A SIMULATION STUDY OF THE EFFECT OF THERAPEUTIC HORSEBACK RIDING - A LOGISTIC REGRESSION APPROACH

by

JEANETTE PAUW

submitted in part fulfilment of the requirements for
the degree of

MASTER OF SCIENCE

in the subject

STATISTICS

at the

UNIVERSITY OF SOUTH AFRICA

SUPERVISOR: PROF V S S YADAVALLI

JOINT SUPERVISOR: PROF F E STEFFENS

NOVEMBER 1998

DECLARATION

Student number: **3144-434-2**

I declare that **A SIMULATION STUDY OF THE EFFECT OF THERAPEUTIC HORSEBACK RIDING - A LOGISTIC REGRESSION APPROACH** is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.

SIGNATURE (Mrs J Pauw)..... Jeanette Pauw

DATE..... 3/3/1999

ACKNOWLEDGEMENTS

I wish to thank the following persons and institutions for their contribution to this dissertation:

UNISA, for financial support to attend the 3rd European Congress of Therapeutic Riding, Munich, September 1998.

My supervisor, Prof Y. S. S. Yadavalli, for his advice and guidance. His support during the preparations for attendance of the 3rd European Congress of Therapeutic Riding, is also appreciated. Professor F. E. Steffens, for his valuable contributions and advice.

My parents for their support and encouragement, especially during the uncertain early stages of the dissertation.

My husband Christiaan, for believing in me and for continued enthusiasm, love and patience.

ABSTRACT[1]

Therapeutic horseback riding (THR) uses the horse as a therapeutic apparatus in physical and psychological therapy. This dissertation suggests a more appropriate technique for measuring the effect of THR. A research survey of the statistical methods used to determine the effect of THR was undertaken. Although researchers observed clinically meaningful change in several of the studies, this was not supported by statistical tests. A logistic regression approach is proposed as a solution to many of the problems experienced by researchers on THR. Since large THR related data sets are not available, data were simulated. Logistic regression and t-tests were used to analyse the same simulated data sets, and the results were compared. The advantages of the logistic regression approach are discussed. This statistical technique can be applied in any field where the therapeutic value of an intervention has to be proven scientifically.

KEYWORDS

change, clinically meaningful, logistic regression, odds ratio, power, qualitative studies, quantitative studies, simulations, statistically significant, therapeutic horseback riding.

[1]The modified version of this dissertation has been presented at the South African Statistical Association (SASA) conference (November 1998).

TABLE OF CONTENTS

SIMULATION 519.536. Pauw

STATUS: COND/PEND 20071015 **OCLC #:**
REQUEST DATE: 20071012 **NEED BEFORE:** 20071115 **SOURCE:** FSISOILL
BORROWER: JNA **RECEIVE DATE:** **DUE DATE:**
RENEWAL REQ: **NEW DUE DATE:** **SPCL MES:**
LENDERS: *OL$, OL$, OL$, OL$, OL$

CALL NUMBER:
AUTHOR: Pauw, Jeanette, M.S
TITLE: A simulation study of the effect of therapeutic horseback riding. A logistic regression approach
IMPRINT: Place of Publication:
DISSERTATION: University of South Africa (South Africa) 1999
VERIFIED: proquest
SHIP TO: Information Delivery Services/Northern Illinois University Libraries/DeKalb, IL 60115
BILL TO: same
SHIP VIA: Ariel, Fax, Mail, ILDS RT1 NI
MAXCOST: IFM - $35 **COPYRIGHT:** **AFFILIATION:** LVIS, ILLINET
FAX: (815) 753-2003 Ariel 131.156.159.48 **EMAIL:** illdept@niu.edu
LEND CHARGES: **SHIPPED DATE:** **SHIP INSURANCE:**
LEND NOTES: Can supply for $35 IFM. Please change max cost if ok.
PATRON: button, sharon

~~1727985~~
1727986

CHAPTER 3

CHAPTER 1

AN INTRODUCTION TO THERAPEUTIC HORSEBACK RIDING (THR)

1.1 The History of Therapeutic Horseback Riding

A Dane, Lis Hartl, won a silver medal in dressage at the 1952 Olympics. What made her achievement extraordinary, was that she had had severe poliomyelitis in the 1940s. She showed courage and endurance in satisfying the very high standards of the Olympics in a sport of perfection, her handicap notwithstanding. After this event, international attention was focused on the therapeutic use of the horse.

In 1964 a Norwegian, Elsbet Bodtker, started to use ponies in her treatment of children with polio and cerebral palsy. Today her pioneer work is seen by many as the beginning of modern-day riding therapy. The Norwegian *Rikstrygdoverket* started to support riding therapy financially in 1971, with the acceptance of this type of therapy as a physiotherapeutic treatment method.

The continued work in Norway went hand in hand with the development of therapeutic riding in other European countries, USA and Canada.

In 1968, in the former German Democratic Republic (GDR), the University Orthopaedic Hospital in Halle started developing methods of therapeutic riding for treating diseases of the musculoskeletal system. In 1971, the GDR's Department of Health accepted therapeutic riding as an offshoot of the physiotherapeutic tree and therefore a reimbursable treatment.

In 1970, the *Kuratorium für Therapeutisches Reiten* was founded in the Federal Republic of Germany. Similar activity was taking place in Switzerland. The *Schweizer Gruppe für Hippotherapie* (Swiss Group for Hippotherapy) was founded in 1976, and the

Stiftung Hippotherapie Zentrum (Hippotherapy Foundation Centre) established in Basel in 1979. The *Austrian Kuratorium für Hippotherapie* was founded in 1977.

NARHA (North American Riding for the Handicapped Association) was founded in 1969 with the purpose of supporting and promoting therapeutic riding in the USA and Canada.

Since the early development in the 1960s, there has been no barrier too challenging for THR (Therapeutic Horseback Riding) enthusiasts. Many countries have developed therapeutic riding programs and the number of THR centres and members is increasing. The first international congress of the *Federation Riding for the Disabled International* was held in Paris in 1974. Several congresses followed: Basel in 1976, Warwick in 1979, Hamburg in 1982, Milan in 1985, Toronto in 1988, Aarhus in 1991, Hamilton in 1994, and the most recent, Denver in 1997. No less than 31 countries were represented at the congress in Denver. The next congress is planned for the year 2000 in Paris.

As a result of the congress in New Zealand, the *Scientific Journal of Therapeutic Riding* saw the light for the first time in 1995.

Recent awareness of the benefits of THR in South Africa, is evident in the publication of several articles on the subject in popular newspapers and magazines (*Rooi Rose*, May 1996; *Sarie*, February 1993; *Rekord-Oos*, November 1995). The South African Riding for the Disabled Association (SARDA) started its first lessons in 1973. SARDA has branches in Cape Town, Port Elizabeth, Durban and Krugersdorp.

THR organisations all over the world are striving to be accredited by national health departments and the resultant reimbursement by social and health insurance carriers. The battle has been won, or at least partly won, in a few countries. In working towards this goal, higher standards are being set for THR, thus distinguishing therapeutic riding as a specialised activity, conducted by professionals. Brown and Tebay (1997) conducted a survey on progress in the education of THR professionals around the world. Twenty countries were represented in this survey. The results showed that, in contrast to only a few years ago, increasingly more THR training courses are in some or other way being associated with universities and colleges.

Therapy with the use of a horse does not appear to be such a new concept. Riede (1988) names several philosophers and physicians, from as early as the 16th century, who mentioned the benefits of riding in promoting and maintaining general health. The benefits of riding were realised centuries ago, but only now are researchers working towards proving these benefits scientifically.

1.2 The Subdivisions of Therapeutic Horseback Riding

Therapy with the use of a horse can be divided into categories. Often the distinctions between categories are not clear, and in fact, can be viewed as a continuum rather than mutually exclusive islands. It is important to note, however, that in dealing with insurance carriers, allied medical and clinical professionals accept only some of the categories.

In **hippotherapy** the patient is passive. A specially trained physiotherapist uses the three-dimensional movement of the horse's back as an "apparatus" to manipulate the patient's body. Hippotherapy is accredited by medical professionals and is prescribed by physicians in some countries.

Remedial riding and vaulting can be conducted by educators, psychologists and psychotherapists. In this category, cognition plays an important role and people with emotional disturbances are treated in this way. Motivation, self-esteem and social skills are often associated with this category.

In **riding for the disabled,** the rider is actively involved in manipulating the horse. The rider works towards the goal of independent riding. This can result in riders participating in competitive sport. Riding for the disabled can take place under the supervision of a qualified riding instructor for handicapped individuals.

Henceforth, in this study, the term "therapeutic horseback riding" will refer to any of the above-mentioned categories.

1.3 The Effects of Riding

The numerous benefits derived from horse riding by handicapped people are described in many articles, reports and books (Riede, 1988; Heipertz, 1989; Biery, 1985; Freeman, 1984; Von Arbin, 1994a). The testimonies of parents, therapists and patients also bear witness to the many benefits.

In 1735, Quellmalz (Riede, 1988) was the first to describe the three-dimensional movement of the horse's back. Researchers are still actively investigating the effect on the rider, of the three-dimensional movement of different horses at different gaits. Riede (1988) documented and analysed these movements in a study also included in his book, *Physiotherapy on the horse.* These simultaneous movements in three dimensions simulate the human gait. As a result, the hemiplegic or paraplegic experiences impulses similar to the demands of walking. The horse, as a living apparatus, constantly demands from the rider adaptive responses to a variety of movements. As a result, riders improve their coordination and balance, and this encourages a better posture. Therapeutic riding is extremely effective in normalisation of muscle tone and is also used to increase joint mobility (Riede, 1988).

Despite the physiological effects, patients also benefit psychologically. The mysterious relationship between human and horse is to many patients a new and enriching experience.

Increased motivation and a better self-esteem are often observed (Von Arbin, 1994a). If a wheelchair-bound paraplegic becomes ambulant astride a horse, one can expect this to be a "spirit lifting" experience! Patients also have the opportunity to socialise with instructors and fellow-riders in a stimulating environment (Koch, 1994; Von Arbin, 1994a; Von Arbin, 1994b). Better concentration and academic performance are sometimes attributed to riding therapy as well as the alleviation of pain and regulation of sleep disturbances (Koch, 1994; Exner et al, 1994). Therapy on a horse can be seen as a preparation for other treatments or an adjunct therapeutic treatment (Koch, 1994; Künzle et al, 1994).

However, there are also contra-indications for riding, and people conducting THR should be aware of them (Heipertz, 1989).

CHAPTER 2

A RESEARCH SURVEY OF THE STATISTICAL METHODS USED IN THR RESEARCH

2.1 Statistical Methods Used in Previous THR Research

A research survey on therapeutic horseback riding showed that the authors of articles on THR were mainly physiotherapists, occupational therapists and psychologists. Special interest was paid to the techniques used to analyse and summarise the effect of therapeutic horseback riding.

(1) Bertoti (1988) conducted a study on 11 children (aged 2 - 9 years) diagnosed with spastic cerebral palsy (CP). The repeated measurement design employed in the Bertoti study consisted of pretest_1, followed by 10 weeks of no riding, pretest_2 followed by a therapeutic riding program of 10 weeks and a post-test. The riding program was conducted twice weekly for one hour. A scale for the assessment of posture in CP children, was developed and used by Bertoti. Each child was evaluated simultaneously, and without verbal discussion, by three pediatric physical therapists. A composite score for each child was calculated at each testing interval, by adding the three scores of the three therapists. The nonparametric Friedman test was used to analyse the data. The results showed a significant improvement (at the 5% level of significance) during the 10 weeks of therapeutic riding. Subjective clinical improvement in self-confidence, muscle tone, weight bearing and sitting balance were noted by the author, the clinical therapists by whom the children were referred and the children's parents.

(2) MacKay-Lyons et al (1988) studied the effects of therapeutic riding on 10 multiple sclerosis (MS) patients who participated in a 9 week, twice weekly riding

program. The age of patients ranged from 25 to 54. Participants were assessed during the week preceding the riding program and in the first week after the termination of the program. Paired t-tests were used to compare the pre- and post-treatment data. Patients were assessed at different levels of the Minimal Record of Disability (MRD) for MS, different dimensions of psychological well-being and several measures of postural sway and gait. Not all the patients were able to complete all the tests. Significant increases were revealed for the relative speed of free speed walking and the average stride length of free speed walking. Also, the psychopathology dimensions of depression and global severity reflected a significant decrease.

(3) The study conducted by Biery and Kauffman (1989) had a similar design to the Bertoti study. The sample consisted of 8 mentally retarded people (aged 12-22 years). The subjects were tested on balance using test items developed by Cratty (1967). After the subjects were tested, six months passed with no intervention whereafter the subjects were again tested on balance using the same procedure followed in the initial stage. The subjects were then subjected to a 6-months, once weekly, therapeutic riding program and on the termination of the program balance was again tested. The non-parametric Wilcoxon matched-pairs signed ranks test was initially used on the data. Paired t-tests were also conducted and yielded the same results as the nonparametric procedure. The results showed a significant improvement in both standing balance and quadruped balance after the 6 months period of riding.

(4) Scheidhacker (1991) selected a control group and an experimental group from the behavioural therapeutic ward in the psychiatric hospital in Haar / Munich. The experimental group consisted of 16 patients who participated in a therapeutic riding program of eight weeks. The control group consisted of 8 patients. Both groups were tested before and after 8 weeks on psychopathology (Brief Psychiatry Scale, BPRS), social behaviour (nurses' Observation Scale for In-patients Evaluation, NOSIE) and minus symptoms (judging the Minus Symptoms According to Andreasen, SANS). After an eight weeks follow-up period, the experimental group was tested again.

T-tests for independent samples were conducted to compare the control group and the experimental group before and after the 8 weeks of therapy. Within BPRS there was not a significant difference between the groups before the riding therapy intervention. After the 8 weeks of riding therapy, there was a significant difference at the 10% level of significance between the two groups. The two groups differed significantly (at the 10% level of significance) within SANS before the intervention, after the 8 weeks period the difference between the groups was significant at the 1% level of significance. Within NOSIE the groups did not show a significant difference either before, or after the 8 weeks riding therapy program. Paired t-tests were conducted to compare the results of the experimental group on each of the three tests (BPRS, SANS, NOSIE) for the three combinations: pretest / post-test, pretest / follow-up and post-test / follow-up. Only the pretest / post-test and the pretest / follow-up differences within BPRS were significant at the 1% level of significance. The pretest / follow-up difference within SANS was significant at the 10% level of significance. The effect of therapeutic riding as subjectively experienced by the experimental group, was described with the aid of a scale for condition (Bf'S). The scale of condition was documented before and after the riding session for each of the 8 weeks of riding therapy. Paired t-tests were used to compare the results of the scale of condition for each of the 8 weeks and for the average. The average result of the scale of condition before the riding therapy session, differed significantly at the 1% level of significance from the average result of the scale of condition after the riding therapy session.

(5) The therapeutic effects of horseback riding were investigated in a study by MacKinnon et al (1995b). Nineteen CP children (aged 4 to 12) were classified into either a mildly or a moderately involved group. Children in each group were then randomly chosen to participate in a riding program. The remaining children acted as controls. The riding program took place once a week for 6 months. The children were pretested and post-tested on several scales to measure gross motor control, posture, fine motor control, activities of daily living and psychosocial changes. Analysis of variance was used to analyse the data, with intervention (experimental / control) and severity (mild / moderate)

acting as the two factors. Only the results of the Peabody fine motor test for grasping, statistically confirmed the proposed benefits of riding therapy. Improvement in posture, trunk control, attention span, pelvic mobility, hand control, social interaction, confidence, balance, flexibility, strength, better muscle tone, enthusiasm etc. were benefits observed by the riding instructor, therapist and parents.

(6) Kulichova et al (1996) included 13 children and young adults (aged 12 - 20 years) with primary or secondary scoliosis in their study on the influence of horseback riding on posture. Vertebrography[2] was used to measure changes of the vertebral column deviation in frontal and sagittal levels. Each subject was measured three times: before riding, after six months of intensive riding (1 - 2 times per week) and after the following six months of sporadic riding (once every two weeks). Paired t-tests were used to analyse the data. The spinal column deviation in the frontal axis decreased significantly after the 6 months of intensive riding. In the following 6 months period, 8 of the 13 patients worsened, but the worsening was not statistically significant. The sagittal axis did not show any statistically significant changes after the first or the second 6 months periods.

(7) The effect of therapeutic riding on behaviour and self-esteem of children with Attention-Deficit / Hyperactivity Disorder (ADHD) was studied by Basile (1997). The study consisted of 13 children between the ages of 9 and 14 years. To assess self-esteem, the Coopersmith Self-Esteem Inventory tool (SEI; Coopersmith, 1984) was used. The Child Behaviour Checklist (CBCL) for parents and the Teacher Report Form (TRF) were used to assess behaviour. The study period was 4 months and riding took place once a week. The children were assessed one week prior to the onset of the program and one week after the completion of the program. Graphs of the test scores were used to investigate general trends in the data. These graphs suggested possible improvements. Paired t-tests were used to analyse the pre- and post-intervention data. Though no control group was used, it is mentioned that a sample size of 61 for both a control group and an experimental

[2]A non-invasive method that enables the continuous monitoring of spinal column deviations.

group was needed to establish a test with power 0.6 for an effect of 0.4. None of the t-tests were significant at the 5 % level of significance. This is probably the result of a too small sample size.

(8) Within South Africa very little quantitative research has been done on therapeutic riding. After a thorough literature survey, only one South African quantitative THR study was found.

Rufus (1997) investigated the value of riding therapy as an adjunctive therapy in the development of a positive self-concept in learning disabled children. 24 learning disabled children enrolled in the Latern Special Education School in Roodepoort were included in the study. The age range of the children was from 7 years to 10 years. Of the 24 children, 12 were randomly assigned to a control group that received no riding therapy and 12 were assigned to an experimental group that participated once a week in a therapeutic riding program for 25 weeks (not necessarily consecutive weeks, due to factors such as school holidays and absenteeism). During the research, 2 children from the experimental group and 2 children from the control group dropped out. Both the control group and the experimental group were assessed before and after the therapeutic riding intervention on the Piers Harris Self-Concept scale (P-H). The experimental group was also assessed on a horse riding ability scale (developed by Rufus) during the fifth riding session and reassessed at the end of the 25 sessions. Since the group sizes were small, both parametric statistical tests and the non-parametric equivalents, were used to analyse the data. The paired t-test and the Wilcoxon signed ranks test were used to test for a significant difference in the P-H score before and after the riding intervention, for both the control and the experimental group. The control group and the experimental group showed a significant improvement (at the 5% level of significance) in self-concept from pretest to post-test on both the parametric and the non-parametric test. A t-test for independent samples and the non-parametric rank sum test, did not reveal a significant difference between the change in self-concept of the control group and the experimental group.

Both the paired t-test and Wilcoxon signed ranks test, showed a significant improvement in horse riding ability for the experimental group.

The studies discussed so far (study 1 to study 8) all involve statistical tests. **Table 2.1** provides a summary of these studies. The numbers in the first column refer to the numbers of the studies. The second column is a description of the subjects included in the sample, the third column indicates the sample size whereas the fourth and fifth columns give the duration and the frequency per week of the riding program. The sixth column describes the experimental design and the seventh gives the statistical technique used to analyse the data. Since it is not always possible to retrieve the exact information from the literature, the last column is only an **indication** of the ratio of statistically significant results (at the 5% level of significance) to the total number of tests conducted.

Table 2.1

Summary of 8 quantitative THR research studies

NO	SUBJECTS	SAMPLE SIZE	DURATION	FREQ (per week)	DESIGN	TECHNIQUE	RESULTS
1	Children with CP	11	10 weeks	2×	◆pretest_1, pretest_2, post-test	◆Friedman test	1/1
2	People with multiple sclerosis	10	9 weeks	2×	◆pretest/post-test	◆paired t-test	4/15
3	People with mental retardation	8	6 months	1×	◆pretest_1/ pretest_2 ◆pretest_2/ post-test ◆pretest_1/ post-test	◆paired t-test, Wilcoxon signed ranks test	0/2 2/2 1/1

NO	SUBJECTS	SAMPLE SIZE	DURATION	FREQ (per week)	DESIGN	TECHNIQUE	RESULTS
4	Chronic schizo phrenic patients	24	8 weeks	1X	◆pretest:control v. exper ◆post-test:control v. exper ◆exper:pretest/ post-test, pretest/follow-up, post-test/follow-up ◆exper:pretest/ post-test	◆t-test for independent samples ◆paired t-test ◆paired t-test	0/3 1/3 2/9 4/9
5	Children with CP	19	6 months	1X	◆2X2 factorial design	◆Analysis of variance	1/20
6	People with primary or secondary scoliosis	13	6 months	1-2X	◆pretest/post-test ◆post-test/ follow-up	◆paired t-test	½ 0/2
7	Children with attention-deficit/hyper activity disorder	13	4 months	1X	◆pretest/post-test	◆paired t-test	0/11
8	Learning disabled children	20	25 weeks	1X	◆control:pretest/ post-test ◆exper.:pretest/ post-test ◆control/exper ◆exper: pretest/post-test	◆paired t-test, Wilcoxon signed ranks test ◆t-test for independent samples, rank sum test ◆paired t-test, Wilcoxon signed ranks test	2/2 2/2 0/2 2/2

Two studies involved very large samples and the results were summarised in

frequency tables or visually presented in bar charts and pie charts.

Exner et al (1994) report on the results of therapeutic riding with paraplegics and tetraplegics in the *Berufsgenossenschaftlichen Unfallkrankenhaus Hamburg*. In total 153 patients were given treatment on a horse, of whom 81 were tetraplegic and 72 were paraplegic. The duration of the effect of therapeutic riding on spasticity, was presented in a frequency table. Of the 60 patients who complained of hip and back pains, therapeutic riding had a pain alleviating effect on 58 patients. The duration of the pain-alleviating effect of therapeutic riding is reported in a frequency table (1 day, 2 days and 1 week). Therapeutic riding improved the joint movability of 37 from the 38 involved patients. Again the duration of the effect (1 day, 2 days and 1 week) is presented in a frequency table. The long lasting effect of therapeutic riding on the reduction of spasticity, seemed to be especially beneficial.

