# DATA ANALYSIS FOR BACK PAIN

## BASED ON THE NATIONAL POPULATION HEALTH SURVEY

XIONG CHEN

# DATA ANALYSIS FOR BACK PAIN[1]

# BASED ON THE NATIONAL POPULATION HEALTH SURVEY

BY

XIONG CHEN, Ph.D. , B. Sc.

A Project
Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

Master of Science

McMaster University

November, 1997

[1] This is a joint work of Dr.N. Balakrishnan and Xiong Chen.

MASTER OF SCIENCE (1997) McMASTER UNIVERSITY
(Statistics) Hamilton, Ontario

TITLE: Data Analysis For Back Pain Based On The National Population Health Survey

AUTHOR: Xiong Chen, Ph.D. (Beijing Normal University, China)
B. Sc. (Sichuan Normal University, China)

SUPERVISOR: Dr. N. Balakrishnan, Department of Mathematics and Statistics, McMaster University

NUMBER OF PAGES:

Remark: This is a joint work of Dr. N. Balakrishnan and Xiong Chen

## ACKNOWLEGEMENTS

Many people have helped me toward the completion of my degree.

I would like to thank my supervisor, Dr. N. Balakrishnan for his patience and knowledge in guiding me through completion of this project. Thanks to Dr. H. S. Shannon for providing me the data set used in the project and some advice he gave on data analysis. Thanks to Dr. P.D.M. Macdonald and Dr. R. Viveros-Aguilera for their suggestions to improve the quality of this report.

Finally, special thanks go to my wife, Louise, who typed most part of this project. Her encouragement and help are well appreciated.

## ABSTRACT

Back pain is an important health and economic problem affecting a significant part of our population. It is of interest to both medical and behavioral professionals concerned with the complex role of the social and psychological factors in the etiology of somatic ailments. Although there has been much written about back injuries in military and industrial settings, little is known about the epidemiological patterns in a general population (Nagi et al., 1973). The objective of this study is to find: a) the major factors connected to back pain, b) whether the general work-stress index is related to back pain, where the general work-stress index is the sum of job stressors including psychological demands, job insecurity, physical exertion, decision latitude and the social support at work, and c) the relationship especially amongst back pain, activity restriction, age, job satisfaction and income.

The National Population Health Survey (NPHS) database is used in this project. Some statistical techniques such as logistic regression and log-linear models are used for data analysis. In this project all explanatory variables in logistic regression models are treated as continuous variables; all variables when used in log-linear models are treated as categorical data. Results are compared between these different methods. They are in close agreement with each other.

We conclude that age has very high impact on back pain with significance level being lower than 1%; activity restriction also has a strong relationship with back pain; chronic stress, childhood and adult stressors all have high association with back pain; job stressor and recent

life bad events are related fairly to back pain at significant level 5%; and income and job satisfaction do not have direct impact on back pain.

Although there is not much that can be done to change the normal aging process of the spinal column, some of the predictors identified such as job stressors are amenable to change.

**Contents**

# Chapter 1

## INTRODUCTION

### 1.1 Background

Back pain is an important health and economic problem affecting a significant part of our population. People spend a lot of money on the treatment of back pain. It is of interest to both medical and behavioral professionals concerned with the complex role of the social and psychological factors in the etiology of somatic ailments. Although there has been much written about back injuries in military and industrial settings, little is known about the epidemiological patterns in a general population (Nagi et al., 1973). The objective of this study is to find: a) the major factors connected to back pain, b) whether the general work-stress index is related to back pain, where the general work-stress index is the sum of job stressors including psychological demands, job insecurity, physical exertion, decision latitude and the social support at work, and c) the relationship especially amongst back pain, activity restriction, age, job satisfaction and income.

### 1.2 Data Source

The National Population Health Survey (NPHS) database is used in this project. The NPHS is designed to collect information related to the health of the Canadian population. The first cycle of data collection began in 1994, and will continue every second year for up to two decades. The survey will collect not only cross–sectional information, but also data from a panel of individuals at two-year intervals.

The target population of the NPHS includes household residents in all provinces, with the principle exclusion of populations on Indian Reserves, Canadian Forces Bases, and some remote areas in Quebec and Ontario. In each household, information was collected from all

household members and one person was randomly selected for a more in-depth interview. In this study, we treat each household as an observation, and assume all observations to be independent of each other.

The database includes components on health, demographic and socio-economic status such as age, education, household income, etc. In addition, a special focus of the first survey was psycho-social factors that may influence health; for example, stress and social support.

The sample size was about 20,000. The number of variables was about forty.

### 1.3 Data Extraction

Three steps were involved in extracting relevant data from the original database due to the fact that the original database is too huge to be studied. Firstly, we select observations by simple random sampling method; secondly, we chose a subset of major variables from the original database; lastly for the purpose of this project, we chose a subset of these variables again in order to limit our analysis to a number of hypotheses.

The sample size of the data set used is 2003. For a given logistic regression model, Cox and others have noted that the corresponding likelihood can be used to generate tests of one or more parameters. These likelihood ratio tests are valid under very general conditions as long as the overall sample size is large. Here, large means that if there are X parameters used in the model, a sample size should be at least 10(X+1) (Freeman, 1987, p.237). In our case, the sample size is large enough for the purpose of statistical analysis in the project.

### 1.4 Variables

There are 19 variables. Some of them are continuous variables; others are treated as categorical variables. We define the "back pain" as a response variable. The variables in the data set are as follows:

b --- chronic stress, with higher number indicating higher stress; Min=0, Max=13

c --- recent life events, with higher number indicating worse events (e.g., physical abuse, unwanted pregnancy, abortion or miscarriage, major financial difficulties, and serious problems at work or in school); Min=0, Max=7

d --- childhood and adult stressors, with higher number indicating more stressors (e.g., traumatic events – parental divorce, a lengthy hospital stay, prolonged parental unemployment, and frequent drug use); Min=0, Max=7

*e* --- general work-stress index; Min=3, Max=39

f --- psychological demands

g --- job insecurity

h --- physical exertion

i --- self- esteem score

j --- mastery score

k --- sense of coherence

l --- social support outside work

m --- age in 5 groups as:

1 --- 18-24

2 --- 25-34

3 --- 35-44

4 --- 45-54

5 --- 55-64

n --- level of education in 4 categories as:

1 --- less than secondary

2 --- secondary complete

3 --- college / university

4 --- diploma/ degree/ M.Sc./ Ph.D

o --- social support at work

p --- job satisfaction in 4 categories as:

1 --- not at all satisfied

2 --- not too satisfied

3 --- somewhat satisfied

4 --- very satisfied

q --- decision latitude

r --- back pain

1 --- yes

2 --- no

s --- activity restriction (any long term disabilities or handicaps)

1 --- yes

2 --- no

t --- income in 4 categories

1 --- lower /lower middle income

2 --- middle income

3 --- upper middle income

4 --- high income

### 1.5 Methods and Models

Some statistical techniques such as logistic regression and log-linear models are used for data analysis. In this project, all explanatory variables in logistic regression models except n, p, t are treated as continuous variables; all variables when used in log-linear models are treated as categorical data. Results are compared between these different methods.

Since the response variable is binary, we cannot use the other well-known multivariate statistical methods such as ANOVA and multiple linear regression which are based on assumptions of normality and homogeneity of variance.

In Chapter 2, we address the following three questions:

(a) What are major factors likely associated with back pain?

(b) Is back pain related to the general work-stress index, where the general work-stress index is the sum of job stressors including psychological demands, job insecurity, physical exertion, decision latitude and the social support at work?

(c) What is the relationship between back pain and activity restriction, age, job satisfaction, and income?

Logistic regression method is used here to address these questions. We conclude that age has very high impact on back pain with significance level being lower than 1%; chronic stress, childhood and adult stressors all have high association with back pain; job stressor and recent life bad events are related fairly to back pain at significance level 5%; and income and job satisfaction do not have direct impact on back pain.

In Chapter 3, for the purpose of confirming the results pertaining to question (c) discussed in Chapter 2, we treat m, p, and t as categorical variables in order to study the association

between r and factors s, m, p, t via log-linear model method. Results similar to those presented in Chapter 2 are obtained.

### 1.6 Limitation of the Study

The data for this study was taken from a survey on health information of a probability sample of national population, the survey was not solely designed for back pain research. Explanatory variables that would provide the 'real' explanation of the differences between different groups of individuals may not have been measured and may even be unknown to the investigators. We suggest that some information such as cause of back pain, if known, anxiety and physical stress be included in the database for a better study.

# Chapter 2

## LOGISTIC REGRESSION

There are 19 variables in the data set as listed in Section 1.4. All of them are defined as continuous variables in this Chapter except r, s, n, p, t. We take "back pain" as the response variable. In this Chapter our aim is to determine: (a) What the major risk factors are to back pain, and (b) Whether the job stressors are related to back pain, where the job stressors include psychological demands, job insecurity, physical exertion, decision latitude and the social support at work. The general work-stress index is the sum of the job stressors. The other variables are considered to be potential confounders. The approach used for analysis is the logistic regression, which we present in the following form:

(i) we outline the basic theory of logistic regression, (ii) we try different models, and carry out goodness-of-fit and regression diagnostics for each model, and (iii) we choose one model as the "best model". From this best model, we perform some statistical analysis such as testing of hypothesis, calculating odds ratio and its 95% confidence interval, and then we interpret these results.

### 2.1 Logistic Regression

A commonly used generalized linear model is called logistic regression based on the binomial distribution. See Hosmer and Lemeshow (1989) for more references.

### 2.2 Model Selection and Diagnostics

In this section, we take r as the response variable.

For simplicity, we take *e* - general work stress index which is the sum of the job stressors (f, g, h, o, q) as the main exposure, and so we should force e into this model.

A proper variable-selection strategy should begin by screening out recorded variables that would be inappropriate candidates for control. In a study of the effect of an exposure on disease, it is recognized that variables influenced by the exposure or disease are inappropriate for control. Since control of such variables may lead to considerable bias, it is therefore essential to exclude such variables from the pool of candidates for control (Greenland, 1989).

