

ISSN 1361-9802



**Manchester Centre
for
Statistical Science**



Technical Report 2001/01

**Diagnosis of Carpal Tunnel Syndrome
using Logistic Regression**

**Lectures given at the Facultad de Matemáticas
Universidad Autónoma de Yucatán
Mérida, Yucatán, Mexico**

12–15 February, 2001

Stephan M. Rudolfer

March 2001

**University of Manchester
Manchester Centre for Statistical Science
Technical Reports**

Available by post from the author
Department of Mathematics, University of Manchester, Manchester M13 9PL, U.K.
or from his webpage
<http://www.maths.man.ac.uk/DeptWeb/Homepages/smr/>

Contents

1	Acknowledgements	0
2	Carpal Tunnel Syndrome (CTS) Dataset	1
2.1	(ORDINAL) Response Variable	2
2.2	Predictor Variables	2
2.2.1	History	2
2.2.2	Clinical Signs	4
2.2.3	Nerve Conduction Studies	5
2.3	Definition of Carpal Tunnel Syndrome	9
2.4	Cause of Carpal Tunnel Syndrome	10
2.5	Computational Aspects	10
2.6	Source of the dataset	10
3	Logistic regression models fitted to the data	11
3.1	Aim of diagnostic (predictive) modelling	11
3.2	Proportional Odds (PO) Models	12
3.2.1	Properties of PO Models	13
3.3	Continuation Ratio (CR) Models	15
3.3.1	Differences between the forward and backward CRs	18
3.4	Comparison of CR and PO Models	18
3.5	Design (Dummy) Variables for Categorical Variables	18
3.6	Classification Procedures	19
3.6.1	Highest Probability (HP)	19
3.6.2	Anderson and Philips (AP) for the PO model	20
4	General methodology for developing clinical prediction models for ordinal data	20
4.1	Preliminary Steps	20
4.1.1	Define the response variable Y and predictor variable \mathbf{x} , after consultation with a medical expert.	20
4.1.2	Choose <i>interactions</i> carefully.	20
4.2	Data Reduction	21
4.2.1	Methods ignoring Y	21
4.2.2	Variable Selection	21
4.3	Verify the Model's Assumptions	21
4.3.1	Linearity of continuous predictors	21
4.3.2	Additivity of predictors (no interaction)	22
4.3.3	Predictors' Distribution	22
4.3.4	Influential Observations	22
4.4	Fit the Model	22

4.5	Compute measures of <i>predictive accuracy</i>	22
4.6	Validate the model	22
4.6.1	Resubstitution	22
4.6.2	Data splitting	23
4.6.3	Cross-validation	23
4.6.4	Bootstrap	23
5	Results	24
5.1	Full Model	24
5.2	Forward Selection of Variables	24
5.3	Backward Elimination of Variables	25
5.4	PCA Selection of Variables	25
5.5	Conclusions	25
6	Open Problems/Further Work	26
7	References	27

1 Acknowledgements

Firstly, I would like to thank Dr. Luis A. Rodriguez Carvajal, Director of the Faculty of Mathematics, for inviting me to give the talks, and the Autonomous University of Yucatan, Merida, for its financial support.

Secondly, I extend my most grateful thanks to Humberto Cervera Rossado for the excellent and difficult job he did of simultaneous translation of my talks.

My thanks also go to Peter Blore, Graphics Designer, and Dr. Chris Paul, Systems Administrator, for their invaluable assistance with the preparation of my talks as a technical report and a web page.