Between 1987 and 1992, the *Swiss Group for Hippotherapy-K* (Künzle et al, 1994) carried out a multi-centre study on the effect of therapeutic riding on 255 patients with multiple sclerosis. The effect of therapeutic riding was rated not only by the patients, but also by the prescribing doctor and the therapist conducting the hippotherapy. The results of these three ratings were visually presented in bar charts and pie charts.

A special test device was designed and built in the Department of Aerospace Engineering Sciences, University of Colorado, to address the problem of limited test instrumentation for use in a variety of rehabilitation programs (Fox et al 1984). The instrument was designed to measure not only sitting balance and coordination, but also posture and hand, hip, knee and ankle strength. The instrument was tested on 19 handicapped children (aged 7 - 14) with a diversity of impairments before and after a 2 hour therapeutic riding session. Each subject acted as his / her own control and group means before and after the riding session were compared. Improvement was expressed as a percentage of the pretest. Balance, coordination, posture and strength revealed an increase after the riding session. The results of the test instrument agreed with the subjective observations of therapists, parents and investigators.

In many studies both quantitative and qualitative data were recorded. Some researchers, however, used only qualitative data, supplemented by audiovisual material. In von Arbin (1994b), the riding and social progress of twenty handicapped children during a Summer Riding Camp are described qualitatively. An ordinal scale, ranging from 0 (no progress) to 3 (very good progress) is used supplementary to the qualitative data to record the progress made by the children. Interviews with friends, parents, physiotherapists and riders themselves, were used in von Arbin (1994a) to assess the development of 51 subjects (ranging from preschool children to young adults) with a variety of disabilities. Videos of all subjects were taken before and after the term. Improvement in balance, self-esteem, better coordination and concentration, decrease in spasticity and a better posture were some of the benefits observed.

Would (1996) made use of an electronic measure device (Penny and Giles electronic goniometer) that measured the pelvic angle and pelvic movements occurring during horse riding. Graphs of disabled riders were compared to graphs of "normal" riders. The development of a better balance in a disabled rider could also be seen on the graphs. All the riders were videotaped.

Several other researchers used graphs to examine the existence of trends in the data, or to compare control and experimental groups.

In addition to quantitative tests (see **Table 2.1 (4)**), Scheidhacker et al (1991) and Scheidhacker (1996) also made use of graphs to compare the experimental group and the control group. Both groups were tested before and after 8 weeks on psychopathology (Brief Psychiatry Scale, BPRS), social behaviour (nurses' Observation Scale for In-patients Evaluation, NOSIE) and minus symptoms (judging the Minus Symptoms According to Anreasen, SANS). After eight weeks follow-up period, the experimental group was again tested. The results at the different levels of the three tests were plotted for both groups at times $t = 0$, 8 and 16 weeks. The graphs of the two groups were compared. Within BPRS the experimental group showed a tendency to improve during the therapy period, compared to the control group. The two groups showed the same tendencies within SANS, and

within NOSIE no changes were found.

In studying the effect of therapeutic riding on balance for children with attention deficit disorder, Yack et al (1997) also made use of graphs to examine the trends in the data. Two children (aged 9 and 10) participated in horse riding for one hour, three times a week for 4 weeks. Standing balance was assessed using the six conditions of the Pediatric Clinical Test of Sensory Interaction for Balance. Measurements of the smoothness of walking patterns using forward / backward, vertical and side-to-side head and trunk accelerations, were used to assess walking balance. The two children were assessed 3 times during the week prior to the beginning of the riding program and twice during each of the 4 weeks that the program was running. Visual analysis was used to investigate whether there were any trends in the data. In some of the graphs an upward trend was recognised, indicating an improvement over time.

Webster et al (1994) recorded electromyograms and acceleration measures near the centre of gravity for 10 healthy people and 9 patients with multiple sclerosis. Measurements were taken while walking before riding, during riding and one hour after riding. The electromyograms and the three-dimensional acceleration graphs for the healthy people and the patients were compared. Though the graphs for the healthy people showed individual differences, there were characteristic signs of a healthy human gait. The graphs of the healthy people were used as the norm with which to compare the graphs of the patients. Improvement after riding therapy was indicated by a graph that resembled more closely those of the healthy people. Of the 9 patients, 6 showed improvement after riding therapy.

Difficulties in finding homogeneous groups when conducting THR research are often experienced and therefore a case study is a reasonable alternative to groups. Van Dyk et al (1994) conducted a case study on a three-year old boy who sustained extensive bilateral brain damage after a tonsillectomy. Intensive therapy followed, but the child did not respond satisfactorily and five months after the operation, horse riding therapy was started as part of an integrated therapeutic program. Thirty minute riding sessions were conducted five times a week. Van Dyk et al describe the physical and mental development

of the child during the first year of riding. Subjective observations of the child's abilities, verbal parent reports and reports from professionals: a paediatrician, an orthopaedic surgeon, neurologists, occupational-, speech- and physiotherapists were used to document the tremendous improvement of the child.

Koch (1994) describes the progress over the period of one year of two boys suffering from dyslexia. Both were submitted to an integrated therapeutic program of which therapeutic riding formed part. Their initial reaction to horses and their behaviour in a group, as well as their physical, social and psychological development, are described. Both boys developed so positively, that the original planned duration of the program was reduced from 2 years to only 1 year.

Freeman (1984) qualitatively describes the physical effects of riding therapy on a 7-year-old quadri paretic girl and a 7-year-old boy who could walk a few steps with crutches. Three weeks after the girl started riding, her balance improved and six months after the commencement of the riding therapy, she rode with her back straight and also with a better head control. After 6 months riding, the boy improved from being able to walk 4 steps, to walking 4 blocks. He also showed psychosocial changes.

In addition to the already mentioned studies, MacKinnon et al (1995a) describe several other studies in a review of the literature. A brief discussion of these studies follows:

Snir et al (1988) studied the effect of therapeutic riding on 4 adolescents with learning disabilities. The participants rode twice a week for the duration of one school year. The effect on biomechanical, physiological and psychomotor variables was investigated before and after the completion of the program. Only 9 of the measured variables showed a significant improvement.

Ricotti et al (1991) included 5 children with brain damage and with an average age of 9 years, in a study to examine the effect of riding therapy on the muscle tone of *triceps surae* and the strength of *tibialis anterior*. The tone and strength of the muscles were tested before and after a 9-month, twice weekly riding program. Electrophysiological techniques were used to measure the tone and the strength of the muscles. The results of the

electrophysiological tests suggested that riding therapy decreases spasticity of the *triceps surae* and increases the strength of the *tibialis anterior*.

In a study conducted by Dismuke (1984), thirty subjects (aged 6 to 10 years) with moderate to severe language disorders accompanied by mild motor impairments, were divided into a control group and an experimental group. The control group received traditional speech therapy, while the experimental group received riding therapy. It was found that after 12 weeks, only the experimental group showed improvement in language skills. Manual muscle testing also revealed significant improvement in muscle strength for the experimental group. The Southern California Sensory Integration Test was conducted for the experimental group before and after the riding program. The results indicated improvement in bilateral motor coordinations, visual perception and left / right discrimination.

From the literature survey, it is clear that THR is a vast research field. The diagnoses of the subjects included in the various studies differ and consequently the variables included for observation (e.g. balance, posture and self-concept) also differ. Even though some of the observed variables coincide between studies, the scales and measurement instruments used to measure changes in the variables, differ from study to study. In addition, the frequency and duration of the therapeutic riding interventions are also not standardised.

The studies were either quantitative or qualitative and very often a combination of the two. Case studies were also conducted. For the quantitative studies, the pretest / post-test design seemed to be the most popular. Where studies included only an experimental group, the subjects were regarded as their own control. In studies that involved only an experimental group, paired t-tests and various nonparametric matched-pairs tests were used in combination with the pretest / post-test design. In studies that include a control group, t-tests for independent samples or non-parametric equivalent tests were often used to analyse the pretest / post-test data. In several studies that involved both quantitative and qualitative data, there was a discrepancy between the results of the quantitative and the

qualitative data. The qualitative data suggested that therapeutic riding had an effect, while the quantitative results were *statistically nonsignificant.*

Altman (1991) describes the power of a test and warns against an *over-reliance on p values*. The power of a test is the probability that the test will detect as statistically significant a real change of a specific magnitude. The power of a paired t-test depends on sample size, variation of the data, the magnitude of the clinically meaningful change and the significance level. For example, if the sample size is too small, the test may fail to detect a real change as *statistically significant.* This illustrates that statistically significant and clinically meaningful cannot be regarded equivalent. Many THR researchers realise that a small sample size is inadequate, but only Basile (1997) calculated the adequate sample size to achieve a power of 0.6 to detect an effect of 0.4. Extensive tables (Machin and Campbell, 1987) are available for the calculation of the sample size for fixed values of the standard deviation, the level of significance, the clinically meaningful change and the required power of the test. If the required sample size cannot be met, the study can either be extended in time, or it can be run at more centres.

The repeated measurement design was also made use of, but with no more than three measurements per subject. In these designs, a subject was usually assessed before the intervention, after the intervention and after a follow-up period. This type of data was often also analysed by applying t-tests.

In the qualitative studies, the subjective assessment of improvement by therapists, parents, or the riders themselves, was recorded. In many studies both the quantitative and the qualitative results were reported.

Graphs were also used to visually assess the improvement of the patients. A control group and an experimental group were observed over a period of time. The mean values for each group were plotted over time. The trends recognisable in the graphs for the experimental group were compared to the trends in the graphs of the control group. In one study, the progress of each individual separately, was plotted over time. Electromyograms

were also used to analyse the data visually. The electromyograms of healthy people show characteristic signs of a healthy human gait. This range of normal electromyograms was used as a norm to judge the electromyograms of handicapped people before riding, during riding and after riding. The disadvantage of specialised measurement tools, is that it is only available to one, or a few centres. If only the centres that have these measurement apparatuses available can be included in research studies, it will be a limitation to the ideal situation where several centres cooperate in research.

2.2 The Need for Further Research in THR

According to Biery (1985), Chassigne was the first to conduct studies on the benefits of horse riding in 1875 in Paris. He concluded that riding was especially effective in treating hemiplegia, paraplegia and other neurological disorders. A variety of physical and psychological benefits derived from riding were mentioned.

Researchers agree unanimously that therapeutic riding is a fallow land for research. *"...empirical evidence supporting the claims that have been made regarding the benefits of therapeutic riding is scarce."* (MacKinnon et al, 1995a).

"Research and scientific studies to support the claim that therapeutic riding is indeed beneficial are almost nonexistent." (Biery, 1985).

"Despite the growing enthusiasm for riding as a form of therapy, research into the efficacy of this intervention is virtually nonexistent". (MacKay-Lyons et al, 1988).

"Further study is needed to isolate additional variables and to examine the effects of therapeutic riding on different disabilities." (Bertoti, 1988).

Biery (1985) describes the different categories of therapeutic riding as well as the benefits thereof and concludes: *"The effects of therapeutic riding have not been subjected to scientific scrutiny"*, and *"More empirical studies are required, using sound research methodology"*.

In reviewing the research on the efficacy of physical therapy, Campbell (1990) inter alia looks at changes in postural control, alignment and stability. Bertoti's study (1988) on the effect of horse riding on the posture of children with cerebral palsy is also mentioned in this context. Campbell comments on the investigated studies: *"What is of more concern...is that these responses have seldom been adequately measured in studies with strong designs, thus bringing into question the validity of the findings"* and *"...it is important to encourage further study in this area..."*

Fox et al (1984) suggest for future research more in-depth, long-term designs using qualitative data as well as case studies. The use of control subjects is also encouraged. The systematic compilation of a medically supported data base wherein riding therapy is

evaluated for different situations, is proposed by van Dyk et al (1994).

There is no way of sidestepping the fact that accreditation of therapeutic riding in the fields of medicine, rehabilitation and education, will have to take place through research.

2.3 Problems in THR Research

An obvious difficulty when conducting THR research is the heterogeneity of subjects regarding variables like age and the nature and severity of handicap. (Van Dyk et al 1994; Fox et al, 1984; MacKinnon et al, 1995a; MacKay-Lyons et al, 1988). Lack of homogeneous subjects leads to problems associated with small samples. The absence of a control group can in many cases be attributed to the lack of homogeneous subjects (MacKay-Lyons et al, 1988; MacKinnon et al, 1995a).

External influences, like other therapies, working in on a patient simultaneously with riding therapy, make it difficult to measure the true influence of riding therapy (MacKinnon et al, 1995a; Bertoti, 1988).

The lack of standardisation of riding therapy method is a potential problem and is evident in the differences between studies regarding duration of riding sessions, frequency of sessions and length of the riding program. Confusion between categories of therapy with the help of a horse (see section 1.2), further impedes standardisation.

Measurements are also not standardised. In almost every study conducted different scales were used to measure physical and psychological variables. Different test instrumentation are utilised, e.g. the Penny and Giles electronic goniometer (Would, 1996), vertebrographic equipment (Kulichova et al, 1996) and a specially designed test instrument (Fox et al 1984). The lack of ability of existing standardised measurement tools to assess small physical or psychological changes in patients is seen as a major difficulty by MacKinnon et al (1995b).

The need for repeated measurements subject to the usual time constraint, was also mentioned by some researchers as a problem (Fox et al, 1984; MacKay-Lyons et al, 1988).

The conclusion reached by MacKinnon et al (1995a) after reviewing eleven THR studies, serves as a summary of the difficulties experienced: *"Methodological problems in this body of research include: the lack of control groups; the failure to measure and control for potential confounders; the use of instruments with unknown psychometric properties; and the use of small samples. There is also a tendency to rely on non-standardised, subjective observations, especially when attempting to assess psychosocial variables."*

It cannot be denied that there are stumbling blocks in the way of THR research. However, the need for well carried out research cannot be denied either.

THR researchers have as a mutual goal to establish therapeutic riding as a credible treatment method. Though many patients, parents and therapists believe, through experience, in the healing value of the horse, an often sceptical outside world needs scientific proof. In this study we want to make a contribution by trying to remove some of the (heavy) stumbling blocks. In approaching THR from a data analytical viewpoint we believe we can contribute to a field dominated by therapists and equestrians.

The aim of the study is to find, among numerous methods of measuring change, a suitable, and at the same time feasible, method for this field of research. This method should be sensitive to clinically meaningful changes and should also take the influence of external variables into account.

2.4 General Remarks on the Measurement of Change

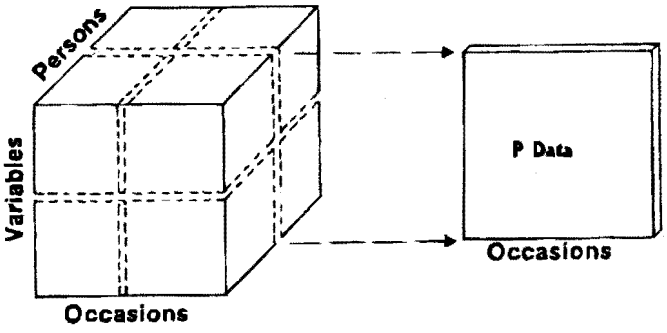
The effect of riding therapy on a person manifests in changes in attributes of that person. Burr et al (1990) describe change as *any variation in quantity or quality of an entity's attributes*. They distinguish between measuring change and structuring change. Change can be measured either on manifest variables or on latent variables. Manifest variables are observed variables, while latent variables are unobserved variables that involve the interrelationships between manifest variables.

The representation and measurement of change is an important and complex field of study and cover a wide range of methods. There is no best way to deal with change and a specific situation will prescribe the method/s most suitable to achieve the objectives at hand.

Cattell (Cattell, 1952) first introduced the *data box* to illustrate six possible covariation techniques. We will use the data box concept to illustrate how different situations and objectives in the study of change will require different slices from the data box. The data box is a cube with the axes defined by persons, variables and occasions (see **Figure 2.1**).

Figure 2.1

Cattell's Data Box



Variables that do not change over time are referred to as static variables, whereas dynamic variables change over time (Collins, 1991). When studying change, variables are assumed to be dynamic. Statistical methods developed to measure interindividual differences in static variables at a particular point in time, are not appropriate for measurement of intra-individual change over time.

The simplest case in the measurement of change is when a slice is cut from the data box for only one variable across persons and on only two occasions. When subjects are observed on the same variable at two points in time, the term *two-wave* data is commonly used. Though two-wave data represent change to some extent, it poorly defines the individual growth curves. The straight line is the most complex curve that can be fitted to two points. The adequacy of straight line growth cannot be supported by the data. Even if the functional form of growth is known, two-wave data does not provide enough information for the estimation of parameters.

The change score is the difference in the observed value of the variable between time 1 and time 2. Change scores are easy to interpret and a popular choice for measuring change. The paired t-test and the nonparametric Wilcoxon signed ranks test can be used to test whether change scores differ significantly from zero.

There is much criticism against change scores. Rogosa (1995) investigated the

basis of the "myths" about the weaknesses of change scores. He found the criticism to be unfounded in some cases.

When a person is observed on only one variable across several points in time, a growth curve can be fitted to the data. The individual's fitted growth curve can be used for prediction by extrapolating the function beyond the range of the data.

Observations of a person on the same variable over time, are serially dependent. This dependence makes ordinary analysis of variance inappropriate. Time series analysis can be used to avoid this problem (Schmitz, 1990). Time series analysis makes it possible to statistically assess the onset and duration of the effect of an intervention. This property of time series is important when the effect of a therapy is investigated.

For analyses of both growth curves and time series, the estimated parameters can be regarded as representative of the person's individual growth pattern. These parameters can be subjected to multivariate analysis of variance when comparing the growth patterns of two or more groups (Thissen et al, 1990).

If a slice is cut for an individual across variables and occasions, P-technique factor analysis (Cattell, 1952) can be used to understand change in the individual. Change on latent variables (factors) can either be assessed in level or in pattern. Shifts in factor scores can be seen as change of a quantitative nature and change in the patterning of relationships between manifest variables can be seen as change of a qualitative nature.

From the few examples mentioned in this section, it is clear that different analyses require different data slices. For the results to be general, however, these slices should be representative of the data box.

To discuss all possible ways to measure change is a vast task and several volumes would be needed to cover the subject. The problems experienced by researchers in the field of THR, as summarised in section 2.3, will serve as guidelines in the suggestion of an alternative, or adjunctive technique which may be used to measure the effect of THR.

2.5 Logistic Regression as a Possible Solution for Some of the Problems Experienced in THR Research

2.5.1 Discrepancy between quantitative and qualitative results

One of the problems of THR research that was mentioned in section 2.3, is the inability of existing standardised measurement tools to assess small physical or psychological changes in patients. Even though qualitative data gathered from therapists, patients and parents support clinically meaningful change in patients, statistical tests "fail" to confirm these findings. "*On the one hand, the qualitative results from parents, attending physiotherapist, and riding instructor generated highly positive, endorsing comments... On the other hand, statistical analyses of the quantified measures produced results that suggest the riding program was unsuccessful in demonstrating therapeutic benefits. This typical contrast is difficult to resolve at this point.*" (MacKinnon et al, 1995b).

The ambiguity of the qualitative results and the quantitative results can, at least to some extent, be attributed to the low power of the studies. The power of a statistical test depends on the sample size, the variation of the data, the level of significance and the clinically meaningful change. Small sample sizes are, almost without exception, mentioned as a limitation in THR research studies. It is not always possible to include in a study subjects that are homogeneous; this can cause large standard deviations in the variables and thus low power. It is therefore my conjecture that the many *nonsignificant* results, are often an indication of studies with low powers rather than the absence of clinically meaningful change. On the other hand, if samples were to become very large (which is quite unlikely), statistical test can become too sensitive for changes of no practical importance.

With logistic regression only two outcomes are possible: improvement and no improvement. The advantage of logistic regression is that therapists, or other specialists, can define a clinically meaningful change. Even if the variable under consideration is measured on a scale, the therapist can define whether the difference in score before riding

therapy and after riding therapy is meaningful. The therapist therefore assigns the status *improvement* or *no improvement* to a patient. This means that a specialised therapist determines what is to be interpreted as clinically meaningful and not a statistical test (with a very low power).

2.5.2 The influence of external variables

The paired t-test, or a non-parametric equivalent test, cannot account for the influence of external variables when the effect of therapeutic riding is investigated. This is another concern of THR researchers. In order to account for influential external variables, researchers tried to use homogeneous experimental and control groups. Often homogeneous groups were difficult to find. This can be seen as one of the reasons why THR sample sizes are so small and perhaps also why control groups are often not included in the study.

Logistic regression makes provision for categorical and continuous external variables. Subjects included in the study need not be homogeneous with regard to the variables included in the model. This means that more subjects will be suitable for inclusion in the study and that a formal *control group* is no longer needed.

2.5.3 Probability of improvement

Logistic regression associates a probability of improvement to a specific profile. The chance of improvement can now be predicted (keeping in mind that it is only a prediction and that a multiple other invisible factors, not included in the model, can also influence the chance of improvement). THR researchers have not even started to exploit the possibilities of prediction. This can perhaps be the beginning of the realisation of the proposal of van Dyk et al (1994), (see also section 2.2) to establish a data base in which riding therapy is evaluated for different situations.

2.5.4 Interpretation of logistic regression results

In addition to the advantage that the probability of improvement can be predicted for a specific profile, the coefficients of the logistic regression model can be interpreted in terms of the odds ratio. The odds ratio and the interpretation of the coefficients will be discussed in more detail in section 3.3.