Severe back pain may well cause activity restriction, and so s should be excluded from the candidates for control. From the18 variables excluding the job stressors f, g, h, o, q, and s, we select four explanatory variables b (chronic stress), c (recent life events), d (childhood and adult stressors), and m (age) as the risk factors, using logistic regression LR forward selection function in SPSS 6.1 for windows.

Now, the objective of the study is to construct a model that can be used to predict the value of the binary response variable r (back pain) on the basis of the above five (*e*, b, c, d, and m) explanatory variables.

Without considering interactions of the explanatory variables, there are 16 possible linear logistic models forcing e as the main exposure that could be used to fit this data set.

The likelihood, for each of the16 possible models, is given in Table 1 of Appendix. The smallest (-2logL) is that which includes all five variables in the model. The deletion of any confounder variable to this basic model increases (-2logL) by an amount large enough to be significant at 5%. The next step is to see if any interaction terms need to be incorporated in the model. Each of the ten second-order terms formed from products of the five explanatory variables is added, in turn, to the basic model. The reduction in

(-2logL) including each of these product terms, from its original value of 1465.23, is then calculated; *e*b, bc, bd, and cd reduce (-2logL) by relatively large amounts on 1 d.f., while no other interactions reduce (-2logL) by more than 1.62 on 1 d.f. Since the reduction in (-2logL) due to cd, adjusted for the other 3 interactions, is not significant at 5% (3.24 on 1 d.f.), we do not have to include cd in the model. To see if *e*b, bc, bd are needed, each is fitted after the other. The reduction in (-2logL) due to *e*b, bc, or bd adjusted for the other two interactions is significant at 5% level at least (see Table 1). Hence, *e*b, bc, and bd are retained in the model. To check that no other two-factor interaction term is needed, the 7 remaining interactions are added to the model with the five basic variables and the 3 important interaction terms. (-2logL) is reduced only by 6.545 on 7 d.f., which confirms that no other two-factor interactions need be included. As a result, the terms included in the model are now *e*, b, c, d, m, *e*b, bc, and bd.

In Figure 1 (see Appendix) of the plot of the influence diagnostics, Cook's distance calls our attention to case 341 and case 2001 with large values. In Figure 2 (see Appendix) of the leverage, values for case 341 and case 1843 are considerably higher than others. Thus, we have three extreme cases.

It is a good idea to fit the model without the three subjects. Since the reduction of (-2logL) equals 6.92, it suggests that the model fit the data better after the three subjects are deleted.

Furthermore, if all three-factor interaction terms are added to the above model the reduction of (-2logL) is only 12.258 on 10 d.f. (for N=2000), which is not significant at 5%. Therefore, no three-factor interaction terms are needed, and the terms in the "best model" for the data are *e*, b, c, d, m, *e*b, bc, and bd, up till now.

**Program listing and output (SPSS)**

a) program listing

```
LOGISTIC REGRESION r
    /METHOD=ENTER e b c d m b*e b*c b*d
    / SAVE PRED PGROUP COOK LEVER DEV
    /CLASSPLOT
    /PRINT=CORR
    /CRITERIA PIN(.05) POUT(.10) ITERATE(20).
```

b) output

Dependent variable – r

Variable(s) Entered:

*e* b c d m b*e b*c b*d

-2log likelihood 1448.376

Classification Table for r:

| | Predicted | | |
|---|---|---|---|
| **Observed** | 1.00 | 2.00 | **Percent Correct** |
| | | | |
| 1.00 | 157 | 96 | 62% |
| 2.00 | 655 | 1092 | 62.5% |
| **Overall** | 19% | 92% | 62.45% |

Table 2.1

Figure 3 indicates that we should take cutoff point around 0.8 for predicted values. We try different cutoff points such as 0.8, 0.87 and 0.9, comparing the classification tables for r in terms of pv+, pv-, sensitivity and specificity. Finally we take 0.87 as cutoff point, from Table 2.1 we see that the sensitivity and specificity are fairly good. But the pv+ is small, hence this analysis is at best

preliminary until more extensive data are available.

Variables in the Equation

| **Variable** | $\hat{B}$ | **S.E.** | **Wald** | **d.f.** | **Sig** |
|---|---|---|---|---|---|
| *e* | - .0616 | .0237 | 6.7804 | 1 | .0092 |
| b | - .3258 | .1045 | 9.7169 | 1 | .0018 |
| c | - .2993 | .1233 | 5.8970 | 1 | .0152 |
| d | - .2895 | .0944 | 9.4146 | 1 | .0022 |
| m | - .2856 | .0618 | 21.3775 | 1 | .0000 |
| b by *e* | .0093 | .0048 | 3.8099 | 1 | .0509 |
| b by c | .0297 | .0222 | 1.8006 | 1 | .1796 |
| b by d | .0236 | .0180 | 1.7185 | 1 | .1899 |
| Constant | 4.8260 | .5589 | 74.5573 | 1 | .0000 |

Table 2.2

## 2.3 Analysis and Results

From SPSS, we give r encoding as following:

| Value | Coding |
|---|---|
| 1 | 0 |
| 2 | 1 |

The final model can be expressed as:

Logit (p(r=1))= $B_0 + B_1 e + B_2 b + B_3 c + B_4 d + B_5 m + B_6 be + B_7 bc + B_8 bd$ .

A. $H_0$: $B_1 = B_6 = 0$

Under $H_0$, $(-2\log \hat{L}_0)$= 1455.369, but for the full model

$(-2\log \hat{L})$ = 1448.376.

Hence $(-2\log \hat{L}_0)$ - $(-2\log \hat{L})$ = 1455.369 – 1448.376 = 6.993 > 5.991 = $\chi^2_{.0.05}(2)$

$\therefore$ We reject $H_0$ at 5% level of significance, which implies that the effect of *e* for back pain is significant at 5%.

B. $H_0: \quad B_2 = B_6 = B_7 = B_8 = 0$

Under $H_0$, $(-2\log \hat{L}_0) = 1464.048$.

Hence, $(-2\log \hat{L}_0) - (-2\log \hat{L}) = 1464.048\text{-}1448.376 = 15.762 > 13.277 = \chi^2_{0.01}(4)$

$\therefore$ We reject $H_0$ at 1% level of significance, which implies that the effect of b for back pain is highly significant at 1%.

C. $H_0: \quad B_3 = B_7 = 0$

Under $H_0$, $(-2\log \hat{L}_0) = 1456.313$.

Hence, $(-2\log \hat{L}_0) - (-2\log \hat{L}) = 1456.313 - 1448.376 = 7.937 > 5.991 = \chi^2_{.05}(2)$

$\therefore$ We reject $H_0$ at 5% level of significance, which implies that the effect of c for back pain is significant at 5%.

D. $H_0: \quad B_4 = B_8 = 0$

Under $H_0$, $(-2\log \hat{L}_0) = 1462.411$.

Hence $(-2\log \hat{L}_0) - (-2\log \hat{L}) = 1462.411 - 1448.378 = 14.035 > 9.210 = \chi^2_{.01}(2)$

$\therefore$ We reject $H_0$ at 1% level of significance, which implies that the effect of d for back pain is highly significant at 1%.

E. $H_0: \quad B_6 = 0$

Under $H_0$, $(-2\log \hat{L}_0) = 1452.259$.

Hence $(-2\log \hat{L}_0) - (-2\log \hat{L}) = 1452.259 - 1448.376 = 3.883 > 3.84 = \chi^2_{.05}(1)$

$\therefore$ The effect of b*e* for back pain is significant at 5% level of significance.

F. $H_0$: $B_7 = 0$

Under $H_0$, $(-2\log \hat{L}_0) = 1450.202$ .

Hence ,

$(-2\log \hat{L}_0) - (-2\log \hat{L}) = 1450.202 - 1448.376 = 1.826 < 3.84 = \chi^2_{.05}(1)$

$\therefore$ We accept $H_0$ at 5% level of significance and conclude that effect of bc for back pain is not significant at 5%.

G. $H_0$: $B_8 = 0$

Under $H_0$, $(-2\log \hat{L}_0) = 1450.117$

Hence,

$(-2\log \hat{L}_0) - (-2\log \hat{L}) = 1450.117 - 1448.376 = 1.741 < 3.84 = \chi^2_{.05}(1)$

$\therefore$ We accept $H_0$ at 5% level of significance and conclude that effect of bd for back pain is not significant at 5%.

H. $H_0$: $B_5 = 0$

Under $H_0$, $(-2\log \hat{L}_0) = 1469.858$ . Hence,

$(-2\log \hat{L}_0) - (-2\log \hat{L}) = 1469.858 - 1448.376 = 21.482 > 7.879 = \chi^2_{.005}(1)$

$\therefore$ We reject $H_0$ at 0.5% level of significance and conclude that effect of m for back pain is very highly significant at 0.5%.

I. Since $\text{logit } p(r=1|m_2) - \text{logit } p(r=1|m_1) = B_5(m_2 - m_1)$,

We have $\log OR_{m_2 v m_1} = -B_5(m_2 - m_1)$,

where OR denotes odds-ratio.

Therefore, $\hat{OR}_{m_2 v m_1} = \exp\left(-\hat{B}_5(m_2 - m_1)\right) = \exp(0.2856(m_2 - m_1)) > 1$,

if $m_2 - m_1 \geq 1$.

A 95% CI for $OR_{m_2 v m_1}$ is then obtained as

$$\exp\left(-\hat{B}_5(m_2 - m_1) \pm 1.96(m_2 - m_1)se\left(\hat{B}_5\right)\right)$$

$$= \exp\left(-\hat{B}_5(m_2 - m_1) \pm 1.96(m_2 - m_1)\right) \times 0.0618$$

$$= \exp((0.2856 \pm 0.1211)(m_2 - m_1))$$

If $m_2 - m_1 = 1$,

$\hat{OR}_{m_2 v m_1} = \exp(0.2856) = 1.33$,

and a 95% CI for $OR_{m_2 v m_1}$ is obtained to be $(1.179, 1.501)$.