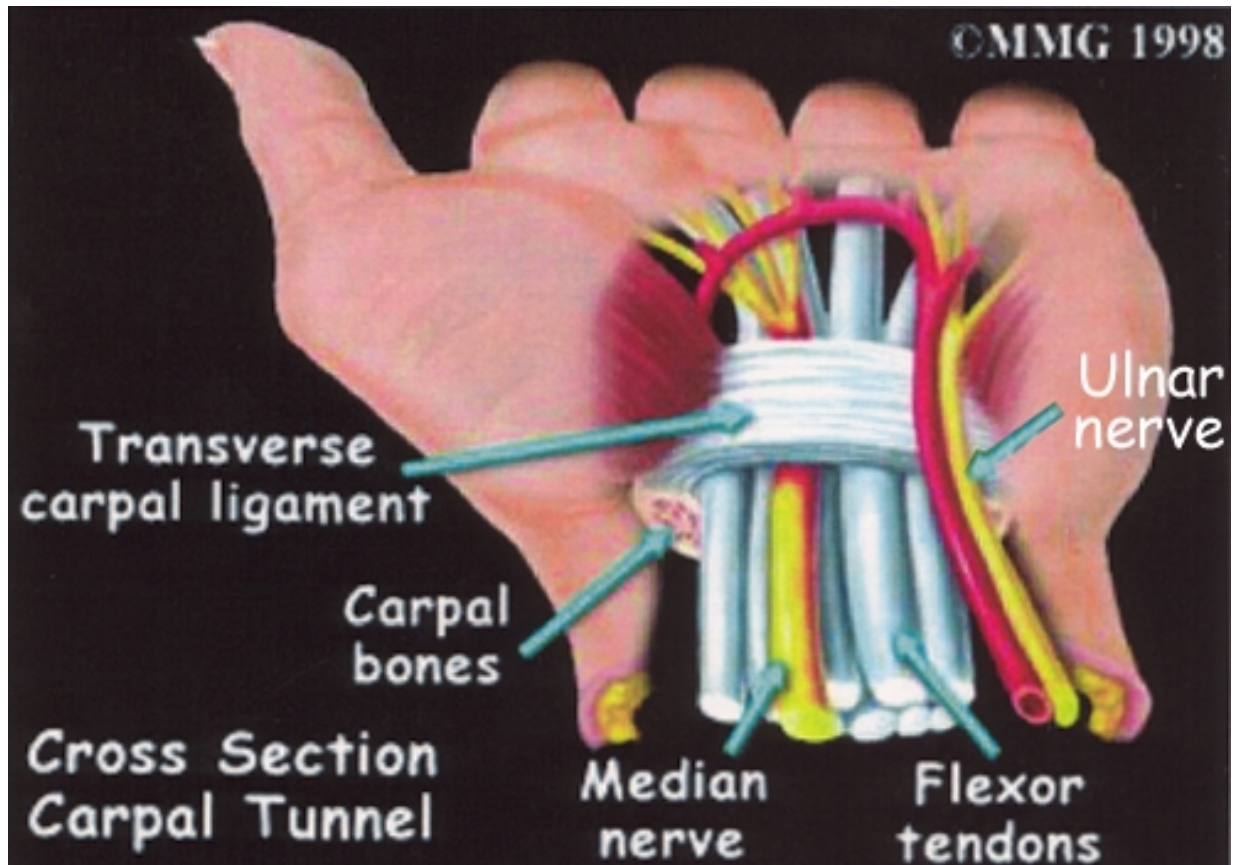
I owe a deep debt of gratitude to the late Dr. John L. James, Consultant Physician, St Luke's Hospital, Huddersfield, Yorkshire, England. He introduced me to Carpal Tunnel Syndrome, and gave consistently helpful and encouraging input to my research in this area with him. The dataset used for these talks would never have seen the light of day without his immense cooperation. I am also profoundly grateful to Dr James' secretary, Mrs Sylvia Hague, and to his technicians, Miss Gill Lockwood, Mrs Ginny Lockwood and Mrs Louise Mullinger, as well as to Martyn Thompson, a more recently appointed technician, for their great patience and cooperation in the acquisition and evaluation of the data.

Lastly, but not leastly, my deepest appreciation goes to my wife, Christine, for her support and encouragement during the preparation and delivery of these lectures.

2 Carpal Tunnel Syndrome (CTS) Dataset

CTS = cluster of certain hand symptoms (to be specified later)

cause = entrapment of the median nerve in the Carpal Tunnel at the wrist



An excellent and comprehensive account of CTS is given in Rosenbaum & Ochoa (1993).

