently,

$$p = \frac{\exp(\beta_0 + \sum_{j=1}^n \beta_j X_j)}{1 + \exp(\beta_0 + \sum_{j=1}^n \beta_j X_j)}$$

With this functional form:

- if we let $\beta_0 + \sum_{j=1}^n \beta_j X_j = 0$, then $p = 0,5$ implying a 50/50 chance for the event to occur.
- as $\beta_0 + \sum_{j=1}^n \beta_j X_j$ gets really big (approaches infinity), p approaches 1, indicating a high likelihood for the event to occur.
- as $\beta_0 + \sum_{j=1}^n \beta_j X_j$ gets really small (negative infinity), p approaches 0.

2.4.2 Likelihood Function

Similarly with the case of the simple logistic regression the likelihood function is obtained as the product of the terms given in (2.9) as

$$l(B) = \prod_{i=1}^n \zeta(x_i) = \prod_{i=1}^n p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}$$

where B is the collection of parameters $\beta_0, \beta_1, \dots, \beta_k$ and $l(B)$ is the likelihood function of B .

The logarithm of the likelihood function, the log likelihood of B is



$$\begin{aligned}
L(B) = \ln l(B) &= \ln \prod_{i=1}^n p(x_i)^{y_i} (1-p(x_i))^{1-y_i} \\
&= \sum_{i=1}^n \left[y_i \ln[p(x_i)] + (1-y_i) \ln[1-p(x_i)] \right]
\end{aligned} \tag{2.11}$$

Differentiate (2.11) with respect to $\beta_0, \beta_1, \dots, \beta_j$ and setting the resulting derivatives equal to zero we determine the value of B that maximizes $L(B)$. These equations are called likelihood equations and there are $j+1$ equations of the following form:

$$\sum_{i=1}^n [y_i - p(x_i)] = 0, \text{ for the intercept } \beta_0$$

and

$$\sum_{i=1}^n x_{ik} [y_i - p(x_i)] = 0, \text{ for } k = 1, 2, \dots, j \text{ for the parameters } \beta_1, \dots, \beta_j$$

The solution to the likelihood equations is the maximum likelihood estimate B. As we have discussed before using computer programs can solve for the solution.

Following the fitting of the model we should begin to evaluate its adequacy.

2.5 Significance of the Model

In practice, several different measures exist for determining the significance, or goodness of fit, of a logistic regression model. According to Hosmer et. al (1991) mention assessing the fit of any logistic regression model it is of great importance . These measures include the G statistic and Hosmer-Lemeshow statistic. In a theoretical sense these measures are equivalent.

2.5.1 Deviance

Before proceeding we have to introduce deviance between two models (Mc Cullagh and Nelder, 1983), an analogous to the residual sum of squares from a linear model, which is defined as:

$$\begin{aligned}
 D &= -2 \ln(\text{likelihood ratio}) = \\
 &= -2 \ln \frac{\text{likelihood of the current model}}{\text{likelihood of the saturated model}} = \quad \quad \quad (2.12) \\
 &= -2 \ln \frac{L_1}{L_s}
 \end{aligned}$$

The reason for using minus twice its natural logarithm (\ln) is to obtain a quantity whose distribution is known and thus can be used for hypothesis testing purposes. Using (2.10) equation (2.12) becomes

$$D = -2 \sum_{i=1}^n y_i \ln \left(\frac{\hat{p}_i}{y_i} \right) + (1 - y_i) \ln \left(\frac{1 - \hat{p}_i}{1 - y_i} \right)$$

where \hat{p}_i the maximum likelihood estimate of the probability $p_i = p(x_i)$.

As a model's ability to predict outcomes improves, the deviance falls. Poorly-fitting models have higher deviance. If a model perfectly predicts outcomes, the deviance will be 0. This is analogous to the situation in linear regression, where the residual sum of squares falls to 0 if the model predicts the values of the dependent variable perfectly.

2.5.2 G statistic /Likelihood ratio test

The most common assessment of model fit in logistic regression is the goodness-of-fit test (G statistic also refer to as the Likelihood ratio test). It is a close analogue to the F statistic for linear regression.

As usual, the likelihood ratio test is based on comparing the quality of the fit without any predictors to the quality of the fit using predictors, as

² A saturated model is one that contains as many parameters as there are data points.

measured by the difference in the deviance for the models with and without predictors, as:

$$G = D(\text{for the reduced model}) - D(\text{for the full model}) \quad (2.13)$$

where reduced is the model without the variable(s) and full the model with the variable(s).

Because the likelihood of the saturated model is common to both values of D being differenced to compute G, it can be expressed as

$$\begin{aligned} G &= -2 \ln \frac{(\text{likelihood of the reduced model})}{(\text{likelihood of the full model})} \\ &= -2 \ln \frac{L_R}{L_F} \end{aligned}$$

Based on G statistic we could determine the overall significance for a model by subtracting the deviance for the model and the deviance for the intercept-only model. The larger the difference, the greater the evidence that the model is significant. The G statistic follows a chi-squared distribution with k- 1 degrees of freedom, where k is the number of parameters in the model.

In our example to test the overall significance of the model fitted with an intercept and a slope, we use results of Table 2.5 and equation (2.13). The -2log likelihood for the model containing only a constant term is 136.664. Fitting a model containing the independent variable Age, along with the constant term results in a -2log likelihood of 107,354. This statistic measures how poorly the model predicts the decisions -- the smaller the statistic the better the model. Adding Age reduced the -2 Log Likelihood statistic by 29,31. This result along with the associated p-value for the chisquare distribution could be obtained form most software packages. Since $P(x^2(1) > 29,31) < 0,01$, the null hypothesis that adding the age variable to the model has not significantly increased our ability to predict the decisions made by our subjects could not be rejected.



In some cases, the traditional goodness-of-fit test (G or the likelihood ratio test) may not be the best assessment of model fit. Some simulations suggest that the deviance statistic is not distributed as chi-square when the data are sparse. The term “sparse” refers to a circumstance in which there are few observed values (and therefore few expected values) in the cells formed by crossing all of the values of all of the predictors.

2.5.3 Hosmer-Lemeshow test

The Hosmer-Lemeshow (Hosmer & Lemeshow, 2000) fit test is designed to correct for this (when there are continuous predictors). The use of this test is not recommend when there is a small sample size (less than 400). In this test the observations are ordered by the fitted probabilities \hat{p}_i . Then the data are formed into G groups (usually the default value for G=10 groups based on percentile ranks) of roughly equal size based on the smallest $1/G$ of the fitted probabilities, the next $1/G$, and so on. For each group, the expected number of success in the group is the sum of the fitted probabilities for that group (Hosmer and Lemeshow, 1989) i.e.

$$\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{p}(x_i)$$

A Pearson-type statistic can be constructed that compares the observed number of success in each group to the expected number, as:

$$HL = \sum_{j=1}^G \frac{(O_j - n_j \bar{p}_j)^2}{n_j \bar{p}_j (1 - \bar{p}_j)}$$

, where O_j is the number of positive observations in group j , \bar{p}_j is the model’s average predicted value in group j and n_j is the size of the group. Lower values (and nonsignificance) indicate a good fit to the data and, therefore, good overall model fit.



In practical terms, the classical approach to model selection usually involves maximising the likelihood function associated with each competing model and then calculating the corresponding criteria value(s). Model choice in logistic regression is based upon the likelihood function, but with the addition of a penalty term for the number of parameters (Burnham and Anderson, 2002). Various criteria such as the Akaike's information criterion (AIC) (Akaike, 1974; Parzen, Tanabe, Kitagawa, 1998; Burnham, Anderson, 1998) or Bayesian information criterion (BIC) (Schwarz, 1976; Raftery, 1995) have been proposed when multiple models need to be compared.

AIC and BIC are defined as:

$$AIC = -2\ln(\text{likelihood}) + 2K$$

$$BIC = -2\ln(\text{likelihood}) + \ln(N)K$$

with K the number of estimable parameters in the model and N the sample size.

The actual values of AIC and BIC are somewhat immaterial; it is the AIC and BIC value for a particular model relative to the AIC and BIC values for the other models in the candidate set that is important. The model with the lowest value of AIC or BIC selected as the preferred model.

2.5.4 *Classification table*

The classification table can be used to evaluate the predictive accuracy of the logistic regression model. It is created by comparing the model-generated predicted probability of event outcome for each observation under analysis to the value of the dependent or response variable. The prior probability of event occurrence is used as criterion to assign an observation as having been predicted as either "event" or "non event".

As the basic idea of any diagnostic test interpretation is to calculate the probability a patient has a disease under consideration given a certain result, a 2x2 table is employed for this purpose. The table is labeled with the test results on the left side and the disease status on the top.



	Disease Present	Disease Absent	
Test Positive	True Positives	False Positives	Total Positive
Test Negative	False Negatives	True Negatives	Total Negatives
	Total with Disease	Total without Disease	Overall Total

Table 2.6 Interpretation of a diagnostic test based on test results and disease status

Classification table is the result of cross-classifying the binary dependent as

$$y = \begin{cases} 0 & \text{if } \hat{p} < c \\ 1 & \text{if } \hat{p} \geq c \end{cases}$$

with c , the classification threshold/cut-point, often taken to be equal to 0.5.

Use 0.5 as a cutoff means that if \hat{p} for a new observation is greater than 0.5, its predicted outcome is $y=1$. Otherwise, it's $y=0$. This approach is reasonable when

- (a) it is equally likely in the population of interest that the outcomes 0 and 1 will occur, and
- (b) the costs of incorrectly predicting 0 and 1 are approximately the same.

Sometimes it is preferred to find the best cutoff for the data set on which the multiple logistic regression model is based. Using this approach, we evaluate different cutoff values and for each cutoff value, calculate the proportion of observations that are incorrectly predicted. We would then select the cutoff value that minimizes the proportion of incorrectly predicted outcomes. A useful tool for this approach is ROC analysis, which we discuss below.

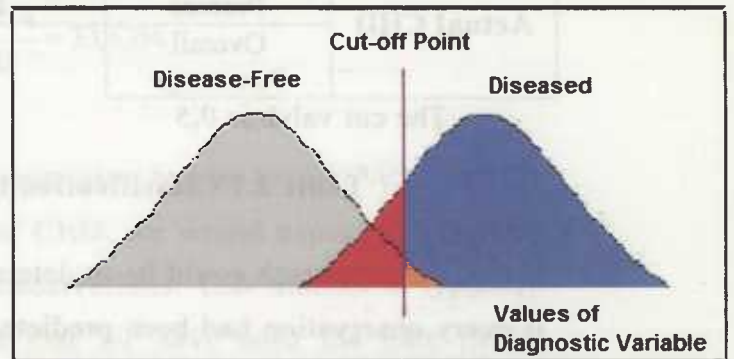
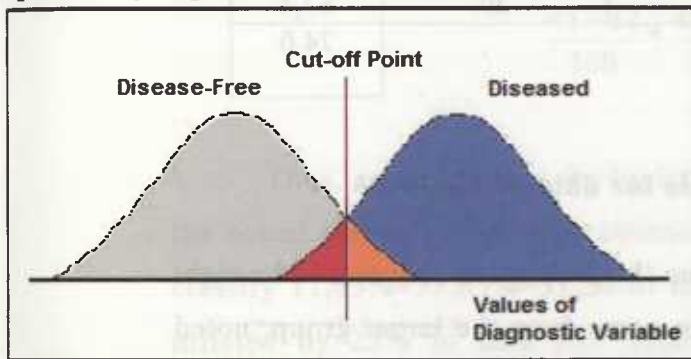
We should mention that the selection of the cut-point, which determines which estimated event probabilities are considered to predict the



event, is very important since each cut-point generates a different classification table, as we could be seen visually at the graph below.

Sensitivity and Specificity Equal

**Increased Specificity
Decreased Sensitivity**



The distribution of a test result described by two Gaussian Curves

Figure 2.8 Relationship between *Sensitivity - Specificity* and *Cut-off Point*

One serious question about classification table is if the observations are correctly classified to the appropriate group. In other words using the logistic model is better in classification that we would expect just by random chance? The answer to this question is not straightforward since the same data are used to build the model and evaluate its ability to do predictions. For this reason, the observed proportion correctly classified can be expected to be biased upwards compared to the situation where the model is applied to completely new data.

The solution to this problem could be to get some new data and apply the fitted model to them to see how well it does, that is to external validate the model on new data. By this way it could be examined the ability of the predictive model to maintain accuracy when applied to new cases and settings different from those on which the models were developed.

Additionally, internal validation methods such as variants of split-sample, cross-validation and bootstrapping methods are available that aim to provide a more accurate estimate of model performance in new subjects.

If no new data are available, two diagnostics have been suggested that can be helpful.



The classification table for data of Table 2.3 has the form:

		Predicted CHD		
		Absent	Present	
Actual CHD	Absent	45	12	78,9
	Present	14	29	67,4
	Overall Percentage			74,0

The cut value is 0,5

Table 2.7 Classification table for data of Table 2.3

One approach could be to determine the proportion that would be right if every observation had been predicted to come from the larger group, noted C_{max} , since performance using such a simplistic strategy would have to be considered a lower bound for performance overall. For the above classification table we have:

$$C_{max} = \max\left(\frac{45 + 12}{100}, \frac{14 + 29}{100}\right) = \max\left(\frac{57}{100}, \frac{43}{100}\right) = 57\%$$

The observed 74% of the observations (74 out of 100) correctly classified is considerably larger than C_{max} , supporting the usefulness of the logistic regression.

Another approach has to do with the argument that if the logistic regression had no power to make predictions, the actual result would be independent of the predicted result. That is,

$$P(\text{actual result CHD and predicted result CHD}) = P(\text{actual result CHD}) \times P(\text{predicted CHD})$$

The right side of the above equation can be estimated using the marginal probabilities from the classification table, yielding:

$$P(\text{actual result CHD and predicted result CHD}) = \frac{14 + 29}{100} * \frac{12 + 29}{100} = 17,63\%$$



That is we would expect to get 17,63% of the observations correctly classified as success just by random chance. A similar calculation for the failures gives

P (actual result a failure and predicted result a failure)=

$$\frac{45 + 12}{100} * \frac{45 + 14}{100} = 33,63\%$$

Thus, assuming that the logistic regression had no predictive power for the actual result, presence or absence of CHD, we would expect to correctly classify 17,63%+33,63%=51,26 of the observations. This number is typically inflated by 25% to take into account that we have used the data twice, resulting the C_{pro} measure, as:

$$C_{pro}=1,25*51,26\%=64,075\%$$

The observed 74% is considerably higher than C_{pro} , which would support the usefulness of the logistic regression as a classifier.

The accuracy of the classification is measured by its sensitivity (the ability to predict an *event* correctly) and specificity (the ability to predict a *nonevent* correctly). Sensitivity is the proportion of *event* responses that were predicted to be *events*. Specificity is the proportion of *nonevent* responses that were predicted to be *nonevents*.

For the discussed example specificity is equal to $29/43=67,44\%$ and sensitivity is equal to $45/57=78,95\%$. Three other conditional probabilities are: *false positive rate*, *false negative rate*, and *rate of correct classification* or *overall classification rate*. The false positive rate is the proportion of predicted *event* responses that were observed as *nonevents*. The false negative rate is the proportion of predicted *nonevent* responses that were observed as *events*. These three probabilities based on the classification Table 2.7, are equal to $12/57=21,05\%$ - *false positive rate* - $14/43=32,56\%$ -*false negative rate* and $74/100=74\%$ -*overall classification rate*, respectively.

Receiver-operating characteristic (ROC) curve (Hanley and McNeil, 1982) plots the false-positive rate (1-specificity) on the x-axis and the true-positive rate (sensitivity) on the y-axis.

The diagonal line represents chance. A curve that is well above the diagonal line means that an indicator is accurate. The closer an ROC curve is to the upper left corner of the graph (as true-positive rate approaches 1 and false-positive rate approaches 0), the larger the area under the curve, and more accurate the prediction model. Each point on the curve represents a cut-off probability. A lower cutoff typically gives more false positive. A high cutoff gives more false negatives, a low sensitivity, and a high specificity.

Graphing an ROC curve gives a good visual representation of accuracy, but often a numerical measure of accuracy is useful as well. Several different measures of accuracy have been developed, but the easiest one is the area under the ROC curve. Area under the ROC curve (AUC), c statistic, is a measure of a model's discriminatory power, which varies between 0.5 and 1.

The values of the AUC have an intuitive interpretation.

An area of	Test is:
0.5-0.60	not better than chance
0.6-0.70	poor
0.7-0.80	fair not perfect but useful
0.8-0.90	good
0.90-1.0	perfect

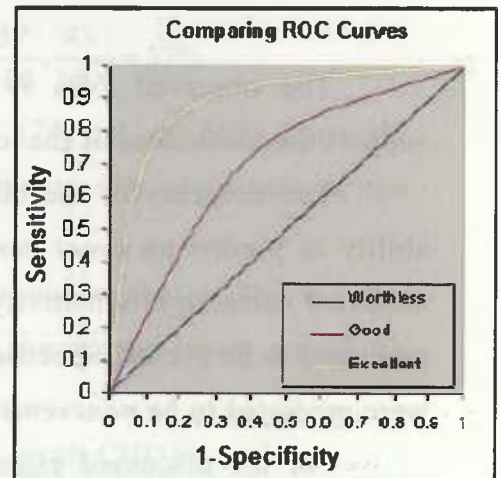
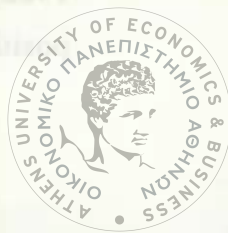


Figure 2.9 Comparing ROC curves

The AUC can be interpreted as the probability that a randomly selected "event" will be regarded with greater suspicion (in terms of its rating or continuous measurement) than a randomly selected "non-event". An AUC of 0,50 means that the diagnostic accuracy in question is equivalent to that which would be obtained by flipping a coin (i.e., random chance).