The above analysis shows that when an age group increases one level (e.g., from level 1 to 2), the odds for back pain increase by at least 17.9%, and by at most 50.1%.

Therefore we can conclude that age has a very high impact on back pain with significance level lower than 1%; chronic stress, childhood and adult stressors all have high association with back pain; job stressor and recent life events are related fairly to back pain at significance level 5%; and income and job satisfaction do not have direct impact on back pain.

The boxplots (see Figures 5-10) also suggest the above results to some extent. For example, the boxplots of back pain vs. age (Figure 6) indicate that the two variables have a strong relationship, at the different levels of back pain the structure of two boxplots are quite different. However, the boxplots of back pain vs. job satisfaction (Figure 9) indicate

that these two variables are not related, as the structure of these two boxplots is alike at the different levels of back pain.

# Chapter 3

## LOG-LINEAR MODEL

In Chapter 3, by treating m, p ,t as categorical variables, we discuss the same problem as considered in Chapter 2 using a different approach via log-linear model. Interestingly, the results obtained are in close agreement with the corresponding ones obtained in Chapter 2.

### 3.1 Log-linear Models

In this chapter a different approach will be considered, namely log-linear models. It is about the analysis of data in which the response and explanatory variables are all categorical, i.e. they are measured on nominal or possibly ordinal scales. For more reference see (Dobson, 1983 pp. 91-104).

### 3.2 Model Selection and Diagnostics

By prior information, we take 5 factors: r, s, m, p and t for the following study. In this Chapter we treat r, s, m, p, t as categorical variables.

1. We have a 5 - way table, far too large to examine all possible relationship models. So the method will be to screen for potentially important effects: to fit the model with these effects, and to add and delete terms to see if the model might be changed.

Since our concern is in the relationship between back pain and these other variables, we can look to see if it is possible to collapse over some variables and refit the model.

Use SPSS log-linear function to look for terms that show significant partial association.

Design 1 has generating class: m*p*r*s*t.

| Test that K-way and higher order effects are zero. | | | | | | |
|---|---|---|---|---|---|---|
| **K** | **DF** | **L.R. Chisq** | **Prob** | **Pearson Chisq** | **Prob** | **Iteration** |
| 5 | 36 | 14.813 | .9993 | 12.056 | .9999 | 4 |
| 4 | 141 | 106.012 | .9876 | 101.045 | .9955 | 6 |
| 3 | 253 | 248.767 | .5634 | 255.864 | .4378 | 5 |
| 2 | 307 | 724.888 | .0000 | 885.418 | .0000 | 2 |
| 1 | 319 | 5335.782 | .0000 | 11191.360 | .0000 | 0 |

Table 3.1

| Tests that K-way effects are zero. | | | | | | |
|---|---|---|---|---|---|---|
| **K** | **DF** | **L.R. Chisq** | **Prob** | **Pearson Chisq** | **Prob** | **Iteration** |
| 1 | 12 | 4710.893 | .0000 | 10305.942 | .0000 | 0 |
| 2 | 54 | 376.121 | .0000 | 629.555 | .0000 | 0 |
| 3 | 112 | 142.755 | .0264 | 154.819 | .0046 | 0 |
| 4 | 105 | 91.200 | .8292 | 88.989 | .8686 | 0 |
| 5 | 36 | 14.813 | .9993 | 12.056 | .9999 | 0 |

Table 3.2

It looks as though we will not need 4 - factors or higher order interactions.

| Tests of PARTIAL Associations. | | | | |
|---|---|---|---|---|
| **Effect Name** | **DF** | **Partial Chisq** | **Prob** | **Iteration** |
| m*p*r*s | 12 | 10.997 | .5292 | 4 |
| m*p*r*t | 36 | 22.413 | .9626 | 4 |
| m*p*s*t | 36 | 19.161 | .9903 | 5 |
| m*r*s*t | 12 | 16.365 | .1751 | 4 |
| p*r*s*t | 9 | 6.114 | .7285 | 5 |
| m*p*r | 12 | 21.325 | .0458 | 5 |
| m*p*s | 12 | 20.468 | .0587 | 5 |
| m*r*s | 4 | 2.930 | .5696 | 6 |
| p*r*s | 36 | 29.544 | .7679 | 5 |
| m*r*t | 12 | 19.181 | .0842 | 6 |
| p*r*t | 9 | 7.756 | .5589 | 4 |
| m*s*t | 12 | 11.192 | .5125 | 6 |
| p*s*t | 9 | 30.758 | .0003 | 4 |
| r*s*t | 3 | 2.446 | .4851 | 6 |
| m*p | 12 | 80.450 | .0000 | 4 |
| m*r | 4 | 16.717 | .0022 | 4 |
| p*r | 3 | 2.342 | .5046 | 4 |
| m*s | 4 | 12.104 | .0166 | 4 |
| p*s | 3 | 7.830 | .0497 | 4 |
| r*s | 1 | 143.370 | .0000 | 5 |
| m*t | 12 | 78.423 | .0000 | 5 |
| p*t | 9 | 13.417 | .1446 | 4 |
| r*t | 3 | 2.950 | .3994 | 5 |
| s*t | 3 | 5.712 | .1265 | 5 |
| m | 4 | 440.683 | .0000 | 5 |
| p | 3 | 1563.856 | .0000 | 2 |
| r | 1 | 1253.873 | .0000 | 2 |
| s | 1 | 1057.658 | .0000 | 2 |
| t | 3 | 394.819 | .0000 | 2 |

Table 3.3

By partial allocation table, the following effects are significant, mpr, pst, mp, mr, ps, rs, mt, m, p, r, s, and t. The terms we need at this stage are: pst, mpr, ms, rs, and mt. These include some lower order terms by the hierarchical nature of this model.

Let's fit this model:

| Model | DF | $\chi^2$ | P |
|---|---|---|---|
| pst, mpr, ms, rs, mt | 235 | 205.57 | .917 |

We can add terms in a stepwise manner.

Table below depicts the models formed by adding terms to model --- pst, mpr, ms, rs, mt.

| Model | DF | $\chi^2_{LR}$ | P |
|---|---|---|---|
| pst, mpr, ms, rs, mt, rt | 232 | 202.564 | 0.919 |
| + rt | 3 | 3.006 | 0.39 |
| pst, mpr, ms, mt, rst | 229 | 198.03 | 0.931 |
| + rst | 6 | 7.54 | 0.27 |
| pst, mpr, mst, rs, | 223 | 192.28 | |
| + mst | 12 | 13.29 | 0.35 |
| pst, mpr, prt, ms, rs, mt | 223 | 197.41 | 0.891 |
| + prt | 12 | 8.16 | 0.77 |
| pst, mpr, mrt, ms, rs | 220 | 184.05 | 0.963 |
| + mrt | 15 | 21.52 | 0.12 |
| pst, mpr, mpt, ms, rs | 199 | 176.562 | 0.872 |
| + mpt | 36 | 29 | 0.79 |
| pst, mpr, prs, ms, mt | 232 | 204.16 | 0.906 |
| + prs | 3 | 1.41 | 0.70 |
| pst, mpr, mrs, mt | 231 | 201.10 | 0.923 |
| + mrs | 4 | 4.47 | 0.35 |
| pst, mpr, mps, mt, rs | 223 | 187.32 | 0.961 |
| + mps | 12 | 18.25 | 0.11 |

Table 3.4

This has added, one at a time, all possible terms to the models. No one term is significant at 5% level.

**Step 1 The best model found is – pst, mpr, ms, mt, and rs.**

We can also try to delete terms. Table below depicts the models formed by deleting terms from model – pst, mpr, ms, mt, and rs.

| Model | DF | $\chi^2_{LR}$ | P |
|---|---|---|---|
| pst, mpr, mt, rs | 239 | 217.89 | 0.833 |
| - ms | 4 | 12.32 | 0.02 |
| pst, mpr, ms, rs | 247 | 282.77 | 0.058 |
| - mt | 12 | 77.2 | 0.00 |
| pst, mpr, ma, mt | 236 | 349.85 | 0.00 |
| - rs | 1 | 144.28 | 0.00 |
| pst, ms, mt, rs, mp, mr, pr | 247 | 224.50 | 0.845 |
| - mpr | 12 | 18.93 | 0.09 |
| mpr, ps, pt, st, ms, mt, rs | 244 | 232.49 | 0.691 |
| - pst | 9 | 26.92 | 0.00 |

Table 3.5

Term mpr is not significant, so we can drop it.

Step 1 best model found is – pst, ms, mt, rs, mp, mr, pr.

| Model | DF | $\chi^2_{LR}$ | P |
|---|---|---|---|
| pst, mt, rs, mp, mr, pr | 251 | 236.68 | 0.733 |
| - ms | 4 | 12.18 | 0.02 |
| pst, ms, rs, mp, mr, pr | 259 | 301.97 | 0.034 |
| - mt | 12 | 77.47 | 0.00 |
| pst, ms, mt, mp, mr, pr | 248 | 368.85 | 0.00 |
| - rs | 1 | 144.35 | 0.00 |
| pst, ms, mt, rs, mr, pr | 259 | 305.20 | 0.026 |
| - mp | 12 | 80.7 | 0.00 |
| pst, ms, mt, rs, mp, pr | 251 | 240.04 | 0.679 |
| - mr | 4 | 15.54 | 0.00 |
| pst, ms, mt, rs, mp, mr | 250 | 226.89 | 0.850 |
| - pr | 3 | 2.389 | 0.50 |
| ms, mt, rs, mp, mr, pr, ps, st | 256 | 251.71 | 0.564 |
| - pst | 9 | 27.21 | 0.00 |

Table 3.6

The term pr is not significant, so we drop it.

**Step 2 The best model found is -- pst, ms, mt, rs, mp, mr.**

Now since we are interested in relationships with back pain, we can overlap p and t, as there are no terms pr and tr in the model.

Using association = r*m*s", we can look at significant interactions by partial association in this reduced situation.