2.5.5 Central data base

Logistic regression cannot be conducted successfully if the sample size is too small, therefore the studies should be extended in time and / or several centres will have to cooperate. The multi-centre study conducted by Künzle et al (1994) between 1987 and 1992, involved 255 patients. This study illustrates clearly the possibility to increase the sample size by extending the time of the study and involving several therapeutic riding centres. If specialised and expensive apparatuses are used, it can limit the time of the study, or impede the number of centres that can cooperate. If measurement of the outcome variable is standardised, yet feasible, several centres can cooperate over a extended period of time. In this way a central data base can be established. If a central data base is well maintained, the logistic regression model can be improved regularly.

The proposal of the establishment of a therapeutic riding data centre, received great support at the 3rd European Congress of Therapeutic Riding (Munich, September 1998). The purpose of the centre will be to organise and make available research data and to support new research projects. The availability of research data will avoid the same studies being conducted repeatedly. Researchers can thus benefit from previously conducted studies.

CHAPTER 3
LOGISTIC REGRESSION

3.1 An Introduction to the Logistic Regression Model

3.1.1 Linear regression versus logistic regression

Linear regression modelling (Draper and Smith, 1981) is a well-known technique among statisticians and users of statistics and does not need much introduction. In linear regression the model describes in as parsimonious as possible, yet interpretable, way the relationship between the continuous dependent variable and the set of independent variables. The analyst usually starts with simple scatter plots to determine the nature and strength of the relationships between the dependent variable and each of the independent variables. One of the basic assumptions in linear regression, is that the relationship between the dependent variable and the independent variable is linear (therefore the term "linear regression").

Logistic regression (Collett, 1991; Hosmer and Lemeshow, 1989) is not as well-known as linear regression, though these two techniques show many resemblances. The primary difference between logistic regression and linear regression, is that the dependent variable in logistic regression is not continuous, but **binary** or **dichotomous** e.g. "yes" / "no" and "success" / "failure". It is customary in logistic regression to represent the binary outcome of the dependent variable by "0"(failure) and "1"(success). A scatter plot of the dependent versus the independent variables will produce a plot where all the points fall on two parallel lines corresponding to the two possible outcomes. This scatter plot is not very informative and does not give an idea of the nature of the relationship between the dependent and the independent variable.

The expected value of Y (dependent variable), given the values of $\boldsymbol{x}$ (the independent variables) can be written as $E(Y|\boldsymbol{x})$. The aim of regression analysis is to model the relationship between the conditional mean of the dependent variable and the independent variables. From the differences in the scatter plots for the linear regression case and the logistic regression case, it is clear that $E(Y|\boldsymbol{x})$ is linear in the independent variables (or transformations of the independent variables) for linear regression, but that it is not the case for logistic regression. For linear regression $E(Y|\boldsymbol{x})$ can range from $-\infty$ to $+\infty$, but for logistic regression this quantity must lie in the interval [0, 1]. In logistic regression, $E(Y|\boldsymbol{x})$ is the probability that Y assumes the value "1" for a given $\boldsymbol{x}$ and we can write this probability as $p(\boldsymbol{x})$.

3.1.2 The logistic function

Many functions have been proposed to describe the relationship between the conditional mean of a binary outcome variable and the independent variables. The most popular choice is the logistic function. The logistic function can be written as :

$$f(z) = \frac{1}{1+e^{-z}}$$

and

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \tag{3.1}$$

where z is an index that combines the X's (the independent variables).

There are several reasons why the logistic function is such a popular choice. Firstly, $f(z)$ ranges between 0 and 1 and is therefore suitable to describe probabilities. Secondly, $f(z)$ approaches zero for z approaching $-\infty$. For z increasing, the function stays close to zero and at a threshold value increases dramatically towards 1, where it levels off and approaches 1 as z approaches $+\infty$. Not only the S-shape, but also the coefficients of the logistic function lend themselves to meaningful interpretation by researchers, especially

epidemiologists (see section 3.3). Thirdly, the logistic function is mathematically flexible and easy to use.

3.1.3 The error term

Another important difference between logistic regression and linear regression lies in the distribution of the error terms. If y is the value of the dependent variable, then we can write $y = E(Y|x) + \epsilon$, where ϵ is the error term. For linear regression the assumption is that ϵ is normally distributed with a mean of zero and a constant variance across the levels of independent variables.

Recalling that we can substitute $E(Y|x)$ by $p(x)$ in the binary situation, we can write

$y = p(x) + \epsilon$ and when $y = 1$, $\epsilon = 1 - p(x)$ with a probability of $p(x)$. If $y = 0$, then $\epsilon = -p(x)$ with a probability of $1 - p(x)$. The error term is therefore binomially distributed with mean

$$\begin{aligned} E(\epsilon) &= [1 - p(x)] \times p(x) + [-p(x)] \times [1 - p(x)] \\ &= 0 \end{aligned} \tag{3.2}$$

and variance

$$\begin{aligned} Var(\epsilon) &= [1 - p(x)]^2 \times p(x) + [-p(x)]^2 \times [1 - p(x)] \\ &= p(x) \times [1 - p(x)] \end{aligned} \tag{3.3}$$

3.1.4 Link function

Link functions are used to transform the logistic function to a function which is linear in the parameters.

The most commonly used link function is the logistic transformation, written as

$$logit(p) = \log_e \frac{p}{(1-p)} \qquad (3.4)$$

$$= \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

where p means the same as $p(\boldsymbol{x})$.

For p in the interval [0, 1], *logit*(p) ranges from $-\infty$ to $+\infty$. Also, *logit*(p) is continuous and linear in the parameters, therefore many of the techniques used in linear regression can also be used, or adapted, for the *logit* function. *Logit*(p) is a sigmoid curve that is symmetrical about $p = 0.5$. Between $p = 0.2$ and $p = 0.8$, the *logit*(p) function is almost linear, but for $p < 0.2$ and $p > 0.8$ the function is distinctly non-linear.

The *probit* function is an example of another link function. The *probit* is defined as $probit(p) = \Phi^{-1}(p) + 5$, where Φ^{-1} is the inverse of the standard normal distribution function. The 5 is added to avoid having to work with negative values. For p in the interval [0, 1], the *probit* function ranges between $-\infty$ and $+\infty$. The *probit* function has the same general form as the *logit* function and is also symmetrical about $p = 0.5$. The *probit* function is, however, not as popular as the *logit* function, since it is more difficult to compute.

The complementary log-log transformation for p is $\log[-\log(1-p)]$. Since this transformation is not symmetric about $p = 0.5$, it is used for situations where the success probability is handled in an asymmetric manner. The log-log transformation is inapplicable to this dissertation, but the interested reader is referred to Collett (1991, paragraph 4.6) where interesting examples for application of the log-log transformation is given.

The *logit* transformation remains the most popular. Not only can the *logit* function be interpreted as the log of the odds in favour of success, but it is also useful in modelling data of a retrospective nature.

3.2 Estimation of the Coefficients and Assessing the Fit of the Coefficients

3.2.1 Methods of estimation

The **least squares method** of estimation of the unknown parameters in a model is associated with linear regression. Least squares estimation can intuitively be understood as the selection of those parameters that will minimise the sum of squared deviations of the observed dependent values from the predicted values. If it can be assumed that the variables are normally distributed, this method will yield estimated coefficients that possess many favourable qualities. When the dependent variable is binomial, however, the estimated coefficients no longer have these qualities and another method of estimation should be used.

A more general method of estimation, is the **maximum likelihood** method. This method can be seen as finding the parameters that have the highest probability of producing the observed data set. Since the exact distributions of the estimated coefficients are not easy to derive, the asymptotic properties of the estimates are used. This causes problems when sample sizes are too small, e.g. maximum likelihood estimates are asymptotically unbiased, but small samples will give biased estimates. When the sample size is large, the maximum likelihood estimate may have a small variance and therefore be a stable estimate. What makes the maximum likelihood estimate even more attractive, is that the estimate is asymptotically normally distributed. This means that, irrespective of the distribution of the original data, one can make inferences about the estimate using standard normal theory. An interesting result is that, if the variables are normally distributed, the least squares method and the maximum likelihood method will both give the same estimates for the means. If the data are normally distributed, the maximum likelihood estimates will therefore be unbiased, even for small samples.

3.2.2 The likelihood function and equations for binary data

To estimate the coefficients with the maximum likelihood method, the **likelihood function** should first be specified. The likelihood function is the joint probability of the observed data and is a function of the unknown parameters. One must assume a distribution function for the data, before the likelihood function can be constructed. The maximum likelihood estimates are the parameters that maximise the likelihood function. Consider a sample of n independent observations of the pair $(\boldsymbol{x}_i, y_i)$ with probability $p(\boldsymbol{x}_i)$ that Y_i is equal to 1, given $\boldsymbol{x}_i$. There are n_i observations with the same values for $\boldsymbol{x}_i$. If $n_i > 1$, we assume a binomial distribution for the data, but for $n_i = 1$, the data have a binary distribution. For binary data, $E(Y|\boldsymbol{x}_i) = p(\boldsymbol{x}_i)$ and $Var(Y|\boldsymbol{x}_i) = p(\boldsymbol{x}_i)[1-p(\boldsymbol{x}_i)]$. A binary distribution for the data will be assumed from now on, but similar results can be derived if the data are binomially distributed.

The likelihood function is the product of n terms:

$$l(\beta) = \prod_{i=1}^{n} p(\boldsymbol{x}_i)^{y_i} [1-p(\boldsymbol{x}_i)]^{1-y_i} \tag{3.5}$$

It is easier to work with the log of the likelihood function, and since the parameters that will maximise $l(\beta)$, will also maximise $\log_e[l(\beta)]$, it is customary to rather use

$$\log_e[l(\beta)] = \sum_{i=1}^{n} \{y_i \log_e[p(x_i)] + (1-y_i)\log_e[1-p(x_i)]\} \tag{3.6}$$

when calculating the maximum likelihood estimates.

The (k + 1) likelihood equations are derived by differentiating the log likelihood function with respect to the (k + 1) coefficients and setting the resulting (k + 1) equations equal to zero.

The likelihood equations are

$$\sum_{i=1}^{n} [y_i - p(x_i)] = 0$$

and

$$\sum_{i=1}^{n} x_{ij}[y_i - p(x_i)] = 0 \quad for\ j = 1, 2, ..., k \tag{3.7}$$

Since these equations are not linear in the unknown parameters, iterative procedures are used to solve for the parameters.

3.2.3 The deviance

The **likelihood ratio** of a given model is the ratio of the likelihood of the given model to the likelihood of the saturated model . The saturated model has as many parameters as there are data. The likelihood function for the saturated model will be

$$\prod_{i=1}^{n} [(y_i^{y_i})(1 - y_i)^{1-y_i}] \tag{3.8}$$

The **deviance** is minus two times the log of the likelihood ratio

$$D = -2\log_e \frac{(likelihood\ of\ the\ current\ model)}{(likelihood\ of\ the\ saturated\ model)} \tag{3.9}$$

For the saturated model of binary data, the log of the likelihood function is

$$\sum_{i=1}^{n} [y_i \log_e y_i + (1 - y_i) \log_e (1 - y_i)]$$

For y_i = 0 and 1, both $y_i \log_e y_i$ and $(1 - y_i) \log_e (1 - y_i)$, are zero and therefore the log likelihood function for the saturated model will be zero and the deviance for binary data becomes

$$D = -2 \sum_{i=1}^{n} [y_i \log_e p(x_i) + (1 - y_i) \log_e (1 - p(x_i))] \tag{3.10}$$

For binomial data, the deviance has an approximate chi-square distribution with $[n - (k + 1)]$ degrees of freedom, where n is the total number of binomial observations and $(k + 1)$ is the total number of unknown parameters in the model. The conditions under which the asymptotic distribution holds, are that the individual binomial denominators n_i must be adequately large and the fitted probabilities under the current model not near zero or one. Therefore, for binary data the deviance does not have an asymptotic chi-square distribution (see paragraph **3.8.2**, Collett, 1991). Since $y_i = p(x_i)$ for binary data, the deviance can anyway not be used as a measure of how well the model fits the data.

3.2.4 Assessing the significance of the coefficients

To assess the significance of a set of independent variables, we calculate the difference in the deviance between the model without the set of independent variables and the model that includes these variables

$$G = D(\textit{model without the variables}) - D(\textit{model with the variables}) \qquad (3.11)$$

$$= -2 \log_e \frac{(\textit{likelihood without the variables})}{(\textit{likelihood with the variables})}$$

Though the deviance for binary data does not have an asymptotic chi-square distribution, G for both binomial and binary data, is asymptotically chi-square distributed. Under the null hypothesis that p linearly independent variables have zero coefficients, G has a chi-square distribution with p degrees of freedom and this is referred to as the **likelihood ratio test**. A special case is when all k linearly independent variables have zero coefficients under the null hypothesis.

3.2.5 The univariate Wald test

If the null hypothesis that the k independent variables have zero coefficients, is rejected by the likelihood ratio test, it means that at least one of the independent variables should be included in the model. The univariate Wald statistic can be used to test the significance of the independent variables individually. The Wald test statistic (W_j) for variable j, is the estimated coefficient, divided by the standard error of the coefficient, i.e.

$$W_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \qquad (3.12)$$

Under the null hypothesis that $\beta_j = 0$, and for a large sample size, the Wald statistic has a standard normal distribution.

A $100 \times (1-\alpha)\%$ confidence interval for the variable β_j is given by

$$\hat{\beta}_j \pm z_{1-\alpha/2} \times SE(\hat{\beta}_j) \tag{3.13}$$

Most statistical packages provide the standard errors of the coefficients. A detailed discussion of the estimation of standard errors is beyond the scope of this study.

It is important to remember that statistical tests are not the alpha and omega and that the research objectives should always be kept in mind. If a specific variable is known to be a meaningful variable in the field under consideration, but is statistically "nonsignificant", it is up to the researcher to retain the variable in the model.

The multivariate Wald test statistic and the multivariate Score test (Cox and Hinkley, 1974), can be used to test the same null hypothesis as the likelihood ratio test, though the likelihood ratio test remains the most popular.

3.3 The Odds Ratio and Interpretation of the Coefficients

3.3.1 The odds ratio

The **odds** for an individual or a group of people with independent variables $x' = (x_1, x_2, ..., x_k)$ to be successful, is

$$\frac{p(x)}{1-p(x)}$$

and when the logistic function is substituted for $p(x)$, the odds is written as

$$e^{(\beta_0 + \sum \beta_i x_i)} \tag{3.14}$$

The **odds ratio** is the ratio of two odds, where the odds are calculated for two different individuals or groups. An odds ratio of $\psi = 2$, for example, means that the odds ratio of a success outcome for group 1, is twice the odds ratio for group 2. On the other hand, $\psi = 0.5$ means that the odds ratio of a success outcome for group 1, is one half of the odds ratio for group 2. When the success probabilities for both groups are small, the odds ratio approximates the relative risk. The relative risk is an indication of how much more likely it is to get a success outcome for the one group, compared to the other.

The odds ratio is the basis for the interpretation of the logistic regression coefficients. The fact that a meaningful interpretation can be linked to the logistic regression coefficients, makes logistic regression a very useful and popular tool.

If the independent variables for groups 1 and 2 are given by $\boldsymbol{x}_1' = (x_{11}, x_{21}, ..., x_{k1})$ and $\boldsymbol{x}_2' = (x_{12}, x_{22}, ..., x_{k2})$, then the odds ratio for group 1 versus group 2 is

$$\psi = \frac{\dfrac{p(\boldsymbol{x}_1)}{1-p(\boldsymbol{x}_1)}}{\dfrac{p(\boldsymbol{x}_2)}{1-p(\boldsymbol{x}_2)}}$$

If we substitute the logistic function for $p(\boldsymbol{x}_i)$, the odds ratio can be written as

$$\psi = \frac{e^{(\beta_0 + \sum \beta_i x_{i1})}}{e^{(\beta_0 + \sum \beta_i x_{i2})}} \tag{3.15}$$

A quantity that is often used, is the log of the odds ratio, which gives the difference between the logit functions for the two groups

$$\begin{aligned} \log_e \psi &= \log_e(e^{(\beta_0 + \sum \beta_i x_{i1})}) - \log_e(e^{(\beta_0 + \sum \beta_i x_{i2})}) \\ &= \beta_0 + \sum \beta_i x_{i1} - [\beta_0 + \sum \beta_i x_{i2}] \\ &= \sum \beta_i (x_{i1} - x_{i2}) \end{aligned} \tag{3.16}$$

The log of the odds ratio is used in the interpretation of the coefficients. If two groups differ only in the value of a dichotomous variable, say $x_{11} = 1$ and $x_{12} = 0$ with the other independent variables (called the **control variables**) fixed to be the same for the two groups, we get $\log_e \psi = \beta_1$. This means that $\exp(\beta_1)$ is the odds ratio.

3.3.2 Interpretation of the coefficient of a dichotomous independent variable

In order to simplify interpretation of the coefficients of the independent variables, we first look at the univariate logistic model and then at a multivariate logistic model. Hence consider the model $logit(p) = \beta_0 + \beta_1 x$.

If the independent variable is dichotomous, the interpretation of the coefficient will depend on the method used to code the variable. The simplest case is to use the (0, 1) coding. If $x = 1$ for group 1 and $x = 0$ for group 2, the log odds ratio of group 1 versus group 2, is β_1. The estimated odds ratio will then be $\exp(\hat{\beta}_1)$, where $\hat{\beta}_1$ is the estimated value of β_1.

If an arbitrary coding is used for a dichotomous variable, say (a, b), the estimated log odds ratio is the difference between the estimated logit functions

$$\log_e[\hat{\psi}(a,b)] = (\hat{\beta}_0 + \hat{\beta}_1 \times a) - (\hat{\beta}_0 + \hat{\beta}_1 \times b) \qquad (3.17)$$

$$= \hat{\beta}_1 \times (a - b)$$

and the estimated odds ratio is $\exp[\hat{\beta}_1(a-b)]$. If the (0, 1) coding is used, a = 1 and b = 0.

A $100 \times (1 - \alpha)\%$ confidence interval for the odds ratio is obtained by taking the exponent of the confidence limits obtained for β_1 and is given by

$$\exp[\hat{\beta}_1 \pm z_{1-\alpha/2} SE(\hat{\beta}_1)] \qquad (3.18)$$

3.3.3 Interpretation of the coefficient of an ordinal independent variable

For the purpose of this dissertation, it will be assumed that the ordinal variables are

on an equally spaced interval scale. Ortogonal polynomials (Hosmer and Lemeshow, 1989) can be used to assess the trend in the relationship between the logit and the ordinal independent variable.

If the independent variable is ordinal with $q > 2$ categories, the two values of the independent variable that is to be compared, must be specified. Suppose the independent variable x is the level of education, and x = (0, 1, 2, 3, 4, 5) with 5 indicating the highest level of education. If the odds of people with the highest level of education is to be compared to the odds of people who have the lowest level of education, the estimated log odds ratio becomes

$$\log_e[\hat{\Psi}(5,0)] = (\hat{\beta}_0 + \hat{\beta}_1 \times 5) - (\hat{\beta}_0 + \hat{\beta}_1 \times 0)$$

$$= 5\hat{\beta}_1$$

and the estimated odds ratio is $\exp(5\hat{\beta}_1)$.

In a similar way, the estimated odds ratio for people with an education level of 4 versus people with an education level of 2 will be $\exp(2\hat{\beta}_1)$. The odds ratio depends on the difference in education level, for example, the estimated odds ratio for people with an education level of 2 versus people with an education level of 0, will also be $\exp(2\hat{\beta}_1)$. In general, the estimated log odds ratio is given by

$$\log_e[\hat{\Psi}(x+c,\ x)] = [\hat{\beta}_0 + \hat{\beta}_1 \times (x+c)] - [\hat{\beta}_0 + \hat{\beta}_1 \times (x)] \qquad (3.19)$$

$$= c\hat{\beta}_1$$

A 100 × (1 - α)% confidence interval for the odds ratio is given by

$$\exp[c\hat{\beta}_1 \pm z_{1-\alpha/2} cSE(\hat{\beta}_1)] \tag{3.20}$$

where c indicates the difference in levels of the ordinal independent variable.

3.3.4 Interpretation of the coefficient of a nominal independent variable

If the independent variable is nominal with q > 2 categories, it is inappropriate to model the variable as if it was ordinal. The proper way to model a nominal independent variable, is to make use of **dummy variables** or **design variables**. If the model contains an intercept term, then in general one will need q - 1 design variables to represent the q categories of the independent variable. If the variable has three categories, for example, one will create two design variables, D_1 and D_2. One of the three categories will serve as the reference category. For a person falling in the reference category, both the design variables will take on the value zero. For a person in category 1, $D_1 = 1$ and $D_2 = 0$ and for a person in category 2, $D_1 = 0$ and $D_2 = 1$.

In general, suppose the design variables for group 1 is indicated by $D_1, D_2, \ldots, D_{q-1}$ and for group 2 by $d_1, d_2, \ldots, d_{q-1}$, the estimated log odds ratio is

$$\log_e \hat{\psi} = (D_1 - d_1)\hat{\beta}_1 + (D_2 - d_2)\hat{\beta}_2 + \ldots + (D_{q-1} - d_{q-1})\hat{\beta}_{q-1} \tag{3.21}$$

The individual coefficient β_i can be interpreted as the log odds ratio of the ith group versus the reference group. It is therefore important to choose a meaningful group as the reference group. In most studies there is a natural reference group, e.g. the control group.

3.3.5 Interpretation of the coefficient of a continuous independent variable

If the independent variable is continuous, the interpretation of the coefficient is

influenced by the scale of the continuous variable. The coefficient gives the change in the log odds ratio for a one unit increase in x. Often a one unit change in x is not meaningful e.g. if x is the metres travelled by car. One would rather be interested in a unit of a 1000 metres. On the other hand, if x is the height of 5-year old children, then an increase of one metre will be too large. It is therefore important to obtain the log odds ratio for a meaningful change of c units.The estimated log odds ratio for a change of c units in x is

$$\log_e[\psi(x+c,\ x)] = [\hat{\beta}_0 + \hat{\beta}_1 \times (x+c)] - [\hat{\beta}_0 + \hat{\beta}_1 \times (x)] \qquad (3.22)$$

$$= c\hat{\beta}_1$$

and the estimated odds ratio is $\exp(c\hat{\beta}_1)$.

