APPLICATION OF LOGISTIC REGRESSION IN BIOSTATISTICS

Yin Li
Department of Mathematics and Statistics
McGill University, Montreal
September 1993

A Thesis submitted to the Faculty of Graduate Studies
and Research
in partial fulfillment of the requirements for the degree of
M.Sc.

Name Yin Li

Dissertation Abstracts International is arranged by broad, general subject categories. Please select the one subject which most nearly describes the content of your dissertation. Enter the corresponding four-digit code in the spaces provided.

Statistics | 0463 | U·M·I

SUBJECT TERM | SUBJECT CODE

Subject Categories

THE HUMANITIES AND SOCIAL SCIENCES

COMMUNICATIONS AND THE ARTS

Architecture	0729
Art History	0377
Cinema	0900
Dance	0378
Fine Arts	0357
Information Science	0723
Journalism	0391
Library Science	0399
Mass Communications	0708
Music	0413
Speech Communication	0459
Theater	0465

EDUCATION

General	0515
Administration	0514
Adult and Continuing	0516
Agricultural	0517
Art	0273
Bilingual and Multicultural	0282
Business	0688
Community College	0275
Curriculum and Instruction	0727
Early Childhood	0518
Elementary	0524
Finance	0277
Guidance and Counseling	0519
Health	0680
Higher	0745
History of	0520
Home Economics	0278
Industrial	0521
Language and Literature	0279
Mathematics	0280
Music	0522
Philosophy of	0998
Physical	0523
Psychology	0525
Reading	0535
Religious	0527
Sciences	0714
Secondary	0533
Social Sciences	0534
Sociology of	0340
Special	0529
Teacher Training	0530
Technology	0710
Tests and Measurements	0288
Vocational	0747

LANGUAGE, LITERATURE AND LINGUISTICS

Language	
General	0679
Ancient	0289
Linguistics	0290
Modern	0291
Literature	
General	0401
Classical	0294
Comparative	0295
Medieval	0297
Modern	0298
African	0316
American	0591
Asian	0305
Canadian (English)	0352
Canadian (French)	0355
English	0593
Germanic	0311
Latin American	0312
Middle Eastern	0315
Romance	0313
Slavic and East European	0314

PHILOSOPHY, RELIGION AND THEOLOGY

Philosophy	0422
Religion	
General	0318
Biblical Studies	0321
Clergy	0319
History of	0320
Philosophy of	0322
Theology	0469

SOCIAL SCIENCES

American Studies	0323
Anthropology	
Archaeology	0324
Cultural	0326
Physical	0327
Business Administration	
General	0310
Accounting	0272
Banking	0770
Management	0454
Marketing	0338
Canadian Studies	0385
Economics	
General	0501
Agricultural	0503
Commerce Business	0505
Finance	0508
History	0509
Labor	0510
Theory	0511
Folklore	0358
Geography	0366
Gerontology	0351
History	
General	0578
Ancient	0579
Medieval	0581
Modern	0582
Black	0328
African	0331
Asia, Australia and Oceania	0332
Canadian	0334
European	0335
Latin American	0336
Middle Eastern	0333
United States	0337
History of Science	0585
Law	0398
Political Science	
General	0615
International Law and Relations	0616
Public Administration	0617
Recreation	0814
Social Work	0452
Sociology	
General	0626
Criminology and Penology	0627
Demography	0938
Ethnic and Racial Studies	0631
Individual and Family Studies	0628
Industrial and Labor Relations	0629
Public and Social Welfare	0630
Social Structure and Development	0700
Theory and Methods	0344
Transportation	0709
Urban and Regional Planning	0999
Women's Studies	0453

THE SCIENCES AND ENGINEERING

BIOLOGICAL SCIENCES

Agriculture	
General	0473
Agronomy	0285
Animal Culture and Nutrition	0475
Animal Pathology	0476
Food Science and Technology	0359
Forestry and Wildlife	0478
Plant Culture	0479
Plant Pathology	0480
Plant Physiology	0817
Range Management	0777
Wood Technology	0746
Biology	
General	0306
Anatomy	0287
Biostatistics	0308
Botany	0309
Cell	0379
Ecology	0329
Entomology	0353
Genetics	0369
Limnology	0793
Microbiology	0410
Molecular	0307
Neuroscience	0317
Oceanography	0416
Physiology	0433
Radiation	0821
Veterinary Science	0778
Zoology	0472
Biophysics	
General	0786
Medical	0760

EARTH SCIENCES

Biogeochemistry	0425
Geochemistry	0996
Geodesy	0370
Geology	0372
Geophysics	0373
Hydrology	0388
Mineralogy	0411
Paleobotany	0345
Paleoecology	0426
Paleontology	0418
Paleozoology	0985
Palynology	0427
Physical Geography	0368
Physical Oceanography	0415

HEALTH AND ENVIRONMENTAL SCIENCES

Environmental Sciences	0768
Health Sciences	
General	0566
Audiology	0300
Chemotherapy	0992
Dentistry	0567
Education	0350
Hospital Management	0769
Human Development	0758
Immunology	0982
Medicine and Surgery	0564
Mental Health	0347
Nursing	0569
Nutrition	0570
Obstetrics and Gynecology	0380
Occupational Health and Therapy	0354
Ophthalmology	0381
Pathology	0571
Pharmacology	0419
Pharmacy	0572
Physical Therapy	0382
Public Health	0573
Radiology	0574
Recreation	0575
Speech Pathology	0460
Toxicology	0383
Home Economics	0386

PHYSICAL SCIENCES

Pure Sciences

Chemistry	
General	0485
Agricultural	0749
Analytical	0486
Biochemistry	0487
Inorganic	0488
Nuclear	0738
Organic	0490
Pharmaceutical	0491
Physical	0494
Polymer	0495
Radiation	0754
Mathematics	0405
Physics	
General	0605
Acoustics	0986
Astronomy and Astrophysics	0606
Atmospheric Science	0608
Atomic	0748
Electronics and Electricity	0607
Elementary Particles and High Energy	0798
Fluid and Plasma	0759
Molecular	0609
Nuclear	0610
Optics	0752
Radiation	0756
Solid State	0611
Statistics	0463

Applied Sciences

Applied Mechanics	0346
Computer Science	0984
Engineering	
General	0537
Aerospace	0538
Agricultural	0539
Automotive	0540
Biomedical	0541
Chemical	0542
Civil	0543
Electronics and Electrical	0544
Heat and Thermodynamics	0348
Hydraulic	0545
Industrial	0546
Marine	0547
Materials Science	0794
Mechanical	0548
Metallurgy	0743
Mining	0551
Nuclear	0552
Packaging	0549
Petroleum	0765
Sanitary and Municipal	0554
System Science	0790
Geotechnology	0428
Operations Research	0796
Plastics Technology	0795
Textile Technology	0994

PSYCHOLOGY

General	0621
Behavioral	0384
Clinical	0622
Developmental	0620
Experimental	0623
Industrial	0624
Personality	0625
Physiological	0989
Psychobiology	0349
Psychometrics	0632
Social	0451

Abstract

The primary objective of this paper is a focused introduction to the logistic regression model and its use in methods for modeling the relationship between a dichotomous outcome variable and a set of covariates. The approach we will take is to develop the model from a regression analysis point of view. Also in this paper, an estimator of the common odds ratio in one-to-one matched case-control studies is proposed. The connection between this estimator and the James-Stein estimating procedure is highlighted through the argument of estimating functions. Comparisons are made between this estimator, the conditional maximum likelihood estimator, and the estimator ignoring the matching.

Résumé

L'objet principal de cet article est une introduction au modèle de régression logistique et de son utilisation dans les méthodes de modélisation des relations entre les conséquences d'une variable dichotomique et un ensemble de covariates. L'approche que nous utiliserons est de développer le modèle à partir du point de vue d'une analyse de régression. Aussi, dans cet article, un estimateur du rapport entre les probabilités et le couplage un à un de l'étude du cas de contrôle est proposé. La connexion entre cet estimateur et la procédure d'estimation de James-Stein est mise en lumière au travers de l'argument des fonctions d'estimation. Les comparaisons faites entre cet estimateur, l'estimateur du maximum de vraisemblance conditionnel et l'estimateur ignorant l'assortiment.

Acknowledgements

I would like to thank my supervisor, Prof. Minggao Gu for the supportive and informative discussions we had during the prepartion of this thesis.

I also want to express my gratitude to Department of Mathematics and Statistics at McGill University for the financial support.

Contents

Chapter 1

The Logistic Regression Model

1.1 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 on two or more possible values. Over the last decade the logistic regression model has become, in many fields, the standard method of analysis in this situation.

What distinguishes a logistic regression model from the linear regression model is that the outcome variable in logistic regression is binary or dichotomous. In this thesis, we express binary variable as present ($y = 1$) and absent ($y = 0$). 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 will motivate our approach to logistic regression.

1.2 The Logistic Regression Model

Consider a collection of p independent variables which will be denoted by the vector $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$. For the moment we will assume that each of these variables is at least interval scaled. Let the conditional probability that the outcome is present be denoted by $P(Y = 1|x) = \pi(x)$. Then the logit of the multiple logistic regression model is given by the equation

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad (1.1)$$

in which case

$$\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}} \qquad (1.2)$$

If some of the independent variables are discrete, nominal scaled variables such as race, sex, treatment group, and so forth, then it is inappropriate to include them in the model as if they were interval scaled. This is because the numbers used to represent the various levels are merely identifiers, and have no numeric significance. In this situation the method of choice is to use a collection of **design variables** (or **dummy variables**). Suppose, for example, that one of the independent variables is race, which has been coded as "white", "black" or "other." In this case two design variables are necessary. One possible coding strategy is that when the respondent is "white", the two design variables, D_1 and D_2, would both be set to zero; when the respondent is "black," D_1 would be set equal to 1 while D_2 would still equal 0; when the race of the respondent is "other," we would use $D_1 = 0$ and $D_2 = 1$. Table 1.1 illustrates this coding of the design variables.

Table 1.1 An Example of the Coding of the Design Variables for Race, Coded at Three Levels.

	Design	Variable
RACE	D_1	D_2
White	0	0
Black	1	0
Other	0	1

Most logistic regression software will generate the design variables, and some programs have a choice of several different methods.

In general, if a nominal scaled variable has k possible values, then k-1 design variables will be needed. This is true since, unless stated otherwise, all of our models have a constant term. The notation to indicate design variables to be used in this text follows.

Suppose that the j^{th} independent variable, x_j has k_j levels. The k_j-1 design variables will be denoted as D_{ju} and the coefficients for these design variables will be denoted as β_{ju}, $u = 1, 2, \ldots, k_j - 1$. Thus, the logit for a model with p variables and the j^{th} variable being discrete would be

$$g(x) = \beta_0 + \beta_1 x_1 + \cdots + \sum_{u=1}^{k_j - 1} \beta_{ju} D_{ju} + \cdots + \beta_p x_p$$

When discussing the multiple logistic regression model we will, in general, suppress the summation and double subscripting needed to indicate when design variables are being used

1.3 Fitting the Logistic Regression Model

Assume that we have a sample of n independent observations of the pair (x_i, y_i), $i = 1, 2, \ldots, n$ Fitting the model requires that we obtain estimates of the vector $\beta' = (\beta_0, \beta_1, \ldots, \beta_p)$. The method of estimation used is maximum likelihood. The likelihood function is

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}$$

The log likelihood is defined as

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^{n}\{y_i \ln[\pi(x_i)] + (1 - y_i)\ln[1 - \pi(x_i)]\} \tag{1.3}$$

There will be p+1 likelihood equations which are obtained by differentiating the log likelihood function with respect to the p+1 coefficients. The likelihood equations that result may be expressed as follows.

$$\sum_{i=1}^{n}[y_i - \pi(\mathbf{x}_i)] = 0$$

and

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(\mathbf{x}_i)] = 0$$

for $j = 1, 2, \ldots, p$.

The solution of the likelihood equations requires special purpose software which may be found in many packaged programs. Let $\hat{\beta}$ denote the solution to those equations. Thus, the fitted values for the multiple logistic regression model are $\hat{\pi}(\mathbf{x_i})$, the value of the expression in equation (1.2) computed using $\hat{\beta}$, and $\mathbf{x_i}$.

Now we will consider the method of estimating the variances and covariances of the estimated coefficients follows from well-developed theory of maximum likelihood estimation. This theory states that the estimators are obtained from the matrix of second partial derivatives of the log likelihood function. These partial derivatives have the following general form

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i (1 - \pi_i) \tag{1.4}$$

and

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_u} = -\sum_{i=1}^{n} x_{ij}^2 x_{iu} \pi_i (1 - \pi_i) \tag{1.5}$$

for $j, u = 0, 1, 2, \ldots, p$ where π_i denotes $\pi(\mathbf{x_i})$. Let the $(p+1)$ by $(p+1)$ matrix containing the negative of the terms given in equations (1.4) and (1.5) be denoted as $\mathbf{I}(\beta)$. This matrix is called the **information matrix**. The variances and covariances of the estimated coefficients are obtained from the inverse of this matrix which we will denote as $\sum(\beta) = \mathbf{I}^{-1}(\beta)$. Except in very special cases it is not possible to write down an explicit expression for the elements in this matrix. Hence, we will use the notation $\sigma^2(\beta_j)$ to denote the j^{th} diagonal element of this matrix, which is the variance of $\hat{\beta}_j$, and $\sigma(\beta_j, \beta_u)$ to denote an arbitrary off-diagonal element, which is the covariance of $\hat{\beta}_j$ and $\hat{\beta}_u$. The estimators of the variances and covariances, which will be denoted by $\hat{\sum}(\hat{\beta})$, are obtained by evaluating $\sum(\beta)$ at $\hat{\beta}$. We will use $\hat{\sigma}^2(\hat{\beta}_j)$ and $\hat{\sigma}(\hat{\beta}_j, \hat{\beta}_u)$, $j, u = 0, 1, 2, \ldots, p$, to denote the values in this matrix. For the most part we will have occasion to use only the estimated standard errors of the estimated coefficients, which we will denote as

$$\hat{SE}(\hat{\beta}_j) = [\hat{\sigma}^2(\hat{\beta}_j)]^{1/2} \tag{1.6}$$

for $j = 0, 1, 2, \ldots, p$.

A formulation of the information matrix which will be useful when discussing model fitting and assessment of fit is $\hat{\mathbf{I}}(\hat{\beta}) = \mathbf{X}'\mathbf{V}\mathbf{X}$ where $\mathbf{X}$ is an n by $p+1$ matrix containing the data for each subject, and $\mathbf{V}$ is an n by n diagonal matrix with general element $\hat{\pi}_i(1 - \hat{\pi}_i)$. That is, the matrix $\mathbf{X}$ is

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ & & \vdots & \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$$

and the matrix $\mathbf{V}$ is

$$\mathbf{V} = \begin{bmatrix} \hat{\pi}_1(1 - \hat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \hat{\pi}_2(1 - \hat{\pi}_2) & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \hat{\pi}_n(1 - \hat{\pi}_n) \end{bmatrix}$$

Now we present an example that will illustrate the formulation of a multiple logistic regression model and the estimation of its coefficients. We use a subset of the variables from the data for the low birth weight study. The goal of this study was to identify risk factors associated with giving birth to a low birth weight baby (weighting less than 2500 grams). In this study data were collected on 189 women, $n_1 = 59$ of which had low birth weight babies and $n_0 = 130$ of which had normal birth weight babies. Four variables which were thought to be of importance were age, weight of the subject at her last menstrual period, race, and number of physician visits during the first trimester of the pregnancy. In this example, the variable race has been recoded using the two design variables shown in Table 1.1. The results of fitting the logistic regression model to these data are given in Table 1.2.

Table 1.2 Estimated Coefficients for a Logistic Regression Model Using the Variables AGE, Weight at last Menstrual Period (LWT), RACE, and Number of First Trimester Physician Visits (FTV) from the Low Birth Weight Data Set

Variable	Estimated Coefficient	Estimated Standard Error	Coeff./SE
AGE	-0.021	0.031	-0.71
LWT	-0.011	0.007	-2.11
RACE(1)	1.001	0.197	2.02
RACE(2)	0.133	0.362	1.20
FTV	-0.019	0.167	-0.30
Constant	1.295	1.069	1.21

Log-likelihood = −111.286

In Table 1.2 the estimated coefficients for the two design variables for race are indicated in the lines denoted by "(1)" and "(2)." The estimated logit is given by the following expression:

$$\begin{aligned}\hat{g}(\mathbf{x}) &= 1.295 - 0.021 \times AGE - 0.011 \times LWT + 1.001 \times D_{31} \\ &\quad +0.433 \times D_{32} - 0.019 \times FTV\end{aligned}$$

where $D_{3i}, i = 1, 2$, denotes the two design variables for RACE. Refer Table 1.2 for coding D_{31} and D_{32}. The fitted values are obtained using the estimated logit, $\hat{g}(\mathbf{x})$.

1.4 Testing for the Significance of the Model

Once we have fit a particular multiple (multivariate) logistic regression model, we begin the process of assessment of the model. The first step in this process is usually assessing the significance of the variables in the model. The test is based on the statistic G

$$G = -2\ln\left[\frac{(\text{likelihood without the variable})}{(\text{likelihood with the variable})}\right] \tag{1.7}$$

Under the null hypothesis that the p "slope" coefficients for the covariates in the model are equal to zero, the distribution of G will be chi-square with p degrees of freedom.

As an example, consider the fitted model whose estimated coefficients are given in Table 1.2. For that model the value of the log likelihood is $L = -111.286$. A second model, fit with the constant term only, yields $L = -117.336$. Hence $G = -2[(-117.336) - (-111.286)] = -2(-6.05) = 12.1$. The p-value for the test is $P[\chi^2(5) > 12.1] = 0.033$ which is significant at the $\alpha = 0.05$ level. Rejection of the null hypothesis in this case has an interpretation analogous to that in multiple linear regression; we may conclude that at least one, and perhaps all p coefficients are different from zero.

Before concluding that any or all of the coefficients are nonzero, we may wish to look at the univariate Wald test statistics, $W_j = \hat{\beta}_j / \hat{SE}(\hat{\beta}_j)$. These are given in the last column in Table 1.2. Under the hypothesis that an individual coefficient is zero, these statistics may give us an indication of which of the variables in the model may or may not be significant. If we use a critical value of 2, which would conclude that the variables LWT and possibly RACE are significant, while AGE and FTV are not significant.

Considering that the overall goal is to obtain the best fitting model while minimizing the number of parameters, the next logical step is to fit a reduced model containing only those variables thought to be significant, and compare it to the full model containing all the variables. The results of fitting the reduced model are given in Table 1.3.

Table 1.3 Estimated Coefficients for a Logistic Regression Model Using the Variables LWT and RACE from the Low Birth Weight Data Set.

Variable	Estimated Coefficient	Estimated Standard Error	Coeff./SE
LWT	-0.015	0.006	-2.37
RACE(1)	1.081	0.487	2.22
RACE(2)	0.481	0.356	1.35
Constant	0.806	0.843	0.96

Log-likelihood = −111.630

The difference between the two models is the exclusion of the variables AGE and FTV from the full model. The likelihood ratio test comparing these two models is obtained using the definition of G given in equation (1.7). It will have a distribution that is chi-square with 2 degrees of freedom under the hypothesis that the coefficients for the variables excluded are equal to zero. The value of the test statistic comparing the models in Table 1.2 and 1.3 is

$$G = -2[(-111.630) - (-111.286)] = 0.688$$

which, with 2 degrees of freedom, has a p-value of $P[\chi^2(2) > 0.688] = 0.71$. Since the p-value is large, exceeding 0.05, we conclude that the reduced model is as good as the full model. Thus there is no advantage to including AGE and FTV in the model. However, we must not base our models entirely on tests of statistical significance. As we will see later, there are numerous other considerations that will influence our decision to include or exclude variables from a model.

Whenever a categorical scaled independent variable is included (or excluded) from a model, all of its design variables should be included (or excluded); to do otherwise implies that we have recoded the variable. For example, if we only include design variable D_1 as defined in Table 1.1, then race is entered into the model as a dichotomous variable coded as black or not black. If k is the number of levels of a categorical variable, then the contribution to the degrees of freedom for the likelihood ratio test for the exclusion of this variable will be k-1. For example, if we exclude race from the model, and race is coded at three levels using the design variables shown in Table 1.1, then there would be 2 degrees of freedom for the test, one for each design variable.

Because of the multiple degrees of freedom we must be careful in our use of the Wald (W) statistics to assess the significance exceed 2, then we could conclude that the design variables are significant. Alternatively, if one coefficient has a W statistic of 3.0 and the other a value of 0.1, then we cannot be sure about the contribution of the variable to the model. The estimated coefficients for the variable RACE in Table 1.3 provide a good example. The Wald statistic for the coefficient for the first design variable is 2.22, and 1.35 for the second. The likelihood ratio test comparing the model containing LWT and RACE to the one containing only LWT yields $G = -2[-111.315 - (-111.630)] = 5.43$ which, with 2 degrees of freedom, yields a p-value of 0.066. Strict adherence to the $\alpha = 0.05$ level of significance would justify excluding RACE from the model. However, RACE is known to be a "biologically important" variable. In this case the decision to include or exclude RACE should be made in conjunction with subject matter experts.

The multivariate analog of the Wald test is obtained from the following vector-matrix calculation

$$\begin{aligned} W &= \hat{\beta}'[\hat{\sum}(\hat{\beta})]^{-1}\hat{\beta} \\ &= \hat{\beta}'(\mathbf{X}'\mathbf{V}\mathbf{X})\hat{\beta} \end{aligned}$$

which will be distributed as chi-square with $p+1$ degrees of freedom under the hypothesis that each of the $p+1$ coefficients is equal to zero. Tests for just the p slope coefficients are obtained by eliminating $\hat{\beta}_0$ from $\hat{\beta}$ and the relevant row (first) and column (first) from $(\mathbf{X}'\mathbf{V}\mathbf{X})$.

1.5 Interpretation of the Coefficients (β's and β_0)

1.5.1 Interpretation of β's

Let's consider an example of a cohort study (Framingham) of 12-year incidence of coronary heart disease (CHD) of 712 men aged 40-49 at start of study.

88 of these men developed CHD within 12 years. Which of the following seven factors measured at initial visit affect the incidence of CHD.

X_1=age (in years)

X_2=cholesterol level

X_3=systolic blood pressure

X_4=relative weight

X_5=hemoglobin level

X_6=smoking(0=none, $1 \leq 1 pack$, $2 = 1 pack$, $3 \geq 1 pack$ per day)

X_7=ECG(0=normal,1=abnormal)

A logistic regression analysis produced:

parameter	estimate	SE
β_0	-13.2573	
β_1	0.1216	0.0137
β_2	0.0070	0.0025
β_3	0.0068	0.0060
β_4	0.0257	0.0091
β_5	-0.0010	0.0098
β_6	0.1223	0.1031
β_7	0.7206	0.1009

[Note that crude $\pi = 88/712 = 0.1186$]

$$\pi = \frac{e^{-13.2573+0.1216X_1+0.0070X_2+\quad+0.7206X_7}}{1+e^{-13.2573+0.1216X_1+\quad+0.7206X_7}}$$

estimates the probability of CHD incidence in the next 12 years for some individual (male) with characteristics $(X_1, X_2, \ldots X_7)$

For example, to estimate the probability of CHD in the next 12 years for a 45 year old man with cholesterol level = 210, SBP = 130, relative weight = 100, hemoglobin level = 120, non smoker ($X_6 = 0$) and normal ECG ($X_7 = 0$), we compute

$$\begin{aligned}
&\hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_7 X_7 \\
&\quad = -13.2753 + .1216(45) + .0070(210) + .. + .1223(0) + .7206(0) \\
&\quad = -2.9813
\end{aligned}$$

Therefore $\hat{\pi} = e^{-2.9813}/(1 + e^{-2.9813}) = .0183$

For a man with the same characteristics as above, but who smokes more than 1 pack per day,

$$\begin{aligned}
&\hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_7 X_7 \\
&\quad = -13.2753 + .1216(45) + .0070(210) + \ldots + .1223(3) + .7206(0) \\
&\quad = -1.7144
\end{aligned}$$

$$\hat{\pi} = e^{-1.7144}/(1 + e^{-1.7144}) = .1526$$

Therefore, measures of association for smoking > 1 pack versus none are

$$\begin{aligned}
\text{RD} &= .1526 - .0483 = .1043 \text{ (risk difference)}\\
\text{RR} &= \frac{.1526}{.0483} = 3.16 \text{ (risk ratio)}\\
\text{OR} &= \frac{.1526/(1-.1526)}{.0483/(1-.0483)} = 3.55 \text{ (odds ratio)}
\end{aligned}$$

(Note that RR = 3.16 $\approx$ OR = 3.55 since $\hat{\pi}$ in the baseline (.0483) is relatively rare)

Notice that

$$\begin{aligned}
e^{\beta_6(3)} &= e^{.4223(3)} = e^{1.27} = 3.55\\
&= \text{OR of disease for smokers of} > 1 \text{ pack.}
\end{aligned}$$

Why is OR = $e^{\beta_6 X_6}$??

Recall that

$$\begin{aligned}
\pi &= \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_7 X_7}}\\
\Longrightarrow\ & 1-\pi = \frac{1}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_7 X_7}}\\
\Longrightarrow\ & \frac{\pi}{1-\pi} = e^{\beta_0+\beta_1 X_1+\ldots+\beta_7 X_7}\\
\Longrightarrow\ & \ln\frac{\pi}{1-\pi} = \beta_0+\beta_1 X_1+\ldots+\beta_7 X_7\\
\Longrightarrow\ & = logit(\pi)
\end{aligned}$$

Then β_6 represents the mean change in $logit(\pi)$ per unit change in X_6 when all other variables are held fixed.

Therefore, for <u>non-smokers</u>,

$$\ln(\frac{\pi_{ns}}{1-\pi_{ns}}) = \beta_0 + \beta_1 X_1 + \ldots\beta_6(0) + \beta_7 X_7$$

and for <u>heavy smokers</u> (> 1 pack)

$$\ln(\frac{\pi_{hs}}{1-\pi_{hs}}) = \beta_0 + \beta_1 X_1 + \ldots\beta_6(3) + \beta_7 X_7$$

Therefore,

$$\ln(\frac{\pi_{hs}}{1-\pi_{hs}}) - \ln(\frac{\pi_{ns}}{1-\pi_{ns}}) = 3\beta_6$$

$$\Longrightarrow \ln(\frac{\pi_{hs}/(1-\pi_{hs})}{\pi_{ns}/(1-\pi_{ns})}) = 3\beta_k$$

$$\Longrightarrow \underbrace{\frac{\pi_{hs}/(1-\pi_{hs})}{\pi_{ns}/(1-\pi_{ns})}}_{"odds-ratio"} = e^{3\beta_k}$$

Note that $e^{3\beta_k}$ is the odds ratio of disease for heavy smokers to non-smokers irrespective of the other characteristics, as long as they are the same. Note that this interpretation assumes no interaction (effect-modification).