Tests that K-way and higher order effects are zero.

| K | DF | L.R. Chisq | Prob | Pearson Chisq | Prob | Iteratio |
|---|---|---|---|---|---|---|
| 3 | 4 | 4.619 | 0.3287 | 4.695 | 0.3201 | 4 |
| 2 | 13 | 175.727 | 0.0000 | 214.720 | 0.0000 | 2 |
| 1 | 19 | 2927.944 | 0.0000 | 4064.540 | 0.0000 | 0 |

Table 3.7

Tests that K-way effects are zero.

| K | DF | L.R. Chisq | Prob | Pearson Chisq | Prob | Iteration |
|---|---|---|---|---|---|---|
| 1 | 6 | 2752.217 | 0.0000 | 3849.820 | 0.0000 | 0 |
| 2 | 9 | 171.108 | 0.0000 | 210.025 | 0.0000 | 0 |
| 3 | 4 | 4.619 | 0.3287 | 4.695 | 0.3201 | 0 |

Table 3.8

It looks like we will not need the 3-factor interaction.

Tests of PARTIAL associations.

| Effect Name | DF | Partial Chisq | Prob | Iteration |
|---|---|---|---|---|
| m*r | 4 | 14.347 | 0.0063 | 2 |
| m*s | 4 | 12.966 | 0.0113 | 2 |
| r*s | 1 | 146.262 | 0.0000 | 2 |
| m | 4 | 440.684 | 0.0000 | 2 |
| r | 1 | 1253.874 | 0.0000 | 2 |
| s | 1 | 1057.695 | 0.0000 | 2 |

Table 3.9

All of the terms are significant by partial association. We can fit a model with them all, and

see if any terms can be eliminated.

**Model 1**

| Model | DF | Likelihood-Ratio Chisq | P |
|---|---|---|---|
| mr, ms, rs | 4 | 4.619 | 0.329 |

Models formed by eliminating terms from model – mr, ms, rs.

| Model | DF | $\chi^2_{LR}$ | P |
|---|---|---|---|
| mr, ms | 5 | 150.88 | 0 |
| -rs | 1 | 146.26 | 0 |
| mr, rs | 8 | 18.96 | 0.015 |
| -mr | 4 | 14.34 | 0.01 |
| mr, rs | 8 | 17.6 | 0.024 |
| ms | 4 | 12.98 | 0.01 |

Table 3.10

No terms can be deleted.

Finally we got the best model as follows – mr, ms, rs, which is the same as the best model got from SPSS 6.1 backward elimination algorithms starting initial model with generation class m*r*s.

### 3.3 Analysis and Results

The best model is as follows:

$$\ln F_{ijk} = c + m_i + r_j + s_k + (mr)_{ij} + (ms)_{ik} + (rs)_{jk} \qquad (3.1)$$

Program listing and output

a) program listing

```
GENLOG
   m r s
   /MODEL=POISSON
```

/PRINT ESTIM CORR COV

/PLOT NORMPROS( DEV )

/CRITERIA= CIN (95) ITERATE(20) CONVERGE (.001) DELTA(.5)

/ DESIGN m r s m*r m*s r*s

/SAVE PRED.

b) output

Goodness-of-fit statistics:

| Likelihood Ratio | Chi-square | DF | p |
|---|---|---|---|
| | 4.619 | 4 | 0.329 |

According to above statistics this model fits well. The p value is well above 0.05. The normal Q-Q plot of deviance residuals also indicates that the model is fairly good (see Figure 4 in appendix).

| Parameter | Estimate | SE | Z-Value |
|---|---|---|---|
| $(r)_1$ | -2.711 ✓ | .2533 | -9.36 |
| $(r)_2$ | 0 | | |
| $(mr)_{11}$ | -0.3901 | 0.3264 | -1.20 |
| $(mr)_{12}$ | 0 | | |
| $(mr)_{21}$ | -0.2524 | 0.2774 | -0.91 |
| $(mr)_{22}$ | 0 | | |
| $(mr)_{31}$ | -0.0010 | 0.2758 | -3.5x10-3 |
| $(mr)_{32}$ | 0 | | |
| $(mr)_{41}$ | -0.0010 | 0.2758 | -3.5x10-3 |
| $(mr)_{42}$ | 0 | | |
| $(mr)_{51}$ | 0 | 0.2856 | 1.46 |
| $(mr)_{52}$ | 0 | | |
| $(rs)_{11}$ | 1.8728 | 0.1500 | 12.49 |
| $(rs)_{12}$ | 0 | | |
| $(rs)_{21}$ | 0 | | |
| $(rs)_{22}$ | 0 | | |

Table 3.11

Variance Matrix of Parameter Estimates

| Parameter | $(mr)_{11}$ | $(mr)_{21}$ | $(mr)_{31}$ | $(mr)_{41}$ | $(rs)_{11}$ |
|---|---|---|---|---|---|
| $(mr)_{11}$ | 0.1065 | | | | |
| $(mr)_{21}$ | 0.0590 | 0.0769 | | | |
| $(mr)_{31}$ | 0.0591 | 0.0590 | 0.0761 | | |
| $(mr)_{41}$ | 0.0594 | 0.0592 | 0.0594 | 0.0816 | |
| $(rs)_{11}$ | 0.0032 | 0.0024 | 0.0032 | 0.0055 | 0.0225 |

Table 3.12

Since $\ln \frac{m_{11k}m_{22k}}{m_{12k}m_{21k}} = (mr)_{11} - (mr)_{12} - (mr)_{21} + (mr)_{22}$

From the estimated parameter table, we have $\ln \frac{m_{11k}m_{22k}}{m_{12k}m_{21k}} = (mr)_{11} - (mr)_{21}$

$= (-0.3901) - (-0.2524)$

$= -0.1377.$

Therefore, $\hat{OR}_{m_1 v m_2} = e^{0.1377} = 1.477$

$v((mr)_{11} - (mr)_{21}) = v((mr)_{11}) + v((mr)_{21}) - 2Cov((mr)_{11}, (mr)_{21})$

$$= 0.1065 + 0.0769 - 2x0.0590 = 0.0654$$

$Se((mr)_{11} - (mr)_{21}) = 0.2557$

95% CI for $\hat{OR}_{m_1 vm_2}$ is $e^{0.1377 \pm 1.96 \times 0.2557}$ or (0.69, 1.894).

Similarly, $\hat{OR}_{m_2 vm_3} = 1.28$

**Remark:** From the logistic model in Chapter 2, where all of explanatory variables are continuous, we have $\hat{OR}_{m_1\ vm_2} = \hat{OR}_{m_2\ vm_3} = 1.33$.

They are similar to the results obtained in the above log-linear models.

Now let's discuss the odds for back pain in favor of activity restriction.

$$\ln \frac{m_{i11} m_{i22}}{m_{i12} m_{i21}} = (rs)_{11} - (rs)_{12} - (rs)_{21} + (rs)_{22}$$

$$= (rs)_{11} - 0 - 0 + 0 = (rs)_{11} = 1.8728$$

$Se((rs)_{11} = 0.15$

$$\therefore \hat{OR}_{s_1 vs_2} = e^{1.8728} = 6.506$$

95% CI for $OR_{s_1 vs_2}$ is $e^{1.8728 \pm 1.96 \times Se((rs)_{11})}$ or (4.84, 8.73)

Thus we get the following result: the odds for back pain in favor of activity restriction is as 6.5 times as that of no activity restriction, even without adjusting age.

**Remark:** log-linear models are equivalent to logistic regression models with categorical variables ( Freeman, Jr. 1987, pp258-260).

Now logistic regression equivalent to the log-linear model (3.1) is as follows:

$$\log it(p(r=1)) = B_0 + B_1 m_{(1)} + B_2 m_{(2)} + B_3 m_{(3)} + B_4 m_{(4)} + B_5 s_{(1)}. \qquad (3.2)$$

| | Value | Coding | | | |
|---|---|---|---|---|---|
| m | | (1) | (2) | (3) | (4) |
| | 1 | 1 | 0 | 0 | 0 |
| | 2 | 0 | 1 | 0 | 0 |
| | 3 | 0 | 0 | 1 | 0 |
| | 4 | 0 | 0 | 0 | 1 |
| | 5 | 0 | 0 | 0 | 0 |
| S | 1 | 1 | | | |
| | 2 | 0 | | | |

Table 3.13

| Variable | B | S.E. | Wald | df | Sig | R | Exp(B) |
|---|---|---|---|---|---|---|---|
| M | | | 14.5973 | 4 | .0056 | .0659 | |
| M(1) | .3899 | .3264 | 1.4273 | 1 | .2322 | .0000 | 1.4769 |
| M(2) | .2524 | .2774 | .8278 | 1 | .3629 | .0000 | 1.2871 |
| M(3) | .0010 | .2758 | .0000 | 1 | .9972 | .0000 | 1.0010 |
| M(4) | -.4167 | .2856 | 2.1292 | 1 | .1445 | -.0092 | .6592 |
| S(1) | -1.8727 | .1499 | 155.9668 | 1 | .0000 | -.3184 | .1537 |
| Constant | 2.3710 | .2533 | 87.6118 | 1 | .0000 | | |

Table 3.14

Comparing Table 3.11 with Table 3.14, We can verify the equivalence of the two models.

# Chapter 4

## CONCLUSIONS AND RECOMMENDATIONS

Back pain is a leading cause of suffering, high medical cost, and loss of productivity in the workplace. The objective of this study is to identify predictors of back pain, and thus lay the groundwork for programs of prevention and control.

From the analysis and results in the chapters 2 and 3, we conclude that age has very high impact on back pain at significant level lower than 1%; when an age group goes up one level (e.g., from level 1 to 2), the odds for back pain increase at least 17.9% and at most 50.1%. The activity restriction also has a strong relationship with back pain. The odds for back pain in favor of activity restriction is about 6.5 times as large as that of no activity restriction, even without adjusting age. Chronic stress, childhood and adult stressors all have high association with back pain; job stressor and recent life events are fairly related to back pain at significant level 5%; and income, job satisfaction do not have direct impact on back pain.

Although there is not much that can be done to change the natural aging process of the spinal column, some of the predictors identified such as job stressors are amenable to change.

# References


Dobson, A. J. (1983). *Introduction to Statistical Modeling*. London: Chapman and Hall.