2.6 Parameter Significance

After estimating the coefficients, our first look at the fitted model commonly concerns an assessment of the significance of the variables in the model. This usually involves formulation and testing of a statistical hypothesis to determine whether the independent variables in the model are significant related to the outcome variable.

2.6.1 *Wald test*

For testing hypotheses about individual coefficients, the Wald test is common. It is analogous to the t-test for individual coefficients in linear regression. The Wald test is commonly used to test the null hypothesis that a coefficient is equal to 0, against the alternative that it is not. In other words, it is used to test

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

The test statistic, a z-statistic, is quite simple:

$$z_j = \frac{\beta_j}{\text{s.e.}(\beta_j)} \quad (2.14)$$

The ratio of B to S.E., squared, equals the Wald statistic seen to Table 2.5. The Wald statistic is significant (i.e., less than 0.05) indicating the covariate Age is useful to the model.

An $(1-\alpha)$ confidence interval for β_j which could help in statistical inference for one model parameter is:

$$\beta_j \pm z_{1-\alpha/2} \text{s.e.}(\beta_j)$$

Confidence intervals whose endpoints do not contain zero indicate a relationship between the predictor x_j and the response after adjusting for any other predictor variables in the model.

To test the null hypothesis that a coefficient is equal to some value k , the test statistic is a simple extension of this:

$$z_j = \frac{\beta_j - k}{\text{s.e.}(\beta_j)}$$

Menard (2002) warns that for large logit coefficients, standard error is inflated, lowering the Wald statistic and leading to Type II errors. That is, there is a flaw in the Wald statistic such that very large effects may lead to large standard errors and small Wald chi-square values. For models with large logit coefficients or when dummy variables are involved, it is better to test the difference in model chi-squares for the model with the independent and the model without that independent, or to consult the Likelihood test discussed below. Also note that the Wald statistic is sensitive to violations of the large-sample assumption of logistic regression.

Hauck & Donner (1977) examined the performance of the Wald test and found that it behaved in an aberrant manner, often failing to reject when the coefficient was significant. They recommended that the likelihood ratio test be used. Jennings (Jennings, 1986) has also looked at the adequacy of inferences in logistic regression based on Wald statistics. His conclusions are similar to those of Hauck & Donner.

2.6.2 *Likelihood ratio test*

The likelihood ratio test in its most basic form, it can also be used to test the hypothesis that all the coefficients in a model are all equal to zero:

$$H_0 : \beta_0 = \beta_1 = \dots = \beta_k = 0$$

If we want to test whether a subset consisting of q coefficients in a model are all equal to zero, the test statistic is the same, except that we use the likelihood from the model without the coefficients and the likelihood from the model with them. This statistic has q degrees of freedom.

In contrast with the case for linear regression, the Wald test for a single coefficient can yield slightly different results from a likelihood ratio

test for the same coefficient. The likelihood-ratio test is preferred, but not always convenient.

There are several measures intended to mimic the R-squared analysis, but none of them are an R-squared. This is because an R^2 measure seeks to make a statement about the "percent of variance explained," but the variance of a dichotomous or categorical dependent variable depends on the frequency distribution of that variable. For a dichotomous dependent variable, for instance, variance is at a maximum for a 50-50 split and the more lopsided the split, the lower the variance. This means that R-squared measures for logistic regressions with differing marginal distributions of their respective dependent variables cannot be compared directly, and comparison of logistic R-squared measures with R^2 from OLS regression is also problematic. Nonetheless, a number of logistic R-squared measures have been proposed.

2.6.3 *Logistic R-squared measures*

Based on the deviance, it is possible to construct an analogue to R^2 for logistic regression, commonly referred to as the Pseudo- R^2 . It is also called McFadden R^2 which is defined as 1 minus the ratio of the likelihood of the full model over the likelihood of the reduced model.

$$\text{McFadden } pseudo - R^2 = 1 - \frac{\ln(\text{likelihood of the full model})}{\ln(\text{likelihood of the reduced model})} = 1 - \frac{\ln L_F}{n L_R}$$

We should note here that models must be nested – the variables in one are a subset of those in the other-in order to be compared. McFadden R^2 does not have a sampling distribution and is not testable; it is simply a descriptive measure of fit. The Pseudo R^2 is a scalar measure which varies between 0 and (somewhat close to) 1. This statistic will equal zero if all coefficients are zero whereas it will come close to 1 if the model is very good. According to McFadden (McFadden, 1979) a rule of thumb for an excellent fit could be

$$0,20 \leq \text{McFadden Pseudo} - R^2 \leq 0,40$$

The Maddala/Cox-Snell R^2 and Cragg-Uhler/Nagelkerke R^2 are two other attempts to provide a logistic analogy to R^2 in OLS regression based on

log-likelihoods (Magee, 1990). The Maddala (Maddala, 1983)/ Cox and Snell (Cox & Snell, 1989) statistic seems to come fairly close to pseudo R^2 . It uses a formula that is equivalent to

$$\text{Maddala/Cox - Snell } R^2 = 1 - \left(\frac{L_0}{L_A} \right)^{\frac{2}{n}}$$

where L_A likelihood of the alternative model and n sample size. As it takes into account sample size, yet cannot achieve a maximum value of 1.

The Cragg –Uhler (Cragg & Uhler, 1970) / Nagelkerke R^2 (Nagelkerke, 1991) adjusts the Maddala/Cox-Snell R^2 to achieve the maximum value of 1.

$$\text{Cragg – Uhler / Nagelkerke } R^2 = \frac{(L_A)^{\frac{2}{n}} - (L_0)^{\frac{2}{n}}}{1 - (L_0)^{\frac{2}{n}}}$$

2.7 Regression Diagnostics

As in the case in any regression model fitting, it is important to check assumptions and assess the adequacy of a logistic regression fit.

McCullagh and Nelder (1989) caution against the use of the deviance alone to assess model fit. However a deviance that is approximately equal to its degrees of freedom is a possible indication of a good model fit.

2.7.1 Residuals

In ordinary least regression, we have several types of residuals and influence measures that help us understand how each observation behaves in the model, such as if the observation is too far away from the rest of the observations, or if the observation has too much leverage on the regression line. Similar techniques have been developed in logistic regression (Jennings, 1986; Lindsey ,1997 ; Hastie and Pregibon 1992). There are two types of residuals- differences between observed and fitted values- in common use.

Pearson residual is defined to be the standardized difference between the observed frequency and the predicted frequency. It measures the relative deviations between the observed and fitted values.

$$\text{Pearson } e_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i(n_i - \hat{y}_i)/n_i}}$$

Deviance residual is another type of residual. It measures the disagreement between the maximum of the observed and the fitted log-likelihood functions. Since the logistic regression uses the maximum likelihood principle, the goal in logistic regression is to minimize the sum of deviance residuals. Therefore, this residual is parallel to the raw residual in ordinary least square regression, where the goal is to minimize the sum of squared residuals.

$$\text{Deviance } e_i = \sqrt{2 \left[y_i \ln \left(\frac{y_i}{\hat{y}_i} \right) + (n_i - y_i) \ln \left(\frac{n_i - y_i}{n_i - \hat{y}_i} \right) \right]}$$

The residuals defined so far are not fully standardized. They take into account the fact that different observations have different variances, but they make no allowance for additional variation arising from estimation of the parameters, in the way studentized residuals in classical linear models do.

Pregibon (1981) has extended to logit models some of the standard regression diagnostics. A key in this development is the weighted *hat* matrix

$$H = W^{1/2} X(X'WX)^{-1} X'W^{1/2},$$

where W is a diagonal matrix of weights with entries $w_i = \hat{y}_i(n_i - \hat{y}_i)/n_i$, evaluated at the maximum likelihood estimation and X the model matrix.

Using this expression it can be shown that the variance of the raw residual is, to a first-order approximation,

$$\text{Var}(y_i - \hat{y}_i) \approx (1 - h_{ii})\text{Var}(y_i)$$

where h_{ii} is the leverage or diagonal element of the weight hat matrix. Thus, an internally studentized residual can be obtained dividing the Pearson residual by the square root of $1 - h_{ii}$, to obtain

$$\text{Pregibon } e_i = \frac{\text{Pearson } e_i}{\sqrt{(1 - h_{ii})}} = \frac{y_i - \hat{y}_i}{\sqrt{(1 - h_{ii})\hat{y}_i(n_i - \hat{y}_i)/n_i}}$$



These three statistics, Pearson residual, deviance residual and Pregibon leverage are considered to be the three basic building blocks for regression diagnostics. A good way of looking at them is to graph them against either the predicted probabilities or simply case numbers (Landwehr et. al. 1984). So, we have two types of plots using these statistics, the plots of the statistics against the predicted values and the plots of these statistics against the index id (it is therefore also called index plot). These two types of plots basically convey the same information. They are useful for assessing lack of fit and conclude if we have a reasonable fitting model. Besides residuals may be plotted to detect outliers visually.

2.7.2 Change in Chisquare (DIFCHISQ) or Deviance (DIFDEV) goodness of fit statistics

DIFDEV and DIFCHISQ are diagnostics for detecting ill-fitted observations; in other words, observations that contribute heavily to the disagreement between the data and the predicted values of the fitted model. DIFDEV is the change in the deviance due to deleting an individual observation while DIFCHISQ is the change in the Pearson chi-square statistic for the same deletion. By using the one-step estimate, DIFDEV and DIFCHISQ for the j th observation are computed as:

$$DIFDEV = d_j^2 + \bar{C}_j \text{ and } DIFCHISQ = \bar{C}_j / h_{jj} .$$

It worth noticing that these statistics are only one-step approximation of the difference, not quite exact the difference, since it would be computationally too extensive to obtain exact difference for every observation.





2.8 Interpretation of the Fitted equation

As we have discussed before, in setting up the logistic regression model, it is assumed that the outcome variable is a linear combination of a set of predictors. For outcome variable Y^c and a set of n predictor variables, X_1, X_2, \dots, X_n we have the following:

$$p' = \text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_j X_j = \beta_0 + \sum_{j=1}^n \beta_j X_j$$

where $\beta_0, \beta_1, \dots, \beta_j$ are the coefficients of the regression equation:

Because of the complicated form of the regression equation the regression coefficients are not as easy to interpret as in linear regression. The coefficient β_j could no longer represent "the change in Y with one unit change in X_j ". Instead, we have to translate using the exponent function. And, as it turns out, we have a type of "coefficient" which is called the odds ratio.

So, the logistic regression coefficients show the change (increase when $\beta_i > 0$, decrease when $\beta_i < 0$) in the predicted logit of having the characteristic of interest for a one-unit change in the independent variables.

When the independent variables are dichotomous variables then the closer the odds ratio is to 1.0 the more the categories of dichotomous variable are independent of the dependent variable. Odds ratio below 1.0 indicate the reference category is associated with greater odds of getting "1" on the independent variable.

By taking the exponential of both sides of the regression equation as given above, the equation can be rewritten as:

$$\text{odds} = \frac{p}{1-p} = e^{\beta_0} \times e^{\beta_1 X_1} \times e^{\beta_2 X_2} \times \dots \times e^{\beta_j X_j}$$

This mean that when a variable X_j increases by 1 unit, with all other factors remaining unchanged, then the odds will increase by a factor e^{β_j} . Here there is the only difference in interpretation for multiple logistic regression



and simple logistic regression. The estimated odds ratio for the predictor variable X_j in the first case assumes that all the other predictor variables are held constant. This factor e^{β_j} is the odds ratio (O.R.) of success relative to failure for the independent variable X_j and it gives the relative amount by which the odds of the outcome increase (O.R. greater than 1) or decrease (O.R. less than 1) when the value of the independent variable is increased by 1 units.

To see this (Neter et. al.), we consider the value of the fitted logit response function (2.8)

at X_i as
$$p'_i = \text{logit}(p_i) = \beta_0 + \beta_1 X_i$$

and at X_{i+1} as
$$p'_{i+1} = \text{logit}(p_{i+1}) = \beta_0 + \beta_1 X_{i+1}$$

The difference between the two fitted values is:

$$p'_{i+1} - p'_i = \beta_1$$

Denoting by $\ln(\text{odds}_1)$ the p'_i , the logarithm of the estimated odds for X_i and $\ln(\text{odds}_2)$ the p'_{i+1} , the logarithm of the estimated odds for X_{i+1} , the difference between the two fitted logit response values can be expressed as

$$\ln(\text{odds}_2) - \ln(\text{odds}_1) = \ln\left(\frac{\text{odds}_2}{\text{odds}_1}\right) = \beta_1$$

Taking antilogs for each side we see that the estimated ratio of the odds equals e^{β_j} :

$$\text{O.R.} = \frac{\text{odds}_2}{\text{odds}_1} = e^{\beta_j}$$

The odds of success when all predictor values equal zero is obtained from the constant term e^{β_0} .

When the predictor variable is dichotomous, there are two values of $p(x)$ and equivalently of $1 - p(x)$, which could be displayed in the following 2 x 2 table:



		Independent Variable X	
		x=1	x=0
Outcome Variable Y	y=1	$p(1) = \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}}$	$p(0) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$
	y=0	$1 - p(1) = \frac{1}{1 + e^{\beta_0 + \beta_1}}$	$1 - p(0) = \frac{1}{1 + e^{\beta_0}}$
Total		1	1

The odds of the outcome being present among individuals with x=1 is defined as $p(1)/[1 - p(1)]$. Similarly, the odds of the outcome being present among individuals with x=0 is defined as $p(0)/[1 - p(0)]$.

The logit could be defined as:

$$\text{logit}(1) = \ln \frac{p(1)}{1 - p(1)}$$

$$\text{logit}(0) = \ln \frac{p(0)}{1 - p(0)}$$

One can compute the odds ratio, based on formula:

$$\begin{aligned} O.R. &= \frac{\text{success vs. failure when } X = 1}{\text{success vs. failure when } X = 0} = \frac{X = 1 \text{ when } Y = 1 \mid X = 1 \text{ when } Y = 0}{X = 0 \text{ when } Y = 1 \mid X = 0 \text{ when } Y = 0} \\ &= \frac{p(1)/1 - p(1)}{p(0)/1 - p(0)} = \frac{\frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}} / 1}{\frac{e^{\beta_0}}{1 + e^{\beta_0}} / 1} = \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} = e^{\beta_1} \end{aligned}$$

Hence, for logistic regression with a dichotomous independent variable

$$O.R. = e^{\beta_1}$$

However, in the case of continuous predictor computing it by hand is too difficult.

Confidence interval on the odds ratio could be helpful in the selection of the appropriate variables for the analysis. A hypothesis test of $O.R.=1$, is



equivalent to a hypothesis test of $\beta_j = 0$. If a 95% confidence interval around the odds ratio includes the value of 1.0, indicates that a change in value of the independent variable is not associated in change in the odds of the dependent variable assuming a given value, then that variable is not considered a useful predictor in the logistic model.

Additionally, the $\text{logit}(p)$ can be backtransformed taking so the probability p of having a positive outcome as

$$p = \frac{1}{1 + e^{-\text{logit}(p)}}$$

Return to the example study that looks at the relationship between age and coronary heart disease, we have:

$$\text{logit}(p) = -5,310 + 0,111 \text{ Age}$$

That is for every unit change in Age, the $\text{logit}(p)$ increases by 0,111.

Take the exponential of the slope we have the estimated odds ratio. Thus $\exp(0,111) = 1.117395$. So, the odds of having CHD multiply by 1.12 for every year of age.



Chapter 3 Data Analysis

One hundred and thirty two Melanocytic Skin Lesion (23 melanomas and 109 melanocytic nevi), from 127 individuals were studied using the specially developed image processing system (Manousaki et.al.,2004). Flat or slightly elevated MSL, measuring between 0.5 up to 3.0 cm in maximal diameter, with clinical features difficult to interpret, suspicious of being melanomas or dysplastic nevi as well as other benign melanocytic nevi, were included. All lesions were photographed using a digital camera. A ruler was placed near the lesion in order to have a scale for the image and resolution was adjusted to 100 pixels /cm, thus making accurate geometric measurements within the lesion feasible ¹². All lesions were initially surgically excised with a safety normal skin margin of 2-3 mm and pathologically examined. Pathology reports clearly stated the diagnosis of melanoma or melanocytic nevi, and gave a description of the location and/or atypia of the cells of interest (nevus cells/ melanocytes and melanoma cells) within the lesion as well as the presence of cytological and architectural atypia for dysplastic nevi ¹³. Additional wide local excision for melanomas was subsequently carried out ^{16,17}. Based on final histology report, two groups of apparently benign common nevi were studied: compound and junctional.

Forty-three (43) variables (Table A. 1) (of geometry, color, fractal geometry for lesion color texture and of sharpness of the lesion border from the surrounding skin) were potential diagnostic parameters of the existence or not of skin melanomas. This outcome is described by the binary response variable, named melanoma for purposes of analysis, with value one (1) to describe the presence of melanoma and zero (0) the absence of melanoma at patient skin.

The purpose of this investigation is to utilize the data collected using an image processing system for a specific disease of the skin to develop, using the statistical tool of logistic regression analysis, a statistical model to assess the relationship between existence of melanoma (the dependent variable) and the associated independent variables.



Since diagnosis has two outcomes, presence or not presence of melanoma each of the forty-three variables was assessed individually and collectively for their predictive power using the statistical tool of logistic regression analysis as a standard statistical method of analysis in such a situation (Hosmer and Lemeshow, 1989). Statistical significance was set at $p < 0.05$ (Van Houwelingen and Le Cessie, 1990). All analyses were performed using the Statistical Analysis System (SAS, Release 8.1).