In order to interpret $\exp(c\hat{\beta}_1)$ correctly, it is important that the variable is linear in the logit. If the independent continuous variable is not linear in the logit, a possible solution is to create design variables by grouping the independent variable. Other possibilities for solving the problem are the inclusion of higher order terms of x (e.g. x^2) or nonlinear transformations of the variable (e.g. $\log_e x$) in the model.

A $100 \times (1 - \alpha)\%$ confidence interval for the odds ratio is given by

$$\exp[c\hat{\beta}_1 \pm z_{1-\alpha/2} c SE(\hat{\beta}_1)] \qquad (3.23)$$

3.3.6 Interpretation of the coefficients for the multivariate case

Up to now the interpretation of the coefficients of a logistic regression model was explained through the use of a univariate model. In a practical situation the model is likely to include various independent variables. The multivariate case is not much different from the nominal independent case, except that in the multivariate case we are working with k distinct variables and not q dummy variables. In general, suppose the independent

variables for group 1 are indicated by $x_1' = (x_{11}, x_{21}, ..., x_{k1})$ and for group 2 by $x_2' = (x_{12}, x_{22}, ..., x_{k2})$, the estimated log odds ratio is

$$\log_e \hat{\psi} = (x_{11} - x_{12})\hat{\beta}_1 + (x_{21} - x_{22})\hat{\beta}_2 + ... + (x_{k1} - x_{k2})\hat{\beta}_k \qquad (3.24)$$

If we calculate the odds ratio for two groups with the values of all the independent variables fixed, except one, say x_j, it is said that the odds ratio is **adjusted** for the control variables. The **estimated adjusted odds ratio** is then $\exp(\hat{\beta}_j)$. The reason for adjustment is to prevent the effect of x_j being incorrectly estimated due to the differences of the distribution in the control variables for different levels of x_j. Hosmer and Lemeshow (1989; paragraph 3.5) illustrate the importance of adjustment. They examine the difference between two groups of 50 men. The response variable indicates whether a subject has seen a physician within the last 6 months (1 = yes, 0 = no). Descriptive statistics for the two groups revealed the difference in the distribution of age for the two groups. The estimated odds ratio (without adjustment for age) was calculated, $\hat{\psi} = 9.33$ and also the age adjusted odds ratio $\hat{\psi} = 4.75$. The effect of group membership on the outcome variable was thus incorrectly estimated the first time, due to the differences in the distribution of age in the two groups.

Certain assumptions must be true for adjustment to be effective: the logit function must be linear in the control variable and the slopes of the logit functions for the different levels of the exposure variable, must be constant. If this slope differs across the levels of the exposure variable, interaction is taking place. The effect of interaction will be discussed in more detail in section 3.4.

If the variable x_i is not involved in an interaction term, then the $100 \times (1 - \alpha)\%$ confidence interval for the odds ratio which adjusts for the independent variables $x_1, x_2, ..., x_{j-1}, x_{j+1}, ..., x_k$ is calculated in a similar way to that of the univariate case.

3.4 Interaction and Confounding

3.4.1 Confounding

A **confounder** is an independent variable that is not only associated with (or a risk factor for) the dependent variable, but is also associated with a main independent variable, though the relationship is not causal. Independent variables that indicate different circumstances that can cause the disease under study, are referred to as **risk factors** or **exposure variables** by epidemiologists. A variable that is caused by, or the cause of an exposure variable, will not be regarded as a confounder. In the model-building stage many variables are included that are **potential confounding** variables. Analysis of the data will clarify whether these variables are true confounders.

It is important to include confounders in a model, since the exclusion of a confounder can lead to a misunderstanding of the true association between the dependent variable and an independent exposure variable. This is clear from the illustrative example provided by Collett (1991; paragraph 7.3). In this example the association between Coronary Heart Disease (CHD) and high alcohol consumption was investigated. The analysis lead to the conclusion that heavy drinking is an important risk factor for CHD. Since smoking is also a risk factor for CHD and it is known that alcohol consumption and smoking are associated, the same analysis was repeated, including smoking as a confounder. The reason for the apparent association between CHD and alcohol consumption in the first analysis, was the fact that a significantly higher proportion of the CHD patients were also smokers. When smoke was included as a confounder, CHD and alcohol consumption were unrelated.

3.4.2 Interaction

If the association between the risk factor and the dichotomous outcome variable is dependent on the confounder, that is, the level or value of the confounder influences the

association between the risk factor and the dependent or outcome variable, **interaction** is taking place. In epidemiology the term **effect modifier** is often used to indicate a variable that interacts with a risk factor. The variables involved in an interaction can either be categorical or numerical.

The concept of interaction can, however, be more easily visualised when the effect modifier is continuous and the risk factor is dichotomous. The logit functions for the two levels of the risk factor can be displayed graphically against the continuous effect modifier. If the two functions are parallel, no interaction is taking place. If the slopes of the two functions differ, however, interaction is taking place and we can say that the association between the risk factor and the outcome variable is influenced by the effect modifier.

Interaction is incorporated into a model via higher order terms, usually the product of the effect modifier and the risk factor. In the model-building stage it is important to assess whether interaction is present in the data and should therefore be included in the model.

3.4.3 Interpretation of the odds ratio when interaction is present

In section 3.3 the odds ratio and the interpretation of the coefficients for the univariate and the multivariate logistic regression models were discussed. The presence of interaction will influence the calculations of the log odds ratio and the odds ratio. Since the effect of the risk factor is influenced by the level of the effect modifier, the odds ratio for two groups will depend on the level of the effect modifier. The effect of interaction on the odds ratio, is illustrated by a simple example: Let E be the risk factor, X_1 and X_2 control variables adjusted for and X_2 also be an effect modifier. The estimated log odds ratio for $E = e_1$ versus $E = e_2$ with X_1 held constant at $X_1 = x_1$ and X_2 held constant at $X_2 = x_2$ is

$$\log_e \hat{\psi} = \hat{\alpha} + \hat{\gamma} e_1 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2 \times e_1 - (\hat{\alpha} + \hat{\gamma} e_2 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2 \times e_2) \qquad (3.25)$$

$$= \hat{\gamma}(e_1 - e_2) + \hat{\beta}_3 x_2 \times (e_1 - e_2)$$

From this expression it is clear that the log odds ratio and the odds ratio will change for different levels of the effect modifier, in this case X_2. The model used here as an example can be extended to a more general form involving several risk factors, confounders and interaction terms.

For the example in this paragraph, let $\hat{l} = \hat{\gamma}(e_1 - e_2) + \hat{\beta}_3 x_2 \times (e_1 - e_2)$. The $100 \times (1 - \alpha)\%$ confidence interval for the odds ratio in the example is

$$\exp[\hat{l} \pm z_{1-\alpha/2} SE(\hat{l})] \tag{3.26}$$

The standard error of $\hat{l}$ can be obtained from the estimated variance-covariance matrix. The estimated variance-covariance matrix is provided by the computer output of statistical packages (e.g. SAS).

3.5 Model-Building Strategies

3.5.1 The purpose of model-building

In the model-building stage of logistic regression, the relevant (statistically significant and / or practically important) variables are selected for inclusion in the model. The result of the model-building stage should be a parsimonious, interpretable model that fits the observed data well. Model-building and the inclusion or exclusion of variables is a subjective process and there is no single way to arrive at a final model, as there is no single "best model". The trend among some epidemiologists to include all scientifically relevant variables in a model, can lead to an unstable model which is very sensitive to the observed data set. When the variables included in the model discriminate the outcome perfectly, that is, the independent variables separate the outcome variable completely, the maximum likelihood estimates do not exist. This numerical problem is referred to as **complete separation** and can usually be recognised by large standard errors and some of the estimated coefficients becoming very large. When the independent variables for the two possible outcomes overlap at a single, or a few tied values, the term **quasicomplete separation** is used by Albert and Anderson (1984). Complete separation is dependent on sample size, the proportion of subjects with the outcome present and the number of variables in the model. The implication is that the more variables are included in the model, the more sensitive the model becomes for complete separation. The term **over fitting** is also used to refer to the problem of complete separation.

3.5.2 How to select variables

Potential independent variables should be derived from the literature and previous experience. Univariate analyses of the potential variables will indicate variables that should be considered for further investigation. The likelihood ratio chi-square test (see section 3.2.4) is used to assess whether an independent variable is significant or not. For a nominal

variable with q > 2 design variables, the test statistic will have a chi-square distribution with (q - 1) degrees of freedom. For nominal independent variables with 2 categories, ordinal and continuous variables, the likelihood ratio test will be approximately chi-square distributed with 1 degree of freedom. A weakness of the univariate analyses, is that a specific variable that shows weak association with the outcome variable, can become an important variable in association with other variables. It is therefore advisable not to judge a variable "nonsignificant" too early in an analysis. If a variable is statistically nonsignificant, but from the literature it is known to be an important variable, it is up to the analyst to include the variable in the model.

After the initial univariate tests, all the selected variables are included in a multivariate model. The importance of each variable in the model can be assessed through the Wald statistics. If a variable seems unimportant, a new model must be fitted, excluding the variable. The "new" model and the "old" model must then be compared through the likelihood test. If there is no significant difference between the two models and the estimated coefficients of the remaining variables do not show large discrepancies between the two models, the "new" model can be adopted. Estimated coefficients that have changed markedly, is a danger sign. It is an indication that one or more variables excluded from the model are important, since the effect of the remaining variables are influenced by these variables. The inclusion, exclusion and refitting process should terminate only when all statistically significant and also practically important variables are included in the model.

The continuous variables included in the model should be checked for linearity in the logit. Various methods exist to verify this assumption. One approach is to divide the range of the continuous variable into meaningful groups and then plot the mean of the outcome variable versus the various group midpoints. Systematic deviations from the linear graph will indicate violation of the linearity assumption. Another approach, the so-called Box-Tidwell approach (Box and Tidwell, 1962), is to add the term $x \log_e(x)$. If this term has a significant coefficient, we can suspect non-linearity. If a variable is not linear in the logit, a transformation or higher order terms can be added to the model (e.g. $\log_e(x)$, x^2).

If all the continuous variables are in the correct scale, interaction terms can be added to the model. If the model contains many variables, it is ineffective to test the significance of all possible interaction terms. Only interaction terms that are suggested in the literature and are interpretable, should be tested for significance. The significance of an interaction term is assessed through the likelihood ratio test, in the same way that the significance of main effect variables is tested. When including interaction terms, it is important that the model is hierarchically well-formulated. For any variable in a **hierarchically well-formulated** model, all lower-order components of the variable are also included in the model.

3.5.3 Stepwise and best subset selection of variables

Similar to linear regression, a stepwise selection procedure can also be used to select variables. The stepwise procedure that starts with the model containing no independent variables, is referred to as the **forward selection procedure**. At each step, the "most important" variable that is not included in the model and is statistically significant, is added to the model. The variables already contained in the model, are tested individually in every step to see whether they can still be regarded as significant, or whether one or more variables should be excluded. The procedure terminates when all the independent variables are added to the model, or when all the variables excluded from the model, are nonsignificant. Backward elimination refers to the stepwise selection procedure that starts with the full model and in every step excludes the variable that produces the smallest change in the log-likelihood statistic.

Since the data are assumed to be normally distributed for linear regression, the F-test is used to test for the significance of a variable. The assumption of normality is not true for the logistic regression model and the log-likelihood test statistic is used to assess the significance of a variable.

The best subset selection procedure selects a few " best models" with only one variable, a few "best models" with only two variables, a few "best models" with only three variables, up to the single model containing all p variables. The "best models" are selected making use of criteria like Mallow's Cp-statistic (Mallows, 1964).

Variable selection procedures, like stepwise selection and best subset selection, should only be seen as aids in the selection procedure. These "automatic" procedures can never replace the role that the subject specialist must play in scrutinising the final model.

3.6 Assessing the Fit of the Model

3.6.1 Goodness-of-fit of a model

In the model-building stage, the significance of the individual variables and interaction terms were assessed and the significant variables were included in the model. It remains to test how well the model can describe the outcome variable, or to assess the **goodness-of-fit** of the model. The goodness-of-fit of a model includes summary measures that give an indication of the distance between the observed outcome values (y) and the predicted outcome values ($\hat{y}$). It also includes measures and graphs of the relative magnitude and pattern in which every individual (y_i, $\hat{y}_i$) pair contribute to the summary statistics. Even though a summary statistic may suggest that the model fits the data, it may not exclude the fact that the model deviates from fit for a few subjects. Therefore it is also important to investigate the individual components of the summary statistics. The investigation of the contribution of the individual values to the fit of the model, is referred to as model diagnostics and will be described in section 3.7. In this section different summary measures are shortly discussed.

3.6.2 The Pearson residuals

We will use the term **covariate pattern** to indicate a single set of values for the independent variables. Let n_i be the number of observations with the same covariate pattern, y_i is the number of successes out of n_i and p_i is the success probability. Further, let J be the number of distinct covariate patterns for a set of data so that $\sum_{i=1}^{J} n_i = n$. The **raw residual** is $y_i - n_i\hat{p}_i$ and the **Pearson residual** is defined as follows

$$P_i = \frac{y_i - n_i\,\hat{p}_i}{\sqrt{n_i\,\hat{p}_i(1-\hat{p}_i)}} \tag{3.27}$$

The denominator is the standard error of y_i. For binary data the Pearson residual is

$$P_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} \tag{3.28}$$

Under the null hypothesis that the model is correct and for the n_i relatively large, $X^2 = \sum P_i^2$, has an asymptotic chi-square distribution with J - (k + 1) degrees of freedom.

The Pearson residual do not have unit variance and therefore the **standardised Pearson residual** is often used. To get the standardised Pearson residual, the raw residuals are divided by their standard errors. The standard error for a raw residual is given by

$$\sqrt{\hat{v}_i(1-h_i)}$$

where $\hat{v}_i$ is the standard error of y_i and h_i is the ith diagonal element of the $n \times n$ **hat matrix** $\boldsymbol{H}$. The hat matrix $\boldsymbol{H}$ is given by $\boldsymbol{H} = \boldsymbol{W}^{1/2}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{1/2}$, where $\boldsymbol{X}$ is the $n \times (k + 1)$ design matrix and $\boldsymbol{X}'$ is the transpose of the design matrix. $\boldsymbol{W}$ is the $n \times n$ diagonal matrix with the ith diagonal term the standard error of y_i. The standardised Pearson residual is

$$r_{P_i} = \frac{y_i - n_i\hat{p}_i}{\sqrt{\hat{v}_i(1-h_i)}} \tag{3.29}$$

3.6.3 The deviance residual

Another commonly used residual is the **deviance residual**, defined as

$$d_i = sgn(y_i - \hat{y}_i)[2y_i \log_e \frac{y_i}{\hat{y}_i} + 2(n_i - y_i)\log_e \frac{(n_i - y_i)}{(n_i - \hat{y}_i)}]^{1/2} \qquad (3.30)$$

where d_i is positive if $y_i \geq \hat{y}_i$ and negative if $y_i < \hat{y}_i$. The sum of the squares of the deviance residuals is the deviance of a logistic model fitted to binomial data, therefore the name "deviance residual". Under the assumption that the model is correct and if the n_i are sufficiently large, $D = \sum d_i^2$ has an asymptotic chi-square distribution with J - (k + 1) degrees of freedom, similar to the sum of squares of the Pearson residuals.

The **standardised deviance residuals** is defined by

$$r_{d_i} = \frac{d_i}{\sqrt{(1 - h_i)}} \qquad (3.31)$$

3.6.4 Likelihood residuals

The **likelihood residuals** measure the change in deviance between a model fitted to the complete set of data and a model fitted to $n - n_i$, $i = 1, 2, ..., J$ observations, where the n_i observations are assumed to have the same covariate pattern. In this way the covariate patterns that are poorly fitted by the model, can be identified. If $J \approx n$, derivation of the likelihood residuals are computationally intensive.

The change in deviance can be approximated by

$$h_i r_{P_i}^2 + (1 - h_i) r_{d_i}^2$$

a computationally less intensive procedure. The signed square root of this expression gives the likelihood residuals

$$r_{l_i} = sgn(y_i - \hat{y}_i)\sqrt{h_i r_{P_i}^2 + (1 - h_i) r_{d_i}^2} \tag{3.32}$$

In a similar way, the difference between the X^2-statistics (see section 3.6.2) can also be used as a residual. The difference between the X^2-statistics can be approximated by the square of the standardised Pearson residual.

3.6.5 The Hosmer-Lemeshow test

The summary measures of fit discussed so far are approximately chi-square distributed if the n_i, $i = 1, 2, ..., J$ are relatively large. If $J \approx n$, Hosmer and Lemeshow (Hosmer and Lemeshow, 1980 and Lemeshow and Hosmer, 1982) proposed grouping of the data, based on the estimated probabilities. The subjects are sorted by estimated probabilities, from the smallest to the largest values and using the percentiles as cut-points, subjects are sorted into g groups.

The Hosmer-Lemeshow goodness-of-fit statistic is given by

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n_k'\bar{p}_k)^2}{n_k'\bar{p}_k(1-\bar{p}_k)} \tag{3.33}$$

where n_k' is the number of subjects in the kth group, o_k is the observed frequency in the kth group and the average estimated probability is

$$\bar{p}_k = \sum_{i=1}^{n_k'} n_i\hat{p}_i / n_k'$$

By means of simulations, Hosmer and Lemeshow (1980) showed that, under the assumption that the model is correct and when J = n, the chi-square distribution with (g - 2) degrees of freedom gives a good approximation of the distribution of $\hat{C}$.

Other cut-points than the percentiles can also be used to define the groups, but the percentiles are the most widely used.

3.6.6. Classification tables

To construct a classification table, the outcome variable is cross-classified with a binary variable that is derived from the predicted probabilities. The binary variable is zero if the predicted probability is less than a cut-point, c (usually 0.5), and equal to one if the predicted probability exceeds c. The percentage correctly classified observations can give an indication of the ability of the model to discriminate between two groups. If the two groups are of unequal sizes, the classification table is biased to classify more observations in the larger group. Classification tables should be used only in conjunction with other measures of goodness-of-fit.

3.7 Diagnostics

3.7.1 Form of the linear predictor

In the previous paragraph summary statistics for the goodness-of-fit for a model were discussed. In this paragraph the individual components of the summary statistics will be briefly looked at. Most of these diagnostics are graphical summaries, rather than formal statistical tests and the experience of the analyst will to a great extent determine the success with which these diagnostics are used.

Plots of the residuals against the linear part of a logistic model can be very informative when binomial data are modelled. If the plot is not random, but rather shows a systematic pattern, it suggests that the model needs to be adjusted in some or other way.

A plot of the residuals against an independent variable not in the model can also be used. If a trend is present in the plot, the excluded variable should be added to the model. A trend in the plot of the residuals against an included variable is an indication that a higher order term, or a transformation of the independent variable, should be added to the model.

Plots of residuals against the linear part of the model, or against independent variables, are not informative when binary data are modelled. Independently of the fit of the model, these plots will form "clouds" on either side of the horizontal line that corresponds to a zero residual. This is the result of the fact that the deviance residual and the Pearson residual are negative if the observed y_i is zero, and positive if the observed y_i is unity.

The **half-normal plot** of residuals is very informative, also for binary data. To construct a half-normal plot, the absolute values of the standardised deviance residuals or the likelihood residuals are ordered in ascending order and plotted against $\Phi^{-1}[(i+n-\frac{1}{8})/(2n+\frac{1}{4})]$ where Φ^{-1} is the inverse of the standard normal distribution function. In addition, simulated envelopes (see Collett (1991); p. 129) are derived to form part of the plot. Residuals falling outside the limits of the constructed envelope, indicate a poor fit of the model.

3.7.2 Outliers

An **outlier** is an observation that is unexpectedly different from the other observations. An outlier can be the result of an observation that was wrongly recorded, but it can also represent a minority group in the population. It is therefore important to investigate the origin of an outlier.

Outliers can be identified in several ways. A plot of the residuals against the observation numbers, or an **index plot**, is one way of detecting observations that deviate from their predicted values under the fitted model. In a half-normal plot, outliers will appear on the top right of the plot and outside the envelope boundaries.

If one or more observations have been identified as outliers, their effect on the model can be assessed by fitting a model to the data after the outliers have been omitted. If this model does not deviate much from the original model, the outliers are not influential. If the outliers appear to be influential, however, the analyst has to go back to the origin of the outliers and decide whether the outliers should be included or excluded from the analysis, or whether the model should be revised to accommodate them. The opinion of experts in the field of application should play an important role in this decision.

3.7.3 Influential observations

An observation is **influential** if the inclusion or exclusion of it in the analysis causes considerable changes to the fit of the logistic regression model. An outlier can, but need not be, influential. An influential observation can influence the form of the fitted model to such an extent that it has a small residual, but with the result that some other observations have larger residuals than when the influential observation is omitted from the analysis.

An observations that is different from the other observation based on the values of the independent variables, is a potential influential observation. The ith diagonal term of the hat matrix $\boldsymbol{H}$, is often referred to as the leverage of observation i, and is a measure of the distance of the ith observation from the other observations. An index plot of the leverage values can indicate possible influential observations. A leverage can be considered high if it is greater than twice the average of the diagonal terms of $\boldsymbol{H}$.

The influence of a single observation can also be assessed through the change in deviance between the model fitted to the full data set and the model fitted to the data set excluding the observation. The square of the likelihood residual (see section 3.6) is an approximate measure of the change in deviance. In a similar way, the square of the standardised Pearson residual is an approximation of the change in the X^2-statistic, if the ith observation is omitted from the analysis. An index plot of either of these residuals will give an indication of the influence of each individual observation.

Prominent differences between the parameter estimates derived from the full data set and the parameter estimates derived from the reduced data set (omitting the ith observation), is an indication of an influential observation.