Freeman, D.H., Jr. (1987). Applied Categorical Data Analysis, New York: Marcel Dekker, Inc.

Greenland, S. (1989). Modeling and Variable Selection in Epidemiologic Analysis *AM J Public Health*. **79**,340-349.

Hosmer, D. W., Lemeshow, S (1989). Applied Logistic Regression, New York: John Wiley & Sons, Inc.

Hosmer, D. W., Lemeshow, S (1991). The Importance of Assessing the Fit of Logistic Regression Models: A Case Study. *Am J Public Health*. **81**,1630-1635.

Nagi, S. Z., Riley, L. E. and Newby, L. G. (1973). A Social Epidemiology of Back Pain in A General Population. *J Chron Dis*, **26**,769-779.

Pregibon, D. (1981). Logistic Regression Diagnostics. *Ann Statistics*. **9**,704-724.

## APPENDIX

| Model | -2LOGLIKELIHOOD |
|---|---|
| e | 1519.79 |
| e+b | 1502.95 |
| e+c | 1502.91 |
| e+d | 1499.38 |
| e+m | 1505.4 |
| e+b+c | 1494.4 |
| e+b+d | 1491.21 |
| e+b+m | 1486.82 |
| e+c+d | 1489.57 |
| e+c+m | 1486.92 |
| e+d+m | 1480.31 |
| e+b+c+d | 1485.3 |
| e+b+c+m | 1477.38 |
| e+b+d+m | 1471.62 |
| e+c+d+m | 1469.76 |
| e+b+c+d+m | 1465.23 |
| e+b+c+d+m+eb | 1458.71 |
| e+b+c+d+m+ec | 1464.58 |
| e+b+c+d+m+ed | 1463.61 |
| e+b+c+d+m+em | 1464.97 |
| e+b+c+d+m+bc | 1461.40 |
| e+b+c+d+m+bd | 1461.57 |
| e+b+c+d+m+bm | 1463.86 |
| e+b+c+d+m+cd | 1458.96 |
| e+b+c+d+m+cm | 1464.51 |
| e+b+c+d+m+dm | 1464.03 |
| e+b+c+d+m+bc+bd | 1459.32 |
| e+b+c+d+m+eb+bd | 1460.87 |
| e+b+c+d+m+eb+bc | 1460.11 |
| e+b+c+d+m+eb+bc+bd | 1455.30 |
| e+b+c+d+m+eb+bc+bd+cd | 1452.06 |
| e+b+c+d+m+all two-order interaction | 1448.75 |