Undoubtedly the choice of independent variables included in a multivariable analysis is not a simple task. The usual approach for development of logistic regression models is a stepwise variable selection method. But there are severe criticisms about this strategy. Judd and McClelland (Judd and McClelland, 1989) say “stepwise methods will not necessarily produce the best model if there are redundant predictors”. Also they mention that “better models and a better understanding of one's data result from focused data analysis, guided by substantive theory” and as Henderson and Velleman (Henderson and Velleman, 1981) say “failure to use that knowledge produces inadequate data analysis”.

Mind this comments we would not select an automatic routine to specify how independent variables are entered into the analysis.

Firstly, the initially proposed variables were examined using univariate logistic regression to reveal, which may or may not correlate with the incidence of the disease.

In order to determine the behavior of each variable in relation to melanoma, univariate logistic regression with p -value for the Wald Chi-Square test and a Wald Confidence Interval (C.I.) of 95% of the Odds Ratio (OR) were used (Colton, 1974). Independent variables characterized as statistically related with melanoma if p -value of the Wald Chisquare test is less than 0.05 or additionally the 95% Confidence Interval of Odd Ratio does not include 1. Both these tests are ways of testing whether the parameter associated with the explanatory variable is zero. If for a particular explanatory variable, the Wald test is significant, or an 95% Confidence Interval of Odd Ratio does not include 1, then we would conclude that the parameter associated with this variable is not zero, and that the variable



should be included in the model. Otherwise, the explanatory variable can be omitted from the model.

Logistic regression analysis revealed that, as reported at Table A. 2, with exception of “Dik” and “CVRM” from the variables represent parameters of geometry, and “Mean Grey”, “Range Grey”, “Mean green”, “Skewness Green”, “Mean blue”, “Skewness blue” and “Min blue” from the variables represent parameters of color, there is evidence that each of the other variables of these groups has some association with the outcome, melanoma or not melanoma diagnosis. Besides that for “lac_grey” from the variables represent the parameters of fractal geometry for lesion color texture and both the two estimates of sharpness of the lesion border from the surrounding skin “sdgrey” and “cv sdgrey” there is evidence that have some association with the outcome, melanoma or not melanoma diagnosis.

The “-2 Log Likelihood” statistic has approximately a chi-square distribution and as we have discussed can be used for assessing the significance of logistic regression, analogous to the use of the sum of squared errors in OLS regression. Poorly-fitting models have higher values of this statistic. Examining the -2LL statistic for each univariate logistic regression presented at Table A. 2 we observe that, variable “rm “ from variables represent parameters of geometry, variable “lac_grey” from variables represent parameters of fractal geometry for lesion color texture, variable “cv_sharp” from variables represent estimates of sharpness of the lesion border and variable “range blue” from variables represent parameters of color, are the set of variables with the smaller values of -2LL statistic. So, we could assume that each of these four variables better contributes (Van Houwelingen and. Le Cessie, 1990) in the proposed model for diagnosis of response variable.

Harell (Harell,2001) mention that any stepwise variable selection has severe problems in the presence of *multicollinearity*. This problem occurs if independent variables have a high correlation with one another. Many of the variables are likely to be highly correlated (e.g., Area and Maximal Diameter). These factors could diminish the statistical power needed to address study questions. Furthermore, with a high correlation the quantitative risk estimates for each variable may be imprecise (Concato et.al., 1993).



The correlation between each of the previous mentioned four variables with any variable belongs to the same group of parameters, was examined (Table A. 3), (Table A. 4) and (Table A. 5). Due to the non-normally distributed data, Spearman’s rank correlation coefficients were used. Significance level $p\text{-value} < 0.05$ for Nonparametric Spearman's rho correlation coefficient indicates not statistically significant correlation. This happens only for parameter “Range_Blue_ and “Mean_Red” from the parameters of color. In the other three groups of parameters there is strong inter-correlations.

We now have an idea, which of the developed by digital analysis factors are good diagnostic predictors and which are not. Based on the above preliminary analysis, the variables selected to include in the multivariate logistic regression analysis are “rm”, “lac_grey”, “cv_sharp”, “range blue” and” mean red”.

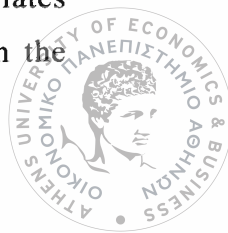
A first look at the selected variables for the proceeding analysis presented at next table:

Variable	Minimum	Maximum	Mean	Std. Deviation
Mean radius	.02	1.76	.3962	.27332
Lacunarity Grey scale	.01	1.13	.2523	.22268
CV(Sharpness)	9.13	56.60	21.1768	7.90961
Mean(red)	34.94	327.00	102.7266	35.87877
range(blue)	70.00	255.00	162.7576	44.03401
Valid N (listwise)	132			

Table 3.1 Descriptives for the five selected covariates

The minimum ratio of valid cases to independent variables for logistic regression is 10 to 1 (Hosmer, et.al, 1989), with a preferred ratio of 20 to 1. In this analysis, there are 132 valid cases and 5 independent variables. The ratio of cases to independent variables is 26.4 to 1, which satisfies the minimum requirement. In addition, the ratio of 26.4 to 1 satisfies the preferred ratio of 20 to 1.

The -2 Log Likelihood statistic for the model with only the intercept has dropped from 122.113 to 55.714 for the model with the five covariates (Table 3.6). To determine if the difference in likelihood scores between the



two models is statistically significant, we must consider the degrees of freedom. The degrees of freedom are five, equal to the number of additional parameters in the more complex model. Using this information we can then determine the critical value of the test statistic from standard statistical tables.

We have a reduction to the -2 Log Likelihood by $122.113 - 55.714 = 66.399$ with five degrees of freedom. The critical value with significance level $\alpha = 0.05$ is 11.07 and so we have an indication that our expanded model fits the data significantly better than the model with only a constant. The existence of a relationship between the independent variables and the dependent variable was supported.

The results of the multivariate logistic regression analysis are presented in the following table:

Parameter	B	S.E.	Wald	Sig.
Intercept	-13.6086	3.3586	16.4181	<.0001
RM	4.6553	1.5384	9.1574	0.0025
LAC GREY	5.5389	1.8241	9.2200	0.0024
CV_SHARP	0.0851	0.0475	3.2051	0.0734
MEAN_RE	0.00685	0.0109	0.3953	0.5295
RANGE BL	0.0298	0.0101	8.8038	0.0030

Table 3.2 Analysis of Maximum Likelihood Estimates for Multivariate Logistic with independent variables “RM”, “LAC_GREY”, “CV_SHARP”, “MEAN_RED”, “RANGE BLUE”

The probability of the Wald statistic for the variable “mean red” and “cv_sharp” for this regression was greater than the level of significance 0.05, indicating that these two variables would be not significant for the model.

We proceed to test the three possible reduced models separately. At first we exclude each of the variables “mean red” and “cv_sharp” from the full model and finally exclude both of them.

At the next three tables we have the results for multiple logistic regression model using



a) variables “Rm”, “Lac_grey”, “ Mean_red” and “range blue”

Parameter	B	S.E.	Wald	Sig.
Intercept	-11.2855	3.0151	14.0097	0.0002
RM	5.3492	1.5602	11.7554	0.0006
LAC GREY	5.4925	1.8676	8.6489	0.0033
MEAN_RE	0.000023	0.0141	0.0000	0.9987
RANGE BL	0.0300	0.00956	9.8748	0.0017

Table 3.3 Analysis of Maximum Likelihood Estimates for Multivariate Logistic with independent variables “RM”, “LAC_GREY”, “MEAN_RED” and “RANGE BLUE”

b) variables “Rm”, “Lac_grey”, “ Cv_sharp” and “range blue”

Parameter	B	S.E.	Wald	Sig.
Intercept	-12.5266	2.7316	21.0291	<.0001
RM	4.5056	1.4917	9.1235	0.0025
LAC GREY	4.9960	1.5585	10.2761	0.0013
CV_SHARP	0.0789	0.0485	2.6458	0.1038
RANGE BL	0.0296	0.00996	8.8572	0.0029

Table 3.4 Analysis of Maximum Likelihood Estimates for Multivariate Logistic with independent variables “RM”, “LAC_GREY”, “CV_SHARP”, and “RANGE BLUE”

and c) only variables “Rm”, “Lac_grey” and “range blue”.

Parameter	B	S.E.	Wald	Sig.
Intercept	-11.2826	2.4072	21.9676	<.0001
RM	5.3486	1.5094	12.5559	0.0004
LAC GREY	5.4907	1.5086	13.2463	0.0003
RANGE BL	0.03	0.00955	9.8829	0.0017

Table 3.5 Analysis of Maximum Likelihood Estimates for Multivariate Logistic with independent variables “RM”, “LAC_GREY”, “RANGE BLUE”

As we could see the two models produced either by excluding the variable “ CV-Sharp” or the variable “Mean Red” could not be regarded as good models. The probability of the Wald statistic for the variable “mean red” at the model presented at Table 3.3 and for “cv_sharp” at the model presented



at Table 3.4 were both greater than the level of significance 0.05, indicating that there is statistical evidence that these two variables do not contribute in the proposed models.

Consequently only the model with the three independent variables “rm”, “lac_grey”, “range blue” could be regarded an adequate model for our analysis.

G- statistic which measures the difference in deviances (Table 3.6) between the two models (G=2,835) follows a chi-squared distribution with 2 degrees of freedom with a *p*-value of 0.24232. This probability *p*-value exceeds significance level 0.05, make us conclude that the reduced model without both “mean red” and “cv_sharp” could be regarded as good as the model including them for the prediction of melanoma (Gail, M. H., 1991; Siminoff, J. S., 1998).

Model	Model Fit Statistics		
	Criterion	Only Intercept	Intercept and Covariates
Full with five covariates	AIC	124.113	67.714
Reduced with three covariates			66.549
Full with five covariates	SC	126.996	85.011
Reduced with three covariates			78.080
Full with five covariates	-2 Log L	122.113	55.714
Reduced with three covariates			58.549

Table 3.6 Model Fit Statistics for models with five & three covariates

A classification table is generated for selected model with three covariates using classification thresholds vary from 0.2 to 0.8 (Table 3.7).

As we have discussed in previous chapter, sensitivity - the measure of accuracy of predicting melanoma, is the ratio of true positives by total actual positives and specificity - accuracy of predicting non-events, is the ratio of true negatives by total actual negatives.



Prob Level	Correct		Incorrect		Correct	Percentages			
	Event	Non-Event	Event	Non-Event		Sensitivity	Specificity	False POS	False NEG
0.2	20	95	14	3	87.1	87.0	87.2	41.2	3.1
0.3	18	99	10	5	88.6	78.3	90.8	35.7	4.8
0.4	15	102	7	8	88.6	65.2	93.6	31.8	7.3
0.5	14	104	5	9	89.4	60.9	95.4	26.3	8.0
0.6	12	105	4	11	88.6	52.2	96.3	25.0	9.5
0.7	9	108	1	14	88.6	39.1	99.1	10.0	11.5
0.8	8	108	1	15	87.9	34.8	99.1	11.1	12.2

Table 3.7 Classification Table for model with three covariates for different cut-off points

Receiver-operating characteristic (ROC) plots the false-positive rate (1-specificity) on the x-axis and the true-positive rate (sensitivity) on the y-axis. Area under the ROC curve is a measure of a model discriminatory power (DeLong E.R., et.al. , 1988). The closer a ROC curve is to the upper left corner of the graph (as true-positive rate approaches 1 and false-positive rate approaches 0), the larger the area under the curve, and more accurate the prediction model.

The area under the ROC curve of the model with five variables as given at the output produced by SAS (Table A. 7) is (0.942) very closed to this of the model with three variables (0.939). This could be also seen at Figure 3.1.

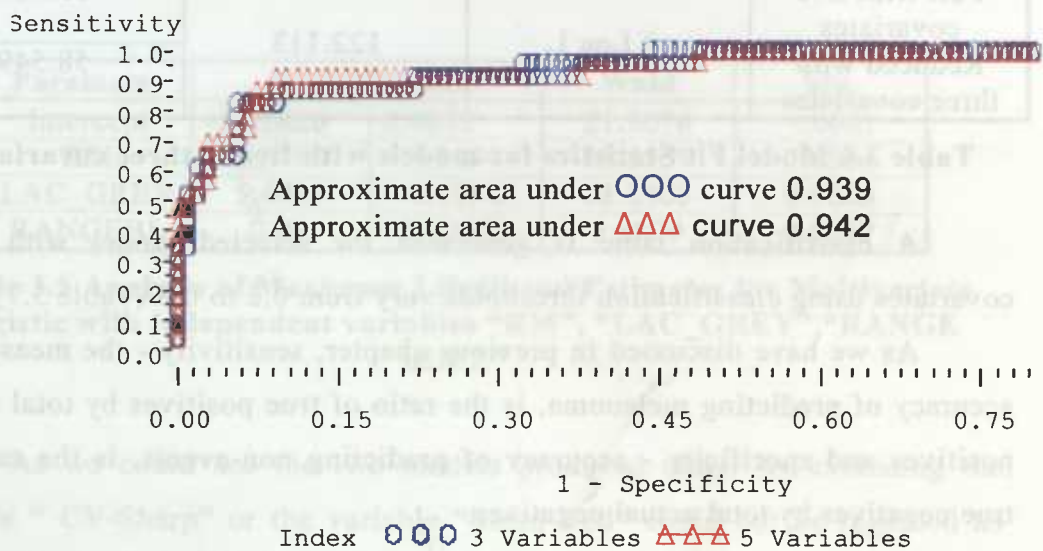


Figure 3.1 ROC curves for models with three and five covariates



Use of the model with three variables does not sacrifice discriminatory power. The three covariates with a value of c-statistic greater than 0.9 would be considered to be "excellent" at separating melanoma from no-melanoma skin disease patients.

Each point on the curve represents a cut-off probability. A lower cut-off typically gives more false positive. A high cut-off gives more false negatives, a low sensitivity, and a high specificity (SAS Institute Inc. ,2001).

The performance of screening test depends on the cut points used to define a positive test. The choice of a higher cut point leaves more cases undetected, and the choice of a lower cut point classifies more healthy individuals as abnormal. Currently, there are no widely accepted or rigorously validated cut points to define positive screening tests for melanocytic skin lesions. Selection of the optimal cut point based either on a receiver-operating characteristic (ROC) curve or on examination of classification table.

At ROC curve for the selected model with three covariates was plotting sensitivity against the false-positive rate (1 - specificity) over a range of cut-point values.

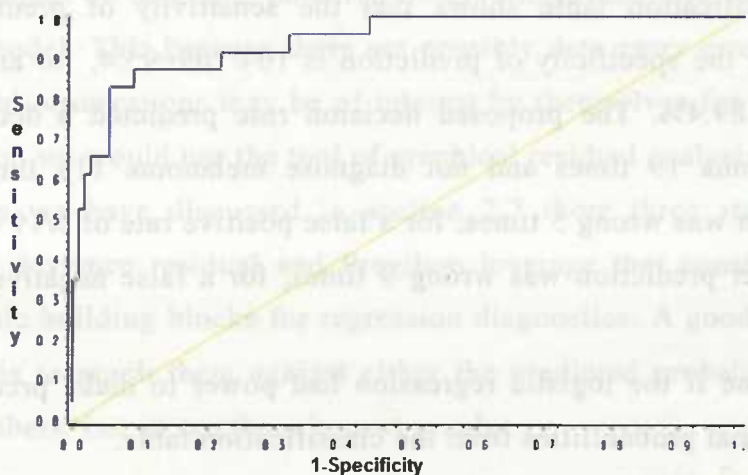
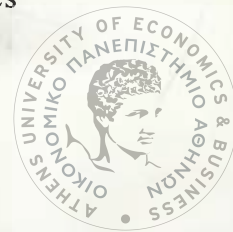


Figure 3.2 ROC curve for the selected model with three covariates

Generally, the best cut point is at or near the shoulder of the ROC curve, where substantial gains can be made in sensitivity with only modest reductions in specificity. So, threshold 0.5 was chosen as probability cut-off point, to balance sensitivity and specificity and to minimize false positives (Table 3.7).



The predictive equation with the three covariates “rm”, “lac-grey” and “range-blue” was calculated with the regression parameters of Table 3.5:

$$\hat{p} = \frac{e^X}{1+e^X} \tag{3.1}$$

where $X = - 11,283 + 5,349*(RM)+5,491*(Lac_grey)+0,030*(Range\ Blue)$.

The classification table for this model and a cut-off point of 0.5 is:

		Predicted Melanoma		
		Absent	Present	
Actual Melanoma	Absent	104	5	95.4
	Present	9	14	60.0
	Overall Percentage			89.4

Table 3.8 Classification Table for model with three covariates for cut-off point of 0.5

The classification table shows that the sensitivity of prediction is $14/ 23 = 60.0\%$, the specificity of prediction is $104/109=95.4$, for an overall success rate of 89.4% . The proposed decision rule predicted a decision to diagnose melanoma 19 times and not diagnose melanoma 113 times. The former prediction was wrong 5 times, for a false positive rate of $5/19 = 26.3\%$ whereas the latter prediction was wrong 9 times, for a false negative rate of $9/113 =8\%$.

To examine if the logistic regression had power to make predictions, we use the marginal probabilities from the classification table.