The influence of the *i*th observation on the parameter estimates can be approximated by

$$D_i = \frac{h_i r_{P_i}^2}{p(1-h_i)} \tag{3.34}$$

with r_{P_i} the *i*th standardised Pearson residual and $p = k+1$ the number of unknown parameters.

Relatively large values on the index plot of D_i correspond to influential observations.

The influence of the *i*th observation on the *j*th parameter estimate is also an important diagnostic, since an observation that influences the *j*th parameter will also influence the estimated odds ratio. The statistic measuring the effect of the *i*th observation on the *j*th parameter is referred to as a **delta-beta** and is given by

$$\Delta_i\hat{\beta}_j = \frac{(X'WX)^{-1}_{j+1}\, x_i(y_i - \hat{y}_i)}{(1-h_i)SE(\hat{\beta}_j)} \tag{3.35}$$

with $(X'WX)^{-1}_{j+1}$ the $(j+1)$th row of the variance-covariance matrix of the parameter estimates and x_i is the vector of independent variables for the *i*th observation.

If influential observations have been identified through the diagnostics, the first step would be to ascertain that it is not the result of measurement errors. If not, the influence of the observations on the model and inferences drawn from the model, should be assessed. Reporting the results of the analyses with and without the influential observation/s, is better practice than discarding the observations automatically.

CHAPTER 4

A SIMULATION STUDY OF THE EFFECT OF THR

Available THR data sets are too small and therefore inadequate for evaluation of the performance of logistic regression analysis with THR data. We therefore had to resort to a simulation study.

This section is divided into Part A and Part B. In Part A the effect of therapeutic riding is simulated for $M = 10$ samples of size $N = 50$, 100 and 200. Since the paired t-test is a popular technique by which therapeutic riding data are analysed, this test is used to analyse homogeneous subgroups in each sample. Further, a logistic regression model is fitted to each sample. The results of the paired t-tests and the logistic regression models are summarised and compared for the $M = 10$ samples.

In Part B a logistic regression model is again fitted to a simulated sample of size $N = 100$. Model selection, the estimated coefficients and regression diagnostics are discussed.

PART A

The Simulation Process and the Paired t-tests and the Logistic Regression Results

4.1 The Simulation Process

It was decided to simulate samples of children having a diagnosis of spastic type cerebral palsy (CP). Important independent variables that can influence a child's improvement, emerged from the literature. These variables are the age of the child, period of participation in therapeutic riding, the severity of the handicap, the intellectual ability

of the child and whether the child receives any other therapies. Though some of these variables might prove unimportant in practice, or perhaps (most probably) important independent variables were not included, it will be assumed for this study that the abovementioned variables are sufficient.

A summary of the basis on which the independent variables were simulated, is given in **Table 4.1**. The variables are described in the first column and the variable names are given in the second. The third and fourth columns respectively give the range of each variable and the associated probabilities by which the variables were simulated. Each of these variables will be discussed in more detail in section 4.4.

Table 4.1

A summary of the basis on which the independent variables were simulated.

DESCRIPTION	VAR NAME	RANGE	PROBABILITIES
Age of child	AGE	0 = 6 years or older, but younger than 9 years, 1 = younger than 6, but older than 4 years	P(AGE = x) = 0.2 x = 4, 5, 6, 7, 8
Intellectual ability of child	IQ	0 = IQ of more than 70, 1 = IQ of 70 or less	P(IQ = x) = 0.5 x = 0, 1
Frequency per week of other therapies received.	THER	frequency: 0, 1, 2, 3, 4, 5, 6, 7	P(THER = x) = 0.08 x = 0, 1, 5, 6, 7; P(THER = x) = 0.2 x = 2, 3, 4
Log_e(THER + 0.5)[5]	LTHER		
Time that child has been subjected to riding therapy	HORSE	0, 3, 6, 9, 12 months	P(HORSE = x) = 0.2 x = 0, 3, 6, 9, 12
Log_e(0.5 × HORSE)[6]	LHORSE		
Severity of handicap	SEV	0 = mildly severe, 1 = moderately severe	P(SEV = x) = 0.5 x = 0, 1
LHORSE × SEV	HSEV		

[5] Since $\log_e(0) = -\infty$, the constant 0.5 is added to THER.

[6] Since $\log_e(0) = -\infty$, the constant 1 is added to HORSE if a child did not ride during the assessment year.

4.2 Assumptions

1. The subjects: It is assumed that the simulated data include children having a diagnosis of spastic type cerebral palsy (CP) and having the ability to sit and stand alone with minimal support. Only children between the ages 4 and 8 are considered. Children with neurological or orthopaedic surgery within the past six months, psychiatric problems and other medical problems are assumed not to be included in the simulated samples. (MacKinnon, 1995b; Bertoti, 1988).

2. The riding program: It is assumed that those children who participate in THR, are receiving riding therapy twice a week for 30 minutes. Though children receive riding therapy at different centres, it is assumed that all the riding therapy centres practise hippotherapy and follow more or less the same program with the same objectives.

3. The assessment period: The improvement of a child one year after the first assessment, is considered. If a child has been doing therapeutic riding for a number of months during the assessment year, the assumption is made that it has been during the last consecutive months of the year.

4.3 Simulation of the Dependent Variables IMPR and DIF

The Bertoti postural scale (Bertoti, 1988) was used to assess improvement. **The aim of this study is not to suggest an adequate scale to be used, but rather to illustrate the use of logistic regression with any adequate scale.** The Bertoti scale was chosen, since it was designed to assess the posture of children with CP. This postural scale was used in both the Bertoti study (1988) and the MacKinnon et al study (1995b) to measure the effect of therapeutic riding on children with CP. Bertoti finds the scale a worthwhile measurement tool, since it seems to adequately reflect the clinical improvement seen and it is easy and quick to use.

From the interpretation of the scores of the 11 children in the Bertoti (1988) study (see section 2.1), it is reasonable to assume that a change in score of 2, is considered clinically meaningful. The change score is the difference in score for an individual before and after the therapeutic riding program. The constant DIFCRIT is used to indicate the magnitude of clinically meaningful change. Only if a child has a change score equal to DIFCRIT or more, will the change be considered meaningful.

Using the values of the simulated independent variables, the probability of improvement (p_j) was calculated for the jth child:

$$p_j = \frac{\exp(x_j)}{1+\exp(x_j)} \tag{4.1}$$

and

$$x_j = -0.75 + 0.7 \times AGE(j) - 0.75 \times IQ(j) + 1.0 \times LTHER(j) + 1.1 \times LHORSE(j)$$
$$-0.54 \times SEV(j) + 1.33 \times HSEV(j) + \epsilon_j$$

$$\epsilon_j \sim N(0,\ (1.5)^2).$$

Note that x_j is associated with the profile of the jth child.

For the simulation it was assumed that the observed change score for child j, indicated by DIF*(j), comes from a normal (DIF(j), 4^2) distribution. The value of σ is a reasonable reflection of the dispersion in the Bertoti (1988) data.

Since the probability of improvement for child j can be written as $p_j = P(DIF^*(j) > DIFCRIT)$ one arrives at $DIF(j) = DIFCRIT - 4 \times \Phi^{-1}(1 - p_j)$, using standardisation arguments and where Φ^{-1} is the inverse of the standard normal distribution function. This relationship is used to simulate the change scores of the children, where the estimated means represent the change scores of the children. To summarise, note that knowledge on a child's probability to improve (as given by (4.1)), was used to find an estimate of the mean change score of children with a profile as expressed by $\boldsymbol{x}_j$. This estimate was then used to represent the change score of the child.

The dichotomous dependent variable *IMPR* (meaning "improvement") assumes the value 0 if $DIF \geq DIFCRIT$ and the value 1 if $DIF < DIFCRIT$.

Keeping THER fixed at 0, **Figure 4.1** is a plot of the probability of improvement (p_j) against the number of months that a child has been riding (HORSE), for all possible combinations of AGE, IQ and SEV. **Figure 4.2** is similar to **Figure 4.1**, except that THER is fixed at 3. **Table 4.2** gives an explanation of the legends P1 - P8 used in the figures.

Table 4.2

An explanation of the legends P1 - P8 used in Figures 4.1 and 4.2.

AGE	IQ	SEV	LEGEND (P)
0	0	0	P1
0	0	1	P2
0	1	0	P3
0	1	1	P4
1	0	0	P5
1	0	1	P6
1	1	0	P7
1	1	1	P8

Figure 4.1

Plot of the probability of improvement against HORSE for THER = 0

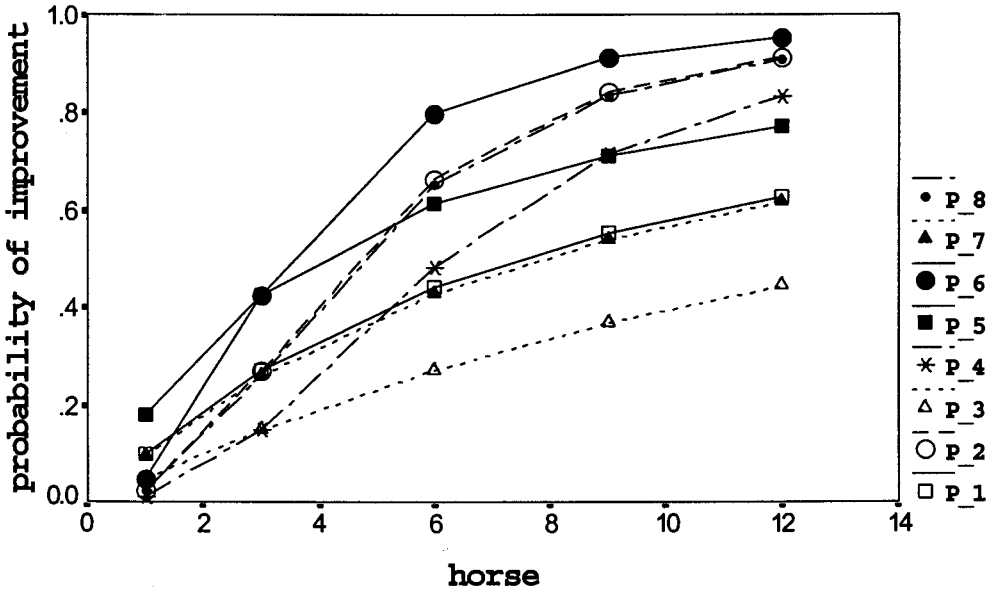
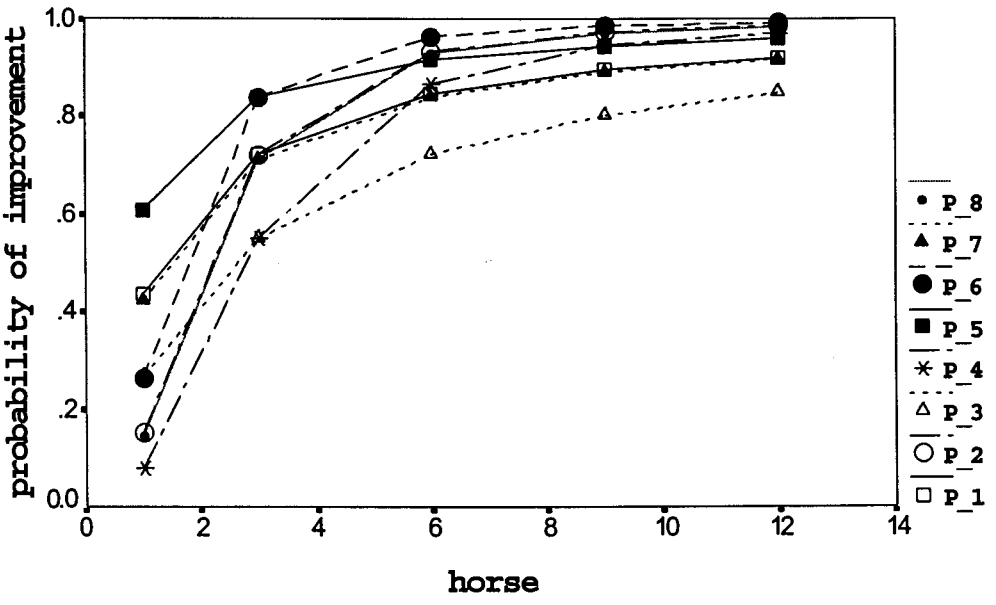


Figure 4.2

Plot of the probability of improvement against HORSE for THER = 3



If the time of riding therapy increases to infinity, the probability of improvement will converge to the maximum of 1 for the 8 combinations of AGE, IQ and SEV for both THER = 0 and THER = 3. From **Figures 4.1** and **4.2** it is clear that convergence takes place more gradually for THER = 0 than for THER = 3.

4.4 Simulation of the Independent Variables and Interpretation of the Logistic Regression Coefficients

4.4.1 Independent variable AGE

Children under the age of six years are probably more likely to improve physically than older children (MacKinnon et al, 1995b). This can perhaps be attributed partly to the fact that older children have longstanding postural habits and compensations. In the study conducted by Bertoti (1988), 11 children with spastic CP were considered. Five of the 6 children under the age of 5 years showed improvement. The youngest child (2 years and 4 months) did not show improvement. The lack of improvement was attributed to the child's fear of the horse and limited active participation in the therapeutic exercises. Of the 5 children older than 5 years, only 3 improved significantly.

It was decided to simulate samples of children between the ages 4 and 8 years only. The ages (in years) were simulated with equal probability of 0.2. The variable AGE is dichotomous, and assumed the value 1 if a child is younger than 6 years, and the value 0 if the child is 6 to 8 years. Age refers to the age of the child at the beginning of the assessment year.

The coefficient of the variable AGE included in the logistic regression model, is $\beta_1 = 0.7$. The interpretation of the coefficient is that, controlled for all other variables, the odds ratio for a child younger than 6 years versus a child of 6 years and older, is

$$\frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} = \exp(0.7)$$

$$= 2.014$$

where p_1 = P(improvement for a child younger than 6 years / **z**) and p_2 = P(improvement for a child older than 6 years / **z**). The vector **z** represents the variables adjusted for: IQ, LTHER, LHORSE and SEV. The odds for improvement for a child younger than 6 years is therefore twice the odds for improvement for a child older than 6 years.

4.4.2 Independent variable IQ

One of the inclusion criteria for the study conducted by MacKinnon et al (1995b), was normal intelligence (above 70 IQ). Bertoti (1988) included in the study only children with "normal intelligence as documented by a psychologist". Though different scales exist to measure intelligence, the IQ scale is well known and will serve the purpose for this study. According to a physiotherapist at the *Nuwe Hoop* School, Pretoria, children with normal intelligence will participate more actively in any therapeutic exercise or activity and will therefore improve more rapidly than children with lower intelligence.

The variable IQ is also a dichotomous variable, where 1 indicates a child with a IQ of 70 or less and 0 indicates a child having normal intelligence (above 70 IQ). The two possible outcomes of the IQ variable were simulated with equal probability, that is P(IQ = 0) = P(IQ = 1) = 0.5. The coefficient for IQ included in the model, is $\beta_2 = -0.75$.

For interpretation of β_2, consider the odds ratio for a child with low intelligence against a child with normal intelligence:

$$\frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} = \exp(-0.75)$$

$$= 0.472$$

where p_1 = P(improvement for a child with low intelligence / **z**) and p_2 = P(improvement for a child with normal intelligence / **z**). The variables AGE, LTHER, LHORSE and SEV are adjusted for and are represented by **z**. The odds of improvement for a child with low intelligence is therefore only half of the odds for a child with normal intelligence.

4.4.3 Independent variable LTHER

It is assumed that the more frequently a child receives any therapy, the more rapid the child's improvement will be. Since the outcome variable for this study concerns physical improvement, namely the improvement in posture, THER indicates the frequency per week of other physical therapies (excluding riding therapy) received by a child. THER was simulated to assume the values 0, 1, 2, 3, 4, 5, 6, 7. Let P(THER = x) indicate the probability that THER assumes the value x. Then P(THER = 0) = P(THER = 1) = P(THER = 5) = P(THER = 6) = P(THER = 7) = 0.08 and P(THER = 2) = P(THER = 3) = P(THER = 4) = 0.2. The log odds of improvement as a function of THER is assumed to rather follow a logistic curve than a linear curve. The log (to the base e) of the frequency is therefore used in the logistic model. Since $\log_e(0) = -\infty$, the constant 0.5 is added to THER to give LTHER = $\log_e$(THER + 0.5). A further assumption to be made, is that the therapies received by a child at the beginning of the assessment year, remain the same for the rest of the year.

The coefficient for LTHER is $\beta_3 = 1$ and p_1=P(improvement of a child receiving therapy $n + c$ times a week / **z**) and p_2 = P(improvement of a child receiving therapy n times a week / **z**) with **z** representing the variables AGE, IQ, LHORSE and SEV which are controlled for. The odds ratio for a child who receives therapy $n + c$ times a week, versus a child who receives therapy n times a week, is:

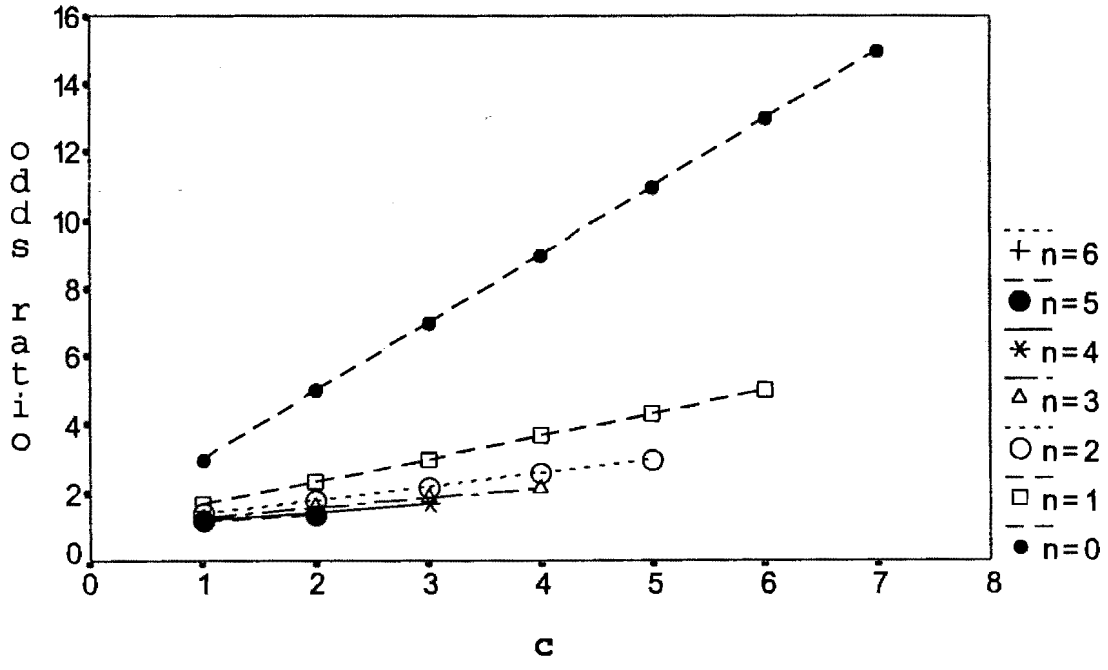
$$\frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} = \exp[\log(n+c+0.5) - \log(n+0.5)]$$

$$= \frac{n+c+0.5}{n+0.5}$$

Figure 4.3 illustrates the implication of $\beta_3 = 1$ for different values of c and n.

Figure 4.3

Plot of the odds ratio for different values of *n* and *c*.



Since $n + c$ ranges between 0 and 7, the range of c decreases for increasing values of n. From **Figure 4.3** it is clear that, for n fixed, the odds ratio is an increasing linear function of c. When n increases, the slope of the linear function of the odds ratio versus c, decreases. If all other variables are adjusted for, one can deduce from the scatter plot, for example, that a child who receives therapy 3 times a week, is 7 times more likely to improve within a year's time than a child who receives no therapy ($n = 0$, $c = 3$). A child who receives therapy 5 times a week, is 2 times more likely to improve than a child who receives therapy 3 times a week ($n = 3$, $c = 2$).

4.4.4 Independent variable HORSE

From the literature, most riding programs run for a fixed period of approximately 3, 6 or 9 months. Since improvement is continuous, a marked improvement can best be expected after a period of time. The variable HORSE that indicates the time (in months) that a child has been riding in the assessment year, is therefore simulated to assume the values 0, 3, 6, 9 and 12 with equal probability. As was the case with THER, the log odds of improvement as a function of HORSE is assumed to rather follow a logistic curve than a linear curve. Since $\log_e(0) = -\infty$, the constant 1 was added to HORSE if a child did not do any riding during the assessment year. The variable LHORSE is $\log_e(0.5 \times HORSE)$, and the coefficient used in the calculation of p_j is $\beta_4 = 1.1$. The interpretation of the coefficient of LHORSE will become clear when the interaction term, HSEV, is discussed later in the chapter.

4.4.5 Independent variables SEV and HSEV

Of the 11 children included in the Bertoti (1988) study, 8 were spastic diplegia and less involved than the 3 spastic qaudriplegia also included in the study. The children were not divided into a *diplegia group* and a *quadriplegia group* and tested separately, but Bertoti does mention that the diplegia showed an overall improvement, whereas the quadriplegia showed improvement only in certain areas of the postural scale. Bertoti

suggests that the quadriplegia might have demonstrated more dramatic and generalised improvement, had the program extended beyond the 10-week period.

The study conducted by MacKinnon et al (1995b) divided a group of 19 children into two groups: a mildly involved group and a moderately involved group. Children who could walk independently were classified as mildly involved, whereas children who used wheelchairs or assistive devices to walk independently, were classified in the moderately involved group. The two groups were then divided into a control group (not riding) and an experimental group (riding for 6 months). MacKinnon et al (1995b) used several measures or scales to measure change after 6 months. Change was measured in gross motor control, fine motor control, posture, activities of daily living and psychosocial changes. Analysis of variance was used to analyse the data, with the riding therapy and the severity (mild / moderate) considered as the factors. The Bertoti scale (1988) was used to assess posture. Though there was no significant difference in posture between the control group and the experimental group, change in posture differed between the mildly involved and the moderately involved groups. The moderately involved group showed positive changes after 6 months, whereas the mildly involved group showed a slight decline. Though the interaction between riding (control vs. experimental) and severity (mildly vs. moderately) was not significant for the Bertoti postural scale, the Peabody fine motor test for grasping, showed significant interaction between the two factors. For the simulation model, it was thus decided to include the interaction between severity and the time subjected to therapeutic riding.