In general, the odds-ratio of disease for an individual with characteristics $x_1^*, x_2^*, ..., x_k^*$ to an individual with characteristics $x_1', x_2', ..., x_k'$ is given by

$$\psi = e^{\beta_1(x_1^*-x_1')+\beta_2(x_2^*-x_2')+ \quad +\beta_k(x_k^*-x_k')}$$

The most common use of this result is when X_k represents a dichotomous "exposure" (1=yes, 0=none) and we are interested in the disease-exposure odds-ratio for two individuals who are differently "exposed" and equal on the remaining variables. This adjusted odds-ratio is then

In our example, the odds-ratio (heavy smoker to non-smoker) is then

$$\psi = e^{\beta_k(1-0)} = e^{\beta_k}$$

In our example, the odds-ratio (heavy smoker to non-smoker) is then

$$\psi = e^{\beta_k(3-0)} = e^{3\beta_k}$$

1.5.2 Interpretation of β_0

The logistic regression approach was developed for cohort studies (see the example). What does β_0 estimate ?

$\beta_0 = logit(\pi)$ where π is the probability of disease when all the X's are 0

Seems uninteresting, but it allows us to estimate probabilities of disease ($\hat{\pi}$'s) for individuals with certain characteristics. From these, we can compute RD, RR, OR.

It has been shown that logistic regression can be used for case control studies with the only difference that β_0 will change; the other $\beta_1, \beta_2, ..., \beta_k$ will be the same as in a cohort study.

In fact for a case-control study,

$$logit(\pi) = \beta_0^* + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

where $\beta_0^* = \beta_0 + \ln(\frac{\theta_1}{\theta_2})$, β_0=β_0 of cohort study, θ_1=sampling fraction for cases and θ_2=sampling fraction for controls

Therefore, because we do not know β_0 exactly (unless we know θ_1 and θ_2), we cannot estimate π's and hence we cannot estimate RD and RR.

However, we know that we do not need π to compute the OR since OR $= e^{\sum \beta_i (x_i^* - x_i')}$, which does not depend on β_0.

It is principally for this reason that we have concentrated our efforts on the odds-ratio as our measure of the exposure-disease association.

Recall:

The logistic model specifies that the probability of disease depends on a set of variables $X_1, X_2, \ldots, X_k$ by

$$\begin{aligned} \pi &= \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k}} \\ \Longrightarrow & \underbrace{\ln\frac{\pi}{1-\pi}}_{logit(\pi)} = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k \end{aligned}$$

where e_i^{β} odds-ratio of disease for a unit change in X_i

We now examine how the logistic model deals with interaction, how the parameters are estimated, tests of significance are conducted and confidence intervals obtained. We will also see how to set-up the computer for logistic regression.

1.6 Interaction

First we must discuss the multiplicative property of the logistic model. Consider the following model

$$logit(\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_k X_k$$

where $X_1 = \begin{cases} 1 & \text{present} \\ 0 & \text{absent} \end{cases}$, $\quad X_2 = \begin{cases} 1 & \text{present} \\ 0 & \text{absent} \end{cases}$

Suppose that X_1 and X_2 are two different agents of exposure (X_1=smoking, X_2=drinking) and we want to assess the effects of X_1 and X_2 separately and jointly for fixed values of $X_3, \ldots, X_k$. We know then that the odds-ratio for X_1^*, X_2^* to X_1', X_2' (other X's remaining the same), is

$$e^{\beta_1(X_1^*-X_1')+\beta_2(X_2^*-X_2')}$$

By making $X_1'=0$ and $X_2'=0$, the referent category (i.e. unexposed by both X_1 and X_2) (non-smoker, non-drinker)
Then

$$e^{\beta_1} \text{ is the odds-ratio for } X_1 \text{ alone}$$
$$e^{\beta_2} \text{ is the odds-ratio for } X_2 \text{ alone}$$

and

$$e^{\beta_1+\beta_2} \text{ is the odds-ratio } X_1 \text{ and } X_2 \text{ jointly}$$

Note: $e^{\beta_1+\beta_2} \neq e^{\beta_1} + e^{\beta_2}$

Unlike in linear regression where the effects are additive, in logistic regression, they are multiplicative in the odds-ratio, i.e.

$$e^{\beta_1+\beta_2} = e^{\beta_1} \quad e^{\beta_2}$$

Example if ψ_{smoke}=3, ψ_{drink}=4, then $\psi_{smoke\ drink}$=12
Note: this is true only if no interaction is present.

Interaction terms in logistic regression are specified in the same way as in linear regression. Consider the familiar context where X_1 represents the binary exposure (1 or 0) under study and $X_2, \ldots, X_k$ are the potential confounders. The model with first-order interactions of X_1 is given by

$$\begin{aligned}
logit(\pi) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \\
&\quad + \alpha_2 X_1 X_2 + \alpha_3 X_1 X_3 + \ldots + \alpha_k X_1 X_k \\
&= \beta_0 + \beta_1 X_1 + \sum_{i=2}^{k} \beta_i X_i + \sum_{i=2}^{k} \alpha_i X_1 X_i
\end{aligned}$$

$$= \beta_0 + \underbrace{(\beta_1 + \sum_{i}^{k} \alpha_i X_i)}_{\substack{\mathit{it\ is\ not\ a\ constant}\\ \mathit{value\ anymore\ but}\\ \mathit{depends\ on\ values}\\ \mathit{of}\ X_2,\ldots,X_k}} X_1 + \sum_{i=2}^{k} \beta_i X_i$$

The odds-ratio for exposure (X_1) is then

$$\psi = e^{\beta_1 + \sum_{i=2}^{k} \alpha_i X_i} = e^{\beta_1} e^{\alpha_2 X_2} \cdots e^{\alpha_k X_k}$$

Note that the X_i's for which α_i is non-zero are called effect-modifiers of the disease-exposure relationship.

Our model can now be specified in the following general way

$$logit(\pi) = \ exposure \ + \ confounders \ + \ effect\text{-}modifiers$$

For a continuous X_2,

$$\psi = e^{\beta_1} e^{\alpha_2 X_2} \quad and \quad \ln \psi = \beta_1 + \alpha_2 X_2$$

i.e., the log of the odds ratio is a linear function of X_2.

1.7 Estimation

The point estimation of the parameters in logistic regression is achieved by the method of maximum likelihood (in contrast with the method of least squares in linear regression). This method is based on the likelihood function of the β's for our sample. This function is the probability of observing the outcome of our sample and this probability (likelihood) is a function of the β's.

Recall that Y is our outcome (dependent) variable. In our sample, we observe $y_1, y_2, y_3, \ldots, y_n$ where each y_i is either a 1 or a 0 with respective probabilities $\pi_1, \pi_2, \cdots, \pi_n$ of being 1.

The likelihood of observing such a sample is

$$
\begin{aligned}
L &= \pi_1^{y_1}(1-\pi_1)^{1-y_1}\pi_2^{y_2}(1-\pi_2)^{1-y_2} \cdots \pi_n^{y_n}(1-\pi_n)^{1-y_n} \\
&= \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}
\end{aligned}
$$

Example of maximum likelihood estimation

$$L = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

<u>Model</u>: $\pi_i = \pi$ for all subjects i=1,....n.

$$
\begin{aligned}
L &= \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \\
&= \pi^{\sum y_i}(1-\pi)^{n-\sum y_i}
\end{aligned}
$$

$$\ln L = \sum y_i \ln(\pi) + (n - \sum y_i)\ln(1-\pi)$$

We wish to find the value of π which maximizes the likelihood function L. This is equivalent to maximising $\ln(L)$. We use derivatives (calculus) of $\ln(L)$ which, when set to 0, produce the maximum likelihood estimator (MLE) of π.

$$
\begin{aligned}
\frac{\partial \ln L}{\partial \pi} &= \frac{\sum y_i}{\hat{\pi}} - \frac{(n-\sum y_i)}{(1-\hat{\pi})} = 0 \\
&\Longrightarrow \sum y_i - \pi \sum y_i = n\hat{\pi} - \hat{\pi}\sum y_i \\
&\Longrightarrow \hat{\pi} = \frac{\sum y_i}{n} \text{ is the MLE of } \pi
\end{aligned}
$$

Variance of $\hat{\pi}$ is obtained from second derivative.

$$\frac{\partial^2 \ln L}{\partial \pi^2} = -\frac{\sum y_i}{\pi^2} - \frac{(n-\sum y_i)}{(1-\pi)^2}$$

We may replace π by its MLE $\hat{\pi}$

$$
\begin{aligned}
&= -\frac{\sum y_i}{\sum(y_i/n)^2} - \frac{(n-\sum y_i)}{(1-\sum y_i/n)^2} \\
&= -\frac{n^2}{\sum y_i} - \frac{n^2}{n-\sum y_i}
\end{aligned}
$$

$$
\begin{aligned}
&= -n\left\{\frac{1}{\hat{\pi}} - \frac{1}{1-\hat{\pi}}\right\} \\
&= -n\left\{\frac{1}{\hat{\pi}(1-\hat{\pi})}\right\} \\
&= \frac{-n}{\hat{\pi}(1-\hat{\pi})}
\end{aligned}
$$

The variance is given from MLE theory by:

$$
\begin{aligned}
V\hat{A}R(\hat{\pi}) &= \frac{-1}{\frac{\partial^2 \ln L}{\partial \pi^2}} \\
&= \frac{\hat{\pi}(1-\hat{\pi})}{n}
\end{aligned}
$$

which is the well-known variance estimator of a binomial proportion.

In the regression context, we also have additional data $X_{1i}, \ldots X_{ki}$ for each subject. We assume the following model between E(Y_i) and X_i:

$$\pi_i = \frac{e^{\beta_0+\beta_1 X_{1i}+ \quad +\beta_k X_{ki}}}{1 + e^{\beta_0+\beta_1 X_{1i}+ \quad +\beta_k X_{ki}}}$$

for the X's of the i^{th} subject in our sample. We can notice from this that L, the likelihood function of our sample is indeed a function of the β's. The maximum likelihood method produces the β's (in fact $\hat{\beta}$'s) which are the most likely to have produced our observed outcomes $y_1, \ldots, y_n$. The computations are much more complex than when $\pi_i = \pi$ all i and require iterative calculations performed by a computer.

In general, MLE's of β's are approximately normal with variances produced directly by the iterative procedure [ML estimation is simply a very powerful tool!].

Therefore, the maximum likelihood procedure produces, for our logistic model, $\hat{\beta}_i$ and $\hat{\sigma}_i$ for each parameter. Because of the approximate normality of MLE's, a (1-α) 100% confidence interval for β_i is given by

$$\hat{\beta}_i \pm Z_{\alpha/2}\hat{\sigma}_i$$

and, therefore, an approximate 100 (1-α)% CI for the odds-ratio ψ_i is

$$e^{\hat{\beta}_i \pm Z_{\alpha/2}\hat{\sigma}_i}$$

More general formulae which involve several β's simultaneously are given in Schlesselman, page 247. These require not only the variances of each $\hat{\beta}$ but also the covariances between β's which are also produced by MLE.

1.8 Discriminant Analysis

Consider two groups of individuals, cases and controls, on which we measure $X_1, \cdots, X_k$ Suppose we want to distinguish between the groups on the basis of <u>one single value</u> $D = \sum \beta_i X_i$, a linear combination of the X's. If $D >$ some D_0 then we say that the individual is diseased (a "case") and if $D < D_0$ is not diseased ("control") By minimizing the probability of misclassifying an individual, we obtain "optimal" β's and $D = \sum \beta_i X_i$ is called the <u>linear discriminant function</u>

Under the assumption of multivariate normality for the $X_1, \cdots, X_k$ (simultaneously) with different means for the two groups but equal covariance matrices, the coefficients β_i's are equivalent to those obtained via logistic regression.

<u>Proof</u>:

Let

$$P = \text{prob of disease} = P(D)$$

Let

$$\begin{aligned} P(X|D) &= f_1(X) = \text{prob of X's among the cases } (D) \\ P(X|D) &= f_0(X) = \text{prob of X's among the controls } (D) \end{aligned}$$

So, by Bayes theorem we get

$$\begin{aligned} P(D|X) &= \frac{f_1(X)P}{f_1(X)P + f_0(X)(1-P)} \\ &= \frac{1}{1 + \frac{f_0(X)(1-P)}{f_1(X)P}} \end{aligned}$$

if X is normal, then

$$\frac{f_0(X)}{f_1(X)} = \frac{e^{-\frac{1}{2\sigma^2}(x^2 - 2x\mu_0 + \mu_0^2)}}{e^{-\frac{1}{2\sigma^2}(x^2 - 2x\mu_1 + \mu_1^2)}}$$

since $X|D \sim N(\mu_0, \sigma^2)$ and $X|D \sim N(\mu_1, \sigma^2)$ (assumption of discriminant analysis) Then

$$\frac{f_0(X)}{f_1(X)} = e^{-\frac{1}{2\sigma^2}[\mu_0^2 - \mu_1^2 + 2x(\mu_1 - \mu_0)]}$$

and

$$P(D|X) = \frac{1}{1 + e^{\left\{-(-\ln\frac{1-p}{p} - \left(\frac{\mu_1^2 - \mu_0^2}{2\sigma^2}\right) + \frac{(\mu_1 - \mu_0)}{\sigma^2} x\right\}}}$$

Let

$$\begin{aligned} \beta_1 &= \frac{\mu_1 - \mu_0}{\sigma^2} \\ \beta_0 &= \ln\frac{p}{1-p} - \beta_1 \left[\frac{\mu_1 + \mu_0}{2}\right] \end{aligned}$$

Then

$$P(D|X) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

This is the logistic model.

Chapter 2

Model Selection for Logistic Regression

Formal model selection methods can be based either on stepwise methods or finding best subsets of variables based on some criterion (e.g., Akaike's information). Fitting lots of models can be very expensive because each fit requires an iterative procedure. Stepwise methods are sequential, hence cheaper than best subset methods Here we only introduce stepwise method.

Stepwise selection of variables has been widely used in linear regression. Most major software packages have either a separate program or an option to perform this type of analysis. At one time, stepwise regression was an extremely popular method for model building. Methodology for performing stepwise logistic regression has been available for much less time. Among major software packages only BMDP offers a program for stepwise logistic regression. We feel that the procedure provides a useful and effective data analysis tool. In particular, there are times when the outcome being studied is relatively new (e.g., AIDS) and the important covariates may not be known and associations with the outcome not well understood. In these instances most studies will collect many possible covariates and screen them for significant associations. Employing a stepwise selection procedure can provide a fast and effective means to screen a large number of variables, and to simultaneously fit a number of logistic regression equations.

Any stepwise procedure for selection or deletion of variables from a model is based on a statistical algorithm which checks for the "importance" of variables, and either includes or excludes them on the basis of a fixed decision rule. The "importance" of a variable is defined in terms of a measure of the statistical significance of the coefficient for the variable. The statistic used depends on the assumptions of the model. In stepwise linear regression an F-test is used since the errors are assumed to be normally distributed. In logistic regression the errors are assumed to follow a binomial distribution, and significance is assessed via the likelihood ratio chi-square test Thus, at any step in the procedure the most important variable, in statistical terms, will be the one that produces the greatest change in the log-likelihood relative to a model not containing the variable (i.e., the one that would result in the largest likelihood ratio statistic, G).

We have pointed out that a polytomous variable with k levels is appropriately modeled through its k-1 design variables. Since the magnitude of G depends on its degrees of freedom, any procedure based on the likelihood ratio test statistic, G, must account for possible differences in degrees of freedom between variables. This is done by assessing significance through the p-value for G.

We will describe and illustrate the algorithm for forward selection followed by backward elimination in stepwise logistic regression. Any variants of this algorithm are simple modifications of this procedure. The method will be described by considering the statistical computations that the computer must perform at each step of the procedure.

Step (0): Suppose we have available a total of p possible independent variables, all of which are judged to be of plausible "biologic" importance in studying the outcome variable. Step (0) begins with a fit of the "intercept only model" and an evaluation of its log-likelihood, L_0. This is followed by fitting each of the p possible univariate logistic regression models and comparing their respective log-likelihoods. Let the value of the log-likelihood for the model containing variable x_j at step zero be denoted by $L_j^{(0)}$. The subscript j refers to that variable which has been added to the model, and the superscript (0) refers to the step. This notation will be used throughout the discussion of stepwise logistic regression to keep track of both step number and variables in the model.

Let the value of the likelihood ratio test for model containing x_j versus the intercept only model be denoted by $G_j^{(0)} = 2(L_j^{(0)} - L_0)$, and its p-value be denoted by $p_j^{(0)}$. Hence, this p-value is determined by the tail probability $Pr[\chi^2(\nu) > G_j^{(0)}] = p_j^{(0)}$, where $\nu = 1$ if

x_j is continuous and $\nu = k - 1$ if x_j is polytomous with k categories.

The most important variable is the one with the smallest p-value. If we denote this variable by x_{e_1}, then $p_{e_1}^{(0)} = min(p_j^{(0)})$, where "min" stands for selecting the minimum of the quantities enclosed in the brackets. The subscript "e_1" is used to denote that the variable is a candidate for entry at step 1. For example, if variable x_2 had the smallest p-value, then $p_2^{(0)} = min(p_j^{(0)})$, and $e_1 = 2$. Just because x_{e_1} is the most important variable, there is no guarantee that it will be "statistically significant." For example, if $p_{e_1}^{(0)} = 0.83$, we would probably conclude that there is little point in continuing this analysis because the "most important" variable is not related to the outcome. On the other hand, if $p_{e_1}^{(0)} = 0.003$, we would like to look at the logistic regression containing this variable and see if there are other variables which are important given that x_{e_1} is in the model.

A crucial aspect of using stepwise logistic regression is the choice of an "alpha" level to judge the importance of variables. Let p_E denote our choice where the "E" stands for entry. The choice for p_E will determine how many variables will eventually be included in the model. Bendel and Afifi(1977) have studied the choice of p_E for stepwise linear regression, and Costanza and Afifi(1979) have studied the choice for stepwise discriminant analysis. The results of this research have shown that the choice of $p_E = 0.05$ is too stringent, often excluding important variables from the model. Choosing a value for p_E in the range 0.15 to 0.20 is more highly recommended. While previous research considered only normal theory models (i.e., linear regession or discriminant analysis), there is reason to believe that use of p_E in the same range would be a suitable criterion for stepwise logistic regession since logistic regression may be viewed as an offshoot of the normal theory discriminant function model. Moreover, use of p_E in this range will provide some assurance that the stepwise procedure will select variables whose coefficients are different from zero.

Sometimes the goal of the analysis may be broader, and models containing more variables are sought to provide a more complete picture of possible models. In these cases use of $p_E = 0.25$ might be a reasonable choice. Whatever the choice for p_E, a variable will be judged important enough to include in the model if the p-value for G is less than p_E. Thus, the program proceeds to step (1) if $p_{e_1}^{(0)} < p_E$; otherwise, it stops.

Step (1): Step (1) commences with a fit of the logistic regression model containing x_{e_1}. Let $L^{(1)}_{e_1}$ denote the log-likelihood of this model. To determine whether any of the remaining p-1 variables are important once the variable x_{e_1} is in the model, we fit p-1 logistic regression momdels containing x_{e_1} and x_j, $j = 1,2,3,\cdot,p$ and $j \neq e_1$. For the model containing x_{e_1} and x_j let the log-likelihood be denoted by $L^{(1)}_{e_1 j}$, and let the likelihood ratio chi-square statistic of this model versus the model containing only x_{e_1} be denoted by $G^{(1)}_j = 2(L^{(1)}_{e_1 j} - L^{(1)}_{e_1})$. The p-value for this statistic will be denoted by $p^{(1)}_j$. Let the variable with the smallest p-value at step (1) be x_{e_2} where $p^{(1)}_{e_2} = min(p^{(1)}_j)$. If this value is less than p_E we proceed to step (2); otherwise we stop.

Step (2): Step (2) begins with a fit of the model containing both x_{e_1} and x_{e_2}. It is possible that once x_{e_2} has been added to the model, x_{e_1} is no longer important. Thus, step (2) includes a check for backward elimination. In general this is accomplished by fitting models that delete one of the variables added in the previous steps and assessing the continued importance of the variable removed. At step (2) let $L^{(2)}_{-e_j}$ denote the log-likelihood of the model with x_{e_j} removed. In similar fashion let the likelihood ratio test of this model versus the full model at step (2) be $G^{(2)}_{-e_j} = 2(L^{(2)}_{e_1 e_2} - L^{(2)}_{-e_j})$ and $p^{(2)}_{-e_j}$ be its p-value.

To ascertain whether a variable should be deleted from the model the program selects that variable which, when removed, yields the maximum p-value. Denoting this variable as x_{r_2}, then $p^{(2)}_{r_2} = max(p^{(2)}_{-e_1}, p^{(2)}_{-e_2})$. To decide whether x_{r_2} should be removed, the program compares $p^{(2)}_{r_2}$ to a second prechosen "alpha" level, p_R, which will indicate some minimal level of continued contribution to the model where "R" stands for remove. Whatever value we choose for p_R, it must exceed the value of p_E to guard against the possibility of having the program enter and remove the same variable at successive steps.

If we do not wish to exclude many variables once they have entered, we might use $p_R = 0.9$. A more stringent value would be used if a continued "significant" contribution were required. For example, if we used $p_E = 0.15$, then we might choose $p_R = 0.20$. If the maximum p-value to remove, $p^{(2)}_{r_2}$, exceeds p_R then x_{r_2} is removed from the model. If $p^{(2)}_{r_2}$ is less than p_R then x_{r_2} remains in the model. In either case the program proceeds to the variable selection phase.

At the forward selection phase each of the p-2 logistic regression models are fit containing x_{e_1}, x_{e_2} and x_j for $j = 1, 2, 3, \cdots, p, j \neq e_1, e_2$. The program evaluates the log-likelihood for each model, computes the likelihood ratio test versus the model containing only x_{e_1} and x_{e_2} and determines the corresponding p-value. Let x_{e_3} denote the variable with the minimum p-value, that is, $p_{e_3}^{(2)} = min(p_j^{(2)})$. If this p-value is smaller than p_E, $p_{e_3}^{(2)} < p_E$, then the program proceeds to step (3); otherwise, it stops.

Step (3): The procedure for step (3) is identical to that of step (2). The program performs a check for backward elimination followed by forward selection. This process continues in this manner until the last step, step (S).

Step (S): This step occurs when: (1) all p variable have entered the model or (2) all variables in the model have p-value to remove which are less than p_R, and the variables not included in the model have p-values to enter which exceed p_E. The model at this step contains those variables that are important relative to the criteria of p_E and p_R. These may or may not be the variables reported in a final model. For instance, if the chosen values of p_E and p_R correspond to our belief for statistical significance, then the model at step S may well contain the significant variables. However, if we have used values for p_E and p_R which are less stringent, then we should select the variables for a final model from a table that summarizes the results of the stepwise procedure.

There are two methods that may be used to select variables from a summary table; these are comparable to methods commonly used in stepwise linear regression. The first method is based on the p-value for entry at each step, while the second is based on a likelihood ratio test of the model at the current step versus the model at the last step.

Let "q" denote an arbitrary step in the procedure. In the first method we compare $p_{e_q}^{(q-1)}$ to a prechosen significance level such as $\alpha = 0.15$. If the value $p_{e_q}^{(q-1)}$ is less than α, then we move to step q. We stop at the step when $p_{e_q}^{(q-1)}$ exceeds α. We consider the model at the previous step for further analysis. In this method the criterion for entry is based on a test of the significance of the coefficient for x_{e_q} conditional on x_{e_1}, x_{e_2}, $\cdots$, $x_{e_{q-1}}$ being in the model. The degrees of freedom for the test are 1 or k-1, depending on whether x_{e_q} is continuous or polytomous with k categories.

In the second method, we compare the model at the current step q, not to the model at the previous step, step q-1, but to the model at the last step, step (S). We evaluate the p-value for the likelihood ratio test of these two models and proceed in this fashion until this p-value exceeds α. This tests that the coefficients for the variables added to the model from step q to step (S) are all equal to zero. At any given step it will have more degrees of freedom than the test employed in the first method. For this reason the second method may possibly select a larger numberof variables than the first method.

It is well known that the p-values calculated in stepwise selection procedures are not p-values in the traditional hypothesis testing context. Instead, they should be thought of as indicators of relative importance among variables. We recommend that one error in the direction of selecting a relatively rich model following stepwise selection. The variables so identified should then be subjected to the more intensive analysis described previously.

A common modification of the stepwise selection procedure just described is to begin with a model at step zero which contains known important covariates Selection is then performed from among other variables. One instance when this approach may be useful is to select interactions from among those possible from a main effects model.

One considerable disadvantage of the stepwise selection procedures just described is that the maximum likelihood estimates for the coefficients of all variables not in the model must be calculated at each step. For large data files with large numbers of variables this can be quite costly both in terms of time and money. Two approximations to this method available in, or could be implemented into. existing programs. One method, available in BMDP, uses a linear approximation to the likelihood ratio test. The resulting test is similar to the one used for variable selection in a two group stepwise discriminant analysis. This is termed the "ACE" method in BMDP. The second procedure selects new variables based on the score tests for the variables not included in the model. A variant of this method using a multivariate Wald statistic has been proposed by Peduzzi, Hardy, and Holford(1980). To date there has been no work published which has compared these different selection methods although it does seem likely that an important variable would be identified, irrespective of method used.

Freedman (1983) urges caution when considering a model with many variables, noting that significant linear regressions may be obtained from variables completely unrelated to the outcome "noise" variable. Flack and Chang (1987) have shown similar results regarding the frequency of selection of "noise" variables. Thus, a thorough analysis that examines statistical and biologic significance is especially important following any stepwise method. As an example, we apply the stepwise variable selection procedure to the low birth weight data. The results of this process are presented in Table 2.1 in terms of the p-values to enter and remove calculated at each step. These p-values are those of the relevant likelihood ratio test described previously. The order of the variables given columnwise in the table is the order in which they were selected. In each row the values to the left of the vertical line are p_R values and values to the right of the vertical lines are p_E values. The program was run using $p_E = 0.15$ and $p_R = 0.20$.

Table 2.1 Results of Applying Stepwise Variable Selection Using the Maximum Likelihood Method to the Low Birth Weight Data Presented at Each Step in Terms of the p-values to Enter, to the Right of the Vertical Line, and the p-Value Remove, to the Left of the Vertical Line in Each Row. The Asterisk Denotes the Maximum p-Value to Remove at Each Step.

Step #	PTL	LWT	HT	RACE	SMOKE	UI	AGE	FTV
0	0.009	0.015	0.015	0.082	0.027	0.021	0.097	0.379
1	0.009	0.031	0.038	0.069	0.078	0.083	0.057	0.441
2	0.022	0.034*	0.006	0.057	0.090	0.139	0.125	0.589
3	0.023*	0.006	0 006	0 078	0.093	0.086	0.162	0.731
4	0.019	0.005	0.009	0.078*	0.015	0.077	0.308	0.873
5	0.067*	0 009	0.010	0.016	0.015	0.088	0.374	0.905
6	0.135*	0.013	0.006	0.016	0.017	0.088	0.155	0.927

Step (0): At step (0) the program selects as a candidate for entry at step (1) the variable with the smallest p-value in the first row of Table 2.1. This is the variable PTL with a p-value of 0.009. Since this p-value is less than 0.15, the program proceeds to step (1).

Step (1): At step (1) the program will not remove the variable just entered since $p_R > p_E$ and the p-value to remove at step (1) is equal to the p-value to enter at step (0). (This is true for the variable entered at any step-not just the first step.) The variable with the smallest p-value to enter at step (1) is LWT with a value of 0.031, which is less than 0.15 so the program moves to step (2).