P (actual result Melanoma and predicted result Melanoma)

$$= \frac{9+14}{132} * \frac{14+5}{132} = 2.51\%$$

That is we would expect to get 4.37% of the observations correctly classified as success just by random chance. A similar calculation for the failures gives

P (actual result No-Melanoma and predicted result No-Melanoma)=



$$\frac{104 + 9}{132} * \frac{104 + 5}{132} = 70.69\%$$

Thus, assuming that the logistic regression had no predictive power for the actual result, presence or absence of Melanoma, we would expect to correctly classify $2.51\%+70.69\%=73.2\%$ of the observations. So the C_{pro} measure, is:

$$C_{pro}=1.25*70.69\%=88.36\%.$$

The observed 89.4% is quite higher than C_{pro} , which would support the usefulness of the logistic regression as a classifier.

Accordingly to Hosmer and Lemeshow goodness of fit test, the p-value of $0.7427 > 0.05$ give insufficient evidence to reject the null hypothesis that the proposed logistic model is appropriate. Additionally model fit could be regarded overly good as measured either by Cox & Snell R- Square (0.3822) or Nagelkerke R- Square (0.6333).

We would try now to detect potential problems with model building. We will focus on detecting potential observations that have significant impact on our model. This because there are possibly data entry errors and secondly influential observations may be of interest by themselves for us to study. For this reason we would use the tool of graphical residual analysis

As we have discussed in section 2.7 there three statistics, Pearson residual, deviance residual and Pregibon leverage that considered to be the three basic building blocks for regression diagnostics. A good way of looking at them is to graph them against either the predicted probabilities or simply case numbers. Let us see them in next graphs.



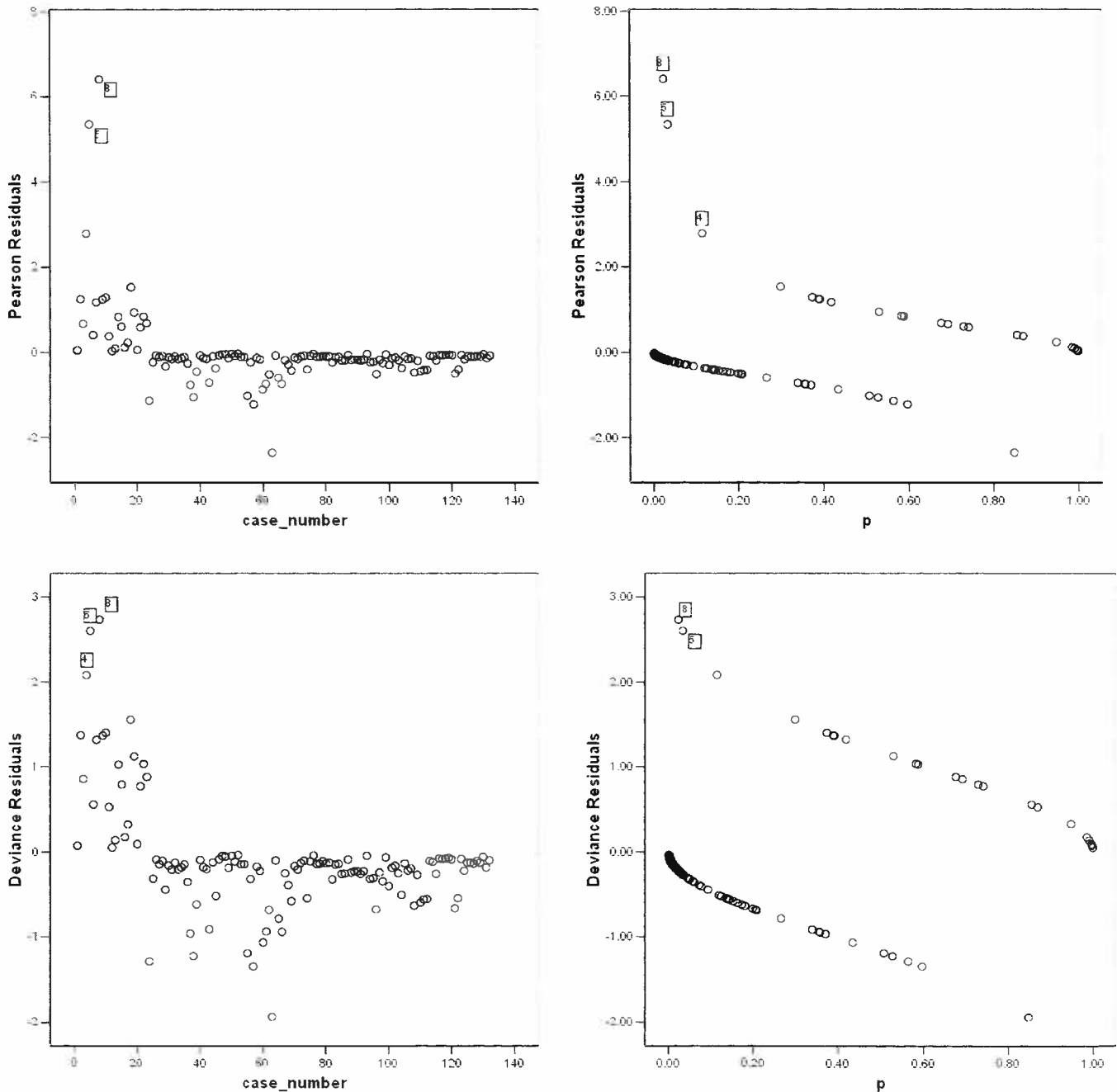


Figure 3.3 Plots of Pearson and Deviance Residuals versus predicted probabilities or case number for the model described by equation (3.1)

At these plots we see observations that are way off from the most of the other observations. The observations with patient number 5 and 8 have high Pearson and deviance residuals. The observed outcome for existence of melanoma is 1 but the predicted probabilities are very very low. This leads to large residuals. These two observations need our particular attention. But as



we could notice at the next two graphs these two observations are not bad in the terms of leverage.

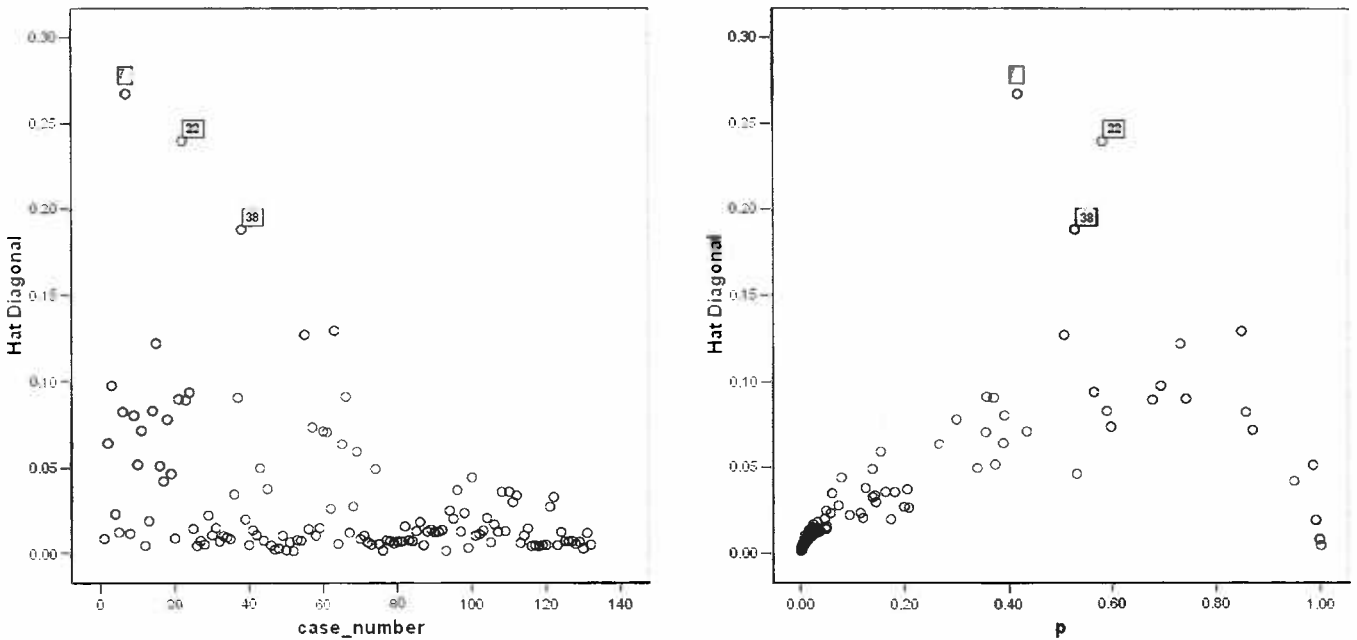


Figure 3.4 Plots of Hat Diagonal versus predicted probabilities or case number for the model described by equation (3.1)

These could make us say that by not including these particular observations our logistic regression estimate won't be too much different from the model includes these observations.

Another important aspect of diagnostic is to identify observations with substantial impact on either the chi-square fit statistic or the deviance statistic. Examine the next two graphs we see that observations 4 and 63 are substantial in terms of chi-square fit statistic and observations 5 and 8 in terms of deviance fit statistic.



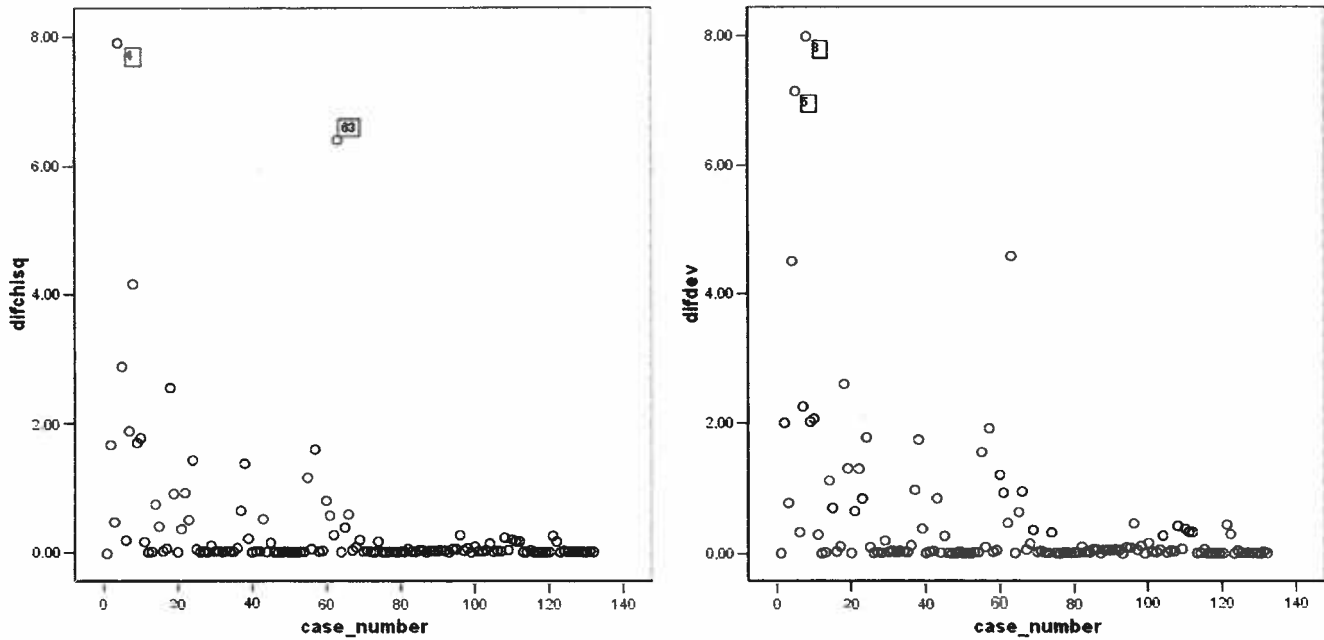


Figure 3.5 Plots of DIFCHISQ and DIFDEV versus case number for the model described by equation (3.1)

This means that when observations 4 and 63 are excluded from the analysis the Pearson chi-square fit statistic will decrease by 8 or 7 units, respectively. The same will happen for the deviance statistic by excluding observations 5 and 8. These four observations have great impact of fit statistics, but not too much impact on parameters estimate, as their leverage are not very large. As we have no special information about these patients we would not exclude them from the analysis but it may be of interest to study them by themselves.

Validation techniques are performed to measure the predictive efficacy of a model, to assess the likely performance of a model on a new series of subjects. In order to validate the prediction rule, the "jackknife" method was performed and the results were presented as misclassification defined as the fraction of patients who were incorrectly classified. The jackknife method is a statistical cross-validation technique that can be employed when it is necessary to use the original study patients to validate a clinical prediction rule (Wasson, J.H., et. al., 1985). In this method, sometimes called the "one-left-out" method, one patient is removed and the rule is rederived and used to classify the excluded patient. The patient's predicted state is compared with



the true state. This process is repeated many times to determine the frequency with which the excluded patient is misclassified. Therefore, in this study, among 132 patients, 1 patient was selected to be excluded. A rule was derived from the remaining 131 patients and applied to the 132th patient. This process was repeated systematically for all 132 patients. The misclassification rate is the fraction of the 132 test patients who were incorrectly classified.

Prob Level	Correct		Incorrect		Correct
	Event	Non-Event	Event	Non-Event	
0.5	15	104	5	8	90.15

Table 3.9 Classification Table for jackknife validation

Cross-validation accuracy was estimated based on the above classification table, as $(104+15)/132=90.15\%$, based on the frequencies close enough to the model accuracy of 89.4%. That both accuracies were high and similar suggests the model is fairly robust and is not overly influenced by characteristics of the model building data.

To identify potential prognostic parameters based on an image processing analysis for this specific disease of the skin, we have established a model to predict the risk for skin melanomas and no-melanomas. We found a statistically significant overall relationship between the combination of independent variables and the dependent variable. There is no evidence of numerical problems in the solution. Moreover, the classification accuracy surpassed the proportional by chance accuracy criteria, supporting the utility of the model.

The predictive equation described by equation (3.1) is

$$\hat{p} = \frac{e^X}{1 + e^X}$$

where $X = - 11,283 + 5,349*(RM)+5,491*(Lac_grey)+0,030*(Range\ Blue)$

The positive coefficient of the covariate “rm” means that increasing values of “rm” tend to occur with increasing probabilities for presence of



melanoma at skin lesion (coded as 1, in our analysis). Likewise the positive coefficients of “lac_grey” and “range-blue” mean that increasing values of these two covariates tend to occur with increasing probabilities for presence of melanoma.

In fact the odds of melanoma increase by a factor of $\exp(5.349)$ for each unit increase in “rm”- Average length of the radii of the lesion, by a factor of $\exp(5.491)$ for each unit increase in “lac_grey”- an expression of the number and size of color voids within the lesion and also increases by a factor of $\exp(0.030)$ for each unit increase in “range-blue”- Range of values of Blue intensity, provided in each case the other two factors held constant.

Base on the logistic regression model we can calculate an individual’s probability of being melanoma positive (or negative as the case may be).

Descriptive statistics were produced to compare probabilities for presence and non-presence of melanoma (Table 3.10). This data indicates that there is a significant difference between melanoma and non-melanoma skin lesions, based on the selected model.

		Number	Mean	Std. Deviation	t-test t/p
Predicted probability	Melanoma	23	.62	.32	7.91/<.001
	Not Melanoma	109	.081	.15	

Table 3.10 Statistics of predictive probability for model described by equation (3.1)

Predictions for individual cases could be obtained by entering the values for the average length of the radii of the lesion, an estimation of the number and size of color voids within the lesion and the range of values of blue intensity from digital image of skin lesions in the formula for the specific case.



Chapter 4 Conclusions

The purpose of the analysis presented at this thesis is, to utilize the data collected using a image processing system for a specific disease of the skin, this of melanoma, to develop, using the statistical tool of logistic regression analysis, a statistical model to assess the relationship between existence of melanoma and the associated parameters.

A predictive model was developed using multiple logistic regression analysis and one hundred and thirty two melanocytic skin lesions, from one hundred twenty seven individuals. Although there were forty-three parameters measured, only three were selected for a parsimonious and best fitting model.

The three selected independent covariates for prediction of melanocytic skin lesion were, the average length of the radii of the lesion, an expression of the number and size of color voids within the lesion and the range of the values of the blue intensity. The sensitivity of prediction is 60.9%, the specificity is 95.4%, for an overall success rate of 89.4% with a cut point of 0.5. The false positive rate is 26.3%, whereas the false negative one is 8%.

The proposed multivariate logistic regression model can be easily implemented in an inexpensive handheld programmable calculator to predict previously undiagnosed skin melanoma that could help on screening for undiagnosed melanocytic skin lesion.