2.1 (ORDINAL) Response Variable

$$Y = \begin{cases} 1 : \text{No Abnormality Detected (NAD)} \\ 2 : \text{Mild CTS} \\ 3 : \text{Moderate CTS} \\ 4 : \text{Severe CTS} \end{cases}$$

Important property of ordinal Y : the event $\{Y \leq j\}$ is defined.

2.2 Predictor Variables

$$\mathbf{x} = \begin{cases} \textit{history} \\ \textit{clinical signs} \\ \textit{nerve conduction studies} \end{cases}$$

2.2.1 History

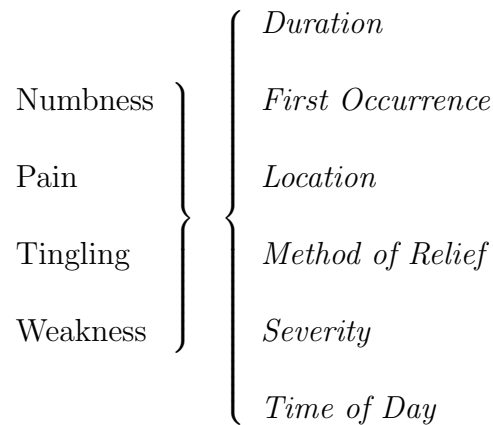
- Age (*in years*)
- Sex (*Male/Female*)

together with the **symptoms**

- Numbness
- Pain
- Tingling
- Weakness

described by their *descriptors* (defined on the next page).

Descriptors of Symptom Variables



Coding of Descriptors for Numbness, Pain, Tingling and Weakness

Descriptor	Coding (0 = symptom absent)
<i>Duration</i>	1 at most 10 minutes
	2 over 10 minutes
<i>First</i>	1 less than 3 months
	2 from 3 months to one year
	3 from 1 to 5 years
	4 from 6 to 10 years
	5 over 10 years
<i>Location</i>	1 first to third fingers
	2 fourth and fifth fingers
	3 all five fingers
	4 other
<i>Relief</i>	1 shaking hand
	2 other
	3 none
<i>Severity</i>	1 mild
	2 moderate
	3 severe
<i>Time</i>	1 daytime episodes
	2 nocturnal episodes
	3 episodes day and night
	4 continuous symptom

2.2.2 Clinical Signs

Coding of Variables

Variable	Coding (0 = symptom absent)
Sensory Loss	
<i>Location</i>	1 first to third fingers
	2 fourth and fifth fingers
	3 all fingers
	4 other
Wasting	
<i>Location</i>	1 Thenar Eminence
	2 Hypothenar Eminence
	3 other
<i>Severity</i>	1 mild
	2 moderate
	3 severe
Weakness	
<i>Location</i>	1 Thenar Eminence
	2 Hypothenar Eminence
	3 other
<i>Severity</i>	1 mild
	2 moderate
	3 severe