Step (2): The p-values to remove appear first in each row. The largest value is indicated with an "*." At step (2) the largest p-value to remove is 0.031, which does not exceed 0.20, thus the program moves to the variable selection phase. The smallest p-value to enter among the remaining variables not in the model is for the variable HT and is 0.006. This value is less than 0.15 so the program proceeds to step (3)

step (3)-step (5): At steps (3) to (5) the program finds that no variable can be removed from the model because each of p-values, indicated with "*" in rows three to five in Table 2.1, less than 20. The program determines, in the selection phase, the variable with the smallest p-value to enter and, since it is less than 0.15, the program proceeds to the next step.

Step (6): At step (6) the program finds that the maximum p value to remove is 0.135 for PTL. This value is less than 0.20, so PTL is not removed from the model. In the selection phase the program finds that the minimum p-value for entry is 0.455 for the variable AGE. Since this value exceeds 0.15, no further variables may be entered into the model, and the program stops.

Since the program was run with $p_E = 0.15$, a value we believe will select variables with significant coefficients, it is not strictly necessary to go to the summary table to select the variables to be used in a final model. We will, however, illustrate the calculations for the two methods of variable selection from the summary table. These are given in Table 2.2.

For method1 we compare the p-value for entry at each step to our chosen level of significance. For purposes of illustration only we will use the value of 0.05, even though we noted earlier in this chapter that it is too stringent for actual practice. The information for method1 is in the second panel of Table 3.2.

Table 2.2 Log-Likelihood for the Model at Each Step and Likelihood Ratio Test Statistics (G), Degrees of Freedom (df), and p-Values for Two Methods of Selecting Variables for a Final Model from a Summary Table.

Variable			Method 1			Method 2		
Step #	Entered	Log-Likelihood	G	df	p-value	G	df	p-value
0		-117.31				32.69	7	< 0.001
1	PTL	-113.95	6.78	1	0.009	25.91	6	< 0.001
2	LWT	-111.70	1.18	1	0.031	21.12	5	0.001
3	HT	-107.98	7.11	1	0.006	13.98	4	0.007
4	RACE	-105.13	5.10	2	0.078	8.86	2	0.012
5	SMOKE	-102.15	5.95	1	0.015	2.91	1	0.088
6	UI	-100.99	2.91	1	0.088			

The value of the likelihood ratio test for the model at step (0) compared to that containing PTL at step (1) is

$$G = 6.78 = 2[-113.916 - (-117.336)]$$

The p-value for G is 0.009 which is less than 0.05 so we conclude that the coefficient for PTL is significant and move to step (2). The p-value for the variable, LWT, entered at step (2) is 0.031. This is the p-value for the likelihood ratio test of the significance of the coefficient for LWT, given that PTL is in the model. The value of the test statistic is

$$G = 1.19 = 2[-111.701 - (-113.916)]$$

Since the p-value for G is less than 0.05 we move to step (3) Calculations proceed in a similar fashion and we compare. at each step, the p-value to 0 05. At step (4) we find that the value of likelihood ratio test of the model at step (4) versus that at step (3) is

$$G = 5.10 = 2[-105.13 - (-107.98)]$$

resulting in a p-value of 0.078. This value is greater than 0.05 so we conclude that RACE does not provide a significant addition to the variables already selected at step (3). Hence, the final model would be the one with all variables entered through step (3) even though the variable entered at step (5), SMOKE, has a p-value of less than 0.05.

The information for method 2 is in the last panel of table 3.2. In the second method the model at each step is compared to the model at the last step via a likelihood ratio test. This is a test of the joint significance of variables added at subsequent steps. We again proceed until the p-value for the test exceeds the chosen significance level. For purposes of illustration only we will use 0.05. The value of G at step (0) is

$$G = 2[-100.993 - (-117.336)] = 32.69$$

with a p-value of < 0.001 based on 7 degrees of freedom. Since this p-value is less than 0.05 we proceed to step (1). At step (1) the test of this model versus that at the last step is

$$G = 2[-100.993 - (-113.916)] = 25.91$$

with a p-value of ¡0.001 based on 6 degrees of freedom. Since the p-value is less than 0.05 we proceed to step (3). We continue in this manner until step (5) The p-value for the likelihood ratio test of the model at step (5) versus that at step (6) is 0.088. This value exceeds 0.05, so we stop and use the variables in the model at step (5).

In this example methods 1 and 2 have identified different sets of variables. Each method provides a test of a different hypothesis at each step. The number of parameters being tested in method 2 is, except for the last step, larger than that for method 1. Thus, method 2 may select, as it does in this example, more variables than method 1. In cases where this occurs, one should carefully examine the additional variables and include them if they seem biologically relevant. In this case we would undoubtably opt for the richer model selected by method 2.

At the conclusion of the stepwise selection process we have only identified a collection of variables which seem to be statistically important. Thus, any known biologically important variables, such as AGE in our example, should be added before proceeding with the steps necessary to obtain the final main effects model. As noted earlier, this should include determining the appropriate scale of continuous covariates

Once the scale of the continuous covariates has been examined, and corrected if necessary, we may consider applying stepwise selection to identify interactions. The candidate interaction terms are those that seem biologically reasonable given the main effects variables in the model. We begin at step (0) with the main effects model and sequentially select from among the possible interactions. We select the significant ones using either method 1 or method 2. Consequently the final model will contain previously identified main effects and significant interaction terms.

The variables identified by the stepwise selection process in the low birth weight data are the same ones identified early by purposeful selection. Therefore, the work necessary to check the scale of continuous covariates is not repeated and we begin stepwise selection of interactions using the model given in Table 2.3 and the interactions listed in Table 2.4. The results of stepwise selection of interactions are given in Table 3.5.

Table 2.3 Estimated Coefficients, Estimated Standard Errors, and Coeff./SE for the Multivariate Model Containing LWD and PTD, Dichotomous Variables Created from LWT and PTL.

	Estimated	Estimated	
Variable	Coefficient	Standard Error	Coeff./SE
AGE	-0.016	0.037	-1.25
LWD	0.842	0.405	2 08
RACE(1)	1 073	0.511	2 09
RACE(2)	0.815	0.441	1.84
SMOKE	0.807	0.404	2.00
PTD	1.282	0.461	2 78
HT	1.435	0.617	2.22
UI	0.658	0.466	1.41
constant	-1.217	0.951	-1.28

Log-likelihood = −98.78

Table 2.4 Log-likelihood, LRT Statistic (G), Degrees of Freedom(df), and p-Value for Possible Interactons of Interest to be Added to the Main Effects Only Model.

Interaction	Log-Likelihood	G	df	p-value
Main Effects Only	-98.78			
AGE×RACE	-98.53	0.50	2	0.78
AGE×SMOKE	-98.51	0.54	1	0.46
AGE×HT	-98.39	0.78	1	0.38
AGE×UI	-98.76	0.04	1	0.84
AGE×LWD	-97.50	2.56	1	0.11
AGE×PTD	-98.36	0.84	1	0.36
RACE×SMOKE	-97.61	2.34	2	0.31
RACE×HT	98.63	0.30	2	0.86
RACE×UI	-97.62	2.32	2	0.31
RACE×LWD	-97.08	3.40	2	0.18
RACE×PTD	-98.50	0.56	2	0.76
SMOKE×HT	-98.71	0.14	1	0.71
SMOKE×UI	-98.12	1.32	1	0.25
SMOKE×LWD	-97.61	2.34	1	0.13
SMOKE×PTD	-98.31	0.94	1	0.33
LWD×HT	-98.22	1.12	1	0.30
AGE×LWD+SMOKE×LWD	96.01	5.54	2	0.06

Table 2.5 Results of Applying Stepwise Variable Selection to Interactions from the Main Effects Model. Using the Maximum Likelihood Method Presented at Each Step in Terms of the p-Values to Enter, to the Right of the Vertical Line, and the p-Values to Remove to the Left of the Vertical Line. The Asterisk Denotes the Maximum p Value to Remove at Each Step.

Step #	AGE×LWD	RACE×LWD	HT×LWD	SMOKE×LWD
0	0.110	0.183	0.291	0.127
1	0.110*	0.081	0.252	0.084
2	0.041	0.081*	0.112	0.615
3	0.029	0.053	0.112*	0.562

Of the 16 possible interactions specified in Table 2.4, only three were chosen. In the last column of Table 2.5 we have given the p-value for entering the SMOKE×LWD interaction term. The results at step (1) indicate that the RACE×LWD interaction is negligibly more significant than the SMOKE×LWD interaction and, once the RACE×LWD interaction is included into the model, there is little additional importance in the SMOKE×LWD interaction. At step (3) we see that the HT×LWD interaction enters the model with p-value of 0.142.

We now face several decisions involving the interactions. We considered this same problem earlier of completeness, we repeat the analysis in the current context. To further explore the tradeoff between including the SMOKE×LWD or the RACE×LWD interaction, a model that forced the SMOKE×LWD interaction into the model and then added the RACE×LWD interaction was fit. The results showed, as expected, that the RACE×LWD interaction was no longer important once the SMOKE×LWD interaction was included in the model. We must, therefore, decide which of these two interactions to include. We choose the SMOKE×LWD interaction as the more important from the biologic standpoint in view of the known relationship between weight and smoking. Potential racial by weight differences are regarded as being of lesser importance to document.

We now must decide if the HT×LWD interaction should be added to the model. The p-value for the inclusion of this interaction after the SMOKE×LWD interaction term is included in the model is 0.160, again a value close to the preferred alpha of 0.15. At this point we must keep in mind that the fundamental reason for developing a model is to provide as clear a description as is possible with the available data of the associations between outcome and covariates. If entering an additional term into the model improves our estimates of the relevant associations then we should put that term into the model regardless of its statistical significance. If a term does not contribute to the overall goal then it may be excluded. In this case we determine that inclusion of the HT×LWD interaction term does not help our understanding of the association between low birth weight and the variables in the model so we choose to leave it out of the model.

In conclusion, stepwise selection identifies variables as candidates for a model solely on statistical grounds. Thus, following stepwise selection of main effects all variables should be carefully scrutinized for biologic plausibility. In general, interactions must attain at least a moderate level of statistical significance to alter the point and interval estimates from a main effects model. Thus, stepwise selection of interactions can provide a valuable contribution to model identification, especially when there are large numbers of biologically plausible interactions generated from the main effects.

Chapter 3

Assessing the Fit of the Model

3.1 Introduction

We begin our discussion of methods for assessing the fit of an estimated logistic regression model with the assumption that we are at least preliminarily satisfied with our efforts at the model building stage. By this we mean that, to the best of our knowledge, the model contains those variables (main effects as well as interactions) that should be in the model and that variables have been entered in the correct functional form. Now we would like to know how effective the model we have is in describing the outcome variable. This is referred to as its **goodness-of-fit.**

If we intend to assess the goodness-of-fit of the model, then we should have some specific ideas about what it means to say that a model fits. Suppose we denote the observed sample values of the outcome variable in vector form as $\mathbf{y}$ where $\mathbf{y}' = (y_1, y_2, y_3, \ldots, y_n)$. We denote the values predicted by the model, or fitted values, as $\hat{\mathbf{y}}$ where $\hat{\mathbf{y}}' = (\hat{y}_1, \hat{y}_2, \hat{y}_3, \ldots, \hat{y}_n)$. We will conclude that the model fits if (1) summary measures of the distance between $\mathbf{y}$ and $\hat{\mathbf{y}}$ are small and (2) the contribution of each pair $(y_i, \hat{y}_i)$, $i = 1, 2, 3, \ldots, n$ to these summary measures is unsystematic and is small relative to the error structure of the model. Thus, a complete assessment of the fitted model will involve both the calculation of summary measures of the distance between $\mathbf{y}$ and $\hat{\mathbf{y}}$, and a thorough examination of the individual components of these measures.

The development of methods for assessment of goodness-of-fit will follow what we feel are the logical steps upon completion of the model building stage. The components of the proposed approach are (1) computation and evaluation of overall measures of fit, (2) examination of the individual components of the summary statistics, and (3) examination of other measures of the difference or distance between the components of $\mathbf{y}$ and $\hat{\mathbf{y}}$.

3.2 The Goodness-of-Fit of the Model

3.2.1 Significance Test

Overall likelihood ratio test (LRT) found in standard printouts verifies null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

H_0 means "None of the independent variable is significant as a risk factor thus information about their values does not improve significantly the prediction of outcome".

Thus, H_0 tested by overall LRT is equivalent to: "The best prediction for all covariate patterns is based on the overall proportion"

$$H_0 : \pi = \frac{e^{\beta_0}}{1 + e^{\beta_0}} = \frac{\sum y_i}{n}$$

$\sum y_i$ = # subjects with $Y_i = 1$, n = total sample size.

"Technically" it is tested by comparing log likelihood $\ln L_k$ obtained by full model using (k+1) parameters (k independent variables) to log likelihood $\ln L_0$ obtained with 1 parameter β_0. Statistic:

$$G = -2[\ln L_0 - \ln L_k] \sim \chi^2_k$$

Note: since the log likelihood is obtained by summing up over all observations the impact of the sample size on LRT is very strong.

Example: assume we have 50 data points and for $k = 5$ we obtain

$$\begin{aligned} G_{50} &= -2[\ln L_0 - \ln L_5] \\ &= -2[-28 - (-30)] = 1 \sim \chi^2_5 \text{ (not significant)} \end{aligned}$$

Now assume our sample is in fact the exact "miniature" of a larger sample of 500 patients (each covariate pattern is repeated 10 times and respective outcomes are the same). Then:

$$G_{500} = -2[10 \times 30 - 10 \times 28] = 10 \sim \chi^2_5 \text{ (highly significant)}$$

The LRT is the test of whether the model using these k covariates is able to reliably differentiate the probability of outcome ($Y = 1$) for different covariate patterns.

3.2.2 Observed/Predicted Discrepancies

Consider the following example:

Covariate	X_1	X_2	X_3	observed proportion of Y=1
1	1	0	1	3/10=0.3
2	1	0	0	2/10=0.2
3	1	0	0	9/10=0.9

observed overall proportion 14/30 = 0.47

Let's assume the logistic model based on X_1 to X_3 (k=3) produces these estimates.

	Pattern	Observed p	Estimated p
10	101	0.3	0.05
10	100	0.2	0.15
10	110	0.9	0.95

mean = 0.47

Clearly the model provides estimates that are much closer to the observed proportions than to the mean $P = 0.47$ (this would be $e^{\beta_0}/(1 + e^{\beta_0})$) for model with the intercept only. Still there are problems with fit to individual cases:

In patterns 1 (101) there is a total of 3 observations for which $Y = 1$ but predicted probability of observed outcome=0.05 only.

The problem of individual values being badly fitted is, however, inherent for these data: whenever for the same covariate patterns different outcomes are observed, some observations will be misfitted. And if there are only few discrepant observations, their prediction will be very poor. The only possible solution is to look for additional covariates which could explain these discrepancies, e.g. for pattern 101 we may hope that there is X_4 such that:

$$\begin{aligned} X_4 &= 0 \text{ for the only observation with } Y = 0 \text{ and} \\ X_4 &= 1 \text{ for 9 other observation with } Y = 1 \end{aligned}$$

This is rarely the case.

The issue of goodness-of-fit is not related directly to

1. significance

2. individual discrepancies within a given covariate pattern

Goodness-of-fit relates to the discrepancies between observed and predicted proportions for "subsets" of observations homogeneous with respect to covariate (independent variables).

These "subsets" are called "cells". If all independent variables are categorical each covariate pattern may be a "cell". Otherwise cells have to be created.

In our example, LRT asks whether predicted proportions (0.05, 0.15, 0.95) are closer to true observed proportions (0.3, 0.2, 0.9) than are constant proportions (0.47 = 0.47 = 0.47 = overall proportion) after having adjusted for degrees of freedom. (LRT tests significance).

Goodness-of-fit tests verify whether predicted proportions (0.05, 0.15, 0.95) are close enough to observed proportions (0.3, 0.2, 0.9).

3.3 Summary Measures of Goodness-of-Fit

We begin with the summary measures of goodness-of-fit, as they are routinely provided as output with any fitted model and give an overall indication of the fit of the model. Because these are summary statistics, they may not be very specific about the individual components. A small value for one of these statistics does not rule out the possibility of some substantial and thus interesting deviation from fit for a few subjects. On the other hand, a large value for one of these statistics is a clear indication of a substantial problem with the model.

3.3.1 Pearson Chi-Square and Deviance

For a given j-th covariate patterns, the Pearson residual is defined as follows:

$$r(y_j, \hat{\pi}_j) = \frac{y_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j(1 - \hat{\pi}_j)}}$$

m_j is the number of observations with patterns j; y_j is the number of observations with $Y = 1$ among m_j observations in pattern j; $\hat{\pi}_j$ is predicted probability of $Y = 1$; thus, $m_j\hat{\pi}_j$ is expected number of observation with $Y = 1$. So that,

$$[r(y_j, \hat{\pi}_j)]^2 = \frac{(observed - expected)^2}{variance(expected)}$$

By summing Pearson squared residuals $[r(y_j, \hat{\pi}_j)]^2$ over all J covariate patterns, we obtain the Pearson chi-square goodness-of-fit statistic

$$X^2 = \sum_{j=1}^{J}[r(y_j, \hat{\pi}_j)]^2 \sim \chi^2_{J-k-1}$$

J is total number of different covariate patterns.

For a given j-th covariate pattern, the deviance residual square is defined as follows:

$$d(y_j, \hat{\pi}_j) = 2\left[y_j \ln\left[\frac{y_j}{m_j\hat{\pi}_j}\right] + (m_j - y_j)\ln\left[\frac{m_j - y_j}{m_j(1 - \hat{\pi}_j)}\right]\right]$$

compares log likelihood for a given model with the log likelihood for a hypothetical "saturated model" which would contain as many parameters as there are distinct covariate patterns (J). Such a "saturated" model would be able to predict exactly each π_j.

The summary statistic based on the deviance residual square is the deviance

$$D = \sum_{j=1}^{J} d(y_j, \hat{\pi}_j) \sim \chi^2_{J-k-1}$$

The problem with (Pearson) X^2 and D is that they work well only if each cell (covariate pattern) has observations larger than 5. Thus if $J \approx n$ (which happens with continuous independent variables) they may be quite unreliable! The most natural solution is then to group the observations!

3.3.2 The Hosmer-Lemeshow Tests

"Goodness-of-fit chi-square (Hosmer-Lemeshow)" is based on discrepancy between observed and expected proportions in artificially created cells-obtained by grouping observations according to estimated probabilities $\hat{\pi}_j$.

Usually $g = 10$ groups are used (in standard packages):

1st Group: contains 10% of observations for which the estimated $\hat{\pi}_j$ is the lowest (1st risk decile)

2nd Group: next 10% with $\hat{\pi}_j$ higher than in 1st group but lower than for any other group.

etc $\cdots$.

Last Group: 10% with highest 10th estimated risk decile.

Example: $N = 50$, then:

1st Group: $\hat{\pi}_j$: 0.02; 0.03; 0.04; 0.07; 0.11;

2nd Group: $\hat{\pi}_j$: 0.13; 0.14; 0.18; 0.18; 0.18;

$\vdots$

10th Group: $\hat{\pi}_j$: 0.88; 0.92; 0.93; 0.93; 0.97;

Hosmer-Lemeshow statistic:

$$\hat{C} = \sum_{i=1}^{g} \frac{(O_i - n'_i\pi_i)^2}{n'_i\pi_i(1-\pi_i)} \sim \chi^2_{g-2}$$

$g = \#$ groups (usually 10)

$O_i =$ observed $\#\ y = 1$ in i-th group

$n'_i = \#$ different covariate patterns in i-th group

$\bar{\pi}_i =$ mean (accross n_i patterns) estimated probability in the group

$= \sum_{j=1}^{n'_i} m_j\hat{\pi}_i/n'_i$

$m_j = \#$ observations in cell j

Note: since all Goodness-of-fit X^2 tests are based on discrepancy measures, large values of X^2 and corresponding small p-values indicate poor fit, i.e. $H_0 =$ "the model fits the data perfectly" and any discrepancies are due to sampling error only.

3.3.3 The Brown's statistic

This statistic compares fitted logistic model with a potentially more complex and more general model thus it is not exactly a test of absolute goodness-of-fit with resoect to the data, but an assessment of the larger logistic assumption. It is used to increase confidence that the logistic model (as a class of models) is reasonable for givem data.

General strategy to use goodness-of-fit:

1. LR-Based ($2 \times O \times \ln(O/E)$) is appropriate only if $J \ll n$ so that each covariate pattern is replicated at least 5 times. If so then it is the best to use.

2. Hosmer-Lemeshow is best with continuous independent variables but in theory weaker than in LR and there is some arbitrariness. It is recommended to investigate individual Pearson residuals since grouping may obsure very poor fit to few cases.

3. Brown may be used as "secondary" statistic to confirm results from 1 or 2.

Chapter 4

Logistic Reg for Matched Case-Control Studies

4.1 Introduction

We are in the context of a matched case-control study where J cases and N controls have been selected. We are interested in the estimation of the relative risk of disease for a (or some) specific exposures while controlling for potential confounders and testing for interaction.

Review of the matched design:

- Each case is matched to M controls based on specific matching variables, e.g. age categories, gender, ethnicity or residence. The case and its controls form a matched set.

- The number of cases and controls are fixed by design.

- The cross-tabulation of the matching variables defines a certain number of strata. There will be few cases and their matched controls in each stratum. A special case of this is when each matched set defines a unique stratum.

For example: Lets say that we are looking at the relationship between death from asthma and use of beta-agonists (drugs used to treat asthma). The matching variables could be:

- age (from 5 to 54): 10 5-year categories
- gender: 2 categories
- residence: 5 categories
- season of the event: 4 categories

There are a total of $10 \times 2 \times 5 \times 4 = 400$ possible strata

Stratum #	Age	Gender	Residence	Season
1	5-9	Male	Big city	Winter
1	10-14	Male	Big city	Winter
1	15-19	Male	Big city	Winter
1	20-24	Male	Big city	Winter
1	25-29	Male	Big city	Winter
1	30-34	Male	Big city	Winter
1	35-39	Male	Big city	Winter
1	40-44	Male	Big city	Winter
1	45-49	Male	Big city	Winter
1	50-54	Male	Big city	Winter
1	5-9	Female	Big city	Winter
1	10-14	Female	Big city	Winter
1	15-20	Female	Big city	Winter
⋮				

4.2 Several Considerations

4.2.1 Why do we need a regression model with a matched sample?

As in the context of a cohort study, modeling in a matched case control study is used to overcome the limitations of a stratified analysis:

- Estimate the effect of a continuous exposure without having to categorise it.
- Some important confounders may not have been considered in the matching.
- To test for interaction between the exposure of interest and some matching variables.

4.2.2 Why do we have to use conditional logistic regression to analyse matched studies?

For purposes of validity (to produce an unbiased estimate of the relative risk), we need to take the matching into account in the analysis.

You could be tempted to use the logistic regression analysis with a model including a parameter for each matched set in order to take into account the design in the analysis (You know that the logistic model can be used with a case-control sample). The model would be:

$$\log(P/(1-P)) = \alpha_1 MS_1 + \alpha_2 MS_2 + \ldots + \alpha_J MS_J + \beta_1 E + \beta_2 C + \beta_3 E * C$$

where → J= number of matched sets

$$\rightarrow MS_j = \begin{cases} 1 & \text{if the subject is in the } j^{th} \text{ matched set} \\ 0 & \text{otherwise} \end{cases}$$

Note: MS stands for Matched Set

BUT

The method of estimation used in the unconditional logistic regression, i.e. the maximum likelihood, **works well** when:

1. The number of subjects in each stratum is large.

   or

2. The number of parameters stays fixed as the sample size increases

In a matched case-control study where each case and its matched controls form a unique stratum, **these assumptions are not respected.**

For example: Consider a 1 to 2 matched case-control design looking at the relationship between lung cancer and cigarette smoking.

- The controls have been matched to the cases by age, gender and environmental exposure.
- 50 cases of lung cancer have been selected.
- We want to estimate the relative risk associated with the number of cigarettes smoked per day while adjusting for 2 potential confounders.

- Lets assume that we want to use the unconditional logistic regression:

    1. To respect the stratified design, we would have to estimate 50 (strata) + 1 (exposure) + 2 (confounders) = 53 parameters. The model would be $\log(P/1-P) = \alpha_1 MS_1 + ... + \alpha_{50} MS_{50} + \beta_1 E + \beta_2 C_1 + \beta_3 C_2$
    2. The sample size is $50 \times 3 = 150$ subjects, but there are only 3 subjects per stratum.
    3. Since the matching is fine, new cases will probably fall in a new stratum, therefore the number of parameters will increase as the sample size increases

4.2.3 What happens if unconditional analysis is used with a matched case-control sample?

First consideraton: The model includes one parameter for each matched set.

In situations where the number of parameters to estimate has the same order of magnitude as the number of subjects, it is known that the technique of maximum likelihood can yield seriously biased estimates.

For the special case of 1 to 1 pair matching with a single binary exposure variable, it can be shown that:

- Unconditional MLE of OR = $(n_{10}/n_{01})^2$

  where n_{10}=number of pairs where the case is exposed and the control is un-exposed and n_{01}=number of pairs where the case is unexposed and the control is exposed. n_{10} and n_{01} are called the discordant pairs.

- Conditional MLE of OR=n_{10}/n_{01} (conditional on the number of discordant pairs)

So an odds ratio of 2 will tend to be estimated as 4 using the unconditional analysis. The bias will persist, to a lesser extent, for other matched designs. The bias will then depend slightly on

- the proportion of the control population exposed

- the true odds ratio
- the number of controls per matched set

For example:

- 1 to 2 matched design, true OR = 2 and prop. of controls exposed is 10%; unconditional OR = 2.9 (bias of 45%)
- 1 to 10 matched design, true OR = 5 and prop. of controls exposed is 10%; unconditional OR = 6.6 (bias of 32%)

Bias increases with the size of the true odds ratio, but decreases with the number of controls per set and the proportion of controls exposed.

<u>Second consideration</u>: The matching is simply ignored.

If someone ignores the matching and uses the unconditional logistic regression (without including in the model a parameter for each matched set) to analyse a matched case-control sample, the estimate may be biased.

Under certain conditions the data across matched sets may be pooled. If the matching variables are either:

- conditionally independent of disease status given the risk factor

or

- conditionally independent of the risk factors given disease status

both pooled and matched analysis provide approximately unbiased estimates of the relative risk for a dichotomous exposure.

In matched studies, the first condition is more relevant since the matching variables are guaranteed to be uncorrelated with disease in the sample as a whole. Of course this does not ensure that they have the same distributions among cases and controls conditionally on categories defined by the risk factors.

When using an unmatched analysis with data collected in a matched design, the estimates will be biased towards the null.

We need then, a special method of analysis which will be able to take the matching into account, but at the same time will only focus on the estimation of the parameters of interest, i.e., the betas associated with the exposure, the confounders and the effect modifiers.