Appendix I





Geometry variables		Color texture variables	
Area	Surface area of the lesion in pixels and cm ²	Fractal dimension of mass of lesion (FDMB)	It is estimated using the Minkowski-Boulingard algorithm. It is an estimate of irregularity in pigment distribution on the lesion surface, as this (irregularity) is perceived by the human eye. It takes values between 2 and 3 (the dimension of an object that is not actually regular in shape and not a cube)
Maximal Diameter	Longest distance between two non adjacent points on the border of the lesion	Fractal dimension of border of lesion (FD)	It is calculated in a 250x250 pixel square and is actually a measure of border indentation in a thresholded image. It takes values between 1 and 2 (the border is extending in two dimensions without totally covering any of the two)
Perimeter	Total length of border of lesion	Gray scale Lacunarity of lesion	It is estimated in gray scale image. It is a color texture parameter, an expression of the number and size of color voids within the lesion.
Circularity (Equivalent Circle Index)	Ratio of the perimeter of the lesion over the perimeter of the circle with the same midpoint (center) and of equal surface area to the studied lesion	Thresholded Lacunarity of lesion.	Thresholded lacunarity in binary image at the 75 th percentile position of histogram of Gray intensity.
Mean radius (Rm)	Average length of the radii of the lesion	Sharpness of border	
Standard deviation of Rm (SDRm)	Standard deviation of the mean radius (range of values for Rm of the lesion)	SD of gray intensity	Intensity of Gray on the border of the lesion. For reasons of accuracy, the area surrounding the lesion and accounting for 10% of the area of the lesion is selected. In gray scale image SD of Gray intensity and mean value of SD of Gray intensity in the selected area of border is calculated. The bigger the value, the more discrete is the lesion from the surrounding normal skin.
Coefficient of variation of Rm (CVRm)	Expresses variability of radius values in relation to its mean value	CV of SD of gray intensity	
Distance of color midpoint from lesion midpoint (Delta, or eccentricity)	Distance between color and geometric mid points within the lesion It is an estimate of abnormal evolution and uneven coloration of the lesion.	Color variables	
Delta to Rm ratio	Distance between midpoint and color midpoint expressed as a fraction of the mean radius	Minimal Gray, Red, Green, Blue	Minimal value of Gray, Red, Green, Blue intensity
		Maximal Gray, Red, Green, Blue	Maximal value Gray, Red, Green, Blue intensity
		Range of Gray, Red, Green, Blue	Range of values of Gray, Red, Green, Blue intensity
		Mean Gray, Red, Green, Blue	Mean value of intensity in Gray scale and, Red, Green, Blue color space
		Standard deviation (SD) of Gray, Red, Green, Blue	Standard deviation of Gray, Red, Green, Blue intensity within the lesion
		Coefficient of variation (CV) of Gray, Red, Green, Blue	Variability of Gray, Red, Green, Blue intensity values in relation to its mean value
		Skewness from Gaussian Curve (Gray, Red, Green, Blue)	Deviation of the histogram of each color from the normal distribution curve.

Table A. 1 The Studied Geometric, Colour, Sharpness and Colour Texture Variables (n=43)



Variable	B	S.E.	Wald	Sig.	Exp(B)	95,0% C.I.for EXP(B)		-2LogLikelihood
						Lower	Upper	
AREA	1,6768	,4399	14,5290	,000	5,348	2,258	12,667	89.409
DIAMETER	2,4341	,5688	18,3113	,000	11,406	3,741	34,779	88.385
PERIMETE	,789	,187	18,2574	,000	2,223	1,541	3,207	87.64
DIK	1,7425	1,680	1,0759	,2996	5,712	,212	153,731	121.12
RM	5,9384	1,3788	18,549	,000	379,325	25,432	>999,99	87.501
SDRM	13,0381	4,5595	8,177	,0042	>999,99	60,440	>999,99	110.720
CVRM	,00918	,0425	,0467	,829	1,009	,929	1,097	122.067
DELTA	58,799	19,6753	8,931	,0028	>999,99	>999,99	>999,99	121.320
DELTARM	,1781	,1955	,8303	,3622	1,195	,815	1,753	106.712
LAC_THRE	-1,308	,6943	3,5503	,0595	,270	,069	1,054	118.637
LAC_GREY	3,4313	,9634	12,6847	,0004	30,917	4,679	204,299	108.286
FDMB	5,5421	2,8553	3,7674	,0523	255,218	,947	>999,99	117.997
FD	-3,3806	2,2562	2,2451	,1340	,034	<,0001	2,833	119.769
SHARP	,1032	,0264	15,2477	<,0001	1,109	1,053	1,168	100.593
CV_SHARP	,1518	,0371	16,7028	<,0001	1,164	1,082	1,252	97.628
MEAN_GRE	-,0172	,0111	2,4091	,1206	,983	,962	1,005	119.572
SDG	,0671	,0257	6,7923	,0092	1,069	1,017	1,125	114.713
CVG	,0367	,0110	11,2455	,0008	1,037	1,015	1,060	109.439
SKEWG	,4563	,6400	,5083	,4759	1,578	,450	5,533	121.595
MING	-,0729	,0261	7,789	,005	,930	,883	,979	107.545
MAXG	,0275	,00754	13,3067	,0003	1,028	1,013	1,043	106.814
RANGEG	,00415	,00301	1,8993	,1682	1,004	,998	1,010	119.276
MEAN_RED	-,0234	,00875	7,1554	,0075	,977	,960	,994	113.677
SDR	,0906	,0260	12,1089	,0005	1,095	1,040	1,152	107.396
CVR	,0513	,0126	16,6762	<,0001	1,053	1,027	1,079	100.686
SKEWR	-,0328	,0991	,1095	,7408	,968	,797	1,175	121.851
MINR	-,0672	,0211	10,1207	,0015	,935	,897	,975	99.762
MAXR	,0267	,00919	8,4579	,0036	1,027	1,009	1,046	112.190
RANGER	,0371	,00838	19,5768	<,0001	1,038	1,021	1,055	93.493
MEAN_GRN	-,0126	,0114	1,2204	,2693	,987	,966	1,010	120.836
SDGRN	,0671	,0257	6,7923	,0092	1,069	1,017	1,125	111.850
CVGRN	,0323	,0103	9,7729	,0018	1,033	1,012	1,054	120.957
SKEWGRN	,6865	,6474	1,1246	,2889	1,987	,559	7,066	121.595
MINGRN	-,0563	,0253	4,9565	,0260	,945	,900	,993	114.560
MAXGRN	,0299	,00746	16,0145	<,0001	1,030	1,015	1,045	102.943
RANGGRN	,0315	,00713	19,5753	<,0001	1,032	1,018	1,047	97.336
MEAN_BLU	,00172	,0118	,0213	,8840	1,002	,979	1,025	122.092
SDBL	,1112	,0296	14,1476	,0002	1,118	1,055	1,184	105.498
CVBL	,0253	,0964	6,8755	,0087	1,026	1,006	1,045	115.157
SKEWBL	,0112	,0135	,6849	,4079	1,011	,985	1,038	121.499
MINBL	-,0533	,0281	3,5973	,0579	,948	,897	1,002	117.219
MAXBL	,0311	,0069	20,3166	<,0001	1,032	1,018	1,046	95.456
RANGBL	,0309	,00652	22,4551	<,0001	1,031	1,018	1,045	92.666

Table A. 2 Analysis of Maximum Likelihood and Odd Ratio Estimates from Univariate Analysis



		Threshold Lacunarity	Lacunarity Grey scale	Fractal Dimension of Mass	Fractal Dimension of Border
Threshold Lacunarity	Coefficient	1,000			
	Sig. (2-tailed)				
Lacunarity Grey scale	Coefficient	0,158	1,000		
	Sig. (2-tailed)	0,070			
Fractal Dimension of Mass	Coefficient	-0,638	-0,027	1,000	
	Sig. (2-tailed)	0,000	0,763		
Fractal Dimension of Border	Coefficient	-0,022	-0,226	0,096	1,000
	Sig. (2-tailed)	0,800	0,009	0,271	

Table A. 3 Spearman's Correlation Coefficient for Parameters of fractal geometry for lesion color texture



		Surface Area	maximal diameter	perimeter	Equivalent Circle Index	Mean radius	SD(R)	CV(R)	Delta	Delta/Rm
Surface Area	Coefficient	1,000								
	Sig. (2-tailed)									
maximal diameter	Coefficient	0,986	1,000							
	Sig. (2-tailed)	0,000								
perimeter	Coefficient	0,987	0,983	1,000						
	Sig. (2-tailed)	0,000	0,000							
Equivalent Circle Index	Coefficient	0,502	0,555	0,544	1,000					
	Sig. (2-tailed)	0,000	0,000	0,000						
Mean radius	Coefficient	0,980	0,970	0,977	0,504	1,000				
	Sig. (2-tailed)	0,000	0,000	0,000	0,000					
SD(R)	Coefficient	0,804	0,869	0,822	0,661	0,785	1,000			
	Sig. (2-tailed)	0,000	0,000	0,000	0,000	0,000				
CV(R)	Coefficient	0,266	0,380	0,305	0,597	0,249	0,719	1,000		
	Sig. (2-tailed)	0,002	0,000	0,000	0,000	0,004	0,000			
Delta	Coefficient	0,351	0,342	0,348	0,100	0,364	0,284	0,042	1,000	
	Sig. (2-tailed)	0,000	0,000	0,000	0,256	0,000	0,001	0,633		
Delta/Rm	Coefficient	-0,254	-0,270	-0,254	-0,262	-0,248	-0,260	-0,190	0,673	1,000
	Sig. (2-tailed)	0,003	0,002	0,003	0,002	0,004	0,003	0,029	0,000	

Table A. 4 Spearman’s Correlation Coefficient for Parameters of Geometry



		Mean Grey	SD Grey	CV Grey	Skewness Grey	Min Grey	Max Grey	Range Grey
Mean Grey	Coefficient	1,000						
	Sig. (2-tailed)							
SD Grey	Coefficient	-0,052	1,000					
	Sig. (2-tailed)	0,551						
CV Grey	Coefficient	-0,771	0,612	1,000				
	Sig. (2-tailed)	0,000	0,000					
Skewness Grey	Coefficient	-0,303	-0,157	0,168	1,000			
	Sig. (2-tailed)	0,000	0,071	0,055				
Min Grey	Coefficient	0,694	-0,514	-0,856	0,041	1,000		
	Sig. (2-tailed)	0,000	0,000	0,000	0,641			
Max Grey	Coefficient	0,342	0,529	0,077	0,094	-0,085	1,000	
	Sig. (2-tailed)	0,000	0,000	0,383	0,282	0,335		
Range Grey	Coefficient	-0,030	0,627	0,407	0,072	-0,478	0,821	1,000
	Sig. (2-tailed)	0,729	0,000	0,000	0,411	0,000	0,000	
Mean Red	Coefficient	0,867	-0,170	-0,764	-0,354	0,704	0,182	-0,155
	Sig. (2-tailed)	0,000	0,051	0,000	0,000	0,000	0,037	0,075
SD Red	Coefficient	-0,289	0,801	0,707	-0,151	-0,640	0,352	0,534
	Sig. (2-tailed)	0,001	0,000	0,000	0,084	0,000	0,000	0,000
CV Red	Coefficient	-0,737	0,578	0,941	0,173	-0,849	0,103	0,434
	Sig. (2-tailed)	0,000	0,000	0,000	0,048	0,000	0,239	0,000
Skewness Red	Coefficient	-0,314	-0,195	0,154	0,883	0,024	0,007	0,021
	Sig. (2-tailed)	0,000	0,025	0,078	0,000	0,783	0,940	0,808
Min Red	Coefficient	0,647	-0,530	-0,828	0,002	0,960	-0,134	-0,504
	Sig. (2-tailed)	0,000	0,000	0,000	0,982	0,000	0,125	0,000
Max Red	Coefficient	0,346	0,410	-0,006	-0,005	-0,017	0,851	0,695
	Sig. (2-tailed)	0,000	0,000	0,942	0,954	0,850	0,000	0,000
Range Red	Coefficient	-0,189	0,680	0,556	-0,016	-0,636	0,699	0,862
	Sig. (2-tailed)	0,030	0,000	0,000	0,857	0,000	0,000	0,000
Mean Green	Coefficient	0,917	-0,086	-0,744	-0,280	0,654	0,303	-0,045
	Sig. (2-tailed)	0,000	0,329	0,000	0,001	0,000	0,000	0,609
SD Green	Coefficient	-0,028	0,863	0,534	-0,179	-0,439	0,472	0,534
	Sig. (2-tailed)	0,747	0,000	0,000	0,040	0,000	0,000	0,000
CV Green	Coefficient	-0,750	0,592	0,966	0,138	-0,816	0,059	0,378
	Sig. (2-tailed)	0,000	0,000	0,000	0,116	0,000	0,503	0,000
Skewness Green	Coefficient	-0,298	-0,169	0,153	0,960	0,043	0,102	0,080
	Sig. (2-tailed)	0,001	0,053	0,080	0,000	0,626	0,245	0,359
Min Green	Coefficient	0,678	-0,486	-0,818	0,061	0,922	-0,063	-0,455
	Sig. (2-tailed)	0,000	0,000	0,000	0,486	0,000	0,471	0,000
Max Green	Coefficient	0,337	0,469	0,042	0,054	-0,064	0,937	0,763
	Sig. (2-tailed)	0,000	0,000	0,632	0,538	0,463	0,000	0,000
Range Green	Coefficient	0,041	0,625	0,346	0,047	-0,398	0,869	0,889
	Sig. (2-tailed)	0,644	0,000	0,000	0,590	0,000	0,000	0,000
Mean Blue	Coefficient	0,764	-0,007	-0,569	-0,223	0,466	0,398	0,084
	Sig. (2-tailed)	0,000	0,940	0,000	0,010	0,000	0,000	0,337
SD Blue	Coefficient	0,047	0,777	0,451	-0,175	-0,409	0,541	0,000
	Sig. (2-tailed)	0,591	0,000	0,000	0,045	0,000	0,000	0,000



		Mean Grey	SD Grey	CV Grey	Skewness Grey	Min Grey	Max Grey	Range Grey
CV Blue	Coefficient	-0,648	0,562	0,865	0,057	-0,733	0,031	0,304
	Sig. (2-tailed)	0,000	0,000	0,000	0,519	0,000	0,728	0,000
Skewness Blue	Coefficient	-0,260	-0,022	0,190	0,805	0,038	0,157	0,121
	Sig. (2-tailed)	0,003	0,801	0,029	0,000	0,663	0,071	0,165
Min Blue	Coefficient	0,612	-0,400	-0,714	0,069	0,809	0,010	-0,331
	Sig. (2-tailed)	0,000	0,000	0,000	0,430	0,000	0,906	0,000
Max Blue	Coefficient	0,190	0,430	0,142	0,041	-0,189	0,821	0,704
	Sig. (2-tailed)	0,029	0,000	0,104	0,641	0,030	0,000	0,000
Range Blue	Coefficient	0,011	0,533	0,334	0,013	-0,399	0,779	0,781
	Sig. (2-tailed)	0,898	0,000	0,000	0,880	0,000	0,000	0,000



		Mean Red	SD Red	CV Red	Skewness Red	Min Red	Max Red	Range Red
Mean Red	Coefficient	1,000						
	Sig. (2-tailed)							
SD Red	Coefficient	-0,299	1,000					
	Sig. (2-tailed)	0,000						
CV Red	Coefficient	-0,822	0,716	1,000				
	Sig. (2-tailed)	0,000	0,000					
Skewness Red	Coefficient	-0,414	-0,209	0,195	1,000			
	Sig. (2-tailed)	0,000	0,016	0,025				
Min Red	Coefficient	0,749	-0,646	-0,889	-0,020	1,000		
	Sig. (2-tailed)	0,000	0,000	0,000	0,820			
Max Red	Coefficient	0,358	0,368	-0,012	-0,088	-0,010	1,000	
	Sig. (2-tailed)	0,000	0,000	0,894	0,318	0,911		
Range Red	Coefficient	-0,233	0,738	0,581	-0,058	-0,643	0,710	1,000
	Sig. (2-tailed)	0,007	0,000	0,000	0,507	0,000	0,000	
Mean Green	Coefficient	0,822	-0,240	-0,692	-0,306	0,612	0,293	-0,186
	Sig. (2-tailed)	0,000	0,006	0,000	0,000	0,000	0,001	0,033
SD Green	Coefficient	-0,081	0,774	0,459	-0,229	-0,418	0,387	0,588
	Sig. (2-tailed)	0,357	0,000	0,000	0,008	0,000	0,000	0,000
CV Green	Coefficient	-0,715	0,649	0,868	0,129	-0,768	0,014	0,527
	Sig. (2-tailed)	0,000	0,000	0,000	0,139	0,000	0,875	0,000
Skewness Green	Coefficient	-0,343	-0,134	0,173	0,813	-0,005	0,045	0,027
	Sig. (2-tailed)	0,000	0,126	0,047	0,000	0,954	0,608	0,757
Min Green	Coefficient	0,624	-0,566	-0,751	0,018	0,848	-0,055	-0,617
	Sig. (2-tailed)	0,000	0,000	0,000	0,838	0,000	0,533	0,000
Max Green	Coefficient	0,209	0,325	0,073	-0,025	-0,111	0,831	0,676
	Sig. (2-tailed)	0,016	0,000	0,404	0,778	0,207	0,000	0,000
Range Green	Coefficient	-0,055	0,528	0,356	-0,017	-0,415	0,767	0,866
	Sig. (2-tailed)	0,529	0,000	0,000	0,843	0,000	0,000	0,000
Mean Blue	Coefficient	0,594	-0,125	-0,468	-0,210	0,386	0,281	-0,050
	Sig. (2-tailed)	0,000	0,152	0,000	0,015	0,000	0,001	0,568
SD Blue	Coefficient	-0,051	0,607	0,392	-0,218	-0,406	0,382	0,545
	Sig. (2-tailed)	0,565	0,000	0,000	0,012	0,000	0,000	0,000
CV Blue	Coefficient	-0,579	0,568	0,721	0,024	-0,659	0,035	0,456
	Sig. (2-tailed)	0,000	0,000	0,000	0,788	0,000	0,688	0,000
Skewness Blue	Coefficient	-0,266	-0,008	0,175	0,587	0,004	0,130	0,104
	Sig. (2-tailed)	0,002	0,932	0,044	0,000	0,967	0,136	0,238
Min Blue	Coefficient	0,518	-0,447	-0,625	0,042	0,724	-0,030	-0,484
	Sig. (2-tailed)	0,000	0,000	0,000	0,635	0,000	0,735	0,000
Max Blue	Coefficient	0,037	0,310	0,194	0,003	-0,251	0,665	0,630
	Sig. (2-tailed)	0,673	0,000	0,026	0,975	0,004	0,000	0,000
Range Blue	Coefficient	-0,109	0,432	0,362	-0,012	-0,437	0,638	0,743
	Sig. (2-tailed)	0,212	0,000	0,000	0,888	0,000	0,000	0,000