The variable SEV is simulated to assume the values 0 and 1 with an equal probability, where 1 indicates a child who is moderately involved and 0 indicates a child who is mildly involved. Let p_1 = P(improvement for a moderately involved child / **z**, riding for n months) and p_2 = P(improvement for a mildly involved child / **z**, riding for n months). The variables AGE, IQ and LTHER are adjusted for and are represented by the vector **z**. From the suggestion of Bertoti (1988), the odds ratio should be smaller than unity for $n < 2.5$ (10 weeks), but might become greater than one after some time.

The odds ratio is indicated by ψ. In the simulated model, it is assumed that

$$\psi < 1 \text{ for } n < 3,$$

$$\psi = 1 \text{ for } n = 3 \text{ and } \psi > 1 \text{ for } n > 3.$$

Controlling for the other independent variables and setting the time of riding to 3 months, the log odds ratio for a child who is moderately involved versus a child who is mildly involved, is zero:

$$\beta_5 + \beta_6 \log(1.5) = 0$$

$$\beta_5 = -\beta_6 \log(1.5)$$

It was decided to choose $\beta_6 = 1.33$, since it gives realistic p_j values. If $\beta_6 = 1.33$ is chosen, then $\beta_5 = -0.54$. Controlling for the independent variables and setting p_1 and p_2 as above, the odds ratio is:

$$\frac{\dfrac{p_1}{1-p_1}}{\dfrac{p_2}{1-p_2}} = \exp(-0.54 + 1.33 \times LHORSE)$$

If n is the number of months a child has been riding with ψ the odds ratio, then **Table 4.3** illustrates that the odds ratio is smaller than unity for $n = 0$, unity for $n = 3$ and the odds ratio is larger than unity for $n = 6$, 9 or 12.

Table 4.3

The odds ratio of a moderately involved child versus a mildly involved child, for different values of *n* and keeping the other independent variables fixed.

n	ψ
0	0.583
3	1.000
6	2.512
9	4.308
12	6.316

Controlling for the other independent variables AGE, IQ and LTHER, represented by **z**, let p_1 = P(improvement for a moderately involved child riding for m months / **z**) and p_2 = P(improvement for a moderately involved child riding for n months / **z**). If n or m is zero, the constant 1 is added (for computational purposes) and it is also assumed that $n \leq m$. The odds ratio is

$$\frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} = \exp(1.1 \times \log_e(\frac{m}{n}) + 1.33 \times \log_e(\frac{m}{n}))$$

Similarly, let p_3= P(improvement for a mildly involved child riding for m months / **z**) and p_4 = P(improvement for a mildly involved child riding for n months / **z**). The definition of **z** is the same as for p_1 and p_2. The odds ratio is given by

$$\frac{\frac{p_3}{1-p_3}}{\frac{p_4}{1-p_4}} = \exp(1.1 \times \log_e(\frac{m}{n}))$$

If a child has not been doing any horse riding, the constant 1 is added to n and / or m. For n = 1 fixed, the odds ratios for moderately involved children and mildly involved children are plotted against m = 1, 3, 6, 9 and 12 in **Figure 4.4**. In **Figure 4.5**, n = 3 is fixed and the odds ratios for moderately involved children and mildly involved children, are plotted against m = 3, 6, 9 and 12.

Figure 4.4

Plot of the odds ratio for mildly involved children and moderately involved children for different values of *m* and *n* = 1 fixed.

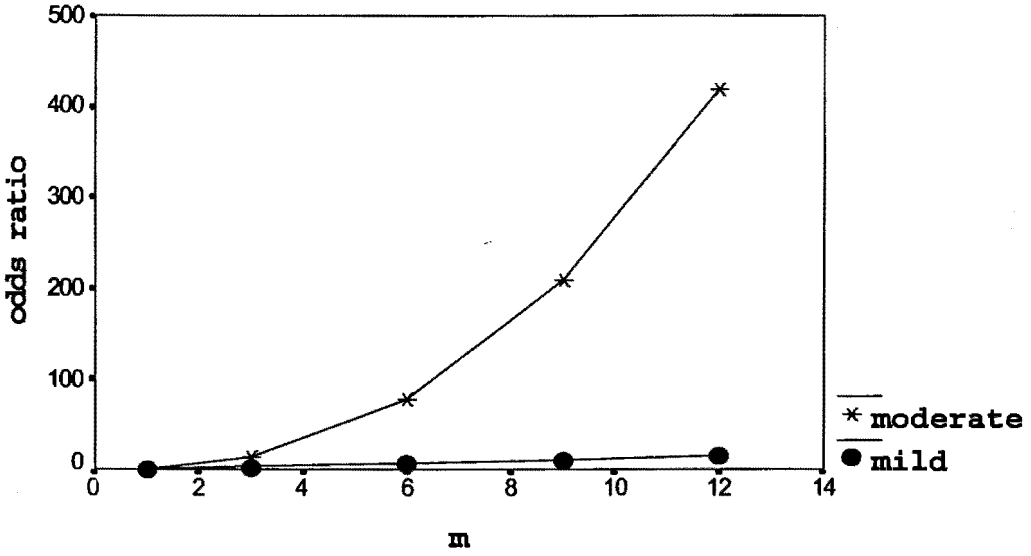
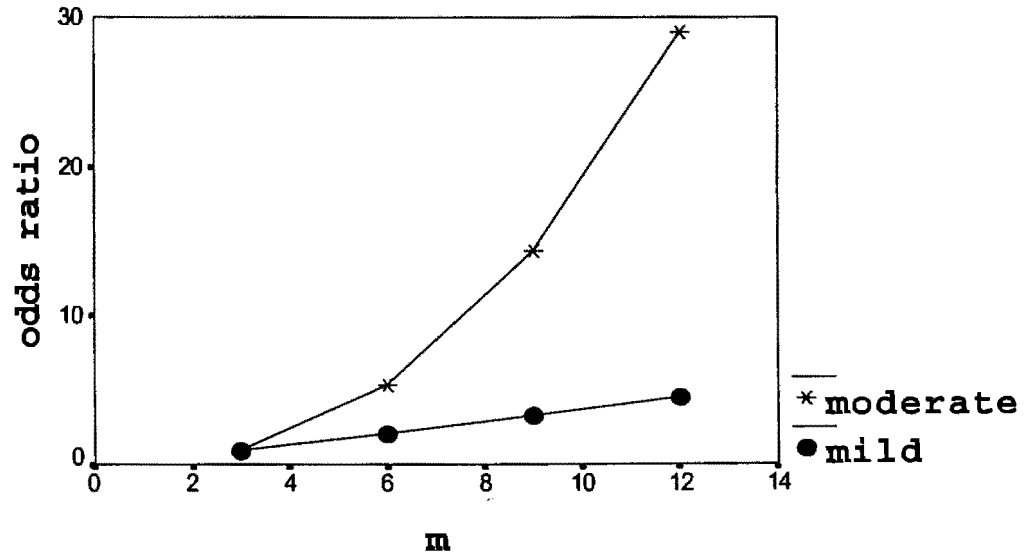


Figure 4.5

Plot of the odds ratio for mildly involved children and moderately involved children for different values of *m* and *n* = 3 fixed.



From **Figure 4.4** and **Figure 4.5** it is clear that the effect of the time a child has been riding on improvement, is influenced by the degree of a child's handicap. The odds ratios for a moderately involved child who has been riding for 9 or 12 months, versus a moderately involved child who has not been riding, are unrealistically high.

4.5 The SAS Program

In the SAS program, $M = 10$ samples of sizes $N = 50$, 100 and 200 were simulated (as described in sections 4.1 - 4.4). Each sample was divided into 10 homogeneous groups for the 10 different combinations of the variables HORSE and SEV. The variables HORSE and SEV were chosen as grouping variables, to simulate a situation similar to the one in the MacKinnon et al (1995b) study. If all the independent variables were used to group the data, it would result in very small homogeneous groups. Since the paired t-test is a popular statistical technique by which therapeutic riding data are analysed, this test was used to test the hypothesis that riding therapy had no effect, versus the alternative hypothesis of a positive effect, for each of the homogeneous groups.

For each of the $M = 10$ samples (of sizes $N = 50$, 100 and 200), a logistic regression model was fitted to the data, with the dichotomous variable, IMPR (see section 4.3), the dependent variable. From this, the estimated probabilities were found through equation (4.1). For each of the 10 estimated probabilities of improvement for a specific profile, the mean was compared to the true probability of improvement (calculated according to equation (4.1)), see **Tables 4.5b - 4.10b**.

A copy of the SAS program (PART A) can be seen in ADDENDUM B, p. II - V.

4.6 Results

The true coefficients and the mean estimated coefficients (for the 10 simulated data sets) for $N = 50$, 100 and 200 are given in **Table 4.4**. The standard deviations of the estimates (calculated over the 10 simulated data sets) are given in brackets next to the mean estimated coefficients. From the large mean estimates and the large standard deviations, it is clear that the estimates are unstable when $N = 50$.

Sample size $N = 50$ thus proved to be too small to fit a logistic regression model with 5 independent variables and one interaction term (see **Table 4.1**). For two of the samples, the SAS program gave the warning: "*There is possibly a quasi complete separation in the sample points. The maximum likelihood estimates may not exist.*" See section 3.5 for an explanation of *quasi complete separation* and *overdispersion*.

In spite of the warning, the logistic procedure continued to process the data and the results were based on the last maximum likelihood iteration. The mean predicted probabilities of improvement (see **Tables 4.5b - 4.10b**) were, however, not so much influenced by the overdispersion, since only two samples were overdispersed.

The mean estimated coefficients and the mean estimated probabilities of improvement for $N = 200$ are very similar to the results for $N = 100$. The gain in precision is therefore not proportional to the sample size.

Table 4.4

True coefficients and mean of the estimated coefficients for the 10 simulated data sets (for $N = 50$, 100 and 200) of the logistic regression model. Standard deviations are given in brackets

VARIABLE NAME	TRUE COEFFICIENT	N=50: MEAN ESTIMATED COEFFICIENT	N=100: MEAN ESTIMATED COEFFICIENT	N=200: MEAN ESTIMATED COEFFICIENT
AGE	0.70	1.002 (2.2)	1.253 (0.9)	0.972 (0.5)
LHORSE	1.10	10.058 (25.8)	1.446(0.5)	1.335 (0.3)
SEV	-0.54	5.272 (23.4)	-0.537(0.9)	-0.727 (0.5)
IQ	-0.75	-3.900 (9.6)	-1.153 (1.0)	-0.969 (0.7)
LTHER	1.00	4.188 (7.2)	1.671 (0.8)	1.363 (0.4)
HSEV	1.33	0.271 (19.1)	1.862 (0.7)	1.667 (0.4)

Two values for DIFCRIT were selected, namely 2 and 0.5. In **Tables 4.5a - 4.10a** the relative frequencies of significant results ($\alpha = 0.05$) for the paired t-tests are reported for several of the HORSE / SEV homogeneous groups and for the four combinations of DIFCRIT and N (DIFCRIT = 0.5; 2 and N = 100; 200).

Within each HORSE / SEV combination, the true probability of improvement (calculated according to equation (4.1)) and the mean estimated probability of improvement for the 8 combinations (AGE = 0, 1; IQ = 0, 1 and THER = 0, 3) are given (see **Tables 4.5b - 4.10b**). The means of the estimated probabilities are compared to the true probabilities. From the way in which the data were simulated, the results for DIFCRIT = 2 and DIFCRIT = 0.5 would be exactly the same. A distinction on DIFCRIT for the logistic regression results was therefore not made in **Tables 4.5b - 4.10b**.

Table 4.5a

Relative frequencies of significant paired t-test results ($\alpha = 0.05$) for the HORSE = 0, SEV = 0 homogeneous group, for the 4 combinations of N and DIFCRIT.

	DIFCRIT = 2	DIFCRIT = 0.5
$N = 100$	1/10	0/10
$N = 200$	2/10	0/10

Table 4.5b

True probabilities of improvement and the means of the estimated logistic regression probabilities for $N = 100$ and $N = 200$, with HORSE = 0 and SEV = 0 fixed, for the 8 combinations (AGE = 0, 1; IQ = 0, 1 and THER = 0, 3).

VARIABLES			PROBABILITIES OF IMPROVEMENT		
AGE	IQ	THER	True p	N=100: Mean estimated p	N=200: Mean estimated p
0	0	0	0.099	0.062	0.073
0	0	3	0.435	0.435	0.474
0	1	0	0.049	0.027	0.042
0	1	3	0.267	0.243	0.277
1	0	0	0.182	0.182	0.164
1	0	3	0.608	0.668	0.693
1	1	0	0.095	0.104	0.091
1	1	3	0.423	0.455	0.467

For $N = 100$, homogeneous subgroups were formed, resulting in sample sizes ranging from 7 to 17, which are comparable with sample sizes of previously conducted therapeutic riding studies. If homogeneous groups with regard to more than 2 independent variables were to be formed, the group sizes will become too small to conduct t-tests. When N was increased to 200, the HORSE = 0, SEV = 0 group sizes ranged from 16 to 27.

For samples of size $N = 100$ and for DIFCRIT = 2, the results of the paired t-test were non-significant ($\alpha = 0.05$) for 9 out of the 10 samples. Keeping $N = 100$ fixed and reducing DIFCRIT to 0.5, none of the tests gave significant results. Increasing N to 200, 2 out of 10 t-tests were significant (probably due to the larger size of homogeneous groups). The result for $N = 200$, DIFCRIT = 0.5 is the same as the result for $N = 100$, DIFCRIT = 0.5.

In contrast to the t-tests, the logistic regression model can accommodate more independent variables. Using logistic regression, the sample as a whole is analysed, and not separately as homogeneous groups. Keeping the HORSE / SEV combination fixed, the logistic regression model estimates the probability of improvement for all AGE × IQ × THER combinations. The estimated probabilities for $N = 100$ and $N = 200$ compare very well with the true probabilities.

Only for a child younger than 6 (AGE = 1), with an IQ of more than 70 (IQ = 0) who receives alternative therapy 3 times a week (THER = 3), was the true and estimated probabilities more than 0.5. For this profile, the probability of improvement is estimated to be larger (but not much larger) than the probability of no improvement. The results of the logistic regression and the t-tests are therefore compatible. The advantage of the logistic regression approach, however, is that the result is not restricted to a significant / nonsignificant outcome, but a probability of improvement is attached to each profile.

Table 4.6a

Relative frequencies of significant paired t-test results ($\alpha = 0.05$) for the HORSE = 0, SEV = 1 homogeneous group, for the 4 combinations of *N* and DIFCRIT.

	DIFCRIT = 2	DIFCRIT = 0.5
$N = 100$	0/10	0/10
$N = 200$	0/10	0/10

Table 4.6b

True probabilities of improvement and the means of the estimated logistic regression probabilities for $N = 100$ and $N = 200$, with HORSE = 0 and SEV = 1 fixed, for the 8 combinations (AGE = 0, 1; IQ = 0, 1 and THER = 0, 3).

VARIABLES			PROBABILITIES OF IMPROVEMENT		
AGE	IQ	THER	True p	N=100: Mean estimated p	N=200: Mean estimated p
0	0	0	0.025	0.013	0.011
0	0	3	0.152	0.144	0.133
0	1	0	0.012	0.005	0.005
0	1	3	0.078	0.051	0.054
1	0	0	0.049	0.060	0.029
1	0	3	0.265	0.375	0.287
1	1	0	0.024	0.020	0.012
1	1	3	0.145	0.163	0.124

Sample sizes for the HORSE = 0, SEV = 1 homogeneous groups ranged between 7 and 14 for $N = 100$ and for $N = 200$, the smallest sample size was 14, the largest was 29. The relative frequencies of significant results ($\alpha = 0.05$) for the t-tests are 0/10 for all four combinations of sample size (N) and DIFCRIT.

The mean estimated probabilities for $N = 100$ and $N = 200$ are in the same order of magnitude as the true probabilities. The probabilities of improvement for moderately handicapped children who have been doing no riding during the assessment year, with different profiles regarding AGE, IQ and THER, are less than the probabilities of no improvement. The logistic regression results thus confirm the t-test results, though the logistic regression results are more informative.

Table 4.7a

Relative frequencies of significant paired t-test results ($\alpha = 0.05$) for the HORSE = 3, SEV = 0 homogeneous group, for the 4 combinations of *N* and DIFCRIT.

	DIFCRIT = 2	DIFCRIT = 0.5
$N = 100$	9/10	5/10
$N = 200$	9/10	8/10

Table 4.7b

True probabilities of improvement and the means of the estimated logistic regression probabilities for $N = 100$ and $N = 200$, with HORSE = 3 and SEV = 0 fixed, for the 8 combinations (AGE = 0, 1; IQ = 0, 1 and THER = 0, 3).

VARIABLES			PROBABILITIES OF IMPROVEMENT		
AGE	IQ	THER	True p	N=100: Mean estimated p	N=200: Mean estimated p
0	0	0	0.270	0.184	0.235
0	0	3	0.721	0.773	0.792
0	1	0	0.148	0.083	0.136
0	1	3	0.550	0.541	0.594
1	0	0	0.426	0.396	0.429
1	0	3	0.839	0.900	0.903
1	1	0	0.260	0.229	0.252
1	1	3	0.711	0.771	0.790

Sample sizes for the HORSE = 3, SEV = 0 homogeneous groups varied between 7 and 20 for $N = 100$. For $N = 200$ the homogeneous group size range is 17 to 37.

The differences between the relative frequencies of significant results for the t-tests are interesting. For $N = 100$ and DIFCRIT = 2, the relative frequency of significant results is 9 out of 10. This frequency decreases to 5 out of 10 when DIFCRIT changes to 0.5. This phenomenon can be explained in terms of the power of the t-test (see section 2.1). The relative frequencies in **Table 4.7a** are approximations of the power of the t-test. For two studies of the same sample size, where the effect of THR is measured on different scales, the power of the test will be smaller for the study where the true change is indicated by a small shift on the scale, than for the study where a larger shift on a different scale indicates true change. One way of improving the power of a test, is to increase the sample size. For DIFCRIT = 2 , the relative frequency was again 9 out of 10 when the sample size was increased from $N = 100$ to $N = 200$. Keeping DIFCRIT = 0.5 fixed and increasing the sample by 100, the relative frequency for significant results increased from 5 out of 10 ($N = 100$ and DIFCRIT = 0.5) to 8 out of 10 ($N = 200$ and DIFCRIT = 0.5). Thus, increasing the sizes of the subgroups resulted in higher power of the t-test.

The results of the logistic regression will not change when the critical value (DIFCRIT) changes. This follows directly from the definition of the variable IMPR. In contrast to the t-tests, the results of the logistic regression approach do not depend directly on the magnitudes of the change scores (DIF), but rather on the magnitudes of the change scores relative to the magnitude of true change (DIFCRIT).

This can be regarded as one of the major advantages of logistic regression in comparison to t-tests. If the scale which is used to assess patients is of such a nature that a clinically meaningful change is indicated by only a small shift on the scale, the power of the t-test will be lower than when a different scale is used where a larger shift on the scale indicates clinically meaningful change. In the case of the first scale, the sample size will have to increase a great deal in order to improve the power of the t-test. When dealing with handicapped children, it is almost impossible to find a homogeneous group of sufficient size. **For the logistic regression approach, the dichotomous independent**

variable depends on the magnitudes of the changes relative to the defined minimum clinically meaningful change. Furthermore, the sample needs not be homogeneous with regard to the included independent variables.

The mean estimated probabilities (for $N = 100$ and $N = 200$) compare well with the true probabilities. Keeping HORSE = 3, SEV = 0 fixed, three profiles show a probability of improvement larger than the probability of no improvement: (AGE = 0, IQ = 0, THER = 3); (AGE = 1, IQ = 0, THER = 3) and (AGE = 1, IQ = 1, THER = 3). Two profiles reveal almost equal probabilities for improvement and no improvement: (AGE = 0, IQ = 1, THER = 3) and (AGE = 1, IQ = 0, THER = 0). Comparing **Table 4.7a** with **Table 4.7b**, one can see that the t-test and the logistic regression approach are compatible. This is evident in a fair mix of improvement probabilities and the estimated power not too close to zero or one.

Table 4.8a

Relative frequencies of significant paired t-test results ($\alpha = 0.05$) for the HORSE = 3, SEV = 1 homogeneous group, for the 4 combinations of *N* and DIFCRIT.

	DIFCRIT = 2	DIFCRIT = 0.5
$N = 100$	8/10	3/10
$N = 200$	9/10	5/10

Table 4.8b

True probabilities of improvement and the means of the estimated logistic regression probabilities for $N = 100$ and $N = 200$, with HORSE = 3 and SEV = 1 fixed, for the 8 combinations (AGE = 0, 1; IQ = 0, 1 and THER = 0, 3).

VARIABLES			PROBABILITIES OF IMPROVEMENT		
AGE	IQ	THER	True p	N=100: Mean estimated p	N=200: Mean estimated p
0	0	0	0.269	0.232	0.219
0	0	3	0.721	0.803	0.777
0	1	0	0.148	0.112	0.115
0	1	3	0.549	0.581	0.583
1	0	0	0.426	0.447	0.414
1	0	3	0.839	0.910	0.890
1	1	0	0.260	0.265	0.235
1	1	3	0.711	0.811	0.781

Homogeneous group sizes for the HORSE = 3, SEV = 1 homogeneous group ranged from 5 to 14 for $N = 100$. The increased homogeneous group sizes for $N = 200$ ranged from 11 to 24.