4.3 Conditional Logistic Regression

Context of the analysis

We are in the context of a matched case-control study where:

- The number of cases and controls are fixed by design
- Each matched set contains exactly 1 case and M controls and each matched set defines a unique stratum
- We observe a vector of independent variables, for each subject

The independent variables $\mathbf{X} = (X_1, \ldots, X_p)$ represent the exposure variables, the confounders and the effect modifiers: there are p independent variables of interest. (This vector does not include the matching variables).

For example: Consider a 1 to 2 matched case-control study with 10 matched sets, looking at the effect of drug A in relation to disease D while controlling for gender ($male = 1, female = 0$). The vector of independent variables $\mathbf{X}_{ji}$, where j stands for the matched set ($j = 1, \ldots, 10$), i stands for the patient within the matched set ($i = 1,2,3$ where $1 = case$ and $2,3 = controls$) can be described as follow:

MATCHED SET j		DRUG	GENDER	$\mathbf{X}_{ji}$
1	Case	1.3	F	(1.3, 0)
	Control	0.7	M	(0.7, 1)
	Control	0.9	M	(0.9, 1)
2	Case	3.6	M	(3.6, 1)
	Control	1.4	M	(1.4, 1)
	Control	1.9	F	(1.9, 0)
3	Case	2.9	F	(2.9, 0)
	Control	1.8	M	(1.8, 1)
	Control	2.3	M	(2.3, 1)
4	Case	1.0	F	(4.0, 0)
	Control	2.1	F	(2.1, 0)
	Control	3.0	F	(3.0, 0)
⋮				
10	Case	2.1	F	(2.1, 0)
	Control	0.9	M	(0.9, 1)
	Control	1.2	M	(1.2, 1)

- We are interested in the estimation of the odds ratio of disease.

- We still assume that the probability of being disease follows a logistic model i.e.

$$P(Y = 1|X) = \frac{e^{\alpha+\sum_{k=1}^{p}\beta_k X_k}}{1 + e^{\alpha+\sum_{k=1}^{p}\beta_k X_k}}$$

where α=intercept (also referred to as β_0)

Note: For simplicity we assume that each matched set contains exactly M+1 subjects, but the theory has been generalized to situations where the number of cases and controls varies across the stratum.

Conditional Likelihood Function

As in the unconditional logistic regression, the method of maximum likelihood is used to estimate the regression parameters. It is precisely here, in defining the likelihood of the data, that the 2 methods (conditional and unconditional) differ.

First, we find the likelihood of observing the data in each matched set separately. The likelihood of the data in the jth matched set is:

$$L_j(\mathbf{X_1},\ldots,\mathbf{X_{M+1}}) = P(\mathbf{X_{j1}}|Y = 1)P(\mathbf{X_{j2}}|Y = 0)\cdots P(\mathbf{X_{jM+1}}|Y = 0)$$

where:

X_{j1} is the vector of independent variables for the case in the jth matched set.

$X_{j2}, \ldots, X_{jM+1}$ are the vectors of independent variables for the M controls in the jth matched set.

Note that in a cohort setting, we observe Y given X.

But even if we observe $P(\mathbf{X_{j1}}|Y)$ we are interested in the estimation of the relative risk (the odds ratio) of disease given the exposure, i.e.

$$\frac{P(Y=1|X)/(1-P(Y=1|X))}{P(Y=1|X')/(1-P(Y=1|X'))}$$

By the rule of conditional probability we can express $P(\mathbf{X_{j1}}|Y)$ in terms of the desired probability:

$$P(X_{ji}|Y) = \frac{P(Y|X_{ji})P(X_{ji})}{P(Y)}$$

where

$$P(Y=1|X_{ji}) = \frac{e^{\alpha_j+\sum_{k=1}^{p}\beta_k X_{jik}}}{1+e^{\alpha_j+\sum_{k=1}^{p}\beta_k X_{jik}}}$$

the logistic model.

Note that the intercept α depends on j, which stands for the matched set but the β's do not. This means that there is a different intercept for each matched set, but the β's are assumed to be constant across the strata.

By conditioning on the unordered observed values of the M+1 vectors **X** in the j^{th} stratum, we will get the following conditional likelihood:

$$\begin{aligned} L_j^*(\beta_1,\ldots,\beta_p) &= \frac{P(X_{j1}|Y=1)P(X_{j2}|Y=0)\cdots P(X_{jM+1}|Y=0)}{\sum_{i=1}^{M+1}P(X_{j1_i}|Y=1)P(X_{j2_i}|Y=0)\cdots P(X_{jM+1_i}|Y=0)} \\ &= \frac{e^{\sum_{k=1}^{p}\beta_k X_{j1k}}}{\sum_{i=1}^{M+1}e^{\sum_{l=1}^{p}\beta_k X_{jil}}} \end{aligned}$$

Where the summation in the denominator is over all possibilities of selecting 1 case among M+1 subjects i.e. M+1 possibilities.

We see that, after simplificaton, the conditional likelihood depends only on the β's parameters and $P_j(\mathbf{X})$, $P_j(Y)$ and the α_j parameters have been eliminated.

The conditional likelihood for the sample is the product of the J stratum specific likelihood:

$$L^*(\beta_1,\ldots,\beta_p) = \prod_{j=1}^{J} L_j^*(\beta_1,\ldots,\beta_p)$$

Example 1: The special case of a 1 to 1 matched design

In this situation there are 2 subjects within each stratum (1 case and 1 control). Let $\mathbf{X}_{j1}$ be the vector of independent variables for the case and $\mathbf{X}_{j0}$ for the control in the j^{th} matched set and $\beta = (\beta_1,\ldots,\beta_p)$ is the vector of regression parameters. For this special design, the conditional likelihood of the j^{th} stratum reduces to:

$$\begin{aligned} L_j^*(\beta_1,\ldots,\beta_p) &= \frac{e^{\sum_{k=1}^{p}\beta_k X_{j1k}}}{e^{\sum_{k=1}^{p}\beta_k X_{j1k}} + e^{\sum_{k=1}^{p}\beta_k X_{j0k}}} \\ &= \frac{e^{\sum_{k=1}^{p}\beta_k (X_{j1k}-X_{j0k})}}{1+e^{\sum_{k=1}^{p}\beta_k (X_{j1k}-X_{j0k})}} \end{aligned}$$

Example 2: The case of a unique dichotomous exposure

Consider the case of a 1 to M matched case-control study with only one dichotomous exposure variable coded $X = 1$ for exposed and $X = 0$ for unexposed. The model is

$$P(Y = 1|X) = \frac{e^{\alpha_j+\beta X}}{1+e^{\alpha_j+\beta X}}$$

The conditional likelihood (in the jth matched set) defined above reduces to:

$$\begin{aligned} L_j^*(\beta) &= \frac{e^{\beta X_{j1}}}{\sum_{\mu=1}^{M+1} e^{\beta X_{j\mu}}} \\ &= \frac{\mathrm{OR}^{X_{j1}}}{\sum_{\mu=1}^{M+1}\mathrm{OR}^{X_{j\mu}}} \end{aligned}$$

where X_{j1} is the case exposure in the j^{th} matched set.

For example, lets take a matched set with 1 case and 3 controls, where the case is exposed ($X = 1$) and only 1 of the 3 controls is exposed. The conditional likelihood of this matched set would be:

$$L_j^*(\beta) = \frac{\mathrm{OR}}{2\mathrm{OR}+2} = \frac{1}{2}\frac{\mathrm{OR}}{\mathrm{OR}+1}$$

The data can also be presented as a series of 2 × 2 tables, i.e. one for each matched set:

Matched set j

	Case	Control	
Exposed	a_j	b_j	m_{1j}
Unexposed	c_j	d_j	m_{0j}
	1	M	$M+1$

The fact of conditioning on the exposure history (unordered **X** vectors of independent variables) in this particular situation requires the knowledge of the total number of exposed in the table (m_{1j}) and thus knowledge of all the marginal totals in the 2 × 2 table (since the number of cases and controls are fixed by design).

If you condition on the number of exposed in a 2 × 2 table like this one, the data in the table are completely defined by the number of exposed cases. The conditional probability of observing a_j exposed cases is:

$$P(A_j = a_j | A_j + B_j = m_{1j}) = \frac{\binom{1}{a_j}\binom{M}{m_{1j}-a_j}\mathrm{OR}^{a_j}}{\sum_u \binom{1}{u}\binom{M}{m_{1j}-u}\mathrm{OR}^{u}}$$

This conditional probability is used in the Fisher's exact test.

lets take the same example as before, where the case is exposed and only 1 of the 3 controls is exposed:

Matched set j

	Case	Control	
Exposed	1	1	2
Unexposed	0	2	2
	1	3	4

$$P(A_i = 1|A_j + B_j = 2) = \frac{\binom{1}{1}\binom{3}{1}\mathrm{OR}}{\binom{1}{1}\binom{3}{1}\mathrm{OR} + \binom{1}{0}\binom{3}{2}}$$

We can see that this conditional probability is proportional to the conditional likelihood we just computed.

Interpretation of the beta coefficients

As in the unconditional logistic regression:

$$\mathrm{OR} = \frac{P(Y = 1|\mathbf{X})/(1 - P(Y = 1|\mathbf{X}))}{P(Y = 1|\mathbf{X}')/(1 - P(Y = 1|\mathbf{X}'))} = exp(\beta_1(X_1 - X_1') + \cdots + \beta_p(X_p - X_p'))$$

Special case: OR for a dichotomous exposure, assuming that all other covariates are equal.

$$\begin{aligned} \mathrm{OR} &= \frac{P(Y = 1|X_1 = 1, X_2, \cdots, X_p)/(1 - P(Y = 1|X_1 = 1, X_2, \cdots, X_p))}{P(Y = 1|X_1 = 0, X_2, \cdots, X_p)/(1 - P(Y = 1|X_1 = 1, X_2, \cdots, X_p))} \\ &= exp(\beta_1) \end{aligned}$$

Estimation of the parameters

The betas are estimated by maximization of the conditional likelihood (CMLE). This is done generally by an iterative process, using a computer, like Newton-Raphson mehtod. This method of estimation (MLE) provides an estimate for each beta, $\hat{\beta}_i$, and an estimate of its variance $\hat{\sigma}_i$. The MLE of β's are, in general, approximately normal.

Inference

- Confidence intervals for β_i

$$\hat{\beta}_i \pm \mathbf{Z}_{\alpha/2}\hat{\sigma}_i$$

- Wald test: $H_0 : \beta_i = 0$ vs $H_1 : \beta_i \neq 0$

$$Z = \hat{\beta}_i/\hat{\sigma}_i \sim N(0, 1)$$

- Likelihood ratio test (LRT)

Suppose we have the model $logit(P) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ and we want to test the hypothesis $H_0 : \beta_i = \beta_{i+1} = \cdots = \beta_p = 0$ (a subset of betas are equal to zero).

The LRT requires the computation of two conditional likelihoods; that of the full model, $L^*(\beta_{FULL})$, and that of the reduced one, $L^*(\beta_{REDUCED})$. The test is given by:

$$X^2 = 2\ln\{L^*(\beta_{FULL})/L^*(\beta_{REDUCED})\} \sim \chi^2_{k-j+1}$$

- G statistic to test the significance of the model (Special case of LRT)

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$X^2 = 2\ln\{L^*(full) - L^*(all\ betas = 0)\}$$

For the second model where each beta equals zero, the conditional likelihood reduces to

$$L^* = 1/(1 + M_1)\ 1/(1 + M_2)\ \cdots\ 1/(1 + M_J)$$

If M, the number of controls, is the same in each matched set then L^* reduces to $(1/(1 + M))^J$.

Consequences of the conditioning

- We can not estimate the probability of disease for a given exposure because their is no estimate of the intercept (α).

- Only the odds ratio of disease for any specified exposure in the model can be estimated.

- We can not estimate the OR associated with the matching variables because their parameters have been eliminated by conditioning. This implies that we can not verify if a matching variable is in fact a confounder. However, you can estimate the interaction between the main exposure and the matching variables.

- The estimation is only based on discordant matched sets. If in a matched set all the subjects (the cases and the controls) are either exposed or un-exposed, then the conditional likelihood of this matched set is equal to one and does not contribute any information to the estimation of the parameters. This can lead to a loss of efficiency.

4.4 Unconditional logistic regression with 1 to 1 match

As we saw previously, the conditional likelihood of the j^{th} stratum for a 1 to 1 matched sample is identical to the unconditional likelihood of a logistic regression model with the intercept equal to zero and the vector of independent variables equal to the value of the case minus the value of the control. The likelihood is:

$$
\begin{aligned}
L_j^*(\beta_1,\ldots,\beta_p) &= \frac{e^{\sum_{k=1}^{p}\beta_k(X_{j1k}-X_{j0k})}}{1+e^{\sum_{k=1}^{p}\beta_k(X_{j1k}-X_{j0k})}} \\
&= \frac{e^{\sum_{k=1}^{p}\beta_k Z_{jk}}}{1+e^{\sum_{k=1}^{p}\beta_k Z_{jk}}}
\end{aligned}
$$

where $Z_{jk} = (X_{j1k} - X_{j0k})$.

This implies that standard logistic regression software can be used to analyse 1 to 1 matched case-control. In order to accomplish this the data must be transformed as follow:

- The sample size is defined as the number of pairs, i.e. each pair becomes one observatoin.

- Each observation has a status of case: the outcome variable is set to one for each observation.

- The vector of independent variables, **Z**, becomes the difference between the case value and the control value.

  If dummy variables have to be used to model a categorical exposure, they have to be constructed for each case and control first and afterwards their differences will be taken. For example, if the exposure is defined in 4 categories, 3 dummy variables will be formed, lets say E_1, E_2 and E_3 where $E_i = 1$ if the subject falls in exposure category 1 and 0 otherwise. Three new variables called Z_1, Z_2 and Z_3 are formed; $Z_i = E_{icase} - E_{icontrol}$. These variables can take 3 possible values $(-1, 0, 1)$. Z_1, Z_2 and Z_3 will be entered in the computer as if they were continuous variables.

- The intercept is set to zero.

Chapter 5

The Use of Concordant Pairs

5.1 Introduction

One-to-one matched designs remain one of the most popular for case-control studies in which the possible association between a disease and a binary risk factor is of interest. Data can be simply summarized in a 2×2 table

		Control		
		+	−	
Case	+	a	b	
	−	c	d	
				n

where, for example, b represents the number among $n = a + b + c + d$ pairs for which cases are exposed and controls are not. The common belief is that matched designs require matched analysis. The preferred estimator of the common odds ratio, ψ, is therefore $\hat{\psi}_1 = b/c$ instead of the pooled estimate, $\hat{\psi}_2 = (a+b)(b+d)/[(a+c)(c+d)]$, which ignores the matching. A primary reason for not using $\hat{\psi}_2$ is that it is biased except when $\psi = 1$ or the matching is indeed unnecessary. Although there is recent research to support the use of $\hat{\psi}_1$, it is understandably frustrating for epidemiologists to use only the discordant pairs b and c given the effort made to collect data on all n pairs.

The distinction between a stratified and pooled analysis is nicely illustrated with data from a matched study of endometrial cancer and oral conjugated estrogen use reported in Schlesselman (1982). The 2×2 table has entries $a = 12$, $b = 43$, $c = 7$, and $d = 121$. Less than one-third of 183 pairs are discordant. The estimates and 95% confidence intervals of ψ for the matched and pooled analysis are $\hat{\psi}_1 = 6.11$ (2.76 to 13.65) and $\hat{\psi}_2 = 3.71$ (2.10 to 6.56). The familiar trade-off between bias and precision is clearly presented in this case. While $\hat{\psi}_1$ may be less subject to bias, it suffers from decreased precision due to the small number of discordant pairs.

A simulation comparing $\hat{\psi}_1$ and $\hat{\psi}_2$ in this case was conducted and the results are summarized in Table 5.1. Data were generated from a distribution with the parameter values observed for the endometrial data. Inferences are reported for $\beta = \ln(\psi)$. The simulation shows that both $\hat{\beta}_1$ and $\hat{\beta}_2$ are subject to serious bias in this case. $\hat{\beta}_2$ is more precise with variance about one-third that of $\hat{\beta}_1$. The negative bias and increased precision, however, result in poor coverage probabilities for $\hat{\beta}_2$. The nominal 2.5% lower and upper intervals for $\hat{\beta}_2$ have actual error rates of 0% and 19% The confidence intervals for $\hat{\beta}_1$ have error rates of 4% and 0.5%.

Table 5.1 Simulation results for comparing the conditional MLE, $\hat{\beta}_1$, pooled estimator, $\hat{\beta}_2$, and James-Stein estimator, $\hat{\beta}_3$, for a population like that of the endometrial cancer example where $\beta = 1.81$, $\phi^* = 1.57$, $\gamma = 0.90$, and the sample size is $n = 183$.

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
E($\hat{\beta}$)	2.01	1.59	1.87
Var($\hat{\beta}$)	.129	.0117	.103
Bias(%)	13	-12	3
MSE	.182	.093	.107
Error rate (%) of nominal 2.5% lower C.I.	1.3	0	1.0
Error rate (%) of nominal 2.5% upper C.I.	.5	19.2	3.0

We will discusses an alternative to $\hat{\beta}_1$ and $\hat{\beta}_2$ that is a compromise between complete stratification and complete pooling. The idea is to use information in the 2×2 table about the heterogeneity among matched pairs to determine the extent to which matching should be retained in the analysis. Recently, Liang (1987a) derived a locally most powerful test for the hypothesis that matching can be ignored. It rejects the null hypothesis if the statistic $S = ad - bc$, after standardization, is sufficiently large. In Section 2, a new estimator, $\hat{\beta}_3$, which incorporates the score statistic, S, is proposed. It uses S to compromise smoothly between $\hat{\beta}_1$ and $\hat{\beta}_2$. When there is little evidence of heterogeneity, $\hat{\beta}_2$ is preferred; when the probability of exposure varies substantially across pairs, $\hat{\beta}_1$ is preferred. The last column of Table 5.1 shows that for the endometrial cancer example, $\hat{\beta}_3$ is nearly unbiased and the performance on coverage probability improves upon both $\hat{\beta}_1$ and $\hat{\beta}_2$. In Section 3, the connection of $\hat{\beta}_3$ with the well-known James-Stein estimating procedure is discussed. More simulation results are presented in Section 4, followed by discussion.

5.2 Proposesd Estimator

5.2.1 The Mixed Model

Let (X_{i1}, X_{i2}) be the binary outcomes for exposure of the ith case and the matched control, $i = 1, \ldots, n$. Consider the model

$$logit[Pr(X_{ij} = 1|\alpha_i)] = \alpha_i + \beta(2 - j) \qquad (i = 1, \ldots, n; j = 1, 2)$$

where $\{\alpha_i\}$ is assumed to be a sequence of unobserved independent and identically distributed random variables which follow an unspecified distribution, F, with mean α abd variance θ. Thus, β is the common log-odds ratio and θ characterizes the variation among strata in probabilities of exposure. When $\theta = 0$, the matching is unnecessary and $\hat{\beta}_2$ in Section 1 is the efficient estimate of β.

The score statistic, $S = ad - bc$, for testing the hypothesis $\theta = 0$. One justification of this statistic which reflects the fact that $\hat{\beta}_2$ is consistent only when $\phi = 1$ or $\theta = 0$. Another justification, which will be useful for later development, is that S is proportional to $ad/(bc) - 1$, which consistently estimates

$$\frac{P_{11}P_{00}}{P_{10}P_{01}} - 1 = \phi - 1$$

where

$$\begin{aligned} P_{lk} &= Pr(X_1 = l, X_2 = k) \qquad (l, k = 0, 1) \\ &= \int e^{\alpha_i(l+k)+\beta l}(1 + e^{\alpha_i+\beta})^{-1}(1 + e^{\alpha_i})^{-1}dF(\alpha_i) \end{aligned}$$

Note that $\phi = 1$ when $\theta = 0$ and $\phi > 1$ when $\theta > 0$, a simple consequence of Hölder's inequality. We also note that because F is not specified, $\mathbf{T} = (a, b, c, d)$ are the minimum sufficient statistics, which have a multinomial distribution of size n and cell probabilities $\mathbf{P} = (P_{11}, P_{10}, P_{01}, P_{00})$.

5.2.2 Estimating Functions for $\hat{\beta}_1$ and $\hat{\beta}_2$

This section develops a common link between $\hat{\beta}_1$ and $\hat{\beta}_2$ that can be exploited to obtain the compromise estimator, $\hat{\beta}_3$. First, $\hat{\beta}_1$ is the solution of the equation

$$\sum_{i=1}^{n}\left(x_{i1} - \frac{x_{i1}e^{\beta x_{i1}} + x_{i2}e^{\beta x_{i2}}}{e^{\beta x_{i1}} + e^{\beta x_{i2}}}\right) = \sum_{i=1}^{n} h_{ii}(\beta) = 0 \qquad (5.1)$$

This is simply the score equation based on the conditional likelihood derived by Breslow et al.(1978) and can be expressed in terms of $\mathbf{T}$ as

$$\frac{b}{e^{\beta}+1} - \frac{ce^{\beta}}{e^{\beta}+1} = 0$$

The pooled estimator, $\hat{\beta}_2$, can be derived as the solution of the following estimating equation

$$\begin{aligned}\sum_{i=1}^{n}\sum_{k=1}^{n}\left(x_{i1} - \frac{x_{i1}e^{\beta x_{i1}} + x_{k2}e^{\beta x_{k2}}}{e^{\beta x_{i1}} + e^{\beta x_{k2}}}\right) &= \sum_{i=1}^{n}\sum_{k=1}^{n} h_{ik}(\beta) \\ &= \sum_{i=1}^{n} h_i^*(\beta) \\ &= \frac{(a+b)(b+d)}{e^{\beta}+1} - \frac{(a+c)(c+d)e^{\beta}}{e^{\beta}+1} \\ &= 0 \end{aligned} \qquad (5.2)$$

To obtain this equation, the conditional probability argument adopted in Breslow et al.(1978) is applied to all possible n^2 case-control combinations regardless of whether they are from the same pair or not. This is consistent with the notion that $\hat{\beta}_2$ is derived by ignoring the matching.

5.2.3 The Proposed Estimator

To obtain $\hat{\beta}_3$, let

$$U_i(\beta) = \frac{h_i^*(\beta)}{n} + [1 - w(\mathbf{T})]\left[h_{ii}(\beta) - \frac{h_i^*(\beta)}{n}\right] \qquad (5.3)$$

be the contribution from the ith pair to a new estimating function for β. Here w(**T**) is a function of **T** that converges to $w(P_{11}.P_{10}.P_{01}.P_{00})$ as $n \to \infty$ in such a way that (i) $0 \leq w \leq 1$; (ii) $w = 1$ when $\theta = 0$, and (iii) $w \to 0$ as $\theta \to \infty$. Equation (3) is introduced to compromise between $\hat{\beta}_1$ and $\hat{\beta}_2$. The use of estmating functions instead of estimators is crucial here because the estimator of β from each pair is undefined for one-to-one matching; no such problem exists when estimating functions are adopted. An estimating function $U(\beta)$ of β is arrived at by summing U_i over pairs, i.e.,

$$U(\beta) = \sum_{i=1}^{n} U_i(\beta)$$

The weight function $w(\mathbf{T})$ we consider is

$$w(\mathbf{T}) = \frac{bc}{ad}$$

It possesses properties (i) (iii) described above For this choice of w, the solution, $\hat{\beta}_3$, of $U(\beta) = 0$ is

$$\hat{\beta}_3 = \ln\left[\frac{bc(a+b)(b+d)/n + (ad-bc)b}{bc(a+c)(c+d)/n + (ad-bc)c}\right]$$

5.2.4 The Asymptotic Distribution of $\hat{\beta}_3$

It can be seen easily that $\hat{\beta}_3$ converges as $n \to \infty$ to

$$\beta_3 = \ln\left[\frac{P_{10}P_{01}(P_{11}+P_{10})(P_{10}+P_{00}+(P_{11}P_{00}-P10P_{01})P_{10}}{P_{10}P_{01}(P_{11}+P_{01})(P_{01}+P_{00}+(P_{11}P_{00}-P10P_{01})P_{01}}\right] = \ln\frac{A}{B} \qquad (5.4)$$

which is identical to β when either $\beta = 0$, in which case $P_{10} = P_{01}$, or when $\theta = 0$, in which case $P_{11}P_{00} = P_{10}P_{01}$.

Because of the multinomial structure of **T**, $\hat{\beta}_3$, has an asymptotically normal distribution with mean β_3 and variance given in Appendix 1. The same argument cna be applied to $\hat{\beta}_1$ and $\hat{\beta}_2$ as special cases when $w = 0$ or 1.

5.3 Connection Between $\hat{\beta}_3$ and James-Stein Procedures

It is of theoretical interest to relate $\hat{\beta}_3$ to the well-known James-Stein (J-S) estimating procedure (James and Stein, 1961). For this reason, we briefly review the J-S procedure for the Gaussian location problem and as a by-product provide a new justification for its use.

Let $\mathbf{x} = (x_1, \ldots, x_n)$ be n independent normal variates with means $(\mu_1, \ldots, \mu_n)$, and common known variance, σ^2. The J-S estimator of μ_i $(i = 1, \ldots, n)$ is

$$\hat{\mu}_i = x + \left[1 - \frac{(n-2)\sigma^2}{\sum_{i=1}^n (x_i - x)^2}\right](x_i - x) = x + [1 - w(x)](x_i - x)$$

One justification for the use of $\hat{\mu}_i$ is given by Efron and Morris (1972), who show that $\hat{\mu}_i$ is approximately the Bayes estimator of $\hat{\mu}_i$ when $\{\mu_i\}$ is assumed to follow a Gaussian prior distribution. We now offer an alternative justification with the normality assumption on $\{\mu_i\}$ relaxed. We first assume that $\{\mu_i\}$ is generated from an unspecified distribution with mean μ and variance θ. Following Liang (1987a), the score statistic for testing $\theta = 0$ is

$$S = \frac{1}{2}\left[\sum_{i=1}^{n} \frac{(x_i - x)^2}{\sigma^4} - \frac{n}{\sigma^2}\right] = S_1 - S_2 \qquad (5.5)$$

Note that

$$\frac{S_2}{S_1} = \frac{n}{n-2} w(x) \rightarrow \frac{1}{1 + \theta/\sigma^2} \quad as \quad n \rightarrow \infty \qquad (5.6)$$

This ratio is equal to 1 when $\theta = 0$ and converges to zero as $\theta \rightarrow \infty$. Either $1 - S_2/S_1$ or $1 - w(x)$ can then be considered as a smooth weight attached to x_i when both x_i and x are combined to estimate μ_i. More detailed derivations of (5) and (6) are given in Appendix 2. The J-S estimate, $\hat{\mu}_i$, can now be written as the solution of the estimating equation

$$\bar{x} - \mu_i + [1 - w(x)][x_i - \mu_i - (x - \mu_i)] = 0$$

which is in direct analogy to the estimating equation for $\hat{\beta}_3$.

There is, however, an intrinsic difference between the J-S and our estimating procedures. While the focus of J-S procedure is on the estimation of $\{\mu_i\}$, our focus is on the estimating fuction of β as $\{\alpha_i\}$ are nuisance parameters.