		Mean Green	SD Green	CV Green	Skewness Green	Min Green	Max Green	Range Green
Mean Green	Coefficient	1,000						
	Sig. (2-tailed)							
SD Green	Coefficient	0,040	1,000					
	Sig. (2-tailed)	0,645						
CV Green	Coefficient	-0,800	0,502	1,000				
	Sig. (2-tailed)	0,000	0,000					
Skewness Green	Coefficient	-0,302	-0,204	0,144	1,000			
	Sig. (2-tailed)	0,000	0,019	0,099				
Min Green	Coefficient	0,723	-0,381	-0,849	0,050	1,000		
	Sig. (2-tailed)	0,000	0,000	0,000	0,571			
Max Green	Coefficient	0,375	0,499	0,010	0,089	-0,015	1,000	
	Sig. (2-tailed)	0,000	0,000	0,912	0,308	0,860		
Range Green	Coefficient	0,079	0,619	0,317	0,080	-0,373	0,910	1,000
	Sig. (2-tailed)	0,366	0,000	0,000	0,360	0,000	0,000	
Mean Blue	Coefficient	0,888	0,125	-0,662	-0,260	0,584	0,478	0,223
	Sig. (2-tailed)	0,000	0,153	0,000	0,003	0,000	0,000	0,010
SD Blue	Coefficient	0,126	0,880	0,409	-0,201	-0,321	0,595	0,661
	Sig. (2-tailed)	0,151	0,000	0,000	0,021	0,000	0,000	0,000
CV Blue	Coefficient	-0,726	0,507	0,930	0,081	-0,788	-0,008	0,265
	Sig. (2-tailed)	0,000	0,000	0,000	0,356	0,000	0,931	0,002
Skewness Blue	Coefficient	-0,252	-0,039	0,192	0,848	0,049	0,175	0,160
	Sig. (2-tailed)	0,004	0,653	0,028	0,000	0,577	0,044	0,067
Min Blue	Coefficient	0,682	-0,313	-0,768	0,045	0,881	0,065	-0,255
	Sig. (2-tailed)	0,000	0,000	0,000	0,610	0,000	0,461	0,003
Max Blue	Coefficient	0,245	0,468	0,098	0,071	-0,105	0,903	0,848
	Sig. (2-tailed)	0,005	0,000	0,265	0,420	0,231	0,000	0,000
Range Blue	Coefficient	0,060	0,551	0,298	0,044	-0,330	0,848	0,900
	Sig. (2-tailed)	0,494	0,000	0,001	0,612	0,000	0,000	0,000



		Mean Blue	SD Blue	CV Blue	Skewness Blue	Min Blue	Max Blue	Range Blue
Mean Blue	Coefficient	1,000						
	Sig. (2-tailed)							
SD Blue	Coefficient	0,274	1,000					
	Sig. (2-tailed)	0,001						
CV Blue	Coefficient	-0,704	0,436	1,000				
	Sig. (2-tailed)	0,000	0,000					
Skewness Blue	Coefficient	-0,264	-0,067	0,181	1,000			
	Sig. (2-tailed)	0,002	0,443	0,038				
Min Blue	Coefficient	0,650	-0,282	-0,816	0,084	1,000		
	Sig. (2-tailed)	0,000	0,001	0,000	0,337			
Max Blue	Coefficient	0,476	0,645	0,030	0,141	0,030	1,000	
	Sig. (2-tailed)	0,000	0,000	0,730	0,107	0,733		
Range Blue	Coefficient	0,290	0,708	0,240	0,105	-0,223	0,950	1,000
	Sig. (2-tailed)	0,001	0,000	0,006	0,232	0,010	0,000	

Table A. 5 Spearman’s Correlation Coefficient Spearman’s Correlation Coefficient for Parameters of color including skewness from the Gaussian curve of normal distribution for the 4 color intensities

		SD(Sharpness)	CV(Sharpness)
SD(Sharpness)	Coefficient	1,000	0,859
	Sig. (2-tailed)		0,000
CV(Sharpness)	Coefficient	0,859	1,000
	Sig. (2-tailed)	0,000	

Table A. 6 Spearman’s Correlation Coefficient for estimates of sharpness of the lesion border from the surrounding skin



Model with	five variables	Percent Concordant	94.1	Somers' D	0.885
	three variables		93.8		0.877
	five variables	Percent Discordant	5.6	Gamma	0.887
	three variables		6.1		0.879
	five variables	Percent Tied	0.3	Tau-a	0.257
	three variables		0.2		0.254
	five variables	Pairs	2507	c	0.942
	three variables		2507		0.939

Table A. 7 Association of Predicted Probabilities and Observed Responses for model with five variables –Rm, Lac_Grey, Cv_Sharp, Mean_Red, Range Blue- and three variables - Rm, Lac_Grey, Range Blue

		N	Mean	Std. Deviation	Std. Error Mean
Predicted probability	Melanoma	23	.6181048	.31861981	.06643682
	Not Melanoma	109	.0805836	.14863698	.01423684

Table A. 8 Group Statistics of predictive probability for model described by equation (3.1)



		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Predicted probability	Equal variances assumed	35.891	.000	12.427	130	.000	.5375212	.04325468	.45194702	.62309546
	Equal variances not assumed			7.911	24.057	.000	.5375212	.06794512	.39730686	.67773562

Table A. 9 Results of independent sample t-test for predictive probability for model described by equation (3.1)





Bibliography

- Abe, M. (1991).** A Moving Ellipsoid Method for Nonparametric Regression and Its Application to Logit Diagnostics With Scanner Data. *Journal of Marketing Research*, 28, 339-346
- Agresti, A. (1990).** *Categorical Data Analysis*. New York: John Wiley
- Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1990).** *Statistical Modelling in GLIM*. Oxford: Oxford University Press
- Akaike, H. (1974).** A new look at the statistical model identification", IEEE Transactions on Automatic Control, 19:716-723
- Alba, R. D. (1988).** Interpreting the Parameters of Log-Linear Models. pp. 258-287 In Scott Long J. (Ed.), *Common Problems/Proper Solutions. Avoiding Error in Quantitative Research*. Newbury Park: Sage
- Aldrich, J.H. and Nelson, F.E. (1984).** *Linear Probability, Logit, and Probit Models*. Newbury Park CA: Sage
- Allison, P. D. (1999).** Comparing logit and probit coefficients across groups. *Sociological Methods and Research*, 28, 2, 186-208
- Allison, P. D. (1987).** Introducing a Disturbance into Logit and Probit Regression Models. *Sociological Methods and Research* 15, 355-374
- Allison, P.D. (1999).** *Logistic Regression Using the SAS® System, Theory and Application*. SAS Institute Inc, Cary, NC
- Altman D.G., Bland M.J.(1994).** Statistics Notes. Diagnostic Tests 1: sensitivity and specificity. *British Medical Journal*, 308,1552
- Amemiya, T. and Frederick N. (1975).** A Modified Logit Model. *Review of Economics and Statistics* 57, 255-257
- Amemiya, T. (1981).** Qualitative Response Models: A Survey. *Journal of Econometric Literature* 19, 1483-1536
- Amemiya, T. (1984).** Tobit Models: A Survey. *Journal of Econometrics* 24, 3-61.
- Anas, A. (1983).** Discrete Choice Theory, Information Theory and the Multinomial Logit and Gravity Models. *Transportation Research B* 17, 13-23
- Anderson, J.A. (1984).** Regression and Ordered Categorical Variables. *Journal of the Royal Statistical Society B* 46, 1-30.



- Anderson, S., Auquier, A., Hauck, W. W., Oakes, D., Vandaele, W. and Weisberg, H. (1980).** *Statistical Methods for Comparative Studies. Techniques for Bias Reduction.* John Wiley, New York
- Andreassi L, Perotti R., Rubegni P., Burroni M., Cevenini G., Biagioli M., Taddeucci P., Dell'Eva G., Barbini P. (1999).** Digital dermoscopy analysis for the differentiation of atypical nevi and early melanoma: a new quantitative semiology. *Arch Dermatol.* 135, 1459-1465
- Arminger, G. (1983).** Multivariate Analyse von qualitativen abhängigen Variablen mit verallgemeinerten linearen Modellen. *Zeitschrift für Soziologie* 12, 49-64
- Armitage, P. (1971).** *Statistical Methods in Medical Research.* Blackwell Scientific Publications, Oxford
- Armstrong B.K., Kricger A. (1994).** Cutaneous Melanoma: Trends in cancer incidence and mortality. *Cancer Surv* 19-20, 219-240
- Arnold, S. F. (1990).** *Mathematical Statistics.* Prentice-Hall, Englewood Cliffs NJ
- Ashton, W. (1972).** *The Logit Transformation. With Special Reference to its Uses in Bioassay.* Griffin, London
- Atkinson, A.C. (1985).** *Plots, Transformations and Regression.* Oxford University Press, Oxford
- Azzalini, A., Bowman, A.W. and Härdle. W. (1989).** On the Use of Nonparametric Regression for Model Checking. *Biometrika* 76, 1-11
- Balch C.M. (1992).** Cutaneous melanomas. Prognosis and treatment results worldwide. *Semin Surgic Oncol* 8, 400-414
- Barnett, V. and Lewis, T. (1994).** *Outliers in Statistical Data (3rd edition).* John Wiley, New York
- Batsell, R. and Lodish, L.M. (1981).** A Model and Measurement Methodology for Predicting Individual Consumer Choice. *Journal of Marketing Research*, 18, 1-12
- Batsell, R.R. ((1980)).** Consumer Resource Allocation Models at the Individual Level. *Journal of Consumer Research*, 7, 78-87
- Becher, H. (1991).** Alternative Parametrization of Polychotomous Models: Theory and Applications to Matched Case-Control Studies. *Statistics in Medicine*, 10, 375-382
- Becher, H. (1992).** The Concept of Residual Confounding in Regression Models and Some Applications. *Statistics in Medicine* 11, 1747-1758



- Bechtel, G. G. (1990).** Share-Ratio Estimation of the Nested Multinomial Logit Model. *Journal of Marketing Research*, 27, 232-237
- Bedrick, E. J. and Hill, J. R. (1990).** Outlier Tests for Logistic Regression: A Conditional Approach. *Biometrika*, 77, 815-827
- Beggs, J. J. (1988).** A Simple Model for Heterogeneity in Binary Logit Models. *Economic Letters*, 27, 245-249
- Ben-Akiva, M. and Lerman, S. R. (1985).** *Discrete Choice Analysis*. The MIT Press, Cambridge MA
- Ben-Akiva, M., Morikawa, T. and Shiroishi, F. (1992).** Analysis of the Reliability of Preference Ranking Data. *Journal of Business Research*, 24, 149-164
- Berndt, E.K., Hall, B.H., Hall, R.E., and Hausman, J.A. (1974).** Estimation and Inference in Nonlinear Structural Models. *Annals of Economic and Social Measurement*, 3, 653-665
- Bhattacharjee, S.K. and Dunsmore, I.R. (1991).** The Influence of Variables in a Logistic Model. *Biometrika*, 78, 851-856
- Bishop, Y.M.M., Fienberg, S.E. & Holland, P. (1975).** *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Boston
- Blundell, R. and Meghir, C. (1987).** Bivariate Alternatives to the Tobit Model. *Journal of Econometrics* 34, 179-200
- Boos, D.B. (1993).** Analysis of Dose-Response Data in the Presence of Extrabinomial Variation. *Applied Statistics*, 42, 173-183
- Breen, R. (1994).** Individual Level Models for Mobility Tables and Other Cross-Classifications. *Sociological Methods and Research*, 23, 147-173
- Breslow, N and Powers W. (1978).** Are There Two Logistic Regressions for Retrospective Studies? *Biometrics*, 34, 100-105
- Breslow, N.E. and Day, N.E. (1980).** *Statistical Methods in Cancer Research. Vol. 1 - The Analysis of Case- Control Studies*. International Agency for Research on Cancer, Lyon
- Brown, C. C. (1982).** On a Goodness of Fit Test for the Logistic Model Based on Score Statistics. *Communications in Statistics*, 11, 1087-1105
- Brownstone, D. and Kenneth A. S. (1989).** Efficient Estimation of Nested Logit Models. *Journal of Business and Economic Statistics*, 7, 67-74
- Buckley, P. G. (1988).** Nested Multinomial Logit Analysis of Scanner Data of a Hierarchical Choice Model. *Journal of Business Research*, 17, 133-154



- Bucklin, R. E. and Sunil G. (1992).** Brand Choice, Purchase Incidence, and Segmentation: An Integrated Modeling Approach. *Journal of Marketing Research*, 29, 201-215
- Bunch, D. S. and Batsell, R. R. (1989).** A Monté Carlo Comparison of Estimators for the Multinomial Logit Model. *Journal of Marketing Research*, 26, 56-68
- Burnham, K. P., and Anderson, D. R. (1998).** *Model selection and inference: a practical information-theoretic approach*. Springer-Verlag, New York, New York, USA
- Bye, B. V., Gallicchio, S. J. and Levy J. M. (1987).** Estimation of Discrete Choice Models in Retrospective Samples. *Sociological Methods and Research*, 15, 467-492
- Carter, W. H. Jr., Chinchilli, V. M., Wilson, J. D., Campbell, E. D., Kessler F. K. and Carchman R. A. (1986).** An Asymptotic Confidence Region for the ED100p From the Logistic Response Surface for a Combination of Agents. *American Statistician*, 40, 124-128
- Caudill, S. B. (1988).** An Advantage of the Linear Probability Model over Probit or Logit. *Oxford Bulletin of Economics and Statistics*, 50, 425-427
- Cessie, S. L. (1991).** *Model Building Techniques for Logistic Regression, With Applications to Medical Data*. PhD Thesis Rijksuniversiteit Leiden
- Chakraborty, G., Woodworth, G. and Gaeth G. (1992).** Screening for Interactions Between Design Factors and Demographics in Choice-Based Conjoint. *Journal of Business Research* 24, 115-133
- Chambers, J., Cleveland, W., Kleiner, B. and Tukey, P. (1983).** *Graphical Methods for Data Analysis*, Duxbury Press, Boston
- Chatterjee S. and Price, B. (1991).** *Regression Analysis by Example 2nd edition*. John Wiley, New York
- Chow, G. C. (1983).** *Econometrics*. McGraw-Hill, Auckland
- Clark, W.A.V. and Onaka, J.L. (1985).** An Empirical Test of a Joint Model of Residential Mobility and Housing Choice. *Environment and Planning, A*, 17, 915-930
- Clayton, D. and Hills, M. (1993).** *Statistical Models in Epidemiology*. Oxford University Press, Oxford
- Cochran, W.G. (1968).** The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. *Biometrics*, 24, 295-313



- Coleman, J. (1981).** *Longitudinal Data Analysis*, Basic Books, New York
- Collett, D. (1991).** *Modelling Binary Data*. Chapman and Hall, London
- Colton, T. (1974).** *Statistics in Medicine*, Little, Brown and Co, Boston
- Copas, J.B. (1983).** Plotting p against x . *Applied Statistics*, 32, 25-31
- Corstjens, M. L. and Gautschi D. A. (1983).** Formal Choice Models in Marketing. *Marketing Science*, 2, 19-56
- Costanzo, C.M., Halperin, W.C., Gale, N.D. and Richardson, G.D. (1982).** An Alternative Method for Assessing Goodness-of-Fit for Logit Models. *Environment and Planning, A*, 14, 963-971
- Cox, D. R. (1970).** *The analysis of binary data*, Methuen, London
- Cox, D.R. & Snell, E.J. (1989).** *The Analysis of Binary Data*, 2nd Ed. Chapman & Hall, London
- Cox, D.R. and Hinkley, D.V. (1974).** *Theoretical Statistics*. Chapman and Hall, London
- Cox, D.R. and Snell, E.J. (1981).** *Applied Statistics. Principles and Examples*. Chapman and Hall, London
- Cragg G. J. and Uhler S. R. (1970).** The demand for automobiles, *Canadian Journal of Economics*, 3, 386-406
- Cramer, J.S. (1986).** Estimation of Probability Models From Income Class Data. *Statistica Neerlandica* 40, 237-247
- Cramer, J.S. (1991).** *Econometric Applications of Maximum Likelihood Methods*. Cambridge University Press, Cambridge
- Cramer, J.S. (1991).** *The LOGIT Model: An Introduction for Economists*. Edward Arnold, London
- Cramer, J.S. (1992).** *Association Among Pairs of Discrete Consumer Choices* (SEO Research Memorandum nr. 9201). SEO, Amsterdam
- Cramer, J.S. and Ridder, G. (1988).** The Logit Model in Econometrics. *Statistica Neerlandica*, 42, 297-314
- Currim, I. S. (1981).** Using Segmentation Approaches for Better Prediction and Understanding from Consumer Mode Choice Models. *Journal of Marketing Research* 18: 301-309.
- Currim, I. S. (1982).** Predictive Testing of Consumer Choice Models Not Subject to Independence of Irrelevant Alternatives. *Journal of Marketing Research*, 19, 208-202