It is again striking that the relative frequencies of significant t-test results change for different combinations of N and DIFCRIT. For $N = 100$ and DIFCRIT = 2, a relative frequency of 8 out of 10 tests gave significant results. It is not surprising that this relative frequency increased to 9 out of 10 for $N = 200$. When DIFCRIT is changed to 0.5, there is an obvious decrease in the relative frequency of significant results. Of the 10 t-tests conducted for this homogeneous group, only 3 test results were significant. The relative frequency of significant results for $N = 200$, DIFCRIT = 0.5 is 5 out of 10 .

Keeping HORSE = 3 and SEV = 1 fixed, the same three profiles that showed improvement probabilities greater than 0.5 in the HORSE = 3, SEV = 0 group, are again showing probabilities greater than 0.5. Similar to the results of the HORSE = 3, SEV = 0 group, the profiles (AGE = 0, IQ = 1, THER = 3) and (AGE = 1, IQ = 0, THER = 0) are both border line cases. The t-test results and the logistic regression results compare well, since the probabilities of improvement are well distributed between zero and one and the estimates of the power of the test are not zero or one.

Table 4.9a

Relative frequencies of significant paired t-test results ($\alpha = 0.05$) for the HORSE = 6, SEV = 0 homogeneous group, for the 4 combinations of N and DIFCRIT.

	DIFCRIT = 2	DIFCRIT = 0.5
$N = 100$	10/10	7/10
$N = 200$	10/10	10/10

Table 4.9b

True probabilities of improvement and the means of the estimated logistic regression probabilities for $N = 100$ and $N = 200$, with HORSE = 6 and SEV = 0 fixed, for the 8 combinations (AGE = 0, 1; IQ = 0, 1 and THER = 0, 3).

VARIABLES			PROBABILITIES OF IMPROVEMENT		
AGE	IQ	THER	True p	N=100: Mean estimated p	N=200: Mean estimated p
0	0	0	0.442	0.340	0.415
0	0	3	0.847	0.900	0.901
0	1	0	0.272	0.167	0.251
0	1	3	0.723	0.751	0.778
1	0	0	0.614	0.568	0.629
1	0	3	0.918	0.956	0.956
1	1	0	0.429	0.376	0.416
1	1	3	0.840	0.904	0.904

With $N = 100$, the minimum homogeneous group size was 7 and the maximum was 14. The minimum group size for $N = 200$ was 13 and the maximum was 24.

For $N = 100$ and DIFCRIT = 2, 10 out of the 10 t-tests conducted, gave significant results. Changing DIFCRIT to 0.5, the relative frequency of significant results reduced by three. For $N = 200$, all the t-test conducted had significant results for both DIFCRIT = 2 and DIFCRIT = 0.5.

The mean estimated probabilities are quite satisfactory for $N = 100$ and $N = 200$. Except for 3 profiles, the probabilities of improvement are larger than the probabilities of no improvement. The trend towards higher probabilities of improvement supports the high relative frequencies of significant results for the t-tests.

Table 4.10a

Relative frequencies of significant paired t-test results ($\alpha = 0.05$) for the HORSE = 6, SEV = 1 homogeneous group, for the 4 combination of N and DIFCRIT.

	DIFCRIT = 2	DIFCRIT = 0.5
$N = 100$	10/10	9/10
$N = 200$	10/10	10/10

Table 4.10b

True probabilities of improvement and the means of the estimated logistic regression probabilities for $N = 100$ and $N = 200$, with HORSE = 6 and SEV = 1 fixed, for the 8 combinations (AGE = 0, 1; IQ = 0, 1 and THER = 0, 3).

VARIABLES			PROBABILITIES OF IMPROVEMENT		
AGE	IQ	THER	True p	N=100: Mean estimated p	N=200: Mean estimated p
0	0	0	0.665	0.612	0.662
0	0	3	0.933	0.976	0.963
0	1	0	0.484	0.410	0.452
0	1	3	0.868	0.912	0.913
1	0	0	0.800	0.774	0.818
1	0	3	0.966	0.992	0.983
1	1	0	0.654	0.619	0.650
1	1	3	0.930	0.974	0.962

Homogeneous group sizes between 5 and 11 were observed for $N = 100$. Though a sample of size 5 is very small to conduct a t-test on, it is not an unrealistic sample size for a therapeutic riding study. This group was therefore also included in the summarised t-test results. Samples of size $N = 200$ resulted in homogeneous groups that ranged in size between 13 and 22.

With the exception of the $N = 100$, DIFCRIT = 0.5 combination, the relative frequencies of significant results are 10 out of 10. The logistic regression results provide only one profile with an improvement probability of less than 0.5, in contrast to the 3 profiles with probabilities less than 0.5 identified in the HORSE = 6, SEV = 0 homogeneous group. This again illustrates the property of logistic regression to provide results containing more information than is possible for the t-tests.

Results were obtained in a similar way for the homogeneous groups HORSE = 9, SEV = 0; HORSE = 9, SEV = 1; HORSE = 12, SEV = 0 and HORSE = 12, SEV = 1. The results for these four groups are very similar and are therefore not provided in tabular format. For all four groups, the t-tests were significant 10 out of the 10 times for the different combinations of N and DIFCRIT. The logistic regression results do not contradict the t-test results, but are more informative. The logistic regression approach makes it possible to identify the single profile (a child who is older than 6 years with an IQ of less than 70 and who receives no additional therapy) in the HORSE = 9, SEV = 0 group with an improvement probability of less than 0.5. The logistic regression results also make it possible to identify the profiles which have an excellent chance to improve.

4.7 Conclusions

From the results in section 4.6, a sample of size $N = 100$ seems to be sufficient to fit a logistic regression model with 5 independent variables and one interaction term. A sample of size $N = 50$ is too small and can result in overdispersion. If the sample size is increased to $N = 200$, the estimated coefficients and the estimated probabilities of improvement are more accurate than the estimates for $N = 100$. The increase in accuracy is, however, not proportional to the increase in sample size. Since it is not always possible to obtain large samples of therapeutic riding data, a sample of size $N = 100$ seems to be a good choice.

The results for the paired t-test and the logistic regression analysis were not contradictory. The t-test results were, however, more influenced by changes in the sample sizes ($N = 100$, $N = 200$) and the magnitude of true change (DIFCRIT = 2, DIFCRIT = 0.5). By increasing the sample size, the group sizes of the homogeneous groups with regard to HORSE and SEV also increased. With increased group sizes, the estimated power of the t-test also increased. This can be seen more clearly in **Table 4.7a** and **Table 4.8a**. Changing the magnitude of true change (DIFCRIT) from 2 to 0.5, also influences the results of the t-tests. The practical interpretation of the DIFCRIT value is the smallest change that can still be regarded as clinically meaningful and this value should be specified by the therapist involved in the research. Decreasing the magnitude of true change thus resulted in smaller estimated power of the t-test (see **Table 4.7a** and **Table 4.8a**).

The advantages of the logistic regression approach is that the therapist, keeping in mind the nature of the measurement tool or scale being used to assess the patient, can specify the smallest change that can be regarded as clinically meaningful. The coding of the dichotomous outcome variable will depend on this critical value. If the assessment scale being used is of such a nature that a small change on the scale depicts meaningful

change, then the observed change, but also the critical value, will be small. The coding of the outcome variable in logistic regression will take the sensitivity of the scale into account, whereas the t-test will not be able to detect a meaningful change as statistically significant (except if the sample size is sufficiently large). In the logistic regression approach the expertise of the therapist is thus taken into account in the analysis.

Another advantage of the logistic regression approach is that it can control for the influence of external variables. Also, subjects included in the sample need not be homogeneous with regard to the independent variables included in the model. Selecting homogeneous groups with regard to only two independent variable (HORSE and SEV), resulted in the drastic decrease in sample sizes. This can be seen in the ranges of sample sizes for the homogeneous groups. The complete sample of size N can be used in the logistic regression analysis, whereas the sample is divided into small homogeneous groups before a t-test is conducted.

The logistic regression results are more informative than the t-test result and a probability of improvement can be attached to a specific profile. It is possible not only to conclude whether the mean change for a group of children was significant or not, but the probability of improvement can be predicted for a child with a specific profile.

PART B

Logistic Regression Results for a Simulated Sample of Size $N = 100$.

A sample of size $N = 100$ was simulated in the same way as was done in PART A. In fitting a logistic regression model, first of all a stepwise selection procedure was used to select from the variables AGE, LHORSE, SEV, IQ and LTHER those variables that should be included in the model. Only the variables LHORSE, LTHER and IQ were included. A second model, including all the independent variables (AGE, IQ, LTHER, LHORSE and SEV) was then fitted to the data. For the third model , the interaction term HSEV (LHORSE × SEV) was added.

4.8 Analysis

For the stepwise selection procedure a significance level of 0.05 was used for entry into the model and the significance level for staying in the model, was set to 0.1. At every step the next variable to be added to the model was determined by the significance of each variable not in the model, adjusted for the variables already included in the model. If an included variable did not meet the 0.1 level of significance to stay in the model after any step in the procedure, it was deleted from the model.

The variables LHORSE, LTHER and IQ were included in the model. No other variables met the 0.05 significance level for entry in the model. **Table 4.11** is a summary of the stepwise procedure.

Table 4.11

Summary of the stepwise procedure.

Step	Variable Entered	Variable Removed	Number In	Score Chi-Square	Pr > Chi-Square
1	LHORSE		1	37.9056	0.0001
2	LTHER		2	5.1595	0.0231
3	IQ		3	4.8637	0.0274

The joint effect of the three variables in the final model is given by the deviance, indicated by $-2\ Log\ L$ in the SAS output. The chi-square statistic and the p-value are also provided. In addition to the deviance, the Akaike Information Criterion (AIC), the Schwartz Criterion (SC) and the Score Statistic (Score) are also printed in the SAS output (see **Table 4.12**). Only the deviance will be considered.

Table 4.12

Model fitting information and testing the global null hypothesis BETA = 0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	118.652	75.586	.
SC	121.257	86.007	.
-2 LOG L	116.652	67.586	49.066 with 3 DF (p=0.0 001)
Score	.	.	43.254 with 3 DF (p=0.0001)

Since 116.652 - 67.586 = 49.066 has a chi-square distribution with 3 degrees of freedom, the hypothesis H_0: $\beta = 0$ is rejected. The three variables improve significantly on the intercept-only model.

Table 4.13 gives the maximum likelihood estimates of the parameters, the standard errors, the Wald chi-square statistics and the p-values of the Wald statistics.

Table 4.13

Analysis of the maximum likelihood estimates for Model 1.

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
INTERCPT	1	-0.7246	0.7231	1.0041	0.3163
LHORSE	1	2.3058	0.4786	23.2160	0.0001
IQ	1	-1.4656	0.6956	4.4388	0.0351
LTHER	1	1.0746	0.4662	5.3137	0.0212

The model selected by the stepwise procedure will be referred to as Model 1. From the literature, the variables AGE and SEV that were not included in the model by the stepwise procedure, are considered important variables. The model containing all 5 independent variables was therefore fitted to the data. This model will be referred to as Model 2. The deviance of Model 2 is 66.66. Under the null hypothesis that Model 1 and Model 2 do not differ significantly, the difference in the deviance for Model 1 and Model 2 will follow a chi-square distribution with 2 degrees of freedom. The calculated value of $G = 67.586 - 66.660 = 0.926$ has a p-value of 0.63 and therefore the likelihood ratio test does not suggest that Model 1 and Model 2 differ significantly. When the biological importance of the variables AGE and SEV is considered, however, Model 2 is preferred to Model 1.

Table 4.14 gives a summary of the maximum likelihood estimates when the variables AGE, IQ, LTHER, LHORSE and SEV are included.

Table 4.14

Analysis of the maximum likelihood estimates for Model 2.

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
INTERCPT	1	-1.2129	0.9457	1.6448	0.1997
AGE	1	0.1673	0.6509	0.0661	0.7971
LHORSE	1	2.2993	0.5112	22.0282	0.0001
SEV	1	0.5986	0.6616	0.8186	0.3656
IQ	1	-1.5690	0.7273	4.6547	0.0310
LTHER	1	1.2025	0.5088	5.5851	0.0181

For Model 3, the interaction term LHORSE × SEV (HSEV) is added. The deviance for Model 3 is 62.708. The likelihood ratio test for the difference between Model 2 and Model 3, yields a value of $G = 66.660 - 62.708 = 3.952$ and $P(\chi^2(1) > 3.952) = 0.047$. This demonstrates that the interaction term is significant and should be added to the model. Model 3 thus gives a significantly better fit to the data than Model 2 and Model 1.

Table 4.15 gives the estimated coefficients for Model 3.

Table 4.15

Analysis of the maximum likelihood estimates for Model 3.

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
INTERCPT	1	-1.3666	1.0462	1.7063	0.1915
AGE	1	0.4237	0.6935	0.3732	0.5413
LHORSE	1	1.7252	0.5729	9.0668	0.0026
SEV	1	-0.1802	0.7973	0.0511	0.8212
IQ	1	-1.6496	0.7557	4.7645	0.0291
LTHER	1	1.6608	0.6633	6.2695	0.0123
HSEV	1	1.8273	0.9872	3.4263	0.0642

The Wald statistics suggest that LHORSE, IQ and LTHER are significant at the 0.05 level of significance. These are the same variables that were selected by the stepwise

procedure. The interaction term is significant at the 0.1 level of significance. For the model to be hierarchically well-formulated (see section 3.5.2) when the interaction term is included, the variables LHORSE and SEV should also be included. Though AGE is not statistically significant, it is considered a biologically important variable and therefore remains in the model.

The importance of the interaction term is also supported by the fact that the estimated coefficients for the other variables changed considerably from Model 2 to Model 3. Especially the estimated coefficients of the two variables involved in the interaction term changed markedly. **Table 4.16** is a summary of the estimated coefficients for the three models.

Table 4.16

Summary of the true coefficients and the estimated coefficients for Models 1 to 3.

MODEL	INTER-CEPT	LHORSE	LTHER	IQ	AGE	SEV	HSEV
1	-0.7246	2.3058	1.0746	-1.4656			
2	-1.2129	2.3993	1.2025	-1.5690	0.1673	0.5986	
3	-1.3666	1.7252	1.6608	-1.6496	0.4237	-0.1802	1.8273
TRUE COEF-FICIENTS	-0.7500	1.1000	1.0000	-0.7500	0.7000	-0.5400	1.3300

When the data were simulated, the assumption was made that the probability of improvement is dependent on the variables AGE, LHORSE, SEV, IQ, LTHER and HSEV. It is therefore expected that Model 3 should give a significantly better fit than the other two models.

Table 4.17 is the classification table for Model 3 when a cut-off value of 0.5 is

used. The overall rate of correct classification is 82%. Of the observed *improvement* cases (outcome 0), 44% were incorrectly classified, while only 8.22% of the *no improvement* cases (outcome 1) were incorrectly classified. This illustrates the sensitivity of classification for the relative group sizes. This method will always favour classification into the larger group.

Table 4.17

The Classification Table.

	CLASSIFIED		
OBSERVED	**0**	**1**	**TOTAL**
0	15	12	**27**
1	6	67	**73**
TOTAL	**21**	**79**	**100**

The SAS LOGISTIC procedure applies a one-step approximation to obtain new estimates of the coefficients to reduce the bias resulting from classifying the same data set that was used to estimate the coefficients.

Hosmer and Lemeshow (1989) state that a classification table is not necessarily an accurate way to assess the fit of a model and that it should be used in addition to other measures of fit.

The diagnostics provided in the SAS output occupy several pages and will not be included in this dissertation, though we will discuss it briefly.

The SAS output produces index plots of the Pearson residuals and the deviance residuals. Very few observations have large residuals. The DIFDEV and DIFCHISQ diagnostics are also given in the SAS output. The DIFDEV diagnostic is the change in deviance between the model fitted to the full data set and the model fitted to the data set excluding one observation. Similarly, the DIFCHISQ diagnostic is the change in the

Pearson chi-square statistic. It is striking that the same observations identified as outliers by the index plots of the residuals, also give large values for the DIFDEV and DIFCHISQ diagnostics and can therefore be considered influential observations. Since J, the number of covariate patterns formed by the independent variables, is in the order of $N = 100$, the distributions of the diagnostics under the hypothesis that the model fits, are unknown. Interpretation of the diagnostics is therefore based on visual assessment rather than on the distributions of diagnostics. The CBAR diagnostic in the SAS output provides a measure of the influence of the individual observations on the estimated coefficients (see equation 3.34). Only observation 82, that was also identified by the DIFDEV and the DIFCHISQ diagnostics, seem to be influential. For this example the observations were simulated and investigation of the origin of these outliers is therefore unnecessary.

In addition, the SAS output also gives an index plot of the diagonal elements of the hat matrix and the DFBETA diagnostics. The DFBETA diagnostic is used to assess the influence of an observation on each estimated coefficient (see equation 3.35).

If the fit of the model has been established, the next step is to draw inferences from the estimated coefficients. The odds ratio for the various independent variables can be estimated from the estimated coefficients. For variables not involved in an interaction, the odds ratios are derived by taking the exponent of the estimated coefficients. **Table 4.18** gives the estimated odds ratios and the 95% confidence intervals of the estimated odds ratios, for the variables AGE, IQ and LTHER. The interval for LTHER is the 95% confidence interval for the estimated odds ratio for a change of 0.2 units in LTHER. The difference in LTHER for a child who receives therapy 5 times a week and a child who receives therapy 4 times a week, is 0.2.

Table 4.18

The estimated odds ratios for the variables AGE, IQ and LTHER.

VARIABLE	ESTIMATED ODDS RATIO	95% CI FOR THE ESTIMATED ODDS RATIO
AGE	1.528	(0.392, 5.947)
IQ	0.192	(0.044, 0.845)
LTHER	1.394	(1.075, 1.808)

The confidence intervals for the estimated odds ratios for LTHER and IQ both exclude one. Though the confidence interval of the estimated odds ratio for AGE includes one, the interval lies heavily to the positive side. We can therefore assume that each of the variables in **Table 4.18** affects the improvement of a child.

The estimation of the odds ratio in the presence of interaction, involves more than taking the exponent of the estimated coefficients.

The estimated odds ratios and the 95% confidence intervals for the odds ratios, for a moderately involved child versus a mildly involved child for HORSE = 0, 3, 6, 9 and 12, are provided in **Table 4.19**. To calculate the estimated odds ratios, we use the formula

$$\exp(\hat{\beta}_5 + \hat{\beta}_6 \times LHORSE) = \exp(-0.1802 + 1.8273 \times LHORSE)$$

The variance of $\hat{\beta}_5 + \hat{\beta}_6 \times LHORSE$ is given by

$$\hat{var}(\hat{\beta}_5) + \hat{var}(\hat{\beta}_6) \times (LHORSE)^2 + 2\hat{cov}(\hat{\beta}_5, \hat{\beta}_6) \times LHORSE$$

$$= 0.6356 + 0.9745 \times (LHORSE)^2 + 2(-0.3584) \times LHORSE$$

The variance / covariance matrix is provided in the SAS output.

The 95% confidence intervals for the estimated odd ratios are calculated as

explained in section 3.4 (see equation 3.26).

Table 4.19

The estimated odds ratios and 95% confidence intervals for the odds ratios, for a moderately involved child versus a mildly involved child, at the 5 levels of HORSE.

NUMBER OF MONTHS THAT CHILD HAS BEEN SUBJECTED TO RIDING	ESTIMATED ODDS RATIO	95% CI FOR THE ESTIMATED ODDS RATIO
0	0.2353	(0.02, 2.81)
3	1.7519	(0.43, 7.06)
6	6.2171	(0.86, 45.19)
9	13.0423	(0.97, 175.90)
12	22.063	(1.01, 483.18)

For HORSE = 0, mildly handicapped children are estimated to have a better chance of improvement than moderately handicapped children. This situation is reversed when the children start with riding therapy.

The effect of SEV on the odds of improving increases exponentially with the number of months that a child has been riding. The growing width of the confidence intervals indicates that there is considerable uncertainty in the estimated odds ratios, if the number of months that a child has been riding is more than six. A larger sample size is needed to estimate the odds ratio accurately.

Table 4.20 provides the estimates of the odds ratios for a child who has been riding for n months, versus a child who has been riding for m months ($n < m$) for both mildly handicapped and moderately handicapped children. The 95% confidence intervals of the odds ratios are given in brackets next to the estimated odds ratios.

Table 4.20

The estimated odds ratios for a child who has been riding for *n* months versus a child who has been riding for *m* months, for both mildly handicapped children and moderately handicapped children.

TIME RIDING (IN MONTHS)		ESTIMATED ODDS RATIO	
n	***m***	**MILDLY HANDICAPPED**	**MODERATELY HANDICAPPED**
0	3	6.65 (1.94, 22.84)	49.54 (6.66, 368.87)
0	6	22.00 (2.94, 164.55)	581.28 (22.00, 15 349.61)
0	9	44.29 (3.76, 522.47)	2454.41 (44.33, 136 000.39)
0	12	72.75 (4.47, 1185.01)	6820.11 (72.80, 639 388.24)
3	6	3.31 (1.25, 7.21)	11.73 (3.30, 41.65)
3	9	6.65 (6.41, 75.52)	49.54 (6.66, 368.87)
3	12	10.93 (2.30, 51.88)	137.66 (10.49, 1 733.47)
6	9	2.01 (1.25, 2.84)	4.22 (2.01, 8.86)
6	12	3.31 (1.52, 7.21)	11.73 (3.30, 41.65)
9	12	1.64 (1.18, 2.19)	2.78 (1.64, 4.70)

None of the estimated odds ratios are less than one and the confidence intervals exclude unity. These are indications that the fitted model confirms that riding therapy is an important factor for the improvement of both mildly and moderately handicapped children.

It is also illustrated by the odds ratios that the effect of riding therapy has a greater influence on moderately involved children than on mildly involved children, though the odds ratios for $n = 0, m = 3, 6, 9$ and 12 seem unrealistically high. Some of the confidence intervals are extremely wide, indicating uncertainty in the estimated odds ratios.