2.2.3 Nerve Conduction Studies

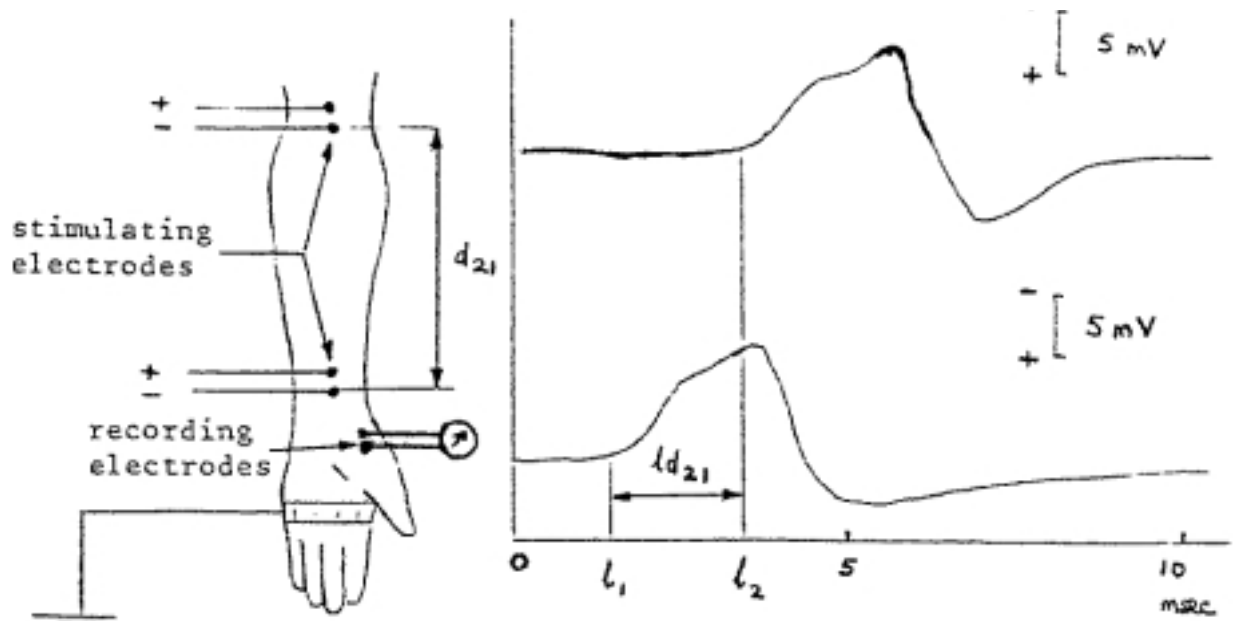
Variables

Nerve	Measurement
Median	Motor Latency at the Wrist
	Motor Latency at the Elbow
	Motor Rate, Elbow to Wrist
	Sensory Latency
	Sensory Amplitude
	Sensory Duration
Ulnar	Motor Latency at the Wrist
	Motor Latency at the Elbow
	Motor Rate, Elbow to Wrist

Units

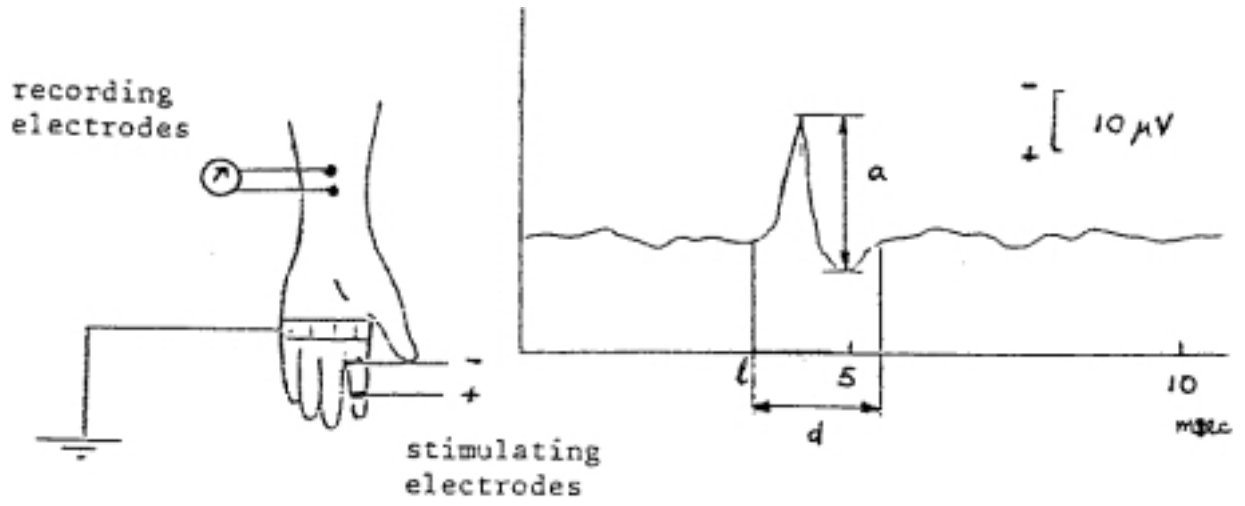
Amplitudes	microvolts
Durations	milliseconds
Latencies	milliseconds
Rates	metres per second