- Cuthbertson, K., Hall, S. G. and Taylor, M. P. (1992). *Applied Econometric Techniques*. Ann Arbor: The University of Michigan Press
- Daganzo, C. F. (1979). The Statistical Interpretation of Predictions with Disaggregate Demand Models. *Transportation Science*, 13, 1-12
- Daganzo, C. (1979). *Multinomial Probit. The Theory and Its Application to Demand Forecasting*. Academic Press, New York
- Dalal, S.R. and Klein, R.W. (1988). A Flexible Class of Discrete Choice Models. *Marketing Science*, 7, 232-251
- Dalal, S. R., Fowlkes, E. B. and Hoadley, B. (1989). Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure. *Journal of the American Statistical Association*, 84, 945-957
- Daly, A. (1987). Estimating "Tree" Logit Models. *Transportation Research, B*, 21, 251-267
- DeLong E.R., DeLong D.M., and Clarke-Pearson D.L. (1988). Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach, *Biometrics* 44, 837-845
- DeMaris, A. (1990). Interpreting Logistic Regression Results: A Critical Commentary. *Journal of Marriage and the Family*, 52, 271-277
- DeMaris, A. (1992). *Logit Modeling. Practical Applications*. Sage, Newbury Park CA
- DeMaris, A. (1993). Odds versus Probabilities in Logit Equations: A Reply to Roncek. *Social Forces*, 71, 1057-1065
- Denk, C.E. and Finkerl, S. E. (1992). The Aggregate Impact of Explanatory Variables in Logit and Linear Probability Models. *American Journal of Political Science* 37, 785-804
- DiPrete, T. A. (1990). Adding Covariates to Loglinear Models for the Study of Social Mobility. *American Sociological Review*, 55, 757-773
- Dobson, A. J. (1990). *An Introduction to Generalized Linear Models*. Chapman and Hall, London
- Draper, N. and Smith, H. (1981). *Applied Regression Analysis (2nd edition)*. John Wiley, New York
- Dubin, J. A. (1986). A Nested Logit Model of Space and Water Heat System Choice. *Marketing Science*, 5, 112-124



- Eliason, S. R. (1993).** *Maximum Likelihood Estimation. Logic and Practice.* Sage, Newbury Park CA
- Elliott, D. and Hollenhorst, J. (1981).** Sequential Unordered Logit Applied to College Selection with Imperfect Information. *Behavioral Science*, 26, 366-378
- Engel, J. (1988).** Polytomous Logistic Regression. *Statistica Neerlandica* 42, 233-252
- Estrella, A. (1998).** A new measure of fit for equations with dichotomous dependent variables. *Journal of Business and Economic Statistics*, 16, (2), 198-205
- Everitt, B.S. (1992).** *The Analysis of Contingency Tables (2nd edition).* Chapman and Hall, London
- Falaris, E. M. (1987).** A Nested Logit Migration Model with Selectivity. *International Economic Review*, 28, 429-443
- Falsh, D. and Leonard E.W. (1979).** A Comparison of Two Logit Models in the Analysis of Qualitative Marketing Data. *Journal of Marketing Research*, 16, 533-538
- Finney, D.J. (1971).** *Probit Analysis (3rd edition).* Cambridge University Press, Cambridge
- Fischer, M. M. and Aufhauser, E. (1988).** Housing Choice in a Regulated Market. A Nested Multinomial Logit Analysis. *Geographical Analysis*, 20, 47-69
- Flath, D. and Leonard, E.W. (1979).** A Comparison of Two Logit Models. *Analysis of Qualitative Marketing Data* 16, 533-538
- Fleiss, J. L. (1981).** *Statistical Methods for Rates and Proportions.* 2nd Ed , John Wiley & Sons, New York
- Follmann, D. A. and Lambert, D. (1989).** Generalizing Logistic Regression by Nonparametric Mixing. *Journal of the American Statistical Association*, 84, 295-300
- Fomby, T. B., Hill, R. C., and Johnson, S. R. (1984).** *Advanced Econometric Methods,* Springer Verlag, New York
- Fotheringham, A.S. and Knudsen, D. C. (1986).** *Goodness-of-Fit Statistics.* Geo Books, Norwich
- Fowlkes, E.B. (1987).** Some Diagnostics for Binary Logistic Regression Via Smoothing. *Biometrika*, 74, 503-515



- Fox, J. (2000). *Multiple and generalized nonparametric regression*. Sage Publications, Thousand Oaks, CA
- Fox, J. (1987). Effect Displays for Generalized Linear Models. *Sociological Methodology*, 17, 347-361
- Gail, M.H., Wieand, S. and Piantadosi, S. (1983). Biased Estimates of Treatment Effect in Randomized Experiments with Nonlinear Regressions and Omitted Covariates. *Biometrika*, 71, 431-444
- Gail, M.H., Tan, W.Y. and Piantadosi, S. (1988). Tests for No Treatment Effect in Randomized Clinical Trials. *Biometrika*, 75, 57-64
- Gail, M. H. (1991). A Bibliography and Comments on the Use of Statistical Models in Epidemiology in the 1980s. *Statistics in Medicine*, 10, 1819-1885
- Geisser, S. (1993). *Predictive Inference: An Introduction*, Chapman & Hall, London
- Gelman A, Carlin, J.B., Stern, H.S., and Rubin D.B. (1995). *Bayesian Data Analysis*. Chapman & Hall, London
- Gensch, D.H. and Recker, W. W. (1979). The Multinomial, Multiattribute Logit Choice Model. *Journal of Marketing Research*, 16, 124-132
- Gilbert, N. (1993). *Analyzing Tabular Data. Loglinear and Logistic Models for Social Researchers*. UCL Press, London
- Goldberg, I. and Nold F. C. (1980). Does Reporting Deter Burglars? An Empirical Analysis of Risk and Return in Crime. *Review of Economics and Statistics* , 62, 424-431
- Goldberger, A. S. (1964). *Econometric Theory*. John Wiley, New York
- Goldberger, A. S. (1973). Correlations Between Binary Outcomes and Probabilistic Predictions. *Journal of the American Statistical Association* 68, 84
- Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield C.D., Lander, E.S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286, 531-537
- Gomulka, J. and Stern, N. (1990). The Employment of Married Women in the United Kingdom 1970-83. *Economica* 57, 171-199
- Gordon, D. V., Lin, Z., Osberg, L. and Phipps, S. (1994). Predicting Probabilities: Inherent and Sampling Variability in the Estimation of Discrete-Choice Models. *Oxford Bulletin of Economics and Statistics* 56, 13-31



- Govindarajuly, Z. (1988).** *Statistical Techniques in Bioassay*. Karger, Basel
- Green, P.J. (1983).** Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and Some Robust and Resistant Alternatives. *Journal of the Royal Statistical Society, B*, 46, 149-192
- Green, P. E., Carmone F. J. and Wachspresss D. P. (1977).** On the Analysis of Qualitative Data in Marketing Research. *Journal of Marketing Research*, 14, 52-59
- Green A, Martin N, Pfitzner J, O'Rourke M, Knight N. (1994).** Computer image analysis in the diagnosis of melanoma. *Journal of the American Academy of Dermatology*, 31, 958-964
- Greene, W. H. (1990).** *LIMDEP Version 6.0: User's Manual and Reference Guide*. Econometric Software, Bellport NY
- Greene, W. H. (1993).** *Econometric Analysis (2nd edition)*. Macmillan, New York
- Guadagni, P. M. and Little, J. D.C. (1983).** A Logit Model of Brand Choice Calibrated on Scanner Data. *Marketing Science* 2, 203-238
- Guilkey, D. K. and Schmidt, P. (1979).** Some Small Sample Properties of Estimators and Test Statistics in the Multivariate Logit Model. *Journal of Econometrics*, 10, 33-42
- Gunderson, M. (1974).** Retention of Trainees. A Study with Dichotomous Dependent Variables. *Journal of Econometrics*, 2, 79-93
- Haberman, S. J. (1982).** Analysis of Dispersion of Multinomial Responses. *Journal of the American Statistical Association*, 77, 568-580
- Haghighi, F., Banerjee, P. and Li, W. (1999).** Application of artificial neural networks in whole-genome analysis of complex diseases" (meeting abstract), *Cold Spring Harbor Meeting on Genome Sequencing & Biology*, 75
- Hagle, T. M. and Mitchell II G. E. (1992).** Goodness-of-Fit Measures for Probit and Logit. *American Journal of Political Science*, 36, 762-784
- Hanley A.J. and McNeil J. B.(1982).** The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve, *Radiology*, 143, 29-36
- Hanushek, E. A. and Jackson, J. E. (1977).** *Statistical Methods for Social Scientists*. Academic Press, New York
- Hartmann, P. H. (1991).** Logistische Regression und Probit-Modelle mit SPSS: Anmerkungen zu zwei sehr unterschiedlichen Prozeduren. *ZUMA-Nachrichten*, 28, 18-28



- Hastie, T.J. and Tibshirani, R.J. (1990).** *Generalized Additive Models*. Chapman and Hall, London
- Hastie, T. J. and Pregibon, D. (1992).** Generalized linear models (Chapter 6). *Statistical Models in S*, ed. J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole, Pacific Grove, California
- Hauck, W.W. & Donner, A. (1977).** Wald's Test as applied to hypotheses in logit analysis, *Journal of the American Statistical Association*, 72, 851-853.
- Hauck, W. W. (1983).** A Note on Confidence Bands for the Logistic Response Curve. *American Statistician* 37, 158-160
- Hauser, J. R. (1978).** Testing the Accuracy, Usefulness, and Significance of Probabilistic Choice Models: An Information-Theoretic Approach. *Operations Research*, 26, 406-421
- Hauser, J. R. and Urban G. L. (1977).** A Normative Methodology for Modeling Consumer Response to Innovation. *Operations Research*, 25, 579-619
- Hausman, J.A. (1978).** Specification Tests in Econometrics. *Econometrica*, 46, 1251-1271
- Hausman, J. and McFadden, D. (1983).** Specification Tests for the Multinomial Logit Model. *Econometrica*, 52, 1219-1240
- Healy, M.J.R. (1988).** *GLIM: An Introduction*. Oxford University Press, Oxford
- Hensher D. A. and Stopher, P. R. (1979).** *Behavioural Travel Modelling*. Croom Helm, London
- Hensher, D. A. (1983).** Achieving Representativeness of the Observable Component of the Indirect Utility Function in Logit Choice Models: An Empirical Evaluation. *Journal of Business*, 57, 265-280
- Hensher, D. A. (1986).** Sequential and Full Information Maximum Likelihood Estimation of a Nested Logit Model. *Review of Economics and Statistics*, 68, 657-667.
- Hensher, D. A. (1992).** The Use of Discrete Choice Models in the Determination of Community Choices in Public Issue Areas Impacting on Business Decision making. *Journal of Business Research*, 24, 165-175
- Hensher, D. A. and Johnson, L. W. (1981).** *Applied Discrete-Choice Modelling*. Halsted Press, New York
- Hensher, David A. and Peter R. Stopher. (1979).** Behavioural Travel Modelling. Pp. 11-51 In David A.



- Hoffman, S. D. and Duncan, G. J. (1988).** Multinomial and Conditional Logit Discrete-Choice Models in Demography. *Demography* 25, 415-427
- Horowitz, J.L. (1987).** Specification Tests for Nested Logit Models. *Environment and Planning, A*, 19, 395- 402
- Horowitz, J.L. (1979).** Confidence Intervals for Choice Probabilities of the Multinomial Logit Model. *Transportation Research Record*, 728, 23-29
- Horowitz, J.L. (1981).** Testing the Multinomial Logit Model against the Multinomial Probit Model without Estimating the Probit Parameters. *Transportation Science*, 15, 153-163
- Hosmer, D. & Lemeshow, S. (1989).** *Applied Logistic Regression*. John Wiley, New York
- Hosmer, D. W. and Lemeshow, S. (1980).** Goodness of Fit Tests for the Multiple Logistic Regression Model. *Communications in Statistics*, 10, 1043-1069
- Hosmer, D. W., Taber, S. and Lemeshow, S. (1991).** The Importance of Assessing the Fit of Logistic Regression Models: A Case Study. *American Journal of Public Health*, 81, 1630-1635
- Jaccard, J. (2001).** *Interaction effects in logistic regression*. Thousand Oaks, CA: Sage Publications. Quantitative Applications in the Social Sciences Series, No. 135
- Jennings, D.E. (1986).** Judging inference adequacy in logistic regression, *Journal of the American Statistical Association*, 81, 471-476
- Jennings, D. E. (1986).** Outliers and Residual Distributions in Logistic Regression. *Journal of the American Statistical Association* 81: 987-990.
- Johnson, W. (1985).** Influence Measures for Logistic Regression: Another Point of View. *Biometrika*, 72, 59-65
- Johnston J. (1991).** *Econometric Methods (3rd edition)*. McGraw-Hill, Auckland
- Jones, J. M. and Zufryden, F. S. (1982).** An Approach for Assessing Demographic and Price Influences on Brand Purchase Behavior. *Journal of Marketing*, 46, 36-46
- Judge, G. G., Hill, R. C., Griffiths, W. E., Lütkepohl, H. and Lee, T. C. (1982).** *Introduction to the Theory and Practice of Econometrics (2nd edition)*. John Wiley, New York



- Judge, G. G., Hill, R. C., Griffiths, W. E., Lütkepohl, H. and Lee, T. C. (1980).** *The Theory and Practice of Econometrics (2nd edition)*. John Wiley, New York
- Kay, R. and Little, S. (1986).** Assessing the Fit of the Logistic Model: A Case Study of Children with the Haemolytic Uraemic Syndrome. *Applied Statistics*, 35, 16-30
- Kay, R. and Little, S. (1987).** Transformations of the Explanatory Variables in the Logistic Regression Model for Binary Data. *Biometrika*, 74, 495-501
- Khuri, A. I. (1993).** *Advanced Calculus with Applications in Statistics*. John Wiley, New York
- King, G. (1989).** *Unifying Political Methodology. The Likelihood Theory of Statistical Inference*. Cambridge University Press, Cambridge
- Kleinbaum, D. G. and Kupper, L. L. (1978).** *Applied Regression Analysis and Other Multivariable Methods*. Wadsworth Publishing, Belmont, CA
- Kleinbaum, D. G. (1994).** *Logistic regression: A self-learning text*. Springer-Verlag, New York
- Kleinbaum, D. G., Kupper, L. L. and Morgenstern, H. (1982).** *Epidemiologic Research. Principles and Quantitative Methods*. Van Nostrand Reinhold, New York
- Koppelman, F. S. (1976).** Methodology for Analyzing Errors in Prediction With Disaggregate Choice Models. *Transportation Research Record*, 592, 17-23
- Kruskal, W. (1987).** Relative Importance by Averaging Over Orderings. *American Statistician*, 41, 6-10
- Kühnel, S. M. (1990).** Lassen sich mit SPSSx-Matrix anwenderspezifische Analyseproblemen lösen? *ZA-Information*, 27, 89-109
- Kühnel, S. M. (1992).** Sparsame Modellierung mit logistischen Zufallsnutzenmodellen. *ZA-Information*, 31, 70-92
- Kühnel, S. M., Jagodzinski W. and Terwey, M. (1989).** Teilnehmen und Boykottieren: Ein Anwendungsbeispiel der binären logistischen Regression mit SPSSx. *ZA-Information*, 25, 44-75
- Künsch, H. R., Stefanski, L. A. and Carroll, R. J. (1989).** Conditional Unbiased Bounded- Influence Estimation in General Regression Models, With Applications to Generalized Linear Models. *Journal of the American Statistical Association*, 84, 460-466



- Lakshmi-Ratan, R. A., Lanning, S. G. and Rotondo, J. A. (1992).** An Aggregate Contextual Choice Model for Estimating Demand for New Products from a Laboratory Choice Experiment. *Journal of Business Research*, 24, 97-114
- Landwehr, J. M., Pregibon, D. and Shoemaker, A. C. (1983).** Graphical Methods for Assessing Logistic Regression Models. *Journal for the American Statistical Association*, 79, 61-71
- Lee, E. T. (1992).** *Statistical Methods for Survival Data Analysis (2nd edition)*. John Wiley, New York
- Lee, H. L. and Cohen M. A. (1985).** A Multinomial Logit Model for the Spatial Distribution of Hospital Utilization. *Journal of Business and Economic Statistics*, 3, 159-168
- Lesaffre, E. and Molenberghs, G. (1991).** Multivariate Probit Analysis: A Neglected Procedure in Medical Statistics. *Statistics in Medicine*, 10, 1391-1403
- Lesaffre, E. and Kaufmann, H. (1992).** Existence and Uniqueness of the Maximum Likelihood Estimator for a Multivariate Probit Model. *Journal for the American Statistical Association*, 87, 805-811
- Li, W. and Yang, Y., (2002).** Zipf's law in importance of genes for cancer classification using microarray data. *Journal of Theoretical Biology*, 219, 539-551
- Li W., Sherriff, A. and Liu, X. (2000).** Assessing risk factors of complex diseases by Akaike information criterion and Bayesian information criterion, *American Journal of Human Genetics*, 67, 222
- Li W, Nyholt, D. (2001).** Marker selection by Akaike information criterion and Bayesian information criterion, *Genetic Epidemiology*, 21, 272-277
- Liang, K.Y., Zeger, S. L and Qaqish, B. (1992).** Multivariate Regression Analysis for Categorical Data. *Journal of the Royal Statistical Society B*, 44, 3-40
- Liao, T. F. (1994).** *Interpreting Probability Models. Logit, Probit, and Other Generalized Linear Models*. Sage, Newbury Park CA
- Linder, A. and Berchtold, W. (1976).** *Statistische Auswertung von Prozentzahlen. Probit- und Logitanalyse mit EDV*. Birkhäuser Verlag, Basel
- Lindsey, J. K., (1995).** *Modelling Frequency and Count Data*. Oxford University Press, Oxford
- Lindsey, J.K. (1997).** *Applying Generalized Linear Models*. Springer, New York