4.9 Conclusions

The analysis of the simulated sample of size $N = 100$ serves as a simplified illustration of how logistic regression can be used to analyse therapeutic horseback riding data.

From the literature five independent variables were identified for possible inclusion in the model. The stepwise procedure selected the variables LHORSE, LTHER and IQ for inclusion in Model 1. The excluded variables, AGE and SEV, are important variables and were therefore included in Model 2. Finally Model 3, including the 5 independent variables and one interaction term (HSEV), was fitted to the data. Model 3 gave a significantly better fit to the data than the other two models. The likelihood ratio test was used to compare the models.

The estimated odds ratios and confidence intervals of the odds ratios for the variables in Model 3, suggest the importance of the variables AGE, IQ and LTHER in the improvement of a child. The odds ratios for the variables which are involved in the interaction term, LHORSE and SEV, are more difficult to interpret. The interpretation is summarised in **Table 4.19** and **Table 4.20**. The wide confidence intervals for the odds ratios in **Tables 4.19** and **4.20** suggest that a larger sample size is needed to estimate the odds ratios more accurately.

CHAPTER 5
CONCLUSIONS

5.1 Conclusions

Research on the benefits of therapeutic riding plays an important role in the quest for recognition by professional, scientific and governmental institutions as a valid therapeutical method. Researchers experience many difficulties in this process.

From a THR research survey we concluded that researchers gave little thought to the appropriateness of the techniques used for analysis of THR data. The ambiguity of the qualitative results and the quantitative results was mentioned in several studies. This paradox can be attributed to the low power of the statistical tests. The power of a statistical test depends on the sample size, the variation of the data, the level of significance and the clinically relevant difference. Small sample sizes are a serious problem in THR research. Also, it is not always possible to include in a study subjects that are homogeneous, a fact which can cause large variation and thus low power.

The objective of this study is not to suggest which parameters should be measured in order to represent change, nor to suggest an adequate measurement tool, but rather to propose logistic regression as an alternative, or adjunctive, technique in the analysis of THR data.

The logistic regression approach has several advantages for the analysis of THR data.

- Logistic regression analysis accounts for the confounding effect of external variables.
- Subjects need not be homogeneous with regard to the variables included in the model (these variables define a subject's profile).
- Sample sizes can be increased if homogeneity is no longer a prerequisite for inclusion in the sample.
- Data from several studies can be pooled together, provided that the same measurement tool or scale was used to assess the patients.
- The dichotomous dependent variable depends on the magnitudes of the changes relative to the defined clinically meaningful change. This means that the expertise of the therapist determines what is to be interpreted as clinically meaningful and not a statistical test.
- The logistic regression coefficients can be interpreted in terms of the odds ratio.
- A probability of improvement is associated with each profile. Profiles with either a poor, or a very good probability to improve, can be identified. THR researchers have not begun to exploit this possibility.

In order to conduct a logistic regression analysis successfully, sample sizes must be adequately large. The subjects included in the sample need not be homogeneous, though. The multi-centre study conducted by Künzle et al (1994) between 1987 and 1992, included 255 patients. This study illustrates clearly the possibility to increase the sample size by extending the time of the study and involving several therapeutic riding centres. In order to enable several centres to cooperate, measurement apparatus should be inexpensive and easy to use.

In chapter 4, samples were simulated and in Part A t-tests and logistic regression analyses were conducted on the same data sets. Though it is not possible to directly

compare t-test results with logistic regression results, **the results of the two approaches were not contradictory.**

The analyses showed how sample sizes decreased when homogeneous groups were needed. **Tables 4.5a** to **4.10a** clearly illustrated how the estimated power of the t-test was influenced by sample size and the defined clinically meaningful change (DIFCRIT). It was not necessary to select homogeneous groups in order to conduct logistic regression analysis. Samples of size $N = 50$ proved to be too small to fit the logistic regression model including 5 independent variables and one interaction term. However, samples of sizes $N = 100$ and $N = 200$ were sufficient. The logistic regression results were influenced only by the magnitudes of the changes relative to the defined clinically meaningful change. The estimated improvement probabilities compared very well with the true probabilities **(Tables 4.5b** to **4.10b).**

In Part B a logistic regression model was fitted to a sample of size $N = 100$. The coefficients were estimated and interpreted in terms of the odds ratio. The analyses served as an example of how the logistic regression approach can be applied in practice.

5.2 Future Research

Time series and **growth curves** are other possible techniques to study intra-individual change. For analyses of both time series and growth curves, the estimated parameters can be regarded as representative of a person's individual growth pattern. These parameters can be subjected to multivariate analysis of variance when comparing the growth patterns of two or more groups (e.g. comparing children with different handicaps). Observing the same patient regularly over time has the advantage that the individual growth pattern can be investigated, though it is an intensive and time-consuming process.

This study focussed on the problems encountered when empirical evidence is needed to support the claims regarding the benefits of therapeutic riding. **The logistic regression approach can, however, be used in any field where the therapeutic value of an intervention has to be proven scientifically**. The remarks and suggestions made in this study thus have a wider application than only riding therapy.

I

ADDENDUM A

Extract from the Simulated Data Set

OBS	NUMBER	AGE	HORSE	LHORSE	IQ	SEV	THER	LTHER	IMPR
1	1	1	12	1.79176	1	0	2.5	0.91629	0
2	1	0	6	1.09861	0	1	6.5	1.87180	0
3	1	0	12	1.79176	0	1	4.5	1.50408	0
4	1	0	9	1.50408	0	0	3.5	1.25276	0
5	1	0	9	1.50408	0	1	0.5	-0.69315	0
6	1	0	9	1.50408	0	1	3.5	1.25276	0
7	1	0	6	1.09861	0	1	4.5	1.50408	0
8	1	0	6	1.09861	0	0	1.5	0.40547	0
9	1	1	12	1.79176	1	1	3.5	1.25276	0
10	1	0	3	0.40547	0	0	0.5	-0.69315	0
11	1	0	9	1.50408	1	0	3.5	1.25276	0
12	1	0	3	0.40547	0	0	4.5	1.50408	0
13	1	1	1	-0.69315	0	1	0.5	-0.69315	1
14	1	0	12	1.79176	0	1	4.5	1.50408	0
15	1	1	9	1.50408	1	1	5.5	1.70475	0
16	1	1	1	-0.69315	1	1	4.5	1.50408	1
17	1	0	3	0.40547	0	0	5.5	1.70475	1
18	1	0	6	1.09861	1	1	7.5	2.01490	0
19	1	1	9	1.50408	1	0	2.5	0.91629	1
20	1	1	9	1.50408	0	0	0.5	-0.69315	0
21	1	1	3	0.40547	1	0	4.5	1.50408	1
22	1	0	3	0.40547	0	0	4.5	1.50408	0
23	1	0	1	-0.69315	1	1	6.5	1.87180	1
24	1	0	9	1.50408	1	0	1.5	0.40547	1
25	1	1	3	0.40547	1	1	4.5	1.50408	1

II

ADDENDUM B

The SAS Program for PART A

```
OPTIONS LS=72;
DATA YZ;
DO J=1 TO 10;
DO I=1 TO 100;
 AGE1=3+RANTBL(15151,0.2,0.2,0.2,0.2,0.2);
  IF AGE1<6 THEN AGE=1;
  ELSE AGE=0;
 HORSE=3*RANTBL(12341,0.2,0.2,0.2,0.2,0.2)-3;
  IF HORSE=0 THEN HORSE=1;
 LHORSE=LOG(0.5*HORSE);
 IQ=RANTBL(15234,0.5,0.5)-1;
 SEV=RANTBL(82764,0.5,0.5)-1;
 THER=RANTBL(24622,0.08,0.08,0.2,0.2,0.2,0.08,0.08,0.08)-0.5;
  LTHER=LOG(THER);
 HSEV=LHORSE*SEV;
 E=1.5*RANNOR(23735);
 DIFCRIT=2;
 NUMBER=J;
 Y=-0.75+0.7*AGE+1.1*LHORSE-0.54*SEV-0.75*IQ+1.0*LTHER+1.33*LHORSE*SEV+E;
 P=EXP(Y)/(1+EXP(Y));
 X=PROBIT(1-P);
 PUT X;
 DIF=DIFCRIT-4*X;
 IF DIF >= DIFCRIT THEN IMPR=0;
 IF DIF < DIFCRIT THEN IMPR=1;
 OUTPUT;
END;
END;
PROC PRINT;
VAR NUMBER AGE HORSE LHORSE IQ SEV THER LTHER IMPR;
PROC SORT DATA=YZ;
BY NUMBER HORSE SEV;
PROC MEANS N MEAN STDERR T PRT;
VAR DIF;
BY NUMBER HORSE SEV;
OUTPUT OUT=T_TOETS MEAN=GEM STD=SF T=CRITVAL PRT=C ;
DATA M;
SET T_TOETS;
```

```
DATA A;
 SET M;
 IF HORSE=1 AND SEV=0;
  IF C<=0.05 AND GEM>0 THEN C=1;
  ELSE C=0;
 PROC FREQ;
 TABLES C;
 TITLE 'HORSE=1 AND SEV=0';
DATA B;
 SET M;
 IF HORSE=1 AND SEV=1;
  IF C<=0.05 AND GEM>0 THEN C=1;
  ELSE C=0;
 PROC FREQ;
 TABLES C;
 TITLE 'HORSE=1 AND SEV=1';
  DATA C;
 SET M;
 IF HORSE=3 AND SEV=0;
  IF C<=0.05 AND GEM>0 THEN C=1;
  ELSE C=0;
 PROC FREQ;
 TABLES C;
 TITLE 'HORSE=3 AND SEV=0';
DATA D;
 SET M;
 IF HORSE=3 AND SEV=1;
  IF C<=0.05 AND GEM>0 THEN C=1;
  ELSE C=0;
 PROC FREQ;
TABLES C;
  TITLE 'HORSE=3 AND SEV=1';
DATA E;
 SET M;
 IF HORSE=6 AND SEV=0;
  IF C<=0.05 AND GEM>0 THEN C=1;
  ELSE C=0;
 PROC FREQ;
TABLES C;
TITLE 'HORSE=6 AND SEV=0';
DATA F;
 SET M;
 IF HORSE=6 AND SEV=1;
  IF C<=0.05 AND GEM>0 THEN C=1;
```

```
 ELSE C=0;
PROC FREQ;
TABLES C;
TITLE 'HORSE=6 AND SEV=1';
DATA G;
SET M;
IF HORSE=9 AND SEV=0;
 IF C<=0.05 AND GEM>0 THEN C=1;
 ELSE C=0;
PROC FREQ;
TABLES C;
TITLE 'HORSE=9 AND SEV=0';
 DATA H;
SET M;
IF HORSE=9 AND SEV=1;
 IF C<=0.05 AND GEM>0 THEN C=1;
 ELSE C=0;
PROC FREQ;
TABLES C;
TITLE 'HORSE=9 AND SEV=1';
DATA K;
SET M;
IF HORSE=12 AND SEV=0;
 IF C<=0.05 AND GEM>0 THEN C=1;
 ELSE C=0;
PROC FREQ;
TABLES C;
TITLE 'HORSE=12 AND SEV=0';
DATA L;
SET M;
IF HORSE=12 AND SEV=1;
 IF C<=0.05 AND GEM>0 THEN C=1;
 ELSE C=0;
PROC FREQ;
TABLES C;
TITLE 'HORSE=12 AND SEV=1';
PROC SORT DATA=YZ;
BY NUMBER;
PROC LOGISTIC  OUTEST=BETAS;
MODEL IMPR=AGE LHORSE SEV IQ LTHER HSEV;
BY NUMBER;
OUTPUT OUT=IMPROVE;
PROC PRINT DATA=BETAS;
DATA N;
```

```
SET BETAS;
PROC UNIVARIATE;
VAR  INTERCEP AGE LHORSE SEV IQ LTHER HSEV;
DATA O;
SET BETAS;
AGE_IN1=0;
HORS_IN1=1;
SEV_IN1=0;
IQ_IN1=0;
THER_IN1=0;
Y1=-0.75+0.7*AGE_IN1+1.1*LOG(0.5*HORS_IN1)-0.54*SEV_IN1-
0.75*IQ_IN1+1.0*LOG(THER_IN1+0.5)+1.33*LOG(0.5*HORS_IN1)*SEV_IN1;
P1=EXP(Y1)/(1+EXP(Y1));
Z1=INTERCEP+AGE*AGE_IN1+LHORSE*LOG(0.5*HORS_IN1)+SEV*SEV_IN1+IQ*IQ_IN1+LTHER*LOG(THER_I
N1+0.5)+HSEV*LOG(0.5*HORS_IN1)*SEV_IN1;
T1=EXP(Z1)/(1+EXP(Z1));
OUTPUT;
PROC PRINT;
VAR P1;
PROC UNIVARIATE;
VAR T1;
RUN;
```

REFERENCES

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

Altman, D.G. (1991). *Practical Statistics for Medical Research*, Chapman and Hall.

Basile, R.B. (1997). The psychological effects of equine facilitated psychotherapy on behaviour and self-esteem in children with attention deficit/hyperactivity disorder (ADHD). *Scientific Journal of Therapeutic Riding*, **3,** 10-15.

Bertoti, D.B. (1988). Effect of therapeutic horseback riding on posture in children with cerebral palsy. *Physical Therapy*, **68:10**, 1505-1512.

Biery, M.J. (1985). Riding and the Handicapped. *Veterinary Clinics of North America: Small Animal Practice*, **15:2**, 345-354.

Biery, M.J. and Kauffman, N. (1989). The effects of therapeutic horseback riding on balance. *Adapted Physical Activity Quarterly*, **6**, 221-229.

Box, G. E. P. and Tidwell, P. W. (1962). Transformation of the independent variables. *Technometrics*, **4**, 531-550.

Brown, J. B. and Tebay, J. M. (1997). International survey of education and training for the Federation Riding for the Disabled International - A summary of responses. In: *Proceedings of the 9th International Therapeutic Riding Congress*; 1997; Denver, USA.

Burr, J. A. and Nesselroade, J. R. (1990). Change measurement. In: Von Eye, A., *Statistical Methods in Longitudinal Research, Volume 1: Principles and Structuring Change*, Academic Press, Inc., San Diego.

Campbell, K. C. (1990). Efficacy of Physical Therapy in Improving Postural Control in Cerebral Palsy. *Pediatric Physical Therapy*, **90:203**, 135-140.

Cattell, R. B. (1952). The three basic factor-analytic research designs - their interrelations and derivatives. *Psychol. Bulletin*, **9**, 99-121.

Collett, D. (1991). *Modelling Binary Data*, Chapman and Hall, London.

Collins, L. M. (1991). Measurement in longitudinal research. In: Collins, L. M. and Horn, J. L., *Best Methods for the analysis of Change*, American Psychological Association, Washington DC.

Coopersmith, S. (1984). *Self-Esteem Inventory*. Palo Alto, CA: Consulting Psychologist Press, Inc.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman Hall, London.

Cratty, B. J. (1967). *Developmental sequences of perceptual-motor tasks: Movement activities for neurologically handicapped and retarded children and youth*, Peek Publishers, Palo Alto.

Dismuke, R. (1984). Handicapped riding. *The Quarter Horse Journal*, 34-37.

Draper, N. and Smith, H. (1981). *Applied Regression Analysis*, John Wiley and Sons Inc., New York.

Exner, G., Engelmann, A., Lange, K. and Wenck, B. (1994). Grundlagen und Wirkungen der Hippotherapie im Konzept der umfassenden Behandlung querschnittgelähmter Patienten. *Rehabilitation*, **33**, 39-43.

Fox, V. M., Lawlor, V. A. and Luttges, M. W. (1984). Pilot study of novel test instrumentation to evaluate therapeutic horseback riding. *Adapted Physical Activity Quarterly*, **1**, 30-36.

Freeman, G. (1984). Therapeutic horseback riding. *Clinical Management*, **4:3**, 20-25.

Heipertz, W. Translated by Takeuchi, M. (1989). *Therapeutic Riding: Medicine, Education, Sports*. Originally published in Germany as *Therapeutisches Reiten*, (1977), National Printers (Ottawa) Inc.

Hosmer, D.W. and Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics*, A10, 1043-1069.

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

Kleinbaum, D.G. (1994). *Logistic Regression: A Self-Learning Text*, Springer-Verlag, New York.

Koch, S. (1994). Therapeutic riding with (secondary neurotic) children suffering from dyslexia. In: *Proceedings of the 8th International Therapeutic Riding Congress*, January 17-20, 1994, New Zealand, 188-190.

Kulichova, J. and Zenklova, J. (1996). The influence of horseback riding under the supervision of a sports medicine doctor, on the posture of children and young adults. *Scientific Journal of Therapeutic Riding*, **2**, 19-24.

Künzle, U., Egli, R. S. and Yasikoff, N. (1994). Hippotherapy-K: The healing rhythmical movement of the horse for patients with multiple sclerosis. In: *Proceedings of the 8th International Therapeutic Riding Congress*, January 17-20, 1994, New Zealand.

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.

Machin, D. and Campbell, M. J. (1987). *Statistical Tables for the Design of Clinical Trials*, Oxford.

MacKay-Lyons, M., Conway, C. and Roberts, W. (1988). Effects of therapeutic riding on patients with multiple sclerosis: a preliminary trial. *Physiotherapy Canada*, **40:2**, 104-109.

MacKinnon, J. R., Noh, S., Laliberte, D., Lariviere, J. and Allen, D. E. (1995a). Therapeutic horseback riding: a review of the literature. *Physical and Occupational Therapy in Pediatrics*, **15:1**, 1-15.

MacKinnon, J. R., Noh, S., Lariviere, J., MacPhail, A., Allen, D. E. and Laliberte, D. (1995b). A study of therapeutic effects of horseback riding for children with cerebral palsy. *Physical and Occupational Therapy in Pediatrics*, **15:1**, 17-34.

Mallows, C. L. (1964). Choosing variables in a linear regression: A graphical aid. *Presented at the Central Regional Meeting of the IMS*, Manhatten, Kansas, May 7-9, 1964.

Quellmalz, D. S. (1735). *Anweisung zu einer der Gesundheit dienlichen neu erfundenen Art der Bewegung*, Teubner, Leipzig.

Riede, D. Translated by Dusenbury, A.C. (1988). *Physiotherapy on the Horse*, Therapeutic Riding Services, Madison. Originally published in Germany as *Therapeutisches Reiten in der Krankengymnastik*, (1986), Richard Pflaum Verlag, Munich.

Ricotti, S., Citterio, D. N., Alfonsi, E., Bilucaglia, E., Carenzio, G. and Dalla Toffol, E. (1991). Horse riding therapy: neuro-motor rehabilitation. In: *Proceedings of the 7th International Therapeutic Riding Congress*, August 12-15, 1991, Denmark, 91-103.

Rogosa, D. (1995). Myths and methods: "Myths about longitudinal research" plus supplemental questions. In: Gottman, J. M., *The Analysis of Change*, Lawrence Erlbaum Associates, Inc., New Jersey.

Rufus, S. D. (1997). The effect of horse riding therapy on the self-concept of learning disabled children. *Dissertation in partial fulfilment of the requirements for the degree Master of Science in Clinical Psychology at the Medical University of South Africa.*

SAS Institute Inc. (1989). *SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 2: The LOGISTIC Procedure,* Cary, NC: SAS Institute Inc.

Scheidhacker, M., Bender, W. and Vaitl, P. (1991). Die Wirksamkeit des therapeutischen Reitens bei der Behandlung chronisch schizophrener Patienten. *Der Nervenarzt*, **62**, 283-287.

Scheidhacker, M. (1996). Not a claim of a cure, but of amelioration of symptoms and an improvement of quality of life. *Scientific Journal of Therapeutic Riding*, **2**, 40-48.

Schmitz, B. (1990). Univariate and multivariate time-series models: The analysis of intraindividual variability and intraindividual relationships. In: Von Eye, A., *Statistical Methods in Longitudinal Research, Volume 2: Time Series and Categorical Longitudinal Data*, Academic Press, Inc., San Diego.

Snir, D., Olin, R., Avalon, A., Yazdi, O. and Inbar, O. (1988). Effects of therapeutic riding on disabled children. In: *Proceedings of the 6th International Therapeutic Riding Congress*, August 23-27, 1988, Canada, 71-88.

Thissen, D. and Bock, R. D. (1990). Linear and nonlinear curve fitting. In: Von Eye, A., *Statistical Methods in Longitudinal Research, Volume 2: Time Series and Categorical Longitudinal Data*, Academic Press, Inc., San Diego.

Van Dyk, E., Odendaal, J. and Botha, L. (1994). Horse riding for cerebral palsy - A case report. *Companion - The Human/Animal Contact Study Group*, **11:3**, 8-17.

Von Arbin, C. (1994a). Evaluation of the importance of riding for disabled children. In: *Proceedings of the 8th International Therapeutic Riding Congress*, January 17-20, 1994, New Zealand, 184-187.

Von Arbin, C. (1994b). Intensive training in riding with disabled children - importance for proficiency in riding and possible integration. In: *Proceedings of the 8th International Therapeutic Riding Congress*; January 17-20, 1994; New Zealand;
191-194.

Webster, A., Pfotenhauer, M., David, E., Leyerer, U., Rimpau, W., Aldridge, D., Reissenweber, J. and Fachner, J. (1995). Registration and evaluation of effects of hippotherapy with patients suffering from multiple sclerosis by means of electromyography and acceleration measurement. *Scientific Journal of Therapeutic Riding*, **1**, 33-46.

Would, J. (1996). Study on posture and the development of balance in disabled riders. *Scientific Journal of Therapeutic Riding*, **2**, 3-18.

Yack, H. J., Bartels, C., Irlmeier, J., Lehan, A., Voyles, H., Haladay, K. and Daly, C. (1997). The effects of therapeutic horseback riding on the quality of balance control in children with attention disorders. *Scientific Journal of Therapeutic Riding*, **3**, 3-9.