Table 1

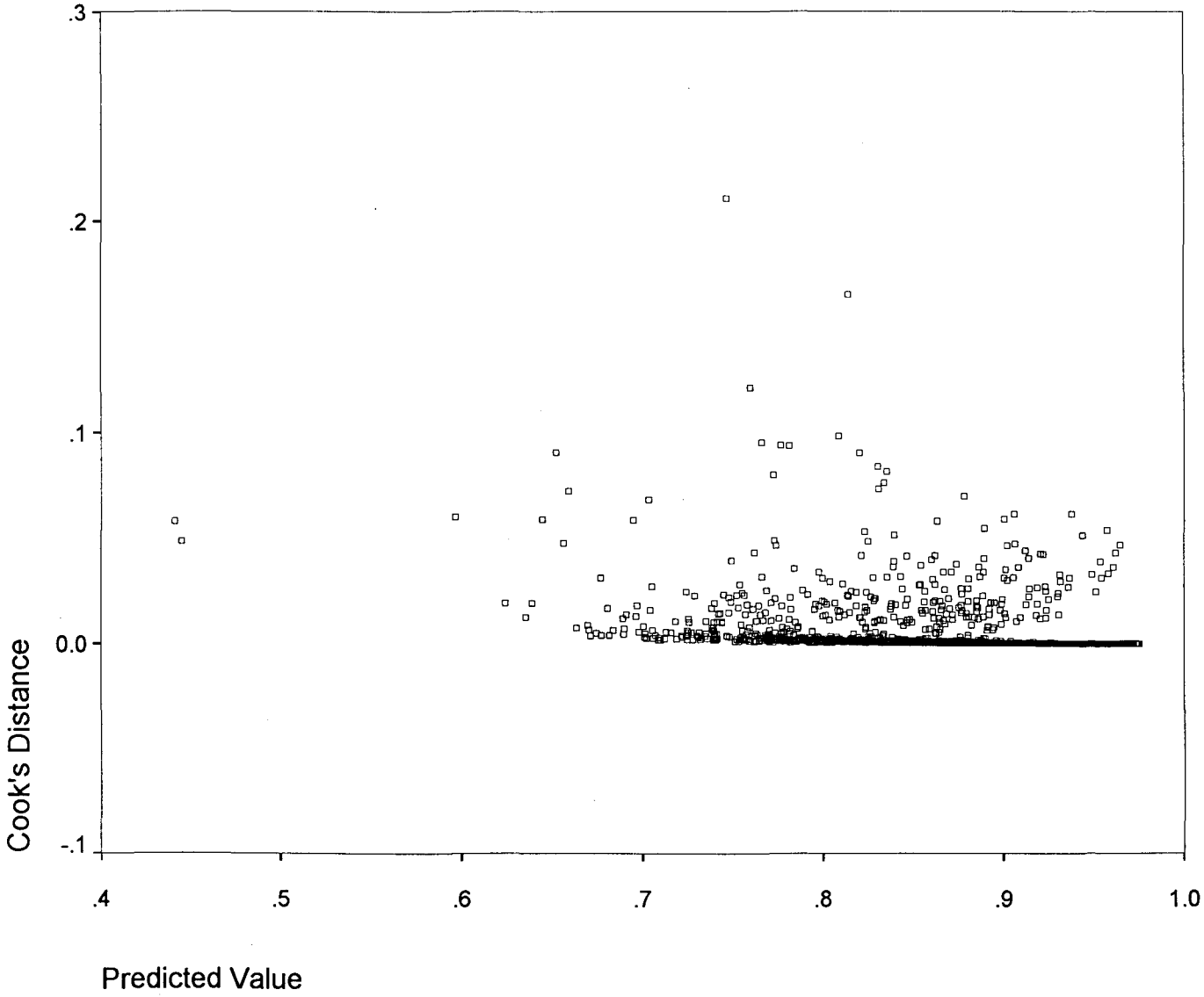


Figure 1

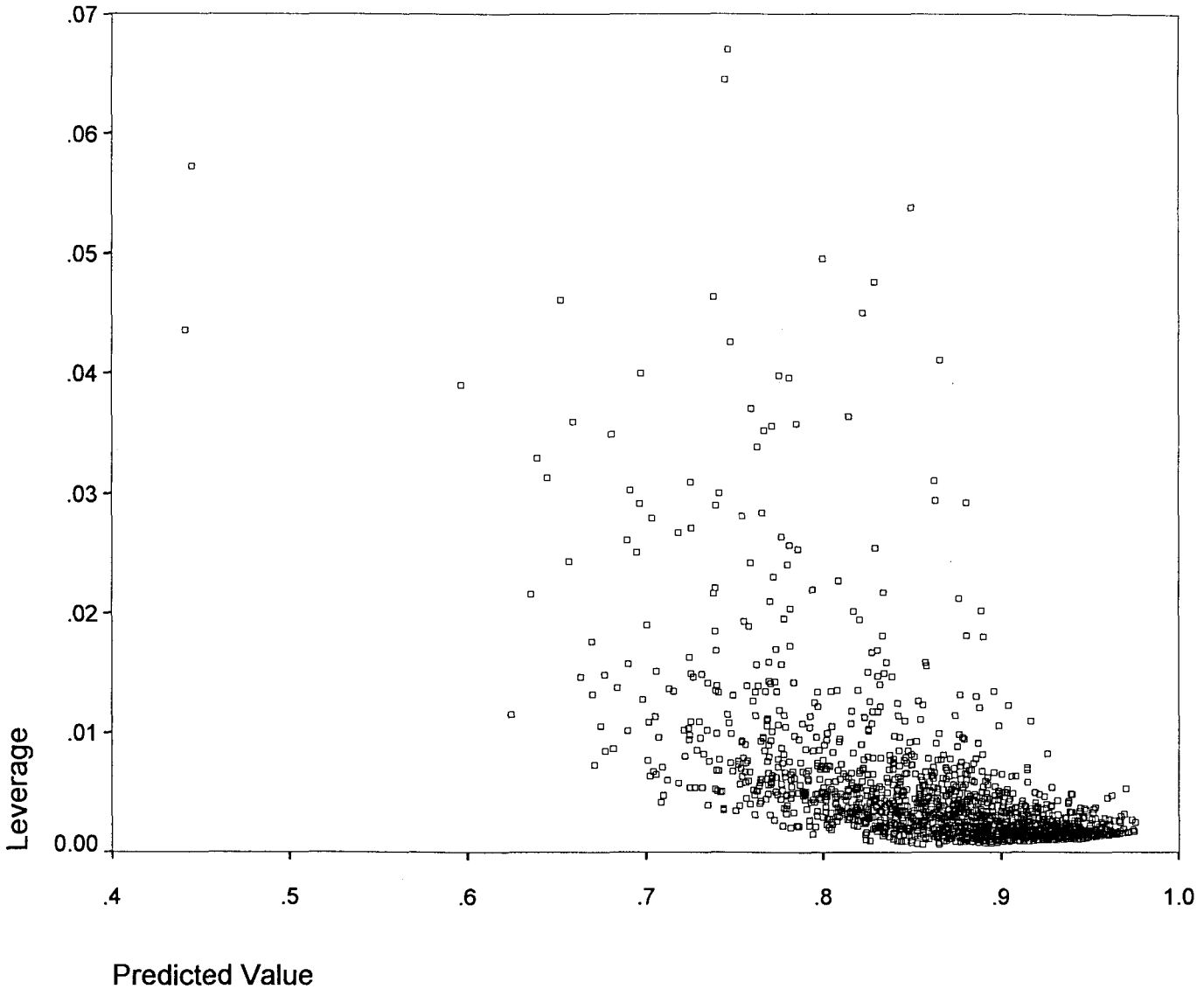


Figure 2

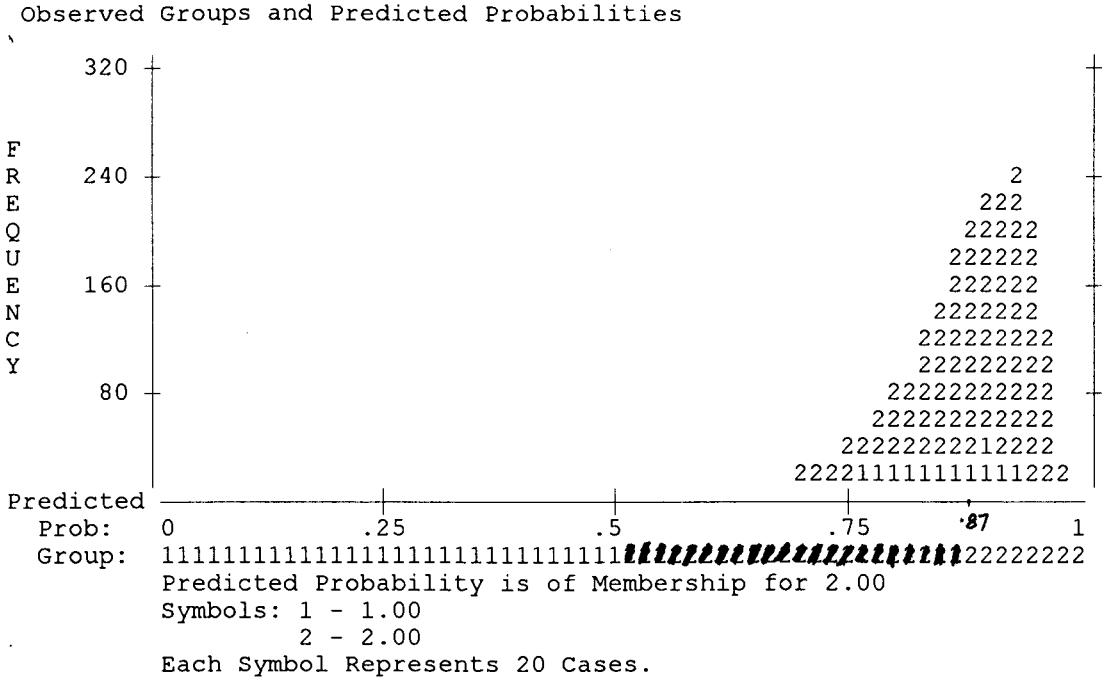
Observed Groups and Predicted Probabilities
FREQUENCY
320
240
160
80
2
222
22222
222222
222222
2222222
222222222
222222222
22222222222
222222222222
22222222212222
222211111111111222
Predicted Prob:
0
.25
.5
.75
.87
1
Group:
11111111111111111111111111111122222222
Predicted Probability is of Membership for 2.00
Symbols: 1 - 1.00
2 - 2.00
Each Symbol Represents 20 Cases.

Figure 3

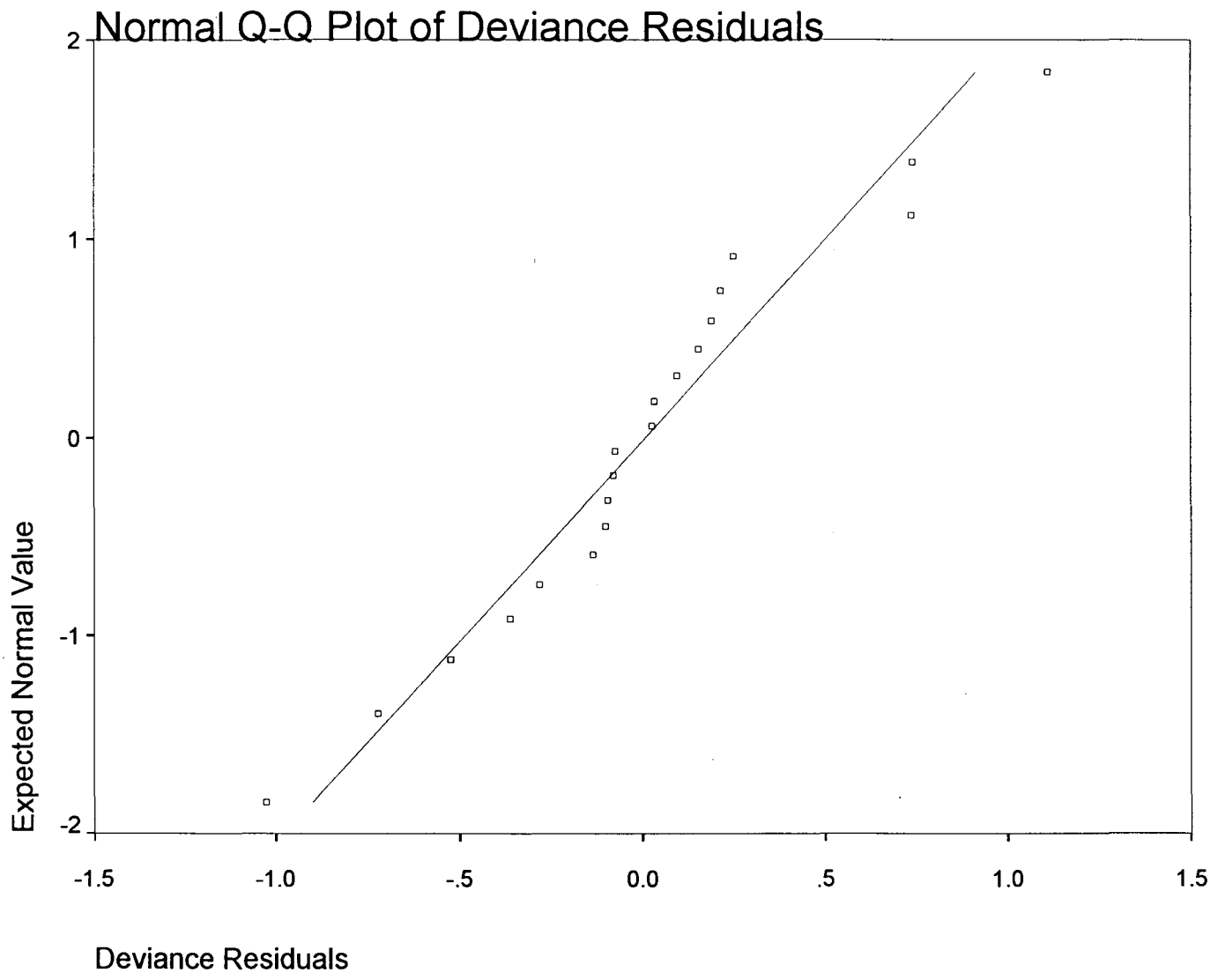
Normal Q-Q Plot of Deviance Residuals
Expected Normal Value
2
1
0
-1
-2
-1.5
-1.0
-.5
0.0
.5
1.0
1.5
Deviance Residuals