- Lindsey, J.K. (1973).** *Inferences From Sociological Survey Data. A Unified Approach.* Elsevier, Amsterdam
- Long, J. S. (1987).** A Graphical Method for the Interpretation of Multinomial Logit Analysis. *Sociological Methods and Research*, 15, 420-446
- Louviere, J. J. (1992).** Experimental Choice Analysis: Introduction and Overview. *Journal of Business Research*, 24, 89-95
- Ludwig-Mayerhofer, W. (1990).** Multivariate Logit-Modelle für ordinalskalierte abhängige Variablen. *ZA-Information*, 27, 62-88
- Lui, K.-J., McGee, D., Rhodes, P. and Pollock, D. (1988).** An Application of a Conditional Logistic Regression to Study the Effects of Safety Belts, Principal Impact Points, and Car Weight on Drivers' Fatalities. *Journal of Safety Research*, 19, 197-203
- Maddala, G.S. (1983).** *Limited-Dependent and Qualitative Variables in Econometrics.* Cambridge University Press, Cambridge UK
- Magee, L. (1990).** R^2 measures based on Wald and likelihood ratio joint significance tests, *The American Statistician*, 44, 250-253
- Magidson, J. (1981).** Qualitative Variance, Entropy, and Correlation Ratios for Nominal Dependent Variables. *Social Science Information*, 10, 177-194
- Maier, G. and Weiss, P. (1990).** *Modelle diskreter Entscheidungen. Theorie und Anwendung in den Sozial- und Wirtschaftswissenschaft.* Springer-Verlag, Wien
- Malhotra, N. K. (1983).** A Comparison of the Predictive Validity of Procedures for Analyzing Binary Data. *Journal of Business and Economic Statistics*, 1, 326-336
- Malhotra, N. K. (1983).** The Use of Linear Logit Models in Marketing Research. *Journal of Marketing Research*, 21, 20-31
- Malhotra, N. K., Jain, A. K. and Lagakos S.W. (1982).** The Information Overload Controversy: An Alternative Viewpoint. *Journal of Marketing*, 46, 27-37
- Manousaki. A., Manios A., Ioannidou D., Panayiotides. J., Tsiftsis D., Tosca A., Tsompanaki. E., Kostaki A. (2004).** A Simple Digital Image Processing System to aid in Melanoma diagnosis In An Everyday Melanocytic Skin Lesion Unit, (to appear)



- Manski, C. F. (1981).** Structural Models for Discrete Data: The Analysis of Discrete Choice. pp. 58- 109 in Samuel Leinhardt (Ed.), *Sociological Methodology 1981*. Jossey-Bass, San Francisco
- Mantel, N. (1989).** Confounding in Epidemiological Studies. *Biometrics*, 45, 1317-1318
- McCullagh, P. & Nelder, J.A. (1983).** *Generalized Linear Models*. Chapman & Hall, London
- McCullagh, P. and Nelder, J.A. (1989).** *Generalized Linear Model (2nd edition)*. Chapman and Hall, London
- McCullagh, P. (1980).** Regression Models for Ordinal Data. *Journal of the Royal Statistical Society, B*, 42, 109-142
- McFadden, D. and Train, K. (1976).** An Application of Diagnostic Tests for the Independence From Irrelevant Alternatives Property of the Multinomial Logit Model. *Transportation Research Board Record*, 637, 39-45
- McFadden, D. L. (1976).** The Mathematical Theory of Demand Models. Pp. 305-314 in Peter R. Stopher and Arnim H. Meyburg (Eds.), *Behavioral Travel-Demand Models*. Lexington MA: Lexington Books.
- McFadden, D. (1974)a.** The Measurement of Urban Travel Demand. *Journal of Public Economics*, 3, 330-328
- McFadden, D. (1974)b.** Conditional Logit Analysis of Qualitative Choice Behavior. pp. 105-152 in Paul Zarembka (Ed.), *Frontiers in Econometrics*. Academic Press, New York
- McFadden, D. (1979).** Quantitative Methods for Analysing Travel Behaviour of Individuals: Some Recent Developments. pp. 279-318 in David A. Hensher and Peter R. Stopher (Eds.), *Behavioural Travel Modelling*. Croom Helm, London
- McFadden, D. (1980).** Econometric Models for Probabilistic Choice Among Products. *Journal of Business*, 53, 13-36
- McFadden, D. (1986).** The Choice Theory Approach to Market Research. *Marketing Science*, 5, 275- 297
- McKelvey, R. and Zavoina, W. (1994).** A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4, 103-120



- McNeil J. B. and Hanley A.J. (1984).** Statistical Approaches to the Analysis of ROC curves, *Medical Decision Making*, 4, 136-149
- Menard, S. (2002).** *Applied logistic regression analysis, 2nd Edition.* Sage Publications, Thousand Oaks, CA
- Messinger, P. R. (1992).** A Nonparametric Test of Attribute Interaction in Consumer Utility Using Graded Comparisons. *Journal of Business Research*, 24, 135-148
- Morgan B.J.T. (1988).** Extended Models for Quantal Response Data. *Statistica Neerlandica*, 42, 253-272
- Morgan, B.J.T. (1992).** *Analysis of Quantal Response Data.* Chapman and Hall, London
- Morgan, S. P. and Teachman, J. D. (1988).** Logistic Regression: Description, Examples, and Comparisons. *Journal of Marriage and the Family*, 50, 929-936
- Morton DL, Wen DR, Foshag LJ, Essner R, Cochran A. (1993).** Intraoperative lymphatic mapping and selective cervical lymphadenectomy for early-stage melanomas of the head and neck. *Journal of Clinical Oncology*, 11, 1751-1756
- Nagelkerke, N. J. D. (1991).** A note on a general definition of the coefficient of determination. *Biometrika*, 78, 3, 691-692
- Nagler, J. (1994).** Scobit: An Alternative Estimator to Logit and Probit. *American Journal of Political Science*, 38, 230-255
- Nakanishi, M. and Cooper, L. G. (1982).** Simplified Estimation Procedures for MCI Models. *Marketing Science*, 1, 314-322
- Nelder, J.A. and Wedderburn, R.W.M. (1972).** Generalised linear models. *Journal of the Royal Statistical Society, A*, 135, 370 -384
- Nerlove, M. and Press, J. (1973).** *Univariate and Multivariate Log-Linear and Logistic Models.* RAND-R1306-EDA/NIH, Santa Monica, California
- Neter, J. and Wasserman, W. (1974).** *Applied Linear Statistical Models. Regression, Analysis of Variance, and Experimental Designs.* Homewood III., Richard D. Irwin
- Nownes, A. J. (1992).** Primaries, General Elections, and Voter Turnout. *American Politics Quarterly*, 20, 205-226
- Pampel, F. C. (2000).** *Logistic regression: A primer.* Sage, Newbury Park



- Parzen E, Tanabe, K., and Kitagawa, G. (1998).** *Selected Papers of Hirotugu Akaike*, Springer-Verlag, New York
- Petersen, T. (1985).** A Comment on Presenting Results from Logit and Probit Models. *American Sociological Review*, 50, 130-131
- Pickles, A. (1985).** *An Introduction to Likelihood Analysis*. Geo Books, Norwich
- Pindyck, R. S. and Rubinfeld D.L. (1991).** *Econometric Models and Economic Forecasts (3rd edition)*. McGraw-Hill, New York
- Pregibon, D. (1980).** Goodness of Link Tests for Generalized Linear Models. *Applied Statistics*, 29, 15- 24
- Pregibon, D. (1981).** Logistic Regression Diagnostics. *Annals of Statistics*, 9, 705-724
- Press, S. J. and Wilson, S. (1978).** Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, Vol. 73: 699-705.
- Pudney, S. (1989).** Modelling Individual Choice. *The Econometrics of Corners, Kinks and Holes*. Blackwell, Cambridge MA
- Raftery, A. E. (1995).** Bayesian model selection in social research. in P. V. Marsden, ed., *Sociological Methodology 1995*: 111-163. Tavistock, London
- Read, T. R. C. and Cressie, N. A. C. (1988).** *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer-Verlag, New York
- Retherford, R. D. and Choe, M. K. (1993).** *Statistical Models for Causal Analysis*. John Wiley, New York
- Rice, J. C. (1994).** Logistic regression: An introduction. in B. Thompson, ed., *Advances in social science methodology*, 3, 191-245, CT: JAI Press, Greenwich
- Ripley, B.D. (1996).** *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge
- Robinson, L. D. and Jewell, N.P. (1991).** Some Surprising Results About Covariate Adjustment in Logistic Regression Models. *International Statistical Review*, 58, 227-240
- Roncek, D. W. (1991).** Using Logit Coefficients to Obtain the Effects of Independent Variables on Changes in Probabilities. *Social Forces*, 70, 509-518



- Roncek, D. W. (1993).** When Will They Ever Learn that First Derivatives Identify the Effects of Continuous Independent Variables or "Officer, You Can't Give Me a Ticket, I Wasn't Speeding for an Entire Hour. *Social Forces*, 71, 1067-1078
- Rosner, B., Spiegelman, D. and Willett, W.C. (1992).** Correction of Logistic Regression Relative Risk Estimates and Confidence Intervals for Random Within-Person Measurement Error. *American Journal of Epidemiology*, 136, 1400-1413
- Ross, S. M. (1993).** *Introduction to Probability Models (5th edition)*. Academic Press, Boston
- Royston, P. (1992).** The Use of Cusums and Other Techniques in Modelling Continuous Covariates in Logistic Regression. *Statistics in Medicine*, 11, 1115-1129
- Salsburg, D. (1986).** *Statistics for Toxicologists*. Marcel Dekker, New York
- SAS Institute Inc. (1988).** *SAS Guide for Personal Computers*. Version 6.03, SAS Institute Inc, Cary, North Carolina
- SAS Institute Inc. (1989).** *SAS/STAT User's Guide* . Vol 2, version 6, 4th ed, SAS Institute Inc, Cary, North Carolina
- SAS Institute Inc. (2001).** *Enterprise Miner Software: Changes and Enhancements, Release 4.1*, SAS Institute Inc, Cary, North Carolina
- Schmidt, P. and Strauss, R. P. (1975).** The Prediction of Occupation Using Multiple Logit Models. *International Economic Review*, 16, 471-486
- Schmitz, P.I.M. (1986).** *Logistic Regression in Medical Decision Making and Epidemiology*. PhD Thesis Erasmus Universiteit Rotterdam.
- Schwarz, G. (1976).** Estimating the dimension of a model, *Annals of Statistics*, 6, 461-464
- Silk, A. J. and Urban G. L. (1978).** Pre-Test-Market Evaluation of New Packaged Goods: A Model and Measurement Methodology. *Journal of Marketing Research*, 15, 171-191
- Silverstone H. (1957).** Estimating the Logistic Curve, *Journal of the American Statistical Association*, 52, 567-577
- Siminoff, J. S. (1998).** Logistic Regression, Categorical Predictors, and Goodness - of-Fit: It Depends on Who You Ask, *The American Statistician*, 52, 1, 10-14



- Sober A.J. and Burstein J.M. (1994).** Computerized digital image analysis: an aid for melanoma diagnosis--preliminary investigations and brief review. *Journal of Dermatology*, 21,885-890
- Soofi, E. S. (1992).** A Generalizable Formulation of Conditional Logit With Diagnostics. *Journal of the American Statistical Association*, 87, 812-816
- Stage, F. K. (1988).** University Attrition: LISREL with Logistic Regression for the Persistence Criterion. *Research in Higher Education*, 29, 343-357
- Steckel, J. H. and Vanhonacker, W. R. (1988).** A Heterogeneous Conditional Logit Model of Choice. *Journal of Business and Economic Statistics*, 6, 391-398
- Stern, S. (1989).** Rules of Thumb for Comparing Multinomial Logit and Multinomial Probit Coefficients. *Economic Letters*, 31, 235-238
- Stinchcombe, A. L. (1983).** Linearity in Log-Linear Analysis. in Samuel Leinhardt (Ed.), *Sociological Methodology 1983-1984*. pp. 104-125, Jossey-Bass, San Francisco
- Stopher, P. R. (1975).** Goodness-of-Fit Measures for Probabilistic Travel Demand Models. *Transportation*, 4, 67-83
- Tabachnick, B.G., and Fidell, L. S. (1996).** *Using multivariate statistics*, 3rd ed. Harper Collins, New York
- Talvitie, A. P. (1976).** Mathematical Theory of Travel Demand. in Peter R. Stopher and Arnim H. Meyburg (Eds.), *Behavioral Travel-Demand Models*. pp. 283-303, Lexington Books, Lexington MA
- Tardiff, T. J. (1976).** A Note on Goodness-of-Fit Statistics for Probit and Logit Models. *Transportation*, 5, 377-388
- Theil, H. and Chung C. F. (1988).** Information-Theoretic Measures of Fit for Univariate and Multivariate Linear Regressions. *American Statistician*, 42, 249-252
- Theil, H. (1969).** A Multinomial Extension of the Linear Logit Model. *International Economic Review*, 10, 251-259
- Theil, H. (1970).** On the Estimation of Relationships Involving Qualitative Variables. *American Journal of Sociology*, 76, 103-154
- Theil, H. (1971).** *Principles of Econometrics*. John Wiley, New York
- Thill, J.C. (1992).** Choice Set Formation for Destination Choice Modelling. *Progress in Human Geography*, 16, 361-382



- Timmermans, H. and Golledge, R. G. (1990).** Applications of Behavioural Research on Spatial Problems II: Preference and Choice. *Progress in Human Geography* 14, 311-354
- Timmermans, H., Borgers, A. and Van Der Waerden, P. (1992).** Mother Logit Analysis of Substitution Effects in Consumer Shopping Destination Choice. *Journal of Business Research*, 24, 177-189
- Train, K. (1986).** *Qualitative Choice Analysis. Theory, Econometrics, and an Application to Automobile Demand.* The MIT Press, Cambridge MA
- Tse, Y.K. (1987).** A Diagnostic Test for the Multinomial Logit Model. *Journal of Business and Economic Statistics*, 5, 283-286
- Tsiatis, A. A. (1980).** A Note on a Goodness-of-Fit Test for the Logistic Regression Model. *Biometrika*, 67, 250-251
- Urban, D. (1993).** *Logit-Analyse. Statistische Verfahren zur Analyse von Modellen mit Qualitativen Response-Variablen.* G. Fischer, Stuttgart
- Van Houwelingen J.C. and Le Cessie, S. (1988).** Logistic Regression. A Review. *Statistica Neerlandica*, 42, 215-232
- Van Houwelingen, J. C. and. Le Cessie S. (1990).** Predictive Value of Statistical Models, *Statistics in Medicine* 1303-1325
- Veall, Michael R. and Zimmermann, K. F. (1992).** Pseudo- R^2 's in the Ordinal Probit Model. *Journal of Mathematical Sociology*, 16, 333-342
- Veall, M. R. and Zimmermann, K. F. (1994)a.** Evaluating Pseudo- R^2 's for Binary Probit Models. *Quality and Quantity*, 28, 151-164
- Veall, M. R. and Zimmermann, K. F. (1994)b.** Goodness of Fit Measures in the Tobit Model. *Oxford Bulletin of Economics and Statistics*, 56, 485-499
- Vidmar, T. J., McKean, J. W. and Hettmansperger, T. P. (1992).** Robust Procedures for Drug Combination Problems with Quantal Responses. *Applied Statistics*, 41, 299-315
- Walsh, A. (1987).** Teaching Understanding and Interpretation of Logistic Regression. *Teaching Sociology*, 15, 178-183
- Wasson, J.H., Sox, H.C., Neff, R.K., Goldman L. (1985).** Clinical prediction rules: applications and methodological standards. *N Engl J Med*, 313, 793-799
- Weiler, W. C. (1987).** An Application of the Nested Multinomial Logit Model to Enrollment Choice Behavior. *Research in Higher Education*, 27, 273-282



- Weinberg, C. R. (1985).** On Pooling Across Strata When Frequency Matching Has Been Followed in a Cohort Study. *Biometrics*, 41, 117-127
- Weisberg, S. (1985).** *Applied Linear Regression*. John Wiley, New York
- Westin, R. B. (1974).** Predictions From Binary Choice Models. *Journal of Econometrics*, 2, 1-16
- Wetherill, G. B. (1986).** *Regression Analysis with Applications*. Chapman and Hall, London
- Williams, D.A. (1987).** Generalized Linear Model Diagnostics Using the Deviance and Single Case Deletions. *Applied Statistics*, 36, 181-191
- Windmeijer, F. A.G. (1992).** *Goodness of Fit in Linear and Qualitative-Choice Models*. Thesis Publishers. Amsterdam
- Wright, R.E. (1995).** Logistic regression. in L.G. Grimm & P.R. Yarnold, eds., *Reading and understanding multivariate statistics*. American Psychological Association, Washington, DC
- Wrigley, N. (1976).** *An Introduction to the Use of Logit Models in Geography*. Geo Books, Norwich
- Wrigley, N. (1982).** Quantitative Methods: Developments in Discrete Choice Modelling. *Progress in Human Geography*, 6, 547-562
- Wrigley, N. (1985).** *Categorical Data Analysis for Geographers and Environmental Scientists*. Longman., London
- Wrigley, N. (1990).** Unobserved Heterogeneity and the Analysis of Longitudinal Spatial Choice Data. *European Journal of Population*, 6, 327-358
- Xie, Y. and Manski, C. F. (1989).** The Logit Model and Response-Based Samples. *Sociological Methods and Research*, 17, 283-302
- Zhang, J. and Hoffman S. D. (1993).** Discrete-Choice Logit Models. Testing the IIA Property. *Sociological Methods and Research*, 22, 193-213
- Zipf G.F. (1949).** *Human Behavior and the Principle of Least Effect* (Addison-Wesley), Reading, MA