Table 5.2 Cases for simulation study. The parameters $\beta = \ln(P_{10}/P_{01})$, $\phi^* = \ln[P_{11}P_{00}/(P_{10}P_{01})]$ and $\gamma = P_{10} + P_{11}$ determine the multinomial probabilities, $P_{11}, P_{10}, P_{01}, P_{00}$

β	ϕ^*	γ	P_{11}	P_{10}	P_{01}	P_{00}
0	.00	.1	.01	.09	.09	.81
0	.00	.3	.09	.21	.21	.19
0	.00	.5	.25	.25	.25	.25
0	.25	.1	.01	.09	.09	.81
0	.25	.3	.10	.20	.20	.50
0	.25	.5	.27	.23	.23	.27
0	1.0	.1	.02	.08	.08	.82
0	1.0	.3	.11	.16	.16	.51
0	1.0	.5	.31	.19	.19	.31
1	.00	.1	.02	.21	.08	.69
1	.00	.3	.16	.38	.11	.32
1	.00	5	.37	.37	.13	.13
1	.25	.1	.03	.20	.07	.70
1	.25	.3	.17	.35	.13	.35
1	.25	.5	.37	.35	.13	.15
1	1.0	.1	.01	.17	.06	.73
1	1.0	.3	.20	.28	.10	.12
1	1.0	.5	.39	.29	.11	.21
2	.00	.1	.05	.11	.05	.49
2	.00	.3	.23	.53	.07	.17
2	.00	.5	.11	.11	.06	.06
2	.25	.1	.05	.38	.05	.52
2	.25	.3	.23	.51	.07	.19
2	.25	.5	.11	.13	.06	.07
2	1.0	.1	.06	.31	.01	.59
2	1.0	.3	.21	.13	.06	.27
2	1.0	.5	.15	.38	.05	.12

5.4 Simulation Results

The finite-sample properties of $\hat{\beta}_1$ and $\hat{\beta}_2$ and the James-Stein estimator, $\hat{\beta}_3$, have been compared in a simulation study. For given values of $\mathbf{P} = (P_{11}, P_{10}, P_{01}, P_{00})$, 1,000 independent realizations of $\mathbf{T} = (a, b, c, d)$ were generated from a multinomial distribution with probabilities **P** and sample size $n = 60, 100, or\ 200$ For each realization, the three estimators of $\beta = \ln(\psi)$ and their asymptotic variances were calculated and stored. The finite-sample expected value and variance of each estimator were estimated by the sample mean and variance of the 1,000 realizations. The coverage probabilities of asymptotic confidence intervals and their mean lengths were also determined.

There are a number of ways to parameterize the true probabilities, **P**. We have chosen the following parameters: $\beta = \ln(P_{10}/P_{01})$, the log-odds ratio; $\phi^* = \ln[P_{11}P_{00}/(P_{10}P_{01})]$, a measure of heterogeneity across pairs, and $\gamma = P_{01} + P_{11}$, the probability of exposure for a control. The simulation includes the cases $\beta = 0, 1, 2$; $\phi^* = 0, .25, 1.0$; and $\gamma = .1, .3$ *and* $.5$. Table 5.2 lists the multinomial parameters, **P**, for the cases srudied.

Table 5.3 present the bias for three estimators. The conditional likelihood estimator, $\hat{\beta}_1$, is the least biased but the James-Stein alternative performs nearly as well. The bias in $\hat{\beta}_2$ is much greater and increases with ϕ^* and γ. Table 5.4 displays the mean squared errors (MSE). The pooled estimator, $\hat{\beta}_2$, is clearly best as bias contributes less than variance at the sample sizes studied except at the largest values of β when $n = 200$. Note, however, that $\hat{\beta}_3$ has smaller MSE than $\hat{\beta}_1$ in most configurations when $\phi^* > 0$ and $\beta > 0$.

Table 5.5 presents coverage probabilities for the nominal 95% interval. Each entry is the difference between the observed and expected coverage (2.5%) in integral standard deviation units. The standard deviation is .5% for this simulation with 1,000 replications. Upper and lower coverages are reported separately. All three estimators perform similarly for the lower limit, tending to be slightly conservative. The probability of failing to cover decreases as β and ϕ^* increase. For the upper limit, $\hat{\beta}_1$ is approximately unbiased for all cases. The pooled estimator, $\hat{\beta}_2$, has grossly incorrect coverage when there is substantial heterogeneity among pairs (e.g., $\theta^* = 1.0$). This results from the negative bias evident in Table 5.3.

Table 5.3 Bias($\times 10$) in $\hat{\beta}_1$, conditional MLE; $\hat{\beta}_2$, pooled estimator; and $\hat{\beta}_3$, James-Stein estimator. A blank entry represents 0.

			$\hat{\beta}_1$			$\hat{\beta}_2$			$\hat{\beta}_3$		
n	β	ϕ^*	$\gamma = .1$	.3	.5	.1	.3	.5	.1	.3	.5
60	0	.00									
		.25									
		1.0									
	1	.00	1		1	1			1		1
		.25	1		1		−1		1		
		1.0	1	1	1	−1	−2	−2			
	2	.00	1	1	1	1	1	1	2	1	2
		.25	1	1	1		−1		1	1	1
		1.0	1	1	1	−3	−1	−1		−1	
100	0	.00									
		.25									
		1.0									
	1	.00	1						1		
		.25	1				−1		1		
		1.0				−2	−2	−2		−1	−1
	2	.00	1	1	1				1	1	1
		.25	1	1	1		−1	−1	1		
		1.0	1	1	1	−3	−5	−1		−1	−1
200	0	.00									
		.25									
		1.0									
	1	.00									
		.25					−1	−1			
		1.0				−1	−2	−2		−1	−1
	2	.00	1						1		
		.25	1			−1	−1	−1			
		1.0	1			−3	−1	−4	−1	−2	−1

The James-Stein estimator, $\hat{\beta}_3$, is nearly as good as $\hat{\beta}_1$ in upper-limit coverage except when ϕ^* is large and $\gamma = .5$. In this case, the pooled estimator, $\hat{\beta}_2$, performs very poorly and the weighting function, $w(\mathbf{T})$, used in $\hat{\beta}_3$ assigns some positive weight to the pooled estimating equation (2). Except in this instance, however, $\hat{\beta}_3$ maintains reasonable coverage.

Table 5.4 Mean squared errors for $\hat{\beta}_1$, conditional MLE; $\hat{\beta}_2$, pooled estimator; and $\hat{\beta}_3$, James-Stein estimator

				$\hat{\beta}_1$			$\hat{\beta}_2$			$\hat{\beta}_3$	
n	β	ϕ^*	$\gamma = .1$	.3	.5	.1	.3	.5	.1	.3	.5
60	0	.00	.52	.18	.15	.19	.16	.11	.55	.18	.15
		.25	.51	.19	.16	.18	.16	.13	.55	.18	.15
		1.0	.58	.25	.21	.12	.13	.11	.57	.21	.17
	1	.00	.11	.19	.19	.35	.16	.15	.10	.18	.17
		.25	.14	.22	.20	.35	.16	.11	.42	.19	.17
		1.0	.43	.29	.26	.30	.17	.15	.39	.21	.21
	2	.00	.11	.33	.10	.30	.17	.27	.36	.21	.33
		.25	.13	35	.10	.30	.16	.21	.36	.21	.31
		1.0	.11	.10	.39	.35	.30	.29	31	.30	.30
100	0	.00	.26	.10	.09	.25	.10	08	.27	.10	.09
		.25	.26	.11	.09	.21	.09	.08	.27	.10	.09
		1.0	.31	.13	.12	.22	.07	.07	.29	.11	.10
	1	.00	.21	.12	.13	.19	.09	.10	.22	.10	.11
		.25	.25	.13	.13	.18	.10	.09	.23	.11	.11
		1.0	.26	.16	.15	.18	.13	.12	.23	.13	.12
	2	.00	.19	.21	.26	.16	.11	.15	23	.15	.20
		.25	.33	.23	.26	.16	.11	.11	25	.16	.19
		1.0	.31	.25	.29	.25	.28	.25	.27	.21	.23
200	0	.00	.12	.05	.04	.12	.05	.01	12	.05	.04
		.25	.12	.05	.01	.12	.05	.01	12	.05	.04
		1.0	.11	.07	.06	.10	.01	.03	.13	.05	.04
	1	.00	.09	.05	.05	.08	.01	.05	.09	.01	.05
		.25	.10	.06	.06	.08	.05	.05	.09	.05	.05
		1.0	.12	.07	.07	.10	.09	.09	.11	.06	.06
	2	.00	.12	.09	.11	.07	.05	.07	.09	.06	.08
		.25	.13	.10	.12	.08	.06	.07	.09	.07	.09
		1.0	.17	11	13	.18	.21	.21	.11	.11	.12

Table 5.5a Actual α level of nominal 2.5% lower confidence limit for β. Entries are the estimated rate at which the lower limit failed to cover the true β divided by 0.5%, the standard deviation of the estimator based on 1,000 replications. Entries are rounded to the nearest integer. Blanks represent 0.

			$\hat{\beta}_1$			$\hat{\beta}_2$			$\hat{\beta}_3$		
n	β	ϕ^*	$\gamma = .1$	.3	.5	.1	.3	.5	.1	.3	.5
60	0	.00	−1		−2			−1	−1		−2
		.25	−2		−2			−1	−1		−2
		1.0	−3		−1	−1	1		−3		−1
	1	.00	−5	−2	−3	−2	−1	−1	−1	−1	−1
		.25	−4	−1	−3	−2	−1	−2	−1	−2	−2
		1.0	−5	−3	−1	−1	−3	−1	−1	−4	−5
	2	.00	−5	−5	−5	−1		−1	−3	−3	−3
		25	−5	−5	−5	−2	−3	−1	−1	−4	−4
		1.0	−5	−5	−5	−1	−5	−1	−5	−5	−5
100	0	.00									
		.25					1				
		1.0					1				
	1	.00	−1		−1				−1	1	1
		.25	−2			−3	−2	−1	−3	−1	−1
		1.0	−5	−1	−3	−5	−4	−4	[illegible]	−2	−4
	2	.00	−5	−3	−1	−1			−3	−1	−3
		.25	−5	−4	−1	−2	−3	−2	−1	−1	−4
		1.0	−5	−1	−5	−1	−5	−5	−5	−5	−5
200	0	.00		1	−1		1	−1		1	−1
		.25		2			2			2	
		1.0		2	−1		3	−1		2	−1
	1	.00	−1	−1			−1				
		.25		−1	−1	−1	−2	−2		−2	−2
		1.0	−1	−2	−1	−5	−4	−4	−3	−3	−3
	2	.00	−1	−1	−2	−1			−1		
		.25	−3	−2		−3	−1	−3	−1	−4	−2
		1.0	−2	−2	−2	−5	−5	−5	−1	−5	−4

Table 5.5b Actual α level of nominal 2.5% upper confidence limit for β. Entries are the estimated rate at which the upper limit failed to cover the true β divided by 0.5%, the standard deviation of the estimate based on 1,000 replications. Blanks represents. Blanks represent 0.

			$\hat{\beta}_1$			$\hat{\beta}_2$			$\hat{\beta}_3$		
n	β	ϕ^*	$\gamma = .1$	.3	.5	.1	.3	.5	.1	.3	.5
60	0	.00	−3			−1		−1	−2		
		.25	−2						−2		
		1.0	−2	−2	1			2	−3	−2	1
	1	.00	2			2		−1	1		−1
		.25	2	3		3	5		2	3	
		1.0	1	1		8	15	11	2	3	2
	2	.00	1				−1		−1	−2	−2
		.25	4	1	1	6	2	1	1		
		1.0		1		20	45	31	4	7	3
100	0	.00	−2	−1		−2			−2		
		.25	−2						−2		
		1.0	−1	−2			−2	1	−4	−2	
	1	.00			1	−1			−1		
		.25		1		1	5	1		2	1
		1.0			2	10	23	25	3	4	5
	2	.00		1					−1	−2	−2
		.25	1	1		6	8	5	2	3	1
		1.0	−1	1	1	30	71	52	3	11	10
200	0	.00		−2			−2			−2	
		.25		−1			−1			−1	
		1.0		−1	1		−1	1		−1	1
	1	.00							1		
		.25			1	1	1	5		1	2
		1.0		1	1	15	15	38	3	8	9
	2	.00		1			1		−1		−2
		.25		3		1	10	6			2
		1.0		2	1	51	121	91	9	17	15

Table 5.6 lists the ratio of the average confidence interval lengths for $\hat{\beta}_1$ and $\hat{\beta}_3$. The James-Stein estimator is more efficient except when $\beta = 0$. In many situations the gain is appreciable. For example, when $\beta = 2$ and $\phi^* = .25$ the average confidence interval lengths for $\hat{\beta}_3$ range from 0% to 50% less than those for $\hat{\beta}_1$. This substantial

gain in efficiency is achieved without serious degradation of the coverage probabilities. In summary, the simulation indicates that gains in efficiency can be achieved by using $\hat{\beta}_3$ without substantial errors in coverage rates. This is because $\hat{\beta}_3$ takes advantage of information in concordant pairs in estimating the log-odds ratio when the data indicate there is little heterogeneity across pairs.

5.5 Discussion

The conditional maximum likelihood estimator, $\hat{\beta}_1$, has long been used to estimate the common odds ratio in case-control studies. Its potential problem is the risk of reducing effective sample size by ignoring concordant pairs. On the other hand, the pooled estimator, $\hat{\beta}_2$, is subject to severe bias though its variance is much lower. The new proposed estimator, $\hat{\beta}_3$, serves as a compromise between bias and precision. The connection between this estimator and the James-Stein estimating procedure is emphasized. The representation of (2) for $\hat{\beta}_2$ through the conditional score argument is new. It serves to link $\hat{\beta}_1$ and $\hat{\beta}_2$ together so that the James-Stein procedure can be adopted in this one-to-one matched setting.

We expect the new estimator, $\hat{\beta}_3$, to be most useful in studies with fewer discordant pairs. When the number of discordant pairs is very large, investigators are unlikely to accept even small amounts of bias to decrease variance. It is in situations where the evidence about β is borderline that trade-off is desirable. An example is in occupational epidemiology, where large, expensive cohort studies are necessary to obtain even 50 or 100 case-control pairs for less prevalent diseases. Here large odds ratio estimates, say between 5 and 10, may have standard errors of the same magnitude when concordant pairs are ignored. The introduction of a small bias is justifiable if a substantial reduction in variance is achieved. Table 5.3 and 5.6 demonstrate that in studies with less than 200 case-control pairs, the bias introduced by using $\hat{\beta}_3$ is small relative to the large reductions in variance.

Table 5.6 Ratio of average confidence interval length for conditional estimator, $\hat{\beta}_1$, to the James-Stein estimator, $\hat{\beta}_3$

				γ	
n	β	σ^*	.1	.3	.5
60	0	.00	1 00	1.00	1.00
		.25	.92	1.09	1.00
		1.0	1.00	1.11	1.15
	1	.00	1.00	1.18	1.09
		.25	1.07	1.15	1.17
		1.0	1.16	1.32	1.29
	2	.00	1.21	1.38	1.36
		.25	1.25	1.53	1.35
		1.0	1.29	1.53	1.11
100	0	.00	.91	1.00	1.00
		.25	1.00	1.09	1.10
		1.0	1.00	1.09	1.15
	1	.00	1.08	1.09	1.09
		.25	1.08	1.08	1.17
		1.0	1.18	1.29	1.31
	2	.00	1.33	1.36	1.29
		.25	1.39	1.11	1.11
		1.0	1.31	1.18	1.15
200	0	.00	1.00	1.00	1.00
		.25	1.00	1.10	1.00
		1.0	1.09	1.08	1.27
	1	.00	1.00	1.10	1 10
		.25	1.09	1.18	1.09
		1.0	1.11	1.27	1.27
	2	.00	1.17	1.27	1.27
		.25	1.23	1.31	1.31
		1.0	1.27	1.11	1.11

The focus has been on one-to-one matching for a single binary exposure variable. The extension of $\hat{\beta}_3$ to more general sparse data is straightforward. Let $X_{i1}, \ldots, X_{im_i}, Z_{i1}, \ldots, Z_{ik_i}$ be the sets of multiple risk outcomes for the m_i cases and k_i controls in the ith stratum, $i = 1, \ldots, n$. If the pairwise argument described in Section 2 for the logistic regression model is adopted, the estimating function for stratified estimator is

$$\sum_{i=1}^{n}\left(\frac{1}{m_i+k_i}\right)\sum_{j=1}^{m_i}\sum_{l=1}^{k_i}\left(x_{ij}-\frac{x_{ij}e^{\beta x_{ij}}+z_{il}e^{\beta z_{il}}}{e^{\beta x_{ij}}+e^{\beta z_{il}}}\right) \tag{5.7}$$

The pooled estimator is the solusion of

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{i'=1}^{n}\sum_{l=1}^{k_i}\left(x_{ij}-\frac{x_{ij}e^{\beta x_{ij}}+z_{i'l}e^{\beta z_{i'l}}}{e^{\beta x_{ij}}+e^{\beta z_{i'l}}}\right)=0 \tag{5.8}$$

A James-Stein estimating function can be derived by combining equations (7) and (8) following the procedures described in Section 2.3.

Note that for a single binary exposure variable, (7), (8) and the weight function, w, reduce to

$$\sum_{i=1}^{n}\left[\frac{x_i(k_i-z_i)}{1+e^{\beta}}-\frac{(m_i-x_i)z_ie^{\beta}}{1+e^{\beta}}\right]/(k_i+m_i)$$

$$\sum_{i=1}^{n}\left[\frac{x_i\sum_l(k_l-z_l)}{1+e^{\beta}}-\frac{(m_i-x_i)\sum_l z_le^{\beta}}{1+e^{\beta}}\right]$$

$$w=\frac{\sum_{i=1}^{n}[x_i+z_i-k_i(\sum x_j)/(\sum k_i)-m_i(\sum z_j)/(\sum m_i)]^2}{(\sum x_i)(\sum k_i-\sum x_i)/(\sum k_i)+(\sum z_i)(\sum m_i-\sum z_i)/(\sum m_i)}$$

where $x_i=\sum_j x_{ij}$ and $z_i=\sum_l z_{il}$. The corresponding James-Stein estimator for ϵ^{β} is then

$$\frac{w(\sum x_i)(\sum k_i-\sum z_i)/n+(1+w)\sum[x_i(k_i-z_i)/(k_i+m_i)]}{w(\sum m_i-\sum x_i)(\sum z_i)/n+(1+w)\sum[(m_i-x_i)z_i/(k_i+m_i)]}$$

Finally, further work is needed to answer two questions: (i) Can some criteria be established to lead us to an "optimal" choice of $w(\mathbf{T})$ and (ii) Does the idea of shrinking estimating function rather than estimators have application in other contexts.

Chapter 6

Appendix

APPENDIX I

The Asymptotic Variances of the $\hat{\beta}$'s

Note that $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ are functions of $\mathbf{T} = (a, b, c, d)$ and that $\sqrt{n}(a - P_{11}, b - P_{10}, c - P_{01}, d - P_{00})'$ converges as $n \to \infty$ to a multivariate normal distribution with mean **0** and covariance

$$\sum = \begin{pmatrix} P_{11}Q_{11} & -P_{11}P_{10} & -P_{11}P_{01} & -P_{11}P_{00} \\ & P_{10}Q_{10} & -P_{10}P_{01} & -P_{10}P_{00} \\ & & P_{01}Q_{01} & -P_{01}P_{00} \\ & & & P_{00}Q_{00} \end{pmatrix}$$

where $\mathbf{Q} = 1 - \mathbf{P}$. The delta method can be applied to obtain the asymptotic distributions of the three estimators. Since both $\hat{\beta}_1$ and $\hat{\beta}_2$ can be regarded as special cases of $\hat{\beta}_3$ with $w(\mathbf{T}) = 0$ and 1, respectively, only the variance of $\hat{\beta}_3$ is presented here. Define

$$\begin{aligned} C_1 &= \frac{\partial \beta_3}{\partial P_{11}} = \frac{P_{10}P_{01}(P_{10} + P_{00}) + P_{10}P_{00}}{A} - \frac{P_{10}P_{01}(P_{01} + P_{00}) + P_{01}P_{00}}{B} \\ C_2 &= \frac{\partial \beta_3}{\partial P_{10}} = \frac{P_{11}P_{00}(P_{01} + 1) + 2(P_{11}P_{01} + P_{01}P_{00} - P_{01})P_{10} + 3P_{01}P_{10}^2}{A} \\ &\quad - \frac{P_{01}(P_{11} + P_{01})(P_{01} + P_{00}) - P_{01}^2}{B} \\ C_3 &= \frac{\partial \beta_3}{\partial P_{01}} = \frac{P_{10}(P_{11} + P_{10})(P_{10} + P_{00}) - P_{01}^2}{A} \\ &\quad - \frac{P_{11}P_{00}(P_{10} + 1) + 2(P_{11}P_{10} + P_{10}P_{00} - P_{10})P_{01} + 3P_{10}P_{01}^2}{B} \end{aligned}$$

$$C_4 = \frac{\partial \beta_3}{\partial P_{00}} = \frac{P_{10}P_{01}(P_{11}+P_{10})+P_{11}P_{10}}{A} - \frac{P_{10}P_{01}(P_{11}+P_{01})+P_{11}P_{01}}{B}$$

where A and B are given in (4). The asymptotic variance of $\hat{\beta}_3$ is

$$var(\hat{\beta}) = \mathbf{C}' \sum \mathbf{C}$$

with $\mathbf{C} = (C_1, C_2, C_3, C_4)'$

APPENDIX II

Derivations of (5) and (6)

For given μ_i, the x_i is normally distributed with density denoted as $f_i(x_i; \mu_i)$. The score statistic S for testing the variance of $\{\mu_i\}$, θ, being zero is

$$S = \frac{1}{2}\sum_{i=1}^{n}\left\{\left[\frac{\partial}{\partial\mu_i}\ln f_i(x_i; \hat{\mu}_i = x)\right]^2 + \frac{\partial^2}{\partial^2\mu_i}\ln f_i(x_i; \hat{\mu}_i = \bar{x})\right\}$$

$$\frac{1}{2}\sum_{i=1}^{n}\left[\frac{(x_i - x)^2}{\sigma^4} - \frac{1}{\sigma^2}\right] = (5)$$

Finally,

$$\frac{S_2}{S_1} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - x)^2/n} \longrightarrow \frac{\sigma^2}{var(x_1)} = \frac{\sigma^2}{\sigma^2 + \theta} = (6)$$

by noting

$$var(x_1) = var[E(x_1|\mu_1)] + E[var(x_1|\mu_1)] = \theta + \sigma^2$$

APPENDIX III

Explanation of the Attached Data Set

The attached data set is collected by Tuyns et al. (1977) in the French department of Ille-et-Vilaine (Brittany). Cases in this study were 200 males diagnosed with oesophageal cancer in one of the regional hospitals between January 1972 and April 1974. Controls were a sample of 778 adult males drawn from electoral lists in each commune, of whom 775 provided sufficient data for analysis. Both types of subject were administered a detailed dietary interview which contained questions about their consumption of tobacco and of various alcoholic beverages in addition to those about foods.

I use SAS and BMDP LR to run this data set in order to demonstrate the application of logistic regression. First, I use SAS to apply classic Mantel-Haenszel methodology to study the joint effects of two risk factors, alcohol and tobacco, on the relative risk of oesophageal cancer in Ille-et-Vilaine. Both factors were partitioned into four levels, actually I transfer alcohol into two levels, yielding 8 risk categories in all. The first approach is to compute separate estimates of the age-adjusted relative risk for each category. Later I estimate relative risks for each alcohol level, simultaneously adjusting for alcohol and age. This procedure requires to construct and summarize several different series of 2 × 2 tables. The relative risks obtained for each alcohol and tobacco level were multiplied together to estimate the joint effect of these two variables. Second, I use BMDP LR to demonstrate unconditional logistic regression analysis with model selection and model assessment. The starting point is the grouping of the cases and 775 controls into 4 × 4 × 6 = 96 cells, each of which represents a combination of the categories of alcohol, tobacco and age. Each cell are treated in the statistical analysis as independent binomial observations, which cases representing the numerator and cases+controls the denominator. The attached computer printouts to this paper give the analysis results of several different models.

Bibliography

[1] Albert, A., and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic models. *Biometrika*, **71**, 1-10.

[2] Beale, E.M.L. (1970). Note on procedures for variable selection in multiple regression. *Technometrics*, **12**, 909-14.

[3] Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York.

[4] BMDP (1992). *BMDP Statistical Software Manual*. Volume 2. University of California Press, Berkeley.

[5] Breslow, N. (1976). Regression analysis of the log odds ratio: a method for retrospective studies. *Biometrics*, **32**, 409-16.

[6] Breslow, N. (1981). Odds ratio estimators when the data are sparse. *Biometrika*, **68**, 73-84.

[7] Breslow, N. and Day, N.E. (1980). *Statistical Methods in Cancer Research*, vol. 1, *The Analysis of Case-Control Studies*, IARC, Lyon.

[8] Breslow, N. E., and Cain, K. C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, **75**, 11-20.

[9] Brown, C. C. (1982). On a goodness-of-fit test for the logistic model based on score statistics. *Communications in Statistics*, **11**, 1087-1105.

[10] Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, **19**, 15-18.

[11] Cook, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association*, **74**, 169-171.

[12] Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*, Chapman and Hall, London.

[13] Cordeiro, G.M. (1983a). Improved likelihood-ratio tests for generalized linear models. *J. R. Statist. Soc.*, **B**, **45**, 401-13.

[14] Day, N. E., and Byar, D. P. (1979). Testing hypotheses in case-control studies-equivalence of Mantel-Haenszel statistics and logit score tests. *Biometrics*, **35**, 623-630.

[15] Dixon, W. J. (1987). *BMDP Statistical Software*. University of California Press, Berkeley.

[16] Draper, N.R. and Smith, H. (1981). *Applied Regression Analysis*, 2nd edn, J.Wiley and Sons, New York.

[17] Fleiss, J. (1979). Confidence intervals for odds ratio in case-control studies: State of the art. *Journal of Chronic Diseases*, **32**, 69-77.

[18] Fleiss, J. (1986). *The Design and Analysis of Clinical Experiments*. Wiley, New York.

[19] Grizzle, J. E., Starmer, C. Frank, and Gary G. (1969). Analysis of categorical data by linear models. *Biometrics*, **25**, 489-504.

[20] Gross, A. J. (1981). A note on "chi-squared tests with survey data." *Journal of the Royal Statistical Society, Series B*, **46**, 270-272.

[21] Hosmer, D. W., Lemeshow, S., and Klar, J. (1988). Goodness-of-fit testing for multiple logistic regression analysis when the estimated probabilities are small. *Biometrical Journal*, **30**, 911-924.

[22] Hosmer, D. W., Wang, C. Y., Lin, I. C., and Lemeshow, S. (1978). A computer program for stepwise logistic regression using maximum likelihood. *Computer Programs in Biomedicine*, **8**, 121-134.

[23] Holford, T.R., White, C. and Kelsey, J.L. (1978). Multivariate analysis for matched case-control studies. *Am. J. Epidemiol.*, **107**, 245-256.

[24] Johnson, W. (1985). Influence measures for logistic regression: Another point of view. *Biometrika*, **72**, 59-65.

[25] Lehmann, E.L. (1986). *Testing Statistical Hypotheses*. Second Edition. John Wiley and Sons, New York.