Figure 4

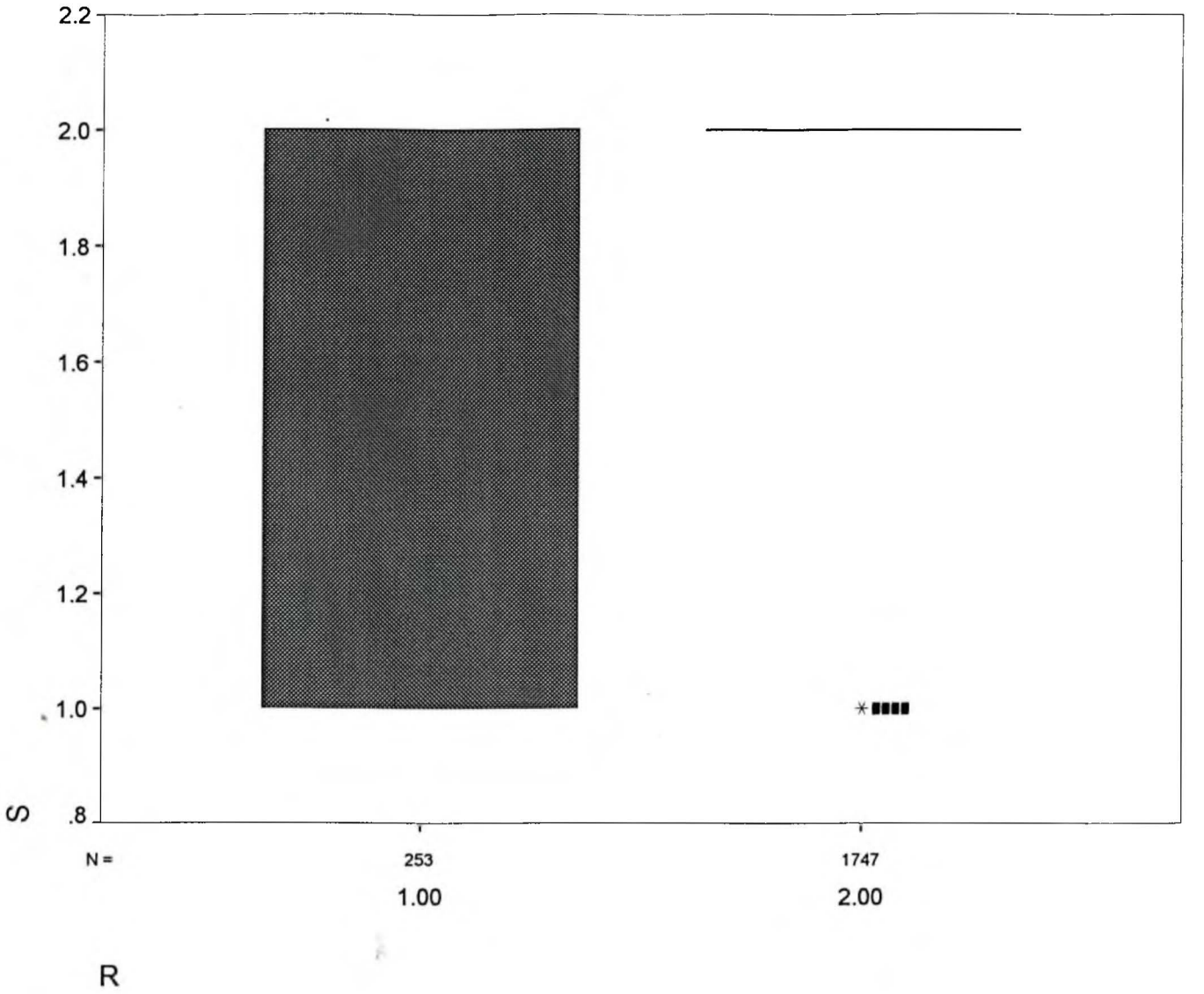


Figure 5

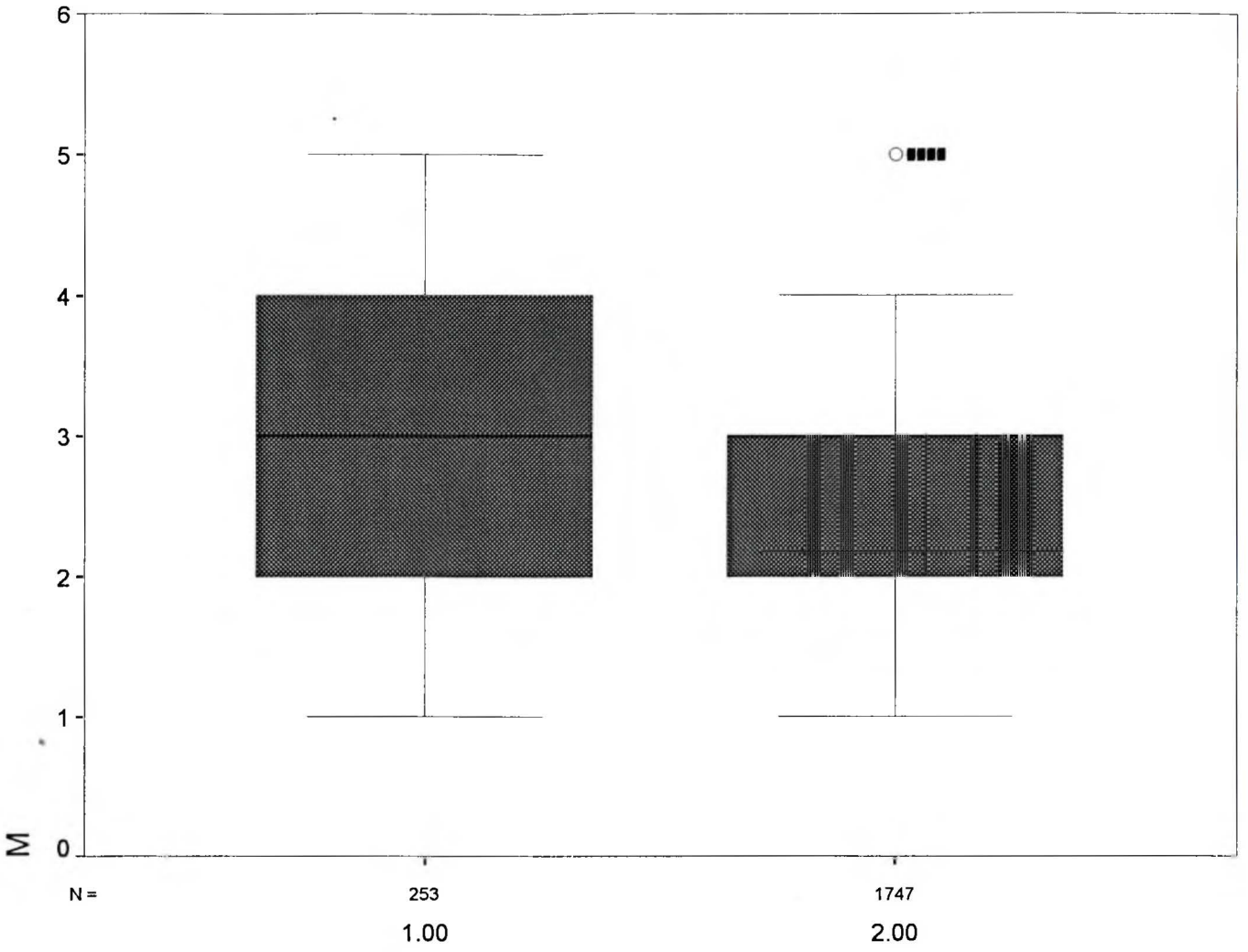


Figure 6

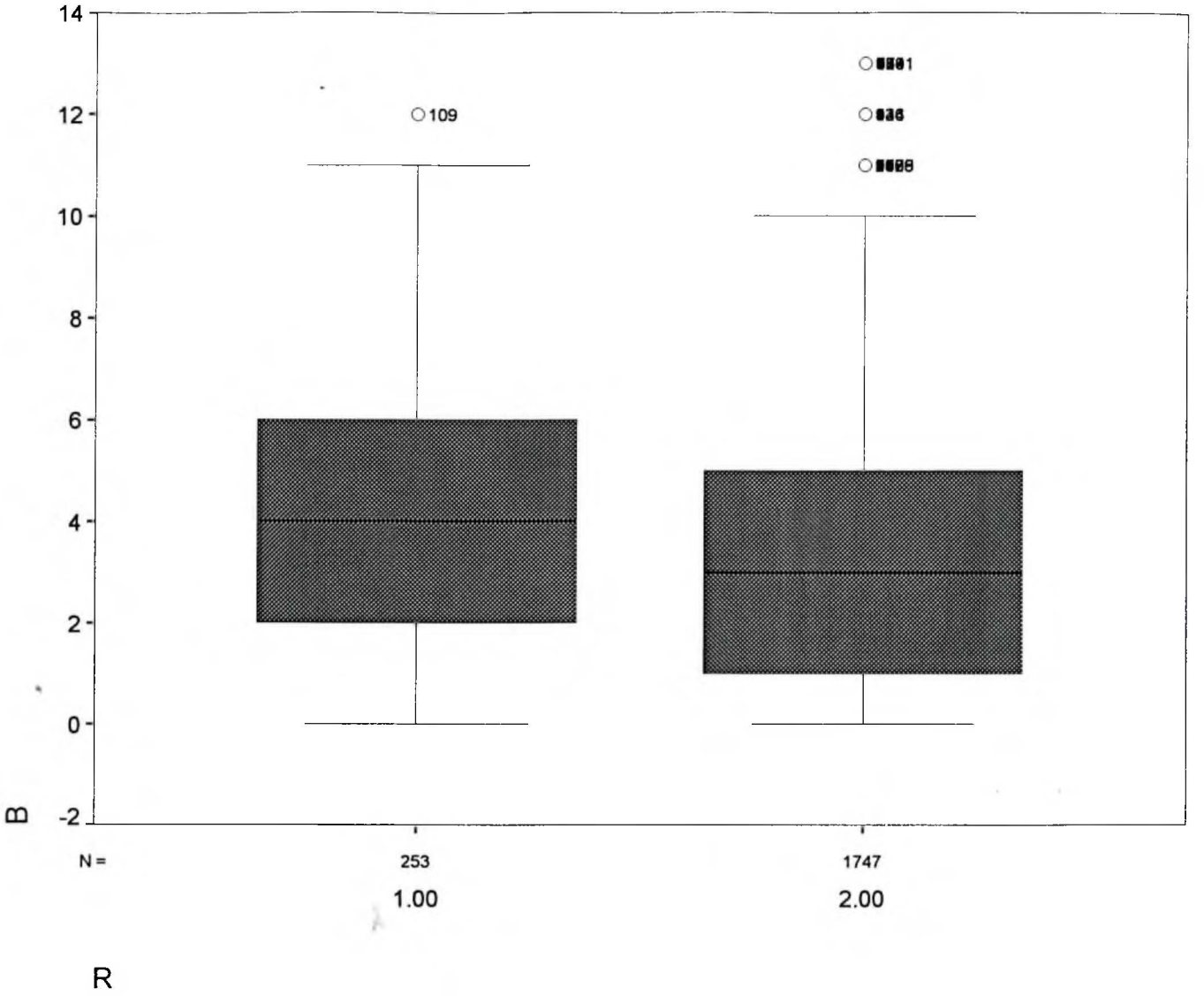


Figure 7

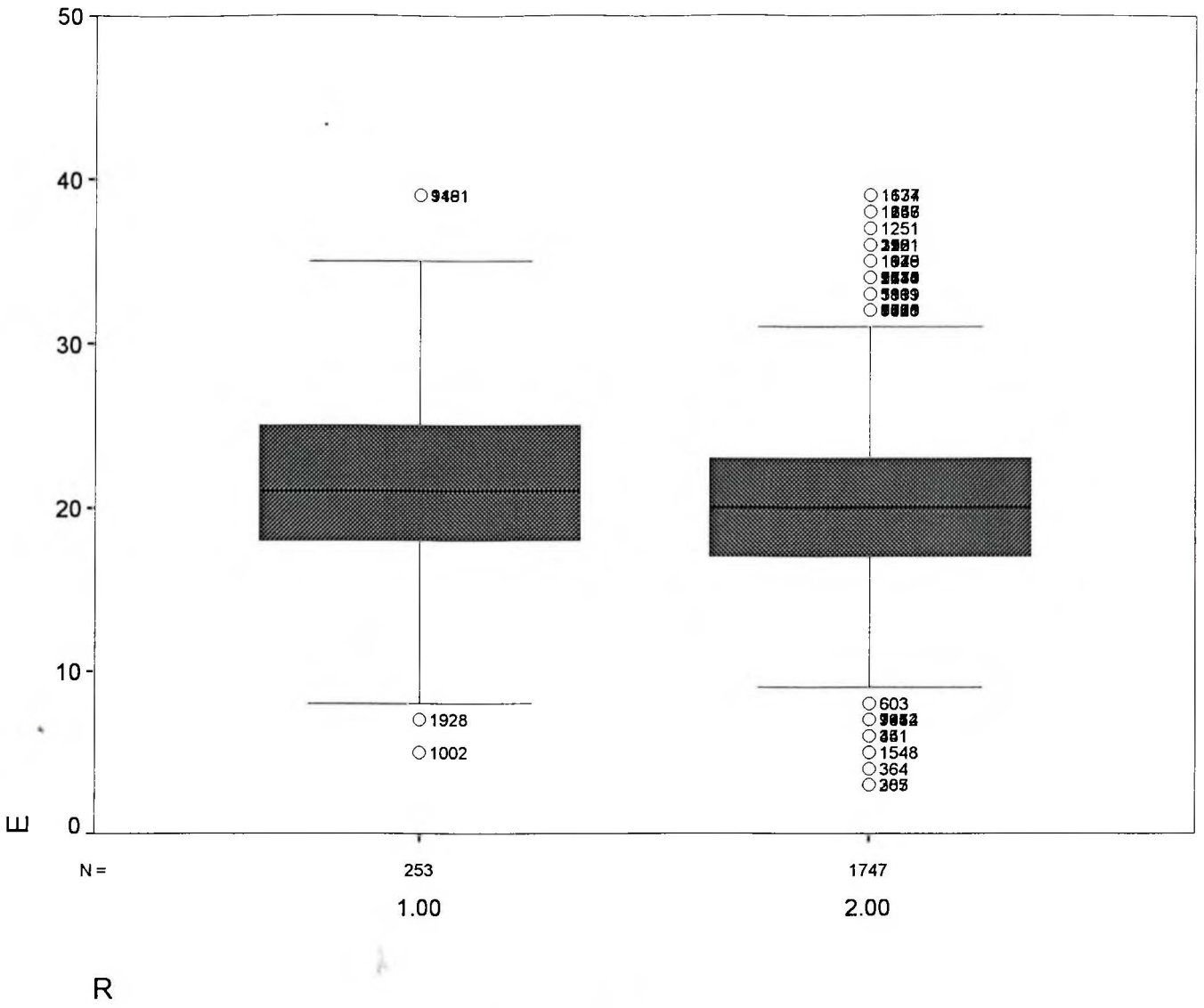