Median Nerve Motor Stimulation



- l_1 = latency at the wrist (milliseconds)
- l_2 = latency at the elbow (milliseconds)
- d_{21} = distance, elbow to wrist (centimeters)
- ld_{21} = $l_2 - l_1$
= latency difference, elbow to wrist
- r_{21} = $10d_{21}/ld_{21}$
= rate, elbow to wrist (meters/second)

Median Nerve Sensory Stimulation

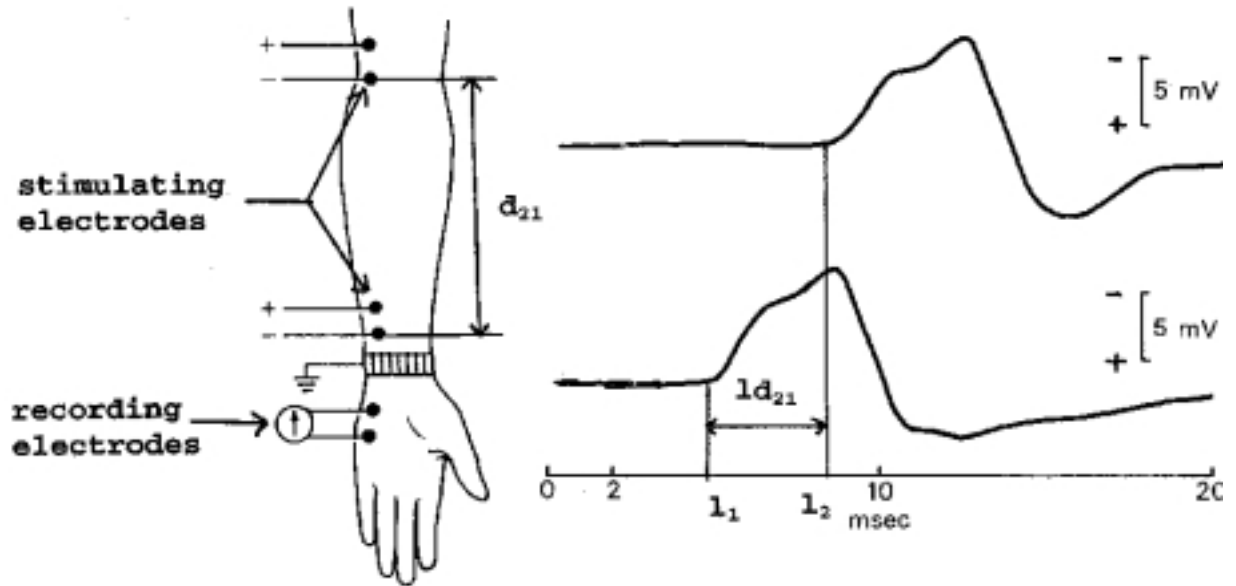


a = median sensory amplitude (microvolts)

d = median sensory duration (milliseconds)

l = median sensory latency (milliseconds)

Ulnar Nerve Motor Stimulation



- l_1 = latency at the wrist (milliseconds)
- l_2 = latency above elbow (milliseconds)
- d_{21} = distance, above elbow to wrist
(centimeters)
- ld_{21} = $l_2 - l_1$
= latency difference, above elbow to wrist
- r_{21} = $10d_{21}/ld_{21}$
= rate, above elbow to wrist (meters/second)

Non-Response to electrical stimulation

occurs in the median motor and sensory measurements

Distribution of non-responses for the whole dataset

	NAD	MILD CTS	MOD CTS	SEV CTS	ALL
Median Motor Wrist	0	0	2	16	18
Median Motor Elbow	0	1 ^a	2 ^a	17	20
Median Sensory	0	17	128	85	230
Total	0	18	132	118	268

^aThese are likely to have been technical errors (inability to elicit a response rather than non-excitability of the nerve).