[26] Lemeshow, S., and Hosmer, D. W. (1982). The use of goodness-of-fit statistics in the development of logistic regression models. *American Journal of Epidemiology*, **115**, 92-106.

[27] Lemeshow, S., and Hosmer, D. W. (1983). Estimation of odds ratios with categorical scaled covariates in multiple logistic regression analysis. *American Journal of Epidemiology*, **119**, 147-151.

[28] Liang, K. Y. (1987a). A locally most powerful test for homogeneity with many strata. *Biometrika*, **74**, 259-261.

[29] Liang, K. Y. (1987b). Extended Mantel-Haenszel estimating procedure for multivariate logistic regression units. *Biometrics*, **43**, 289-299.

[30] Mantel, N. and Hankey, W. (1975). The odds ratio of a 2 × 2 contingency table. *Am. Statistician*, **29**, 143-5.

[31] McCullagh, P. and Nelder, J.A. (1988). *Generalized Linear Models*. Chapman and Hall, London New York.

[32] McCullagh, P. (1986). The conditional distribution of goodness-of-fit statistics for discrerte data. *Journal of the American Statistical Association*, **81**, 104-107.

[33] Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models.*Journal of the Royal Statistical Society, Series A*, **135**, 370-384.

[34] Nordberg, L. (1981). Stepwise selection of explanatory variables in the binary logit model. *Scandinavian Journal of Statistics*, 8, 17-26.

[35] Nordberg, L. (1982). On variable selection in generalized linear and related regression models. *Communications in Statistics*, A, **11**, 2427-2449.

[36] Peduzzi, P. N., Hardy, R. J., and Holford, T. R. (1980). A stepwise selection procedure for nonlinear regression models. *Biometrics*, **36**, 511-516.

[37] Pike, M.C., Hill, A.P. and Smith, P.G.. (1980). Bias and efficiency in logistic analysis of stratified case-control studies. *Int. J. Epidemiol.*, **9**, 89-95

[38] Pregibon, D.(1981). Logistic regression diagnostics. *Annals of Statistics*, **9**, 705-724.

[39] Pregibon, D.(1984). Data analytic methods for matched case-control studies. *Biometrics*, **40**, 639-651.

[40] Rao, C. R. (1973). *Linear Statistical Inference and Its Application*. Second Edition. Wiley, New York.

[41] Richards, F.S.G. (1961). A methods of maximum likelihood estimation. *J. R. Stat. Soc. B.*, **23**, 469-475.

[42] Ronald, C. *Log-Linear Models*. Springer-Verlag.

[43] SAS Institute Inc. (1988). *SAS Guide for Personal Computer, Version 6.04*. SAS Institute Inc., Cary, NC.

[44] Seber, G.A.F. (1977). *Linear Regression Analysis*. J. Wiley and Sons, New Youk.

[45] Schlesselman, J. J. (1982). *Case-Control Studies: Design, Conduct, Analysis*. Oxford: Oxford University Press.

[46] Tsiatis, A. A. (1980). A note on a goodness-of-fit test for the logistic regression model. *Biometrika*, **67**, 250-251.

[47] Weisberg, S. (1985). *Applied Linear Regression*, Second Edition. New York: John Wiley and Sons.

[48] Wermuth, N. (1976). Model search among multiplicative models. *Biometrics*, **32**, 253-263.

ALCOHOL-OESOPHAGEAL CANCER DATA: INDIVIDUAL OUTCOME FORMAT

OBS	AGE	ALCOHOL	DALCOHOL	TOBACCO	STATUS	COUNT
1	1	0	0	0	0	40
2	1	0	0	1	0	10
3	1	0	0	2	0	6
4	1	0	0	3	0	5
5	1	1	0	0	0	27
6	1	1	0	1	0	7
7	1	1	0	2	0	4
8	1	1	0	3	0	7
9	1	2	1	0	0	2
10	1	2	1	1	0	1
11	1	2	1	3	0	2
12	1	3	1	0	0	1
13	1	3	1	1	1	1
14	1	3	1	2	0	1
15	1	3	1	3	0	2
16	2	0	0	0	0	60
17	2	0	0	1	1	1
18	2	0	0	1	0	13
19	2	0	0	2	0	7
20	2	0	0	3	0	8
21	2	1	0	0	0	35
22	2	1	0	1	1	3
23	2	1	0	1	0	20
24	2	1	0	2	1	1
25	2	1	0	2	0	13
26	2	1	0	3	0	8
27	2	2	1	0	0	11
28	2	2	1	1	0	6
29	2	2	1	2	0	2
30	2	2	1	3	0	1
31	2	3	1	0	1	2
32	2	3	1	0	0	1
33	2	3	1	1	0	3
34	2	3	1	2	1	2
35	2	3	1	2	0	2
36	3	0	0	0	1	1
37	3	0	0	0	0	45
38	3	0	0	1	0	18
39	3	0	0	2	0	10
40	3	0	0	3	0	4
41	3	1	0	0	1	6
42	3	1	0	0	0	32
43	3	1	0	1	1	4
44	3	1	0	1	0	17
45	3	1	0	2	1	5
46	3	1	0	2	0	10
47	3	1	0	3	1	5
48	3	1	0	3	0	2
49	3	2	1	0	1	3
50	3	2	1	0	0	13
51	3	2	1	1	1	6
52	3	2	1	1	0	8
53	3	2	1	2	1	1
54	3	2	1	2	0	4
55	3	2	1	3	1	2
56	3	2	1	3	0	2
57	3	3	1	0	1	4
58	3	3	1	1	1	3
59	3	3	1	1	0	1
60	3	3	1	2	1	2
61	3	3	1	2	0	1
62	3	3	1	3	1	4
63	4	0	0	0	1	2
64	4	0	0	0	0	47
65	4	0	0	1	1	3
66	4	0	0	1	0	19
67	4	0	0	2	1	3
68	4	0	0	2	0	9
69	4	0	0	3	1	4

ALCOHOL-OESOPHAGEAL CANCER DATA: INDIVIDUAL OUTCOME FORMAT

OBS	AGE	ALCOHOL	DALCOHOL	TOBACCO	STATUS	COUNT
70	4	0	0	3	0	2
71	4	1	0	0	1	9
72	4	1	0	0	0	31
73	4	1	0	1	1	6
74	4	1	0	1	0	15
75	4	1	0	2	1	4
76	4	1	0	2	0	13
77	4	1	0	3	1	3
78	4	1	0	3	0	3
79	4	2	1	0	1	9
80	4	2	1	0	0	9
81	4	2	1	1	1	8
82	4	2	1	1	0	7
83	4	2	1	2	1	3
84	4	2	1	2	0	3
85	4	2	1	3	1	4
86	4	3	1	0	1	5
87	4	3	1	0	0	5
88	4	3	1	1	1	6
89	4	3	1	1	0	1
90	4	3	1	2	1	2
91	4	3	1	2	0	1
92	4	3	1	3	1	5
93	4	3	1	3	0	1
94	5	0	0	0	1	5
95	5	0	0	0	0	43
96	5	0	0	1	1	4
97	5	0	0	1	0	10
98	5	0	0	2	1	2
99	5	0	0	2	0	5
100	5	0	0	3	0	2
101	5	1	0	0	1	17
102	5	1	0	0	0	17
103	5	1	0	1	1	3
104	5	1	0	1	0	7
105	5	1	0	2	1	5
106	5	1	0	2	0	4
107	5	2	1	0	1	6
108	5	2	1	0	0	7
109	5	2	1	1	1	4
110	5	2	1	1	0	8
111	5	2	1	2	1	2
112	5	2	1	2	0	1
113	5	2	1	3	1	1
114	5	3	1	0	1	3
115	5	3	1	0	0	1
116	5	3	1	1	1	1
117	5	3	1	1	0	1
118	5	3	1	2	1	1
119	5	3	1	3	1	1
120	6	0	0	0	1	1
121	6	0	0	0	0	17
122	6	0	0	1	1	2
123	6	0	0	1	0	4
124	6	0	0	3	1	1
125	6	0	0	3	0	2
126	6	1	0	0	1	2
127	6	1	0	0	0	3
128	6	1	0	1	1	1
129	6	1	0	1	0	2
130	6	1	0	2	0	3
131	6	1	0	3	1	1
132	6	2	1	0	1	1
133	6	2	1	1	1	1
134	6	3	1	0	1	2
135	6	3	1	1	1	1

ALCOHOL-OESOPHAGEAL CANCER DATA: CASE-CONTROL FORMAT

OBS	AGE	ALCOHOL	DALCOHOL	TOBACCO	CASES	CONTROLS
1	1	0	0	0	0	40
2	1	0	0	1	0	10
3	1	0	0	2	0	6
4	1	0	0	3	0	5
5	1	1	0	0	0	27
6	1	1	0	1	0	7
7	1	1	0	2	0	4
8	1	1	0	3	0	7
9	1	2	1	0	0	2
10	1	2	1	1	0	1
11	1	2	1	2	0	0
12	1	2	1	3	0	2
13	1	3	1	0	0	1
14	1	3	1	1	1	0
15	1	3	1	2	0	1
16	1	3	1	3	0	2
17	2	0	0	0	0	60
18	2	0	0	1	1	13
19	2	0	0	2	0	7
20	2	0	0	3	0	8
21	2	1	0	0	0	35
22	2	1	0	1	3	20
23	2	1	0	2	1	13
24	2	1	0	3	0	8
25	2	2	1	0	0	11
26	2	2	1	1	0	6
27	2	2	1	2	0	2
28	2	2	1	3	0	1
29	2	3	1	0	2	1
30	2	3	1	1	0	3
31	2	3	1	2	2	2
32	2	3	1	3	0	0
33	3	0	0	0	1	45
34	3	0	0	1	0	18
35	3	0	0	2	0	10
36	3	0	0	3	0	4
37	3	1	0	0	6	32
38	3	1	0	1	4	17
39	3	1	0	2	5	10
40	3	1	0	3	5	2
41	3	2	1	0	3	13
42	3	2	1	1	6	8
43	3	2	1	2	1	4
44	3	2	1	3	2	2
45	3	3	1	0	4	0
46	3	3	1	1	3	1
47	3	3	1	2	2	1
48	3	3	1	3	4	0
49	4	0	0	0	2	47
50	4	0	0	1	3	19
51	4	0	0	2	3	9
52	4	0	0	3	4	2
53	4	1	0	0	9	31
54	4	1	0	1	6	15
55	4	1	0	2	4	13
56	4	1	0	3	3	3
57	4	2	1	0	9	9
58	4	2	1	1	8	7
59	4	2	1	2	3	3
60	4	2	1	3	4	0
61	4	3	1	0	5	5
62	4	3	1	1	6	1
63	4	3	1	2	2	1
64	4	3	1	3	5	1
65	5	0	0	0	5	43
66	5	0	0	1	4	10
67	5	0	0	2	2	5
68	5	0	0	3	0	2
69	5	1	0	0	17	17