50
40
30
20
10
0
E
9481
1928
1002
1251
603
1548
364
N =
253
1747
1.00
2.00
R

Figure 8

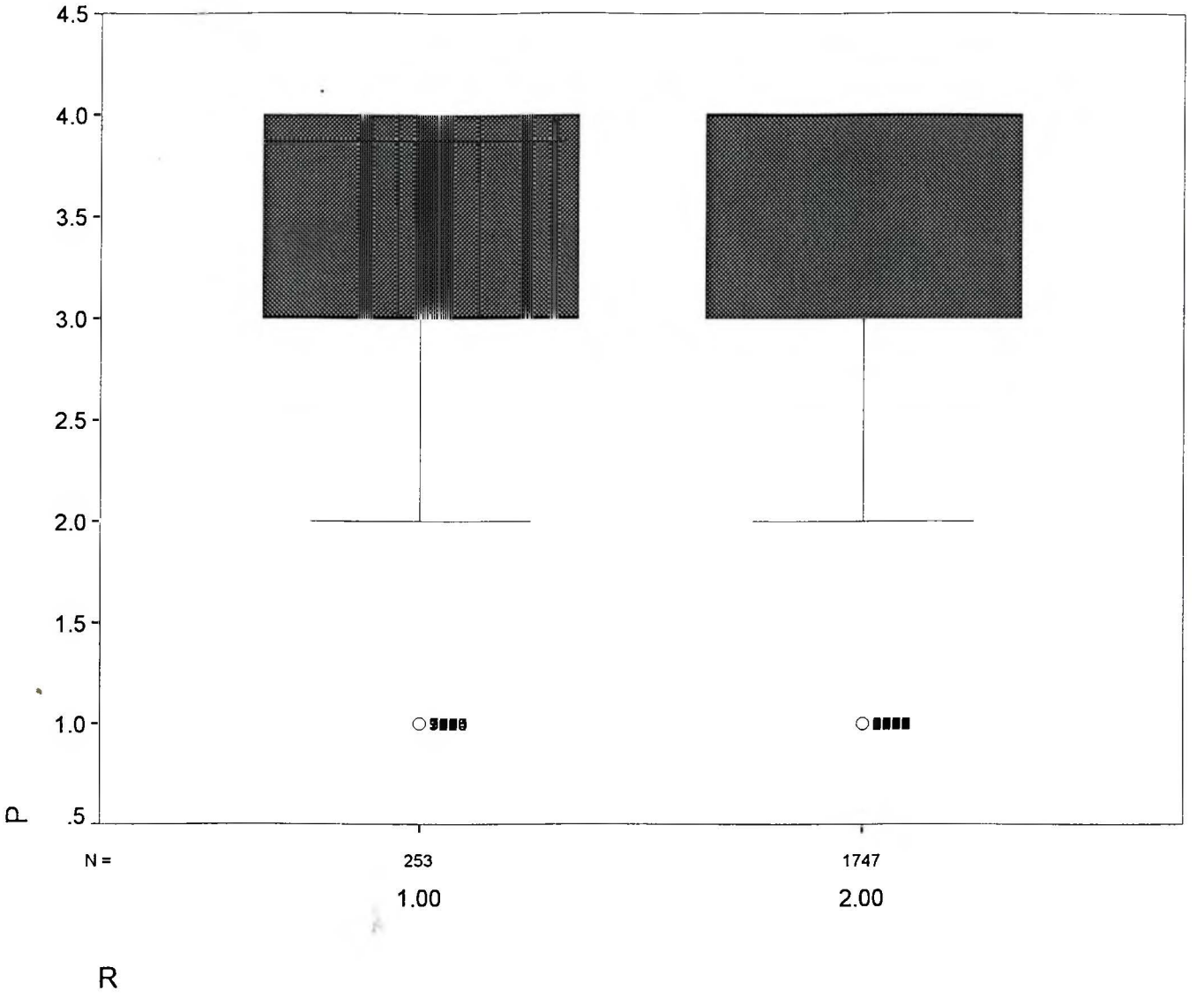


Figure 9

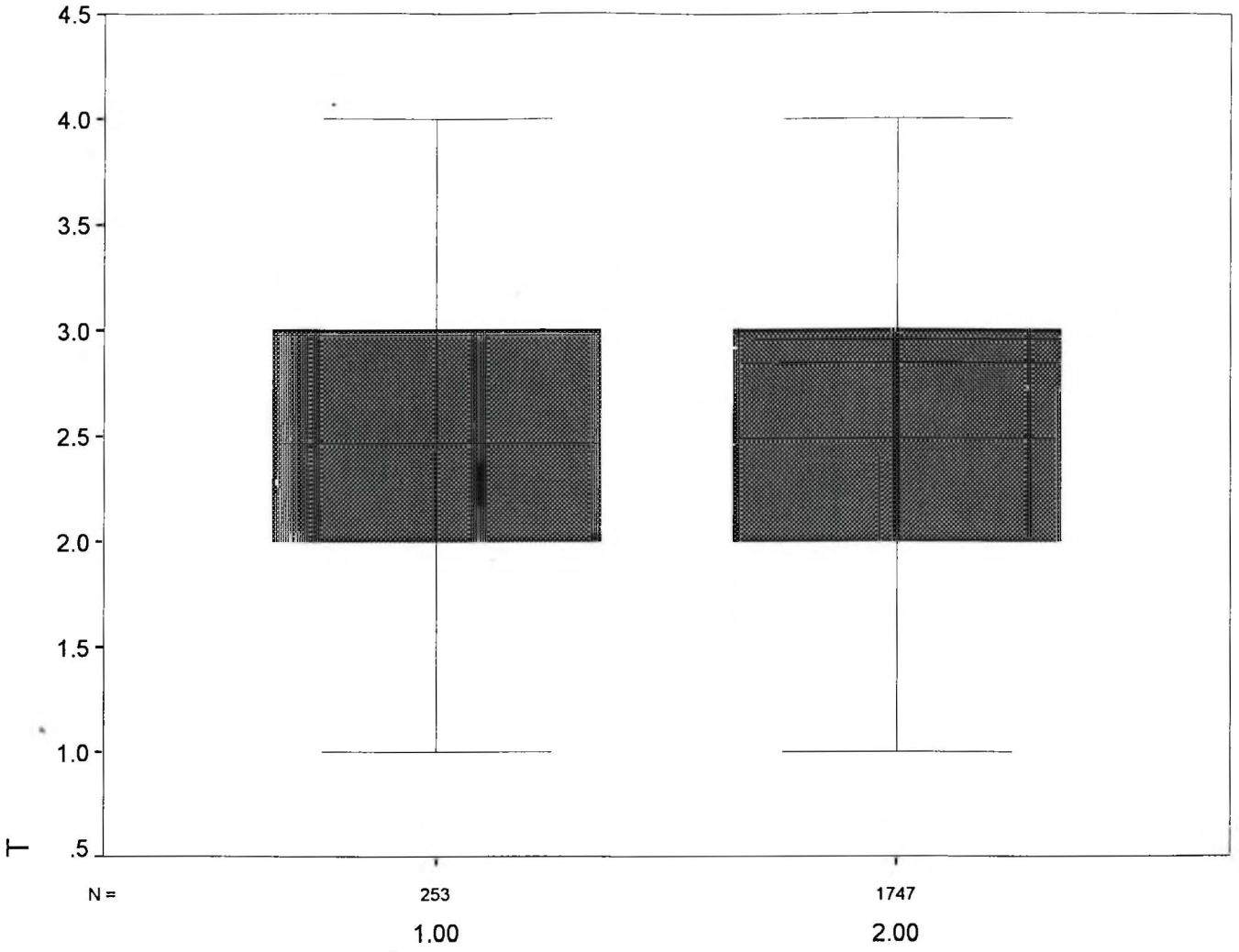


Figure 10