ALCOHOL-OESOPHAGEAL CANCER DATA: CASE-CONTROL FORMAT

OBS	AGE	ALCOHOL	DALCOHOL	TOBACCO	CASES	CONTROLS
70	5	1	0	1	3	7
71	5	1	0	2	5	4
72	5	1	0	3	0	0
73	5	2	1	0	6	7
74	5	2	1	1	4	8
75	5	2	1	2	2	1
76	5	2	1	3	1	0
77	5	3	1	0	3	1
78	5	3	1	1	1	1
79	5	3	1	2	1	0
80	5	3	1	3	1	0
81	6	0	0	0	1	17
82	6	0	0	1	2	4
83	6	0	0	2	0	0
84	6	0	0	3	1	2
85	6	1	0	0	2	3
86	6	1	0	.	1	2
87	6	1	0	2	0	3
88	6	1	0	3	1	0
89	6	2	1	0	1	0
90	6	2	1	1	1	0
91	6	2	1	2	0	0
92	6	2	1	3	0	0
93	6	3	1	0	2	0
94	6	3	1	1	1	0
95	6	3	1	2	0	0
96	6	3	1	3	0	0

```
                         SAS     11:02 Thursday, September 30, 1993

                    TABLE OF DAL BY GROUP

              DAL        GROUP

              Frequency|
              Col Pct  |cas      |con      |  Total
              ---------+---------+---------+
              hig      |      96 |     109 |    205
                       |   48.00 |   14.06 |
              ---------+---------+---------+
              low      |     104 |     666 |    770
                       |   52.00 |   85.94 |
              ---------+---------+---------+
              Total          200       775      975


              STATISTICS FOR TABLE OF DAL BY GROUP

         Statistic                        DF      Value         Prob
         ------------------------------------------------------------
         Chi-Square                        1    110.255        0.000
         Likelihood Ratio Chi-Square       1     96.433        0.000
         Continuity Adj. Chi-Square        1    108.221        0.000
         Mantel-Haenszel Chi-Square        1    110.142        0.000
         Fisher's Exact Test  (Left)                           1.000
                              (Right)                       1.03E-22
                              (2-Tail)                      1.08E-22
         Phi Coefficient                          0.336
         Contingency Coefficient                  0.319
         Cramer's V                               0.336


         Statistic                                Value         ASE
         ------------------------------------------------------------
         Gamma                                    0.699        0.045
         Kendall's Tau-b                          0.336        0.036
         Stuart's Tau-c                           0.221        0.027

         Somers' D C|R                            0.333        0.037
         Somers' D R|C                            0.339        0.037

         Pearson Correlation                      0.336        0.036
         Spearman Correlation                     0.336        0.036

         Lambda Asymmetric C|R                    0.000        0.000
         Lambda Asymmetric R|C                    0.000        0.000
         Lambda Symmetric                         0.000        0.000

         Uncertainty Coefficient C|R              0.097        0.020
         Uncertainty Coefficient R|C              0.096        0.020
         Uncertainty Coefficient Symmetric        0.097        0.020

              Estimates of the Relative Risk (Row1/Row2)

                                                     95%
         Type of Study             Value      Confidence Bounds
         ------------------------------------------------------------
         Case-Control              5.640       4.001        7.951
         Cohort (Col1 Risk)        3.467       2.753        4.367
         Cohort (Col2 Risk)        0.615       0.539        0.701

         Sample Size = 975
```

```
                                SAS     11:02 Thursday, September 30, 1993

                    SUMMARY STATISTICS FOR DAL BY GROUP

   Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic    Alternative Hypothesis      DF        Value       Prob
-------------------------------------------------------------------
    1        Nonzero Correlation          1      110.142      0.000
    2        Row Mean Scores Differ       1      110.142      0.000
    3        General Association          1      110.142      0.000



        Estimates of the Common Relative Risk (Row1/Row2)
                                                          95%
Type of Study     Method                 Value    Confidence Bounds
-------------------------------------------------------------------
Case-Control      Mantel-Haenszel        5.640        4.083   7.791
  (Odds Ratio)    Logit                  5.640        4.001   7.951

Cohort            Mantel-Haenszel        3.467        2.749   4.373
  (Col1 Risk)     Logit                  3.467        2.753   4.367

Cohort            Mantel-Haenszel        0.615        0.561   0.673
  (Col2 Risk)     Logit                  0.615        0.539   0.701

The confidence bounds for the M-H estimates are test-based.

Total Sample Size = 975
```

```
                 TABLE 1 OF DAL BY GROUP
                 CONTROLLING FOR TOB-0

          DAL        GROUP

          Frequency|
          Col Pct  |cas     |con     |  Total
          ---------+--------+--------+
          hig      |     35 |     50 |     85
                   |  44.87 |  11.19 |
          ---------+--------+--------+
          low      |     43 |    397 |    440
                   |  55.13 |  88.81 |
          ---------+--------+--------+
          Total          78      447      525


           STATISTICS FOR TABLE 1 OF DAL BY GROUP
                  CONTROLLING FOR TOB 0

Statistic                                 Value       ASE
------------------------------------------------------------
Gamma                                     0.732     0.063
Kendall's Tau-b                           0.325     0.054
Stuart's Tau-c                            0.170     0.033

Somers' D C|R                             0.314     0.055
Somers' D R|C                             0.337     0.058

Pearson Correlation                       0.325     0.054
Spearman Correlation                      0.325     0.054

Lambda Asymmetric C|R                     0.000     0.000
Lambda Asymmetric R|C                     0.000     0.000
Lambda Symmetric                          0.000     0.000

Uncertainty Coefficient C|R               0.101     0.031
Uncertainty Coefficient R|C               0.096     0.030
Uncertainty Coefficient Symmetric         0.098     0.031

         Estimates of the Relative Risk (Row1/Row2)

                                               95%
Type of Study              Value       Confidence Bounds
------------------------------------------------------------
Case-Control               6.463       3.787      11.028
Cohort  (Col1 Risk)        4.213       2.878       6.167
Cohort  (Col2 Risk)        0.652       0.544       0.781

Sample Size = 525
```

```
                TABLE 2 OF DAL BY GROUP
                 CONTROLLING FOR TOB=1

          DAL       GROUP

          Frequency|
          Col Pct  |cas     |con     |  Total
          ---------+--------+--------+
          hig      |     31 |     36 |     67
                   |  53.45 |  20.22 |
          ---------+--------+--------+
          low      |     27 |    142 |    169
                   |  46.55 |  79.78 |
          ---------+--------+--------+
          Total          58      178      236


          STATISTICS FOR TABLE 2 OF DAL BY GROUP
                  CONTROLLING FOR TOB=1

     Statistic                                Value      ASE
     ------------------------------------------------------
     Gamma                                    0.638     0.096
     Kendall's Tau-b                          0.317     0.068
     Stuart's Tau-c                           0.246     0.057

     Somers' D C|R                            0.303     0.067
     Somers' D R|C                            0.332     0.072

     Pearson Correlation                      0.317     0.068
     Spearman Correlation                     0.317     0.068

     Lambda Asymmetric C|R                    0.000     0.000
     Lambda Asymmetric R|C                    0.060     0.110
     Lambda Symmetric                         0.032     0.060

     Uncertainty Coefficient C|R              0.084     0.036
     Uncertainty Coefficient R|C              0.079     0.034
     Uncertainty Coefficient Symmetric        0.082     0.035

          Estimates of the Relative Risk (Row1/Row2)

                                                 95%
     Type of Study             Value     Confidence Bounds
     ------------------------------------------------------
     Case-Control              4.529     2.406      8.524
     Cohort (Col1 Risk)        2.896     1.881      4.458
     Cohort (Col2 Risk)        0.639     0.507      0.806

     Sample Size = 236
```

```
                 TABLE 3 OF DAL BY GROUP
                  CONTROLLING FOR TOB=2

            DAL        GROUP

            Frequency|
            Col Pct  |cas     |con     |  Total
            ---------+--------+--------+
            hig      |     13 |     15 |     28
                     |  39.39 |  15.15 |
            ---------+--------+--------+
            low      |     20 |     84 |    104
                     |  60.61 |  84.85 |
            ---------+--------+--------+
            Total          33       99     132


           STATISTICS FOR TABLE 3 OF DAL BY GROUP
                  CONTROLLING FOR TOB=2

   Statistic                              Value       ASE
   ------------------------------------------------------
   Gamma                                  0.569     0.153
   Kendall's Tau-b                        0.257     0.095
   Stuart's Tau-c                         0.182     0.072

   Somers' D C|R                          0.272     0.102
   Somers' D R|C                          0.242     0.092

   Pearson Correlation                    0.257     0.095
   Spearman Correlation                   0.257     0.095

   Lambda Asymmetric C|R                  0.000     0.000
   Lambda Asymmetric R|C                  0.000     0.000
   Lambda Symmetric                       0.000     0.000

   Uncertainty Coefficient C|R            0.054     0.039
   Uncertainty Coefficient R|C            0.058     0.042
   Uncertainty Coefficient Symmetric      0.056     0.040

         Estimates of the Relative Risk (Row1/Row2)

                                              95%
   Type of Study            Value    Confidence Bounds
   ------------------------------------------------------
   Case-Control             3.640       1.497     8.850
   Cohort  (Col1 Risk)      2.414       1.379     4.226
   Cohort  (Col2 Risk)      0.663       0.464     0.948

   Sample Size = 132
```

```
                              SAS    11:02 Thursday, September 30, 1993

                    TABLE 4 OF DAL BY GROUP
                     CONTROLLING FOR TOB=3

              DAL        GROUP

              Frequency|
              Col Pct  |cas     |con     |  Total
              ---------+--------+--------+
              hig      |     17 |      8 |     25
                       |  54.84 |  15.69 |
              ---------+--------+--------+
              low      |     14 |     43 |     57
                       |  45.16 |  84.31 |
              ---------+--------+--------+
              Total          31       51       82


              STATISTICS FOR TABLE 4 OF DAL BY GROUP
                     CONTROLLING FOR TOB=3

         Statistic                                  Value       ASE
         ------------------------------------------------------------
         Gamma                                      0.734      0.122
         Kendall's Tau-b                            0.412      0.105
         Stuart's Tau-c                             0.368      0.099

         Somers' D C|R                              0.434      0.109
         Somers' D R|C                              0.392      0.103

         Pearson Correlation                        0.412      0.105
         Spearman Correlation                       0.412      0.105

         Lambda Asymmetric C|R                      0.290      0.136
         Lambda Asymmetric R|C                      0.120      0.209
         Lambda Symmetric                           0.214      0.154

         Uncertainty Coefficient C|R                0.127      0.066
         Uncertainty Coefficient R|C                0.137      0.071
         Uncertainty Coefficient Symmetric          0.132      0.068

               Estimates of the Relative Risk (Row1/Row2)

                                                 95%
         Type of Study            Value     Confidence Bounds
         ------------------------------------------------------
         Case-Control             6.527      2.320     18.362
         Cohort (Col1 Risk)       2.769      1.632      4.697
         Cohort (Col2 Risk)       0.424      0.235      0.765

         Sample Size = 82
```

```
                                SAS    11:02 Thursday, September 30, 1993


                   SUMMARY STATISTICS FOR DAL BY GROUP
                          CONTROLLING FOR TOB


  Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

 Statistic   Alternative Hypothesis    DF       Value       Prob
 --------------------------------------------------------------
     1        Nonzero Correlation       1       96.922      0.000
     2        Row Mean Scores Differ    1       96.922      0.000
     3        General Association       1       96.922      0.000



      Estimates of the Common Relative Risk (Row1/Row2)
                                                         95%
 Type of Study     Method               Value    Confidence Bounds
 --------------------------------------------------------------
 Case-Control      Mantel-Haenszel      5.257       3.778     7.315
   (Odds Ratio)    Logit                5.313       3.747     7.533

 Cohort            Mantel-Haenszel      3.181       2.526     4.004
   (Col1 Risk)     Logit                3.190       2.537     4.012

 Cohort            Mantel-Haenszel      0.628       0.572     0.689
   (Col2 Risk)     Logit                0.636       0.559     0.724

 The confidence bounds for the M-H estimates are test-based.



      Breslow-Day Test for Homogeneity of the Odds Ratios

 Chi-Square =    1.617          DF =    3          Prob = 0.656

 Total Sample Size = 975
```

```
      TABLE 1 OF DAL BY GROUP
       CONTROLLING FOR AGE=1

DAL       GROUP

Frequency|
Col Pct  |cas     |con     |  Total
---------+--------+--------+
hig      |      1 |      9 |     10
         | 100.00 |   7.83 |
---------+--------+--------+
low      |      0 |    106 |    106
         |   0.00 |  92.17 |
---------+--------+--------+
Total           1      115      116


      TABLE 2 OF DAL BY GROUP
       CONTROLLING FOR AGE=2

DAL       GROUP

Frequency|
Col Pct  |cas     |con     |  Total
---------+--------+--------+
hig      |      4 |     26 |     30
         |  44.44 |  13.68 |
---------+--------+--------+
low      |      5 |    164 |    169
         |  55.56 |  86.32 |
---------+--------+--------+
Total           9      190      199


      TABLE 3 OF DAL BY GROUP
       CONTROLLING FOR AGE=3

DAL       GROUP

Frequency|
Col Pct  |cas     |con     |  Total
---------+--------+--------+
hig      |     25 |     29 |     54
         |  54.35 |  17.37 |
---------+--------+--------+
low      |     21 |    138 |    159
         |  45.65 |  82.63 |
---------+--------+--------+
Total          46      167      213
```

```
               TABLE 4 OF DAL BY GROUP
                CONTROLLING FOR AGE=4

DAL        GROUP

Frequency|
Col Pct  |cas     |con     |  Total
---------+--------+--------+
hig      |     42 |     27 |     69
         |  55.26 |  16.27 |
---------+--------+--------+
low      |     34 |    139 |    173
         |  44.74 |  83.73 |
---------+--------+--------+
Total          76      166      242


               TABLE 5 OF DAL BY GROUP
                CONTROLLING FOR AGE=5

DAL        GROUP

Frequency|
Col Pct  |cas     |con     |  Total
---------+--------+--------+
hig      |     19 |     18 |     37
         |  34.55 |  16.98 |
---------+--------+--------+
low      |     36 |     88 |    124
         |  65.45 |  83.02 |
---------+--------+--------+
Total          55      106      161


               TABLE 6 OF DAL BY GROUP
                CONTROLLING FOR AGE=6

DAL        GROUP

Frequency|
Col Pct  |cas     |con     |  Total
---------+--------+--------+
hig      |      5 |      0 |      5
         |  38.46 |   0.00 |
---------+--------+--------+
low      |      8 |     31 |     39
         |  61.54 | 100.00 |
---------+--------+--------+
Total          13       31       44
```

```
                          SAS     11:02 Thursday, September 30, 1993

               SUMMARY STATISTICS FOR DAL BY GROUP
                      CONTROLLING FOR AGE

         Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

     Statistic    Alternative Hypothesis    DF       Value      Prob
     -------------------------------------------------------------
         1        Nonzero Correlation        1      85.009     0.000
         2        Row Mean Scores Differ     1      85.009     0.000
         3        General Association        1      85.009     0.000



           Estimates of the Common Relative Risk (Row1/Row2)
                                                            95%
     Type of Study     Method              Value     Confidence Bounds
     -------------------------------------------------------------
     Case-Control      Mantel-Haenszel     5.158       3.639     7.310
       (Odds Ratio)    Logit *             5.100       3.512     7.407

     Cohort            Mantel-Haenszel     2.888       2.305     3.618
       (Col1 Risk)     Logit *             2.947       2.371     3.663

     Cohort            Mantel-Haenszel     0.644       0.586     0.707
       (Col2 Risk)     Logit *             0.780       0.708     0.859

     The confidence bounds for the M-H estimates are test-based.

     * denotes that the logit estimators use a correction
          of 0.5 in every cell of those tables that contain a zero.


           Breslow-Day Test for Homogeneity of the Odds Ratios

     Chi-Square =     9.323          DF =    5          Prob = 0.097

     Total Sample Size = 975
```

```
                                      SAS     11:02 Thursday, September 30, 1993


                              TABLE OF GROUP BY ALC

                 GROUP      ALC

                 Frequency|       0|       1|       2|       3|  Total
                 ---------+--------+--------+--------+--------+
                 cas      |     29 |     75 |     51 |     45 |    200
                 ---------+--------+--------+--------+--------+
                 con      |    386 |    280 |     87 |     22 |    775
                 ---------+--------+--------+--------+--------+
                 Total         415      355      138       67      975



                              TABLE OF GROUP BY TOB

                 GROUP      TOB

                 Frequency|       0|       1|       2|       3|  Total
                 ---------+--------+--------+--------+--------+
                 cas      |     78 |     58 |     33 |     31 |    200
                 ---------+--------+--------+--------+--------+
                 con      |    447 |    178 |     99 |     51 |    775
                 ---------+--------+--------+--------+--------+
                 Total         525      236      132       82      975



                              TABLE OF GROUP BY AGE

        GROUP      AGE

        Frequency|       1|       2|       3|       4|       5|       6|  Total
        ---------+--------+--------+--------+--------+--------+--------+
        cas      |      1 |      9 |     46 |     76 |     55 |     13 |    200
        ---------+--------+--------+--------+--------+--------+--------+
        con      |    115 |    190 |    167 |    166 |    106 |     31 |    775
        ---------+--------+--------+--------+--------+--------+--------+
        Total         116      199      213      242      161       44      975
```

```
BMDPLR - STEPWISE LOGISTIC REGRESSION

Copyright 1977, 1979, 1981, 1982, 1983, 1985, 1987, 1988
  BMDP Statistical Software, Inc.

  BMDP Statistical Software, Inc.| BMDP Statistical Software
  1440 Sepulveda Blvd            | Cork Technology Park, Model Farm Rd
  Los Angeles, CA 90025 USA      | Cork, Ireland
  Phone (213) 479-7799           | Phone +353 21 542722
  Fax   (213) 312-0161           | Fax +353 21 542822
  Telex 4972934 BMDP UI          | Telex 75659 SSWL EI

Version: 1988 (IBM PC/DOS) No Math Coprocessor Required.

Manual : BMDP Manual Vol. 1 and Vol. 2 .

Digest : BMDP User's Digest .

Updates: State NEWS. in the PRINT paragraph for summary of new features.

09/20/93     AT 13:36:03

PROGRAM INSTRUCTIONS

/problem title is 'alcohol-oesophageal cancer: logistic regression'.
/input variables = 6.
       format = free.
       file = 'scc.dat'.

/variable names = age, alcohol, dalcohol, tobacco, cases, controls.

/regress scount=cases.
         fcount=controls.
         model=dalcohol.
         dvar=part.
         start=out.
         move=1.
         method=mlr.

/end

PROBLEM TITLE IS
alcohol-oesophageal cancer: logistic regression

NUMBER OF VARIABLES TO READ IN. . . . . . . . .        6
NUMBER OF VARIABLES ADDED BY TRANSFORMATIONS. .        0
TOTAL NUMBER OF VARIABLES . . . . . . . . . . .        6
CASE FREQUENCY VARIABLE . . . . . . . . . . . .
CASE LABELING VARIABLES . . . . . . . . . . . .
NUMBER OF CASES TO READ IN. . . . . . . . . . . TO END
MISSING VALUES CHECKED BEFORE OR AFTER TRANS. . NEITHER
BLANKS ARE. . . . . . . . . . . . . . . . . . . MISSING
INPUT FILE. . .scc.dat
REWIND INPUT UNIT PRIOR TO READING. . DATA. . .     YES
NUMBER OF WORDS OF DYNAMIC STORAGE. . . . . . .   16298

VARIABLES TO BE USED
     1 age     2 alcohol   3 dalcohol    4 tobacco     5 cases
     6 controls

INPUT FORMAT IS
FREE

MAXIMUM LENGTH DATA RECORD IS 80 CHARACTERS.

DEPENDENT VARIABLE. . . . . . . . . . . . . . . . 0
    COUNT VARIABLE. . . . . . . . . . . . . . . . 0
   SCOUNT VARIABLE. . . . . . . . . . . . . . . . 5 cases
   FCOUNT VARIABLE. . . . . . . . . . . . . . . . 6 controls
METHOD TO SELECT NEXT TERM TO REMOVE OR ENTER . mlr
HIERARCHICAL TERM INCLUSION RULE USED . . . . . SING
REMOVE LIMIT (P-VALUE MUST BE GREATER). . . . . 0.1500   0.1500
```

```
ENTER LIMIT (P-VALUE MUST BE LESS). . . . . . .  0.1000  0.1000
TOLERANCE . . . . . . . . . . . . . . . . . . .  0.0001000
CONVERGENCE CRITERION . . . . . . . . . . . . .  0.0000010
MAXIMUM NUMBER OF ITERATIONS. . . . . . . . . .      10
                  STEP HALVINGS . . . . . . . .       5

NUMBER OF CASES TO BE PRINTED . . . . .    10

 CASE    1       2        3        4        5         6
  NO.  age  alcohol  dalcohol  tobacco   cases    controls
----- --------- -------- -------- -------- -------- --------
   1     1       0        0        0        0        40
   2     1       0        0        1        0        10
   3     1       0        0        2        0         6
   4     1       0        0        3        0         5
   5     1       1        0        0        0        27
   6     1       1        0        1        0         7
   7     1       1        0        2        0         4
   8     1       1        0        3        0         7
   9     1       2        1        0        0         2
  10     1       2        1        1        0         1

*** DATA ERROR ***  CASE NO.  97 WILL BE DELETED.
WHILE READING VARIABLE     1, 2 RECORD(S) WOULD BE READ.
AS DEFINED BY CASE ONE, THERE MUST BE 1 RECORD(S) PER CASE.

NUMBER OF CASES READ. . . . . . . . . . . . . .       97
   CASES WITH USE SET TO NEGATIVE VALUE . . . .        1
      REMAINING NUMBER OF CASES . . . . . . . .       96

  TOTAL NUMBER OF RESPONSES USED IN THE ANALYSIS      975.
                                 cases     . . . . . .   200.
                                 controls . . . . . .   775.

  NUMBER OF DISTINCT COVARIATE PATTERNS . . . . .        2

DESCRIPTIVE STATISTICS OF INDEPENDENT VARIABLES
-------------------------------------------

 VARIABLE     GROUP              DESIGN VARIABLES
NO. N A M E   INDEX     FREQ        (  1)

  3 dalcohol    0        770           0
                1        205           1

STEP NUMBER    0
--------------

                         LOG LIKELIHOOD = -494.744
GOODNESS OF FIT CHI-SQ   (2*O*LN(O/E)) =    96.433   D.F.= 1  P-VALUE= 0.000
GOODNESS OF FIT CHI-SQ   ( C.C.BROWN ) =     0.000   D.F.= 0  P-VALUE= 1.000

                             STANDARD
    TERM     COEFFICIENT      ERROR      COEFF/S.E.   EXP(COEFFICIENT)

CONSTANT       -1.3545     0.7931E-01   -17.08          0.2581

STATISTICS TO ENTER OR REMOVE TERMS
---------------------------------

                APPROX.           APPROX.
     TERM       CHI-SQ.  D.F.     CHI-SQ. D.F.                LOG
                ENTER             REMOVE        P-VALUE   LIKELIHOOD

dalcohol        96.43     1                      0.0000    -446.5278
CONSTANT                          362.15    1    0.0000    -675.8185
CONSTANT                              IS IN        MAY NOT BE REMOVED.

STEP NUMBER    1                  dalcohol           IS ENTERED
--------------

                          LOG LIKELIHOOD =   -446.528
IMPROVEMENT CHI-SQUARE   ( 2*(LN(MLR) ) =      96.433  D.F.=  1 P-VALUE= 0.000
                                   STANDARD
        TERM      COEFFICIENT       ERROR         COEFF/S.E.      EXP(COEFFICIENT)

dalcohol          1.7299          0.1752             9.872           5.640
CONSTANT         -1.8569          0.1054           -17.61            0.1562

CORRELATION MATRIX OF COEFFICIENTS
----------------------------------

              dalcohol      CONSTANT

dalcohol      1.000
CONSTANT     -0.602         1.000

STATISTICS TO ENTER OR REMOVE TERMS
-----------------------------------
                      APPROX.               APPROX.
        TERM          CHI-SQ.   D.F.        CHI-SQ.   D.F.                  LOG
                      ENTER                 REMOVE           P-VALUE    LIKELIHOOD

dalcohol                                     96.43     1      0.0000     -494.7442
dalcohol                                        IS IN            MAY NOT BE REMOVED.
CONSTANT                                    457.76     1      0.0000     -675.4060
CONSTANT                                        IS IN            MAY NOT BE REMOVED.

NO TERM PASSES THE REMOVE AND ENTER LIMITS (  0.1500  0.1000 ) .

SUMMARY OF STEPWISE RESULTS

STEP          TERM                     LOG      IMPROVEMENT       GOODNESS OF FIT
NO.   ENTERED REMOVED   DF    LIKELIHOOD CHI-SQUARE P-VAL CHI-SQUARE       P-VAL
--- ------------------------ --- ---------- ---------- ----- ---------- -----
  0                              -494.744                      96.433         0.000
  1 dalcohol             1      -446.528     96.433    0.000    0.000         1.000

NUMBER OF INTEGER WORDS OF STORAGE USED IN PRECEDING PROBLEM     6976

/regress scount=cases.
         fcount=controls.
         model=dalcohol,age.
         dvar=part.
         start=in,in.
         move=0,0.
         method=mlr.

  TOTAL NUMBER OF RESPONSES USED IN THE ANALYSIS        975.
                                   cases    . . . . . .      200.
                                   controls . . . . . .      775.

  NUMBER OF DISTINCT COVARIATE PATTERNS . . . . .         12


DESCRIPTIVE STATISTICS OF INDEPENDENT VARIABLES
-----------------------------------------------


 VARIABLE     GROUP              DESIGN VARIABLES
NO. N A M E   INDEX    FREQ     (  1)    (  2)    (  3)    (  4)    (  5)

 3 dalcohol     0       770       0
                1       205       1

 1 age          1       116       0        0        0        0        0
                2       199       1        0        0        0        0
                3       213       0        1        0        0        0
                4       242       0        0        1        0        0
                5       161       0        0        0        1        0
                6        44       0        0        0        0        1

 STEP NUMBER   0
 --------------
```

```
                        LOG LIKELIHOOD    =    -394.461
GOODNESS OF FIT CHI-SQ   (2*O*LN(O/E))    =      11.041   D.F.=  5 P-VALUE=  0.051
GOODNESS OF FIT CHI-SQ (HOSMER-LEMESHOW)          3.650   D.F.=  7 P-VALUE=  0.819
GOODNESS OF FIT CHI-SQ   ( C.C.BROWN )    =       0.697   D.F.=  2 P-VALUE=  0.706

                            STANDARD
   TERM      COEFFICIENT    ERROR      COEFF/S.E.   EXP(COEFFICIENT)

dalcohol      1.6699        0.1896       8.807         5.312
age      (1)  1.5423        1.066        1.447         4.675
         (2)  3.1988        1.023        3.126         24.50
         (3)  3.7135        1.019        3.646         41.00
         (4)  3.9669        1.023        3.877         52.82
         (5)  3.9622        1.065        3.720         52.57
CONSTANT     -5.0543        1.009       -5.007         0.6382E-02


CORRELATION MATRIX OF COEFFICIENTS
----------------------------------


           dalcohol  age (1)  age (2)  age (3)  age (4)  age (5)  CONSTANT

dalcohol     1.000
age   (1)   -0.019   1.000
age   (2)   -0.018   0.931    1.000
age   (3)   -0.009   0.935    0.974    1.000
age   (4)    0.010   0.931    0.970    0.974    1.000
age   (5)    0.033   0.894    0.931    0.936    0.932    1.000
CONSTANT    -0.060  -0.942   -0.982   -0.987   -0.984   -0.946    1.000


STATISTICS TO ENTER OR REMOVE TERMS
-----------------------------------
                     APPROX.          APPROX.
    TERM             CHI-SQ. D.F.     CHI-SQ.  D.F.             LOG
                     ENTER            REMOVE        P-VALUE   LIKELIHOOD

dalcohol                               79.52     1  0.0000     -434.2220
dalcohol                                  IS IN         MAY  NOT BE REMOVED.
age                                   104.13     5  0.0000     -446.5278
age                                       IS IN         MAY  NOT BE REMOVED.
CONSTANT                              169.44     1  0.0000     -479.1807
CONSTANT                                  IS IN         MAY  NOT BE REMOVED.


NO TERM PASSES THE REMOVE AND ENTER LIMITS (  0.1500   0.1000 ) .

/regress  scount=cases.
          fcount=controls.
          model=dalcohol,age,dalcohol*age.
          dvar=part.
          start=in,in,in.
          move=0,0,0.
          method=mlr.

  TOTAL NUMBER OF RESPONSES USED IN THE ANALYSIS        975.
                                 cases    . . . . . .   200.
                                 controls . . . . . .   775.

  NUMBER OF DISTINCT COVARIATE PATTERNS . . . . .        12


DESCRIPTIVE STATISTICS OF INDEPENDENT VARIABLES
-----------------------------------------------


 VARIABLE     GROUP                DESIGN VARIABLES
NO. N A M E  INDEX     FREQ     (  1)    (  2)    (  3)    (  4)    (  5)
```

```
   3 dalcohol    0         770         0
                 1         205         1

   1 age         1         116         0         0         0         0         0
                 2         199         1         0         0         0         0
                 3         213         0         1         0         0         0
                 4         242         0         0         1         0         0
                 5         161         0         0         0         1         0
                 6          44         0         0         0         0         1

DESIGN VARIABLES FOR INTERACTION TERMS ARE GENERATED
FROM THE DESIGN VARIABLES OF MAIN EFFECTS.
FOR EXAMPLE WITH TWO VARIABLES, VARIABLE U HAVING 3 DESIGN
VARIABLES (NAMED U(1), U(2) AND U(3)) AND VARIABLE V HAVING
2 DESIGN VARIABLES (NAMED V(1) AND V(2)), THEIR INTERACTION
U*V WILL HAVE 6 DESIGN VARIABLES   U*V (1) = U(1) * V(1) ,
                                   U*V (2) = U(2) * V(1) ,
                                   U*V (3) = U(3) * V(1) ,
                                   U*V (4) = U(1) * V(2) ,
                                   U*V (5) = U(2) * V(2) ,
                                   U*V (6) = U(3) * V(2) .

AFTER 10 ITERATIONS CONVERGENCE CRITERION= 0.1098E-05 .
YOU MAY NEED TO INCREASE THE NUMBER OF ITERATIONS OR INCREASE
THF CONVERGENCE CRITERION IN THE REGRFSSION PARAGRAPH.

STEP NUMBER   0
---------------


                  LOG LIKELIHOOD -   -388.951
GOODNESS OF FIT CHI-SQ  (2*O*LN(O/E))  -     0.022 D.F.  1 P-VALUE  0.883
GOODNESS OF FIT CHI-SQ (HOSMER-LEMESHOW)-    0.011 D.F.  8 P-VALUE  1.000
GOODNESS OF FIT CHI-SQ ( C.C.BROWN )         0.000 D.F.  0 P-VALUE  1.000

                           STANDARD
     TERM    COEFFICIENT    ERROR   COEFF/S.E.  EXP(COEFFICIENT)

dalcohol      7.0107        1.131      6.200     1108.
age  (1)      5.7270       0.6029      9.499     307.0
     (2)      7.3369       0.4603      15.94     1536.
     (3)      7.8115       0.4400      17.75     2469.
     (4)      8.3258       0.4429      18.80     4129.
     (5)      7.8650       0.0000     0.0000     2605.
      THE ABOVE TERM DID NOT PASS THE TOLERANCE TEST.
d*a  (1)     -5.3869        1.331     -4.046     0.4576E-02
     (2)     -5.2764        1.187     -4.447     0.5111E-02
     (3)     -5.1608        1.173     -4.400     0.5737E-02
     (4)     -6.0628        1.194     -5.077     0.2328E-02
     (5)      4.5467        73.47     0.6189E-01  94.32
CONSTANT     -9.2196       0.3962     -23.27     0.9908E-04

CORRFLATION MATRIX OF COEFFICIENTS
----------------------------------


        dalcohol age(1) age(2) age(3) age(4) age(5) d*a(1) d*a(2) d*a(3) d*a(4) d*a(5)

dalcohol  1.000
age  (1)  0.230  1.000
age  (2)  0.302  0.566  1.000
age  (3)  0.316  0.592  0.775  1.000
age  (4)  0.314  0.588  0.770  0.806  1.000
age  (5)  0.000  0.000  0.000  0.000  0.000  0.000
d*a  (1) -0.849 -0.453 -0.256 -0.268 -0.266  0.000  1.000
d*a  (2) -0.953 -0.219 -0.388 -0.301 -0.299  0.000  0.809  1.000
d*a  (3) -0.964 -0.222 -0.291 -0.375 -0.302  0.000  0.819  0.919  1.000
d*a  (4) -0.947 -0.218 -0.286 -0.299 -0.371  0.000  0.804  0.902  0.933  1.000
d*a  (5) -0.014 -0.000 -0.000 -0.000 -0.000  0.000  0.011  0.013  0.013  0.013  1.000
CONSTANT -0.350 -0.657 -0.861 -0.901 -0.895  0.000  0.298  0.334  0.338  0.332  0.000

          CONSTANT

CONSTANT  1.000
```

```
STATISTICS TO ENTER OR REMOVE TERMS
-----------------------------------
                   APPROX.             APPROX.
     TERM          CHI-SQ. D.F.        CHI-SQ. D.F.                 LOG
                   ENTER               REMOVE          P-VALUE   LIKFLIHOOD

dalcohol                                  IS IN         MAY NOT BE REMOVED.
age                                       IS IN         MAY NOT BE REMOVED.
d*a                                    11.02      5    0.0510     -394.4609
d*a                                       IS IN         MAY NOT BE REMOVED.
CONSTANT                              146.93      1    0.0000     -462.4142
CONSTANT                                  IS IN         MAY NOT BE RFMOVED.


NO TERM PASSES THE REMOVE AND ENTER LIMITS (  0.1500   0.1000 ) .

AFTER  10 ITERATIONS CONVERGENCE CRITERION= 0.1098E-05 .
YOU MAY NEED TO INCREASE THE NUMBER OF ITERATIONS OR INCREASE
THE CONVERGENCE CRITERION IN THE REGRESSION PARAGRAPH.

/regress scount=cases.
         fcount=controls.
         interval=age.
         model=dalcohol,age,dalcohol*age.
         dvar=part.
         start=in,in,in.
         move=1,1,1.
         method=mlr.

  TOTAL NUMBER OF RESPONSES USED IN THE ANALYSIS       975.
                               cases     . . . . . .    200.
                               controls . . . . . .    775.

  NUMBER OF DISTINCT COVARIATE PATTERNS . . . . .       12


DESCRIPTIVE STATISTICS OF INDEPENDENT VARIABLES
-----------------------------------------------


 VARIABLE                                     STANDARD
NO. N A M E  MINIMUM  MAXIMUM    MEAN    DEVIATION   SKEWNESS  KURTOSIS

  1 age       1.0000   6.0000    3.2718    1.3867     0.0170   -0.9050


 VARIABLE    GROUP           DESIGN VARIABLES
NO. N A M E  INDEX    FREQ     (  1)

  3 dalcohol   0       770        0
               1       205        1

DESIGN VARIABLES FOR INTERACTION TERMS ARE GENERATED
FROM THE DESIGN VARIABLES OF MAIN EFFECTS.
FOR EXAMPLE WITH TWO VARIABLES, VARIABLE U HAVING 3 DESIGN
VARIABLES (NAMED U(1), U(2) AND U(3)) AND VARIABLE V HAVING
2 DESIGN VARIABLES (NAMED V(1) AND V(2)), THEIR INTERACTION
U*V WILL HAVE 6 DESIGN VARIABLES    U*V (1) = U(1) * V(1) ,
                                    U*V (2) = U(2) * V(1) ,
                                    U*V (3) = U(3) * V(1) ,
                                    U*V (4) = U(1) * V(2) ,
                                    U*V (5) = U(2) * V(2) ,
                                    U*V (6) = U(3) * V(2) .
STEP NUMBER   0
--------------

                         LOG LIKELIHOOD = -404.905
GOODNESS OF FIT CHI-SQ  (2*O*LN(O/E))  =   31.929  D.F.= 8   P-VALUE= 0.000
GOODNESS OF FIT CHI-SQ (HOSMER-LEMESHOW)= 19.558  D.F.= 8   P-VALUE= 0.012
GOODNESS OF FIT CHI-SQ   ( C.C.BROWN )    = 17.856  D.F.= 2  P-VALUE= 0.000

                               STANDARD
   TERM    COEFFICIENT          ERROR      COEFF/S.E.  EXP(COEFFICIENT)

dalcohol    1.7510            0.6384        2.743          5.761
age        0.61368            0.8531E-01    7.193          1.847
d*a        0.77896E-02        0.1642        0.4743E-01     1.008
CONSTANT   -4.0913            0.3611       -11.33          0.1672E-01


CORRELATION MATRIX OF COEFFICIENTS
----------------------------------


         dalcohol   age    d*a    CONSTANT

dalcohol  1.000
age       0.539    1.000
d*a      -0.956   -0.519   1.000
CONSTANT -0.566   -0.952   0.495   1.000


STATISTICS TO ENTER OR REMOVE TERMS
-----------------------------------
                APPROX.           APPROX.
  TERM          CHI-SQ.  D.F.     CHI-SQ.   D.F.              LOG
                ENTER             REMOVE          P-VALUE   LIKELIHOOD

dalcohol                             IS IN           MAY NOT BE REMOVED.
age                                  IS IN           MAY NOT BE REMOVED.
d*a                                0.00       1    0.9626    -404.9061
CONSTANT                         219.51       1    0.0000    -514.6605
CONSTANT                             IS IN           MAY NOT BE REMOVED.


STEP NUMBER   1         d*a       IS REMOVED
--------------

                              LOG LIKELIHOOD = -404.906
IMPROVEMENT CHI-SQUARE    ( 2*(LN(MLR) )   =   0.002    D.F.= 1   P-VALUE= 0.963
GOODNESS OF FIT CHI-SQ    (2*O*LN(O/E))    =  31.931    D.F.= 9   P-VALUE= 0.000
GOODNESS OF FIT CHI-SQ   (HOSMER-LEMESHOW) =  19.620    D.F.= 8   P-VALUE= 0.012
GOODNESS OF FIT CHI-SQ    ( C.C.BROWN )    =   7.072    D.F.= 2   P-VALUE= 0.029

                              STANDARD
  TERM     COEFFICIENT         ERROR      COEFF/S.E.  EXP(COEFFICIENT)

dalcohol    1.7800           0.1871        9.514         5.930
age        0.61579           0.7291E-01    8.446         1.851
CONSTANT   -4.0998           0.3141      -13.05          0.1658E-01


CORRELATION MATRIX OF COEFFICIENTS
----------------------------------


         dalcohol    age     CONSTANT

dalcohol  1.000
age       0.170     1.000
CONSTANT -0.365    -0.936     1.000


STATISTICS TO ENTER OR REMOVE TERMS
-----------------------------------
                APPROX.          APPROX.
   TERM         CHI-SQ.   D.F.   CHI-SQ.  D.F.               LOG
                ENTER            REMOVE          P-VALUE   LIKELIHOOD

dalcohol                         92.38      1    0.0000    -451.0978
age                              83.24      1    0.0000    -446.5278
d*a             0.00       1                     0.9626    -404.9050
```

```
d*a           IS OUT                           MAY NOT BE ENTERED.
CONSTANT                     283.55    1    0.0000   -546.6801
CONSTANT                        IS IN          MAY NOT BE REMOVED.

NO TERM PASSES THE REMOVE AND ENTER LIMITS (  0.1500  0.1000 ) .

SUMMARY OF STEPWISE RESULTS

STEP         TERM                LOG         IMPROVEMENT    GOODNESS OF FIT
NO.   ENTERED REMOVED  DF LIKELIHOOD CHI-SQUARE P-VAL CHI-SQUARE P-VAL
--- ---------------------- --- ---------- ---------- ----- ---------- -----
  0                           -404.905                       31.929   0.000
  1             d*a      1    -404.906   0.002      0.963   31.931   0.000

NUMBER OF INTEGER WORDS OF STORAGE USED IN PRECEDING PROBLEM    6906

/regress scount=cases.
         fcount=controls.
         interval=age,tobacco.
         model=dalcohol,age,tobacco,dalcohol*age,dalcohol*tobacco.
         dvar=part.
         start=in,in,in,in,in.
         move=0,0,0,1,1.
         method=mlr.

  TOTAL NUMBER OF RESPONSES USED IN THE ANALYSIS      975.
                                   cases  . . . . . .    200.
                                   controls . . . . . .    775.

  NUMBER OF DISTINCT COVARIATE PATTERNS . . . . .       48


DESCRIPTIVE STATISTICS OF INDEPENDENT VARIABLES
-----------------------------------------------


 VARIABLE                                    STANDARD
NO. N A M E  MINIMUM  MAXIMUM    MEAN    DEVIATION   SKEWNESS KURTOSIS

 1 age       1.0000   6.0000    3.2718    1.3867      0.0170  -0.9050
 4 tobacco   0.0000   3.0000    0.7651    0.9778      1.0223  -0.1576


 VARIABLE    GROUP            DESIGN VARIABLES
NO. N A M E  INDEX    FREQ     (  1)

 3 dalcohol    0      770       0
               1      205       1

DESIGN VARIABLES FOR INTERACTION TERMS ARE GENERATED
FROM THE DESIGN VARIABLES OF MAIN EFFECTS.
FOR EXAMPLE WITH TWO VARIABLES, VARIABLE U HAVING 3 DESIGN
VARIABLES (NAMED U(1), U(2) AND U(3)) AND VARIABLE V HAVING
2 DESIGN VARIABLES (NAMED V(1) AND V(2)), THEIR INTERACTION
U*V WILL HAVE 6 DESIGN VARIABLES    U*V  (1)  = U(1)  * V(1)  ,
                                    U*V  (2)  = U(2)  * V(1)  ,
                                    U*V  (3)  = U(3)  * V(1)  ,
                                    U*V  (4)  = U(1)  * V(2)  ,
                                    U*V  (5)  = U(2)  * V(2)  ,
                                    U*V  (6)  = U(3)  * V(2)  .

STEP NUMBER   0
---------------

                         LOG LIKELIHOOD = -391.123
GOODNESS OF FIT CHI-SQ  (2*O*LN(O/E))   =   69.442  D.F.= 40  P-VALUE= 0.003
GOODNESS OF FIT CHI-SQ (HOSMER-LEMESHOW)=   12.592  D.F.=  8  P-VALUE= 0.127
GOODNESS OF FIT CHI-SQ  ( C.C.BROWN )   =    9.596  D.F.=  2  P-VALUE= 0.008

                         STANDARD
  TERM    COEFFICIENT     ERROR      COEFF/S.E.   EXP(COEFFICIENT)
```

```
dalcohol  1.7331        0.7230        2.397         5.658
age       0.66621       0.8964E-01    7.432         1.947
tobacco   0.49397       0.1101        4.485         1.639
d*a       0.17661E-01   0.1704        0.1036        1.018
d*t      -0.73182E-01   0.1889       -0.3874        0.9294
CONSTANT -4.7075        0.4120       -11.43         0.9027E-02


CORRELATION MATRIX OF COEFFICIENTS
----------------------------------


          dalcohol   age   tobacco   d*a    d*t   CONSTANT

dalcohol  1.000
age       0.530    1.000
tobacco   0.252    0.215    1.000
d*a      -0.933   -0.526   -0.113   1.000
d*t      -0.443   -0.125   -0.583   0.217  1.000
CONSTANT -0.570   -0.930   -0.443   0.489  0.258   1.000


STATISTICS TO ENTER OR REMOVE TERMS
-----------------------------------
              APPROX.            APPROX.
  TERM        CHI-SQ.  D.F.      CHI-SQ.  D.F.              LOG
              ENTER              REMOVE         P-VALUE  LIKELIHOOD

dalcohol                           IS IN         MAY NOT BE REMOVED.
age                                IS IN         MAY NOT BE REMOVED.
tobacco                            IS IN         MAY NOT BE REMOVED.
d*a                              0.01      1    0.9177  -391.1279
d*t                              0.15      1    0.6991  -391.1973
CONSTANT                       237.63      1    0.0000  -509.9365
CONSTANT                           IS IN         MAY NOT BE REMOVED.


STEP NUMBER   1        d*a      IS REMOVED
---------------

                                    LOG LIKELIHOOD = -391.128
IMPROVEMENT CHI-SQUARE    ( 2*(LN(MLR) )   =    0.011  D.F.=  1  P-VALUE= 0.918
GOODNESS OF FIT CHI-SQ    (2*O*LN(O/E))    =   69.453  D.F.  41  P-VALUE  0.004
GOODNESS OF FIT CHI-SQ  (HOSMER-LEMESHOW)=   12.484  D.F.=  8  P-VALUE= 0.131
GOODNESS OF FIT CHI-SQ    ( C.C.BROWN )    =    5.851  D.F.=  2  P-VALUE= 0.054

                              STANDARD
   TERM    COEFFICIENT         ERROR      COEFF/S.E.  EXP(COEFFICIENT)

dalcohol    1.8030          0.2609       6.911        6.068
age         0.67113         0.7627E-01   8.800        1.956
tobacco     0.49527         0.1095       4.521        1.641
d*t        -0.77418E-01     0.1842      -0.4203       0.9255
CONSTANT   -4.7285          0.3599     -13.14         0.6840E-02

CORRELATION MATRIX OF COEFFICIENTS
----------------------------------

           dalcohol    age   tobacco    d*t  CONSTANT

dalcohol  1.000
age       0.130     1.000
tobacco   0.411     0.184   1.000
d*t      -0.684    -0.014  -0.577   1.000
CONSTANT -0.364    -0.906  -0.448   0.180  1.000

STATISTICS TO ENTER OR REMOVE TERMS
-----------------------------------
               APPROX.          APPROX.
   TERM        CHI-SQ. D.F.     CHI-SQ. D.F.              LOG
               ENTER            REMOVE       P-VALUE LIKELIHOOD
```

```
dalcohol                      IS IN     MAY NOT BE REMOVED.
age                         92.77    1  0.0000   -437.5125
age                           IS IN     MAY NOT BE REMOVED.
tobacco                       IS IN     MAY NOT BE REMOVED.
d*a          0.01    1                  0.9177   -391.1226
d*a         IS OUT                      MAY NOT BE ENTERED.
d*t                          0.18    1  0.6748   -391.2159
CONSTANT                   302.43    1  0.0000   -542.3448
CONSTANT                      IS IN     MAY NOT BE REMOVED.

STEP NUMBER   2       d*t      IS REMOVED
--------------

                               LOG LIKELIHOOD = -391.216
IMPROVEMENT CHI-SQUARE   ( 2*(LN(MLR) )   =  0.176  D.F.=  1  P-VALUE= 0.675
GOODNESS OF FIT CHI-SQ  (2*O*LN(O/E))     = 69.629  D.F.= 42  P-VALUE= 0.005
GOODNESS OF FIT CHI-SQ (HOSMER-LEMESHOW)= 11.717  D.F.=  8  P-VALUE= 0.164
GOODNESS OF FIT CHI-SQ  ( C.C.BROWN )     =  6.627  D.F.=  2  P-VALUE= 0.036

                               STANDARD
   TERM    COEFFICIENT     ERROR      COEFF/S.E.   EXP(COEFFICIENT)

dalcohol    1.7283       0.1907        9.061         5.631
age         0.67085      0.7619E-01    8.805         1.956
tobacco     0.46882      0.8980E-01    5.220         1.598
CONSTANT   -4.7024       0.3536      -13.30          0.9074E-02

CORRELATION MATRIX OF COEFFICIENTS
---------------------------------

          dalcohol    age   tobacco    CONSTANT

dalcohol   1.000
age        0.165     1.000
tobacco    0.022     0.214     1.000
CONSTANT  -0.335    -0.920    -0.425      1.000

STATISTICS TO ENTER OR REMOVE TERMS
-----------------------------------
                APPROX.           APPROX.
  TERM          CHI-SQ.  D.F.     CHI-SQ.  D.F.            LOG
                ENTER             REMOVE          P-VALUE  LIKELIHOOD

dalcohol                          83.76     1     0.0000   -433.0976
dalcohol                           IS IN            MAY NOT BE REMOVED.
age                               92.84     1     0.0000   -437.6351
age                                IS IN            MAY NOT BE REMOVED.
tobacco                           27.38     1     0.0000   -404.9061
tobacco                            IS IN            MAY NOT BE REMOVED.
d*a              0.04     1                       0.8467   -391.1973
d*a             IS OUT                              MAY NOT BE ENTERED.
d*t              0.18     1                       0.6748   -391.1279
d*t             IS OUT                              MAY NOT BE ENTERED.
CONSTANT                         309.85     1     0.0000   -546.1429
CONSTANT                           IS IN            MAY NOT BE REMOVED.


NO TERM PASSES THE REMOVE AND ENTER LIMITS (  0.1500  0.1000 ) .

SUMMARY OF STEPWISE RESULTS

STEP       TERM                   LOG        IMPROVEMENT    GOODNESS OF FIT
NO.   ENTERED REMOVED  DF    LIKELIHOOD CHI-SQUARE P-VAL CHI-SQUARE P-VAL
--- ---------------------- --- ---------- ---------- ----- ---------- -----
  0                          -391.123                      69.442   0.003
  1             d*a      1   -391.128    0.011      0.918  69.453   0.004
  2             d*t      1   -391.216    0.176      0.675  69.629   0.005

NUMBER OF INTEGER WORDS OF STORAGE USED IN PRECEDING PROBLEM     7016

/regress scount=cases.
```

```
          fcount=controls.
          model=alcohol.
          dvar=part.
          rtart=in.
          move=0.
          method=mlr.

   TOTAL NUMBER OF RESPONSES USED IN THE ANALYSIS       975.
                                       cases    . . . . . .    200.
                                       controls . . . . . .    775.

        NUMBER OF DISTINCT COVARIATE PATTERNS . . . . . 4


DESCRIPTIVE STATISTICS OF INDEPENDENT VARIABLES
-----------------------------------------------


 VARIABLE     GROUP               DESIGN VARIABLES
NO. N A M E  INDEX      FREQ     (  1)     (  2)     (  3)

 2 alcohol     0        415       0         0         0
               1        355       1         0         0
               2        138       0         1         0
               3         67       0         0         1

STEP NUMBER    0
---------------

                   LOG LIKELIHOOD =  -421.495

                                 STANDARD
  TERM       COEFFICIENT      ERROR    COEFF/S.E.   EXP(COEFFICIENT)

alcohol(1)     1.2712       0.2323      5.472        3.565
       (2)     2.0545       0.2611      7.868        7.803
       (3)     3.3042       0.3237      10.21        27.23
CONSTANT      -2.5885       0.1925     -13.44        0.7513E-01

CORRELATION MATRIX OF COEFFICIENTS
----------------------------------

         alcoh(1) alcoh(2)  alcoh(3)  CONSTANT

alcoh(1)  1.000
alcoh(2)  0.611    1.000
alcoh(3)  0.493    0.439     1.000
CONSTANT -0.829   -0.737    -0.595     1.000

STATISTICS TO ENTER OR REMOVE TERMS
-----------------------------------
             APPROX.        APPROX.
  TERM       CHI-SQ. D.F.   CHI-SQ. D.F.              LOG
             ENTER          REMOVE        P-VALUE   LIKELIHOOD

alcohol                     146.50   3    0.0000   -494.7442
alcohol                       IS IN          MAY NOT BE REMOVED.
CONSTANT                    365.05   1    0.0000   -604.0208
CONSTANT                      IS IN          MAY NOT BE REMOVED.

NO TERM PASSES THE REMOVE AND ENTER LIMITS (  0.1500  0.1000 ) .

NUMBER OF INTEGER WORDS OF STORAGE USED IN PRECEDING PROBLEM     6976

/regress scount=cases.
         fcount=controls.
         model=tobacco.
         dvar=part.
         start=in.
         move=0.
         method=mlr.
  TOTAL NUMBER OF RESPONSES USED IN THE ANALYSIS      975.
                            cases    . . . . . .      200.
                            controls . . . . . .      775.

       NUMBER OF DISTINCT COVARIATE PATTERNS . . . . . 4

DESCRIPTIVE STATISTICS OF INDEPENDENT VARIABLES
-----------------------------------------------

 VARIABLE    GROUP              DESIGN VARIABLES
NO. N A M E  INDEX   FREQ    (  1)    (  2)    (  3)

 4 tobacco    0       525     0        0        0
              1       236     1        0        0
              2       132     0        1        0
              3        82     0        0        1

STEP NUMBER   0
---------------

                    LOG LIKELIHOOD =  -480.821

                                STANDARD
  TERM       COEFFICIENT         ERROR    COEFF/S.E.  EXP(COEFFICIENT)

tobacco(1)   0.62451          0.1947      3.207        1.867
       (2)   0.64724          0.2355      2.748        1.910
       (3)   1.2480           0.2587      4.824        3.483
CONSTANT    -1.7458           0.1227    -14.23         0.1745

CORRELATION MATRIX OF COEFFICIENTS
----------------------------------

       tobac(1)  tobac(2)  tobac(3)  CONSTANT

tobac(1)  1.000
tobac(2)  0.328    1.000
tobac(3)  0.299    0.247     1.000
CONSTANT -0.630   -0.521    -0.474     1.000

STATISTICS TO ENTER OR REMOVE TERMS
-----------------------------------
           APPROX.       APPROX.
  TERM     CHI-SQ. D.F.  CHI-SQ. D.F.          LOG
           ENTER         REMOVE       P-VALUE   LIKELIHOOD

tobacco                   27.85    3    0.0000  -494.7442
tobacco                      IS IN         MAY NOT BE REMOVED.
CONSTANT                 286.57    1    0.0000  -624.1059
CONSTANT                     IS IN         MAY NOT BE REMOVED.

NO TERM PASSES THE REMOVE AND ENTER LIMITS (  0.1500  0.1000 ) .

NUMBFR OF INTEGER WORDS OF STORAGE USED IN PRECEDING PROBLEM     6976

/regress scount=cases.
         fcount=controls.
         model=age.
         dvar=part.
         start=lr.
         move=0.
         method=mlr.

  TOTAL NUMBER OF RESPONSES USED IN THE ANALYSIS      975.
                            cases    . . . . . .      200.
                            controls . . . . . .      775.

       NUMBER OF DISTINCT COVARIATE PATTERNS . . . . . 6

DESCRIPTIVE STATISTICS OF INDEPENDENT VARIABLES
-----------------------------------------------
```

```
VARIABLE    GROUP              DESIGN VARIABLES
NO. N A M E  INDEX   FREQ  (  1)  (  2)  (  3)  (  4)  (  5)

  1 age       1      116    0      0      0      0      0
              2      199    1      0      0      0      0
              3      213    0      1      0      0      0
              4      242    0      0      1      0      0
              5      161    0      0      0      1      0
              6       44    0      0      0      0      1

STEP NUMBER   0
---------------

                   LOG LIKELIHOOD =   -434.222

                              STANDARD
    TERM     COEFFICIENT      ERROR    COEFF/S.E.    EXP(COEFFICIENT)

  age(1)       1.6951         1.061      1.598        5.447
     (2)       3.4556         1.018      3.394        31.68
     (3)       3.9637         1.014      3.910        52.65
     (4)       4.0888         1.018      4.017        59.67
     (5)       3.8759         1.057      3.666        48.23
CONSTANT      -4.7449         1.004     -4.724        0.8696E-02

CORRELATION MATRIX OF COEFFICIENTS
----------------------------------

            age(1)   age(2)   age(3)   age(4)   age(5)   CONSTANT

age   (1)   1.000
age   (2)   0.934    1.000
age   (3)   0.938    0.977    1.000
age   (4)   0.934    0.973    0.977    1.000
age   (5)   0.899    0.937    0.941    0.937    1.000
CONSTANT  -0.947   -0.987   -0.991   -0.987   -0.950    1.000

STATISTICS TO ENTER OR REMOVE TERMS
-----------------------------------
           APPROX.          APPROX.
   TERM    CHI-SQ. D.F.     CHI-SQ. D.F.              LOG
           ENTER            REMOVE        P-VALUE   LIKELIHOOD

age                         121.04    5     0.0000   -494.7442
age                            IS IN          MAY NOT BE REMOVED.
CONSTANT                    149.31    1     0.0000   -508.8777
CONSTANT                       IS IN          MAY NOT BE REMOVED.

NO TERM PASSES THE REMOVE AND ENTER LIMITS (  0.1500  0.1000 ) .

NUMBER OF INTEGER WORDS OF STORAGE USED IN PRECEDING PROBLEM     6976

/regress scount=cases.
         fcount=controls.
         interval=alcohol,tobacco,age.
         model=alcohol,tobacco,age,alcohol*tobacco,alcohol*age.
         dvar=part.
         start=in,in,in,in,in.
         move=0,0,0,0,0.
         method=mlr.

  TOTAL NUMBER OF RESPONSES USED IN THE ANALYSIS        975.
                                 cases     . . . . . .   200.
                                 controls . . . . . .    775.

  NUMBER OF DISTINCT COVARIATE PATTERNS . . . . .        96

DESCRIPTIVE STATISTICS OF INDEPENDENT VARIABLES
---------------------------------------------- ----

 VARIABLE                               STANDARD
NO. N A M E  MINIMUM  MAXIMUM  MEAN    DEVIATION  SKEWNESS  KURTOSIS
```

```
2 alcohol    0.0000    3.0000    0.8533  0.9063     0.8461      -0.1396
4 tobacco    0.0000    3.0000    0.7651  0.9778     1.0223      -0.1576
1 age        1.0000    6.0000    3.2718  1.3867     0.0170      -0.9050

SINCE THE FIRST CHARACTERS OF VARIABLES NAMES ARE NOT
UNIQUE THE CHARACTERS A,B,... WILL BE USED TO INDICATE
ELEMENTS OF AN INTERACTION TERM.
   A  INDICATES VARIABLE   2  alcohol
   B  INDICATES VARIABLE   4  tobacco
   C  INDICATES VARIABLE   1  age

DESIGN VARIABLES FOR INTERACTION TERMS ARE GENERATED
FROM THE DESIGN VARIABLES OF MAIN EFFECTS.
FOR EXAMPLE WITH TWO VARIABLES, VARIABLE U HAVING 3 DESIGN
VARIABLES (NAMED U(1), U(2) AND U(3)) AND VARIABLE V HAVING
2 DESIGN VARIABLES (NAMED V(1) AND V(2)), THEIR INTERACTION
U*V WILL HAVE 6 DESIGN VARIABLES    U*V (1) = U(1) * V(1) ,
                                    U*V (2) = U(2) * V(1) ,
                                    U*V (3) = U(3) * V(1) ,
                                    U*V (4) = U(1) * V(2) ,
                                    U*V (5) = U(2) * V(2) ,
                                    U*V (6) = U(3) * V(2) .

STEP NUMBER   0
---------------

                          LOG LIKELIHOOD = -364.321
GOODNESS OF FIT CHI-SQ  (2*O*LN(O/E))    =  107.107  D.F.= 82  P-VALUE= 0.033
GOODNESS OF FIT CHI-SQ (HOSMER-LEMESHOW)=    12.426  D.F.=  8  P-VALUE= 0.133
GOODNESS OF FIT CHI-SQ  ( C.C.BROWN )    =    9.677  D.F.=  2  P-VALUE= 0.008

                         STANDARD
   TERM   COEFFICIENT    ERROR        COEFF/S.E.  EXP(COEFFICIENT)

alcohol     1.2674       0.3638        3.484        3.552
tobacco     0.58056      0.1490        3.896        1.787
age         0.75595      0.1246        6.065        2.130
A*B        -0.12321      0.9520E-01   -1.294        0.8841
A*C        -0.11182E-01  0.8333E-01   -0.1342       0.9889
CONSTANT   -5.8243       0.5866       -9.929        0.2955E-02

CORRELATION MATRIX OF COEFFICIENTS
----------------------------------

          alcohol  tobacco   age     A*B    A*C   CONSTANT

alcohol    1.000
tobacco    0.359    1.000
age        0.746    0.218   1.000
A*B       -0.463   -0.783  -0.179  1.000
A*C       -0.922   -0.165  -0.753  0.223  1.000
CONSTANT  -0.789   -0.441  -0.933  0.353  0.685  1.000

STATISTICS TO ENTER OR REMOVE TERMS
-----------------------------------
            APPROX.           APPROX.
   TERM     CHI-SQ.  D.F.     CHI-SQ.  D.F.             LOG
            ENTER             REMOVE        P-VALUE  LIKELIHOOD

alcohol                          IS IN          MAY NOT BE REMOVED.
tobacco                          IS IN          MAY NOT BE REMOVED.
age                              IS IN          MAY NOT BE REMOVED.
A*B                           1.64      1     0.1998   -365.1428
A*B                              IS IN          MAY NOT BE REMOVED.
A*C                           0.02      1     0.8933   -364.3300
A*C                              IS IN          MAY NOT BE REMOVED.
CONSTANT                    174.77      1     0.0000   -451.7037
CONSTANT                         IS IN          MAY NOT BE REMOVED.

NO TERM PASSES THE REMOVE AND ENTER LIMITS (  0.1500  0.1000 ) .

NUMBER OF INTEGER WORDS OF STORAGE USED IN PRECEDING PROBLEM    7016
```

```
/regress scount=cases.
         fcount=controls.
         interval=alcohol,tobacco,age.
         model=alcohol,tobacco,age.
         dvar=part.
         start=in,in,in.
         move=0,1,1.
         method=mlr.

  TOTAL NUMBER OF RESPONSES USED IN THE ANALYSIS      975.
                                   cases    . . . . . .    200.
                                   controls . . . . . .    775.

  NUMBER OF DISTINCT COVARIATE PATTERNS . . . . .      96

DESCRIPTIVE STATISTICS OF INDEPENDENT VARIABLES
-----------------------------------------------

 VARIABLE                                   STANDARD
NO. N A M E  MINIMUM  MAXIMUM    MEAN    DEVIATION  SKEWNESS KURTOSIS

 2 alcohol   0.0000   3.0000    0.8533   0.9063     0.8461  -0.1396
 4 tobacco   0.0000   3.0000    0.7651   0.9778     1.0223  -0.1576
 1 age       1.0000   6.0000    3.2718   1.3867     0.0170  -0.9050

STEP NUMBEP   0
---------------

                                LOG LIKELIHOOD = -365.157
GOODNESS OF FIT CHI-SQ  (2*O*LN(O/E))    -  108.779  D.F.=  84  P-VALUE  0.036
GOODNESS OF FIT CHI-SQ (HOSMER-LEMESHOW) =   15.358  D.F.=   8  P-VALUE  0.053
GOODNESS OF FIT CHI-SQ  ( C.C.BROWN )    -    8.462  D.F.=   2  P-VALUE  0.015

                              STANDARD
   TERM    COEFFICIENT         ERROR      COEFF/S.E.   EXP(COEFFICIENT)

alcohol     1.1026       0.1032        10.69       3.012
tobacco     0.43085      0.9394E-01     4.587      1.539
age         0.74375      0.8179E-01     9.094      2.104
CONSTANT   -5.6305       0.4083       -13.79       0.3587E-02

CORRELATION MATRIX OF COEFFICIENTS
----------------------------------

          alcohol  tobacco  age   CONSTANT

alcohol   1.000
tobacco   0.011    1.000
age       0.264    0.210   1.000
CONSTANT -0.517   -0.384  -0.905  1.000

STATISTICS TO ENTER OR REMOVE TERMS
-----------------------------------
           APPROX.           APPROX.
   TERM    CHI-SQ. D.F.      CHI-SQ. D.F.            LOG
           ENTER             REMOVE        P-VALUE   LIKELIHOOD

alcohol                      135.88   1     0.0000   -433.0976
alcohol                        IS IN           MAY NOT BE REMOVED.
tobacco                       21.04   1     0.0000   -375.6745
age                          102.39   1     0.0000   -416.3496
CONSTANT                     374.60   1     0.0000   -552.4568
CONSTANT                       IS IN           MAY NOT BE REMOVED.

NO TERM PASSES THE REMOVE AND ENTER LIMITS (   0.1500   0.1000 ) .

NUMBER OF INTEGER WORDS OF STORAGE USED IN PRECEDING PROBLEM     6998

/variable names = age, alcohol, dalcohol, tobacco, status, count.
          freq=count.

/transform
           alcogm=20+alcohol*40.
```

```
            tobagm=5+tobacco*10.
            agey=20+age*10.
            agey2=agey**2.

/regress dependent=status.
         interval=alcogm,tobagm,agey,agey2.
         model=alcogm,tobagm,agey,agey2.
         dvar=part.
         start=in,in,in,in.
         move=0,0,0,1.
         remove=.000002.
         enter=.000001.
         method=mlr.

  TOTAL NUMBER OF RESPONSES USED IN THE ANALYSIS       975.
                              SUCCESS . . . . . .      200.
                              FAILURE . . . . . .      775.

  NUMBER OF DISTINCT COVARIATE PATTERNS . . . . .       88

DESCRIPTIVE STATISTICS OF INDEPENDENT VARIABLES
-----------------------------------------------

 VARIABLE                                       STANDARD
NO. N A M E   MINIMUM   MAXIMUM    MEAN      DEVIATION   SKEWNESS  KURTOSIS

  7 alcogm    20.0000  140.0000   54.1333    36.2521     0.8461   -0.1396
  8 tobagm     5.0000   35.0000   12.6513     9.7779     1.0223   -0.1576
  9 agey      30.0000   80.0000   52.7179    13.8671     0.0170   -0.9050
 10 agey2    900.0000 6400.0000 2971.2820 1479.1560     0.4509   -0.6011

STEP NUMBER    0
---------------

                           LOG LIKELIHOOD =  -357.353
GOODNESS OF FIT CHI-SQ   (2*O*LN(O/E))    =    93.172  D.F.= 83  P-VALUE= 0.209
GOODNESS OF FIT CHI-SQ  (HOSMER-LEMESHOW)=    11.749  D.F.=  8  P-VALUE= 0.163
GOODNESS OF FIT CHI-SQ   ( C.C.BROWN )    =     4.753  D.F.=  2  P-VALUE= 0.093

                           STANDARD
  TERM      COEFFICIENT     ERROR    COEFF/S.E.   EXP(COEFFICIENT)

alcogm      0.26628E-01   0.2614E-02    10.18       1.027
tobagm      0.43951E-01   0.9559E-02    4.598       1.045
agey        0.34424       0.7551E-01    4.559       1.411
agey2      -0.23417E-02   0.6402E-03   -3.658       0.9977
CONSTANT   -15.298        2.219        -6.895       0.2270E-06

CORRELATION MATRIX OF COEFFICIENTS
----------------------------------

           alcogm  tobagm  agey   agey2  CONSTANT

alcogm     1.000
tobagm     0.027   1.000
agey      -0.003   0.059   1.000
agey2      0.032  -0.037  -0.993  1.000
CONSTANT  -0.110  -0.145  -0.982  0.955   1.000

STATISTICS TO ENTER OR REMOVE TERMS
-----------------------------------

         APPROX.         APPROX.
  TERM   CHI-SQ.  D.F.   CHI-SQ.  D.F.               LOG
         ENTER           REMOVE        P-VALUE   LIKELIHOOD

alcogm                   122.79   1    0.0000    -418.7488
alcogm                     IS IN          MAY NOT BE REMOVED.
tobagm                    21.17   1    0.0000    -367.9383
tobagm                     IS IN          MAY NOT BE REMOVED.
agey                      26.22   1    0.0000    -370.4630
agey                       IS IN          MAY NOT BE REMOVED.
agey2                     15.61   1    0.0001    -365.1567
CONSTANT                  75.62   1    0.0000    -395.1629
CONSTANT                    IS IN        MAY NOT BE REMOVED.


STEP NUMBER   1         agey2     IS REMOVED
---------------

                           LOG LIKELIHOOD =  -365.157
IMPROVEMENT CHI-SQUARE   ( 2*(LN(MLR) )  =     15.607   D.F.    1   P-VALUE  0.000
GOODNESS OF FIT CHI-SQ   (2*O*LN(O/E))   =    108.779   D.F.   84   P-VALUE  0.036
GOODNESS OF FIT CHI-SQ  (HOSMER-LEMESHOW)=     15.358   D.F.    8   P-VALUE  0.053
GOODNESS OF FIT CHI-SQ   ( C.C.BROWN )   =      8.462   D.F.    2   P-VALUE  0.015

                         STANDARD
  TERM      COEFFICIENT   ERROR       COEFF/S.E.  EXP(COEFFICIENT)

alcogm      0.27564E-01  0.2579E-02    10.69        1.028
tobagm      0.43085E-01  0.9394E-02    4.587        1.044
agey        0.74375E-01  0.8179E-02    9.094        1.077
CONSTANT  -7.8848        0.6029       -13.08        0.3764E-03

CORRELATION MATRIX OF COEFFICIENTS
----------------------------------

          alcogm    tobagm    agey    CONSTANT

alcogm    1.000
tobagm    0.011     1.000
agey      0.264     0.210     1.000
CONSTANT -0.508    -0.396    -0.923    1.000

STATISTICS TO ENTER OR REMOVE TERMS
-----------------------------------
           APPROX.          APPROX.
  TERM     CHI-SQ.  D.F.   CHI-SQ.  D.F.             LOG
           ENTER           REMOVE         P-VALUE  LIKELIHOOD

alcogm                     135.88    1    0.0000  -433.0976
alcogm                        IS IN          MAY NOT BE REMOVED.
tobagm                      21.04    1    0.0000  -375.6745
tobagm                        IS IN          MAY NOT BE REMOVED.
agey                       102.39    1    0.0000  -416.3496
agey                          IS IN          MAY NOT BE REMOVED.
agey2      15.61    1                     0.0001  -357.3533
agey2       IS OUT                           MAY NOT BE ENTERED.
CONSTANT                   305.63    1    0.0000  -517.9704
CONSTANT                      IS IN          MAY NOT BE REMOVED.

NO TERM PASSES THE REMOVE AND ENTER LIMITS (  0.0000   0.0000 ) .

SUMMARY OF STEPWISE RESULTS

STEP         TERM                LOG         IMPROVEMENT      GOODNESS OF FIT
NO.   ENTERED REMOVED   DF  LIKELIHOOD  CHI-SQUARE  P-VAL  CHI-SQUARE P-VAL
-----------------------------------------------------------------------------
  0                         -357.353                          93.172   0.209
  1            agey2     1  -365.157        15.607 0.000     108.779   0.036

NUMBER OF INTEGER WORDS OF STORAGE USED IN PRECEDING PROBLEM     7140
```