Note: sensory fibres are thinner than motor fibres, hence are damaged more easily.

Coding of non-responses: *Pseudo-values*

These were taken as 99.9 for latencies and durations, 0 for amplitudes and motor rates. These aren't actual physical measurements, but represent plausible codings: if there is no response, then the amplitude of the waveform will be zero, as will be the motor rate; similarly, the time to response (latency) will be "infinite". We have taken the largest possible number available in the given format (F4.1 in FORTRAN notation), namely, 99.9. The choice of 99.9 for non-response duration is somewhat arbitrary, but seems to fit in with the overall pattern.

2.3 Definition of Carpal Tunnel Syndrome

- varies among doctors
- "typical" hand symptoms: some of *numbness, pain, tingling, weakness*
 - lasting at most ten minutes
 - at night (waking patient)
 - in the first to third fingers
 - relieved by shaking hand

2.4 Cause of Carpal Tunnel Syndrome

is agreed by most doctors to be the entrapment of (pressure on) the median nerve at the wrist, with resulting damage to the nerve at that point.

There are many and varied reasons for this entrapment: for example,

- fracture of the wrist
- rheumatoid arthritis of the wrist
- fluid retention, as in pregnancy

The above symptoms can also be caused by damage to

- median nerve at the elbow or shoulder
- nerves in the neck

ONLY NERVE CONDUCTION STUDIES CAN FIND OUT THE EXACT LOCATION OF THE NERVE DAMAGE

2.5 Computational Aspects

Statistical Software used: SAS Version 8

Computing Problem:

Floating point zero division overflow when median motor latencies at the wrist or elbow or median sensory latency or duration equal their pseudo-value 99.9 (occurs in computing Pearson residuals).

Solution:

- (1) The 20 hands with median motor latencies equal to 99.9 were omitted.
- (2) Median sensory latency and duration were not fitted.

SAS are aware of the problem, which is solved in Version 8.2 (March, 2001).

2.6 Source of the dataset

Patients referred with suspected CTS to

- the Electromyography Clinics of
 - the late Dr John L James, Consultant Physician, St Luke's Hospital, Huddersfield, Yorkshire, England
- between March 1991 and March 1994

Dr James' diagnoses of the examined hands:

- NAD
- Mild CTS
- Moderate CTS
- Severe CTS
- Non-CTS Abnormality (possibly with some severity of CTS)

Non-CTS Abnormality class was omitted from the study, since it was very *inhomogeneous*.

Distribution of hand diagnoses for the whole dataset

NAD	MILD CTS	MOD CTS	SEV CTS	TOTAL
777	620	292	85	1774

3 Logistic regression models fitted to the data

3.1 Aim of diagnostic (predictive) modelling

is to allocate an individual to one of J diagnostic groups on the basis of his/her covariate (predictor) vector $\mathbf{x} = (x_1, \dots, x_p)^T$, using the probability of being in group j given \mathbf{x} , $P(Y = j|\mathbf{x})$; the response variable Y take values $1, 2, \dots, J$, where the outcome

- $Y = 1$: “best” (NAD here)
- $Y = J$: “worst” (severe CTS here, $J = 4$)

For $j = 1, \dots, J$, let the probability that $Y = j$ given \mathbf{x} be

$$\pi_j(\mathbf{x}) = P(Y = j|\mathbf{x}),$$

and the cumulative probability of $Y \leq j$ given \mathbf{x} be

$$\gamma_j(\mathbf{x}) = P(Y \leq j | \mathbf{x}).$$

The *log odds*, $LO_j(\mathbf{x})$, of $Y \leq j$ given \mathbf{x}

$$LO_j(\mathbf{x}) = \log[P(Y \leq j | \mathbf{x})/P(Y > j | \mathbf{x})]$$

can be expressed in terms of the *logit function*

$$\text{logit}(\pi) = \log \left(\frac{\pi}{1-\pi} \right) \quad (0 < \pi < 1)$$

by

$$LO_j(\mathbf{x}) = \text{logit}(\gamma_j(\mathbf{x})).$$

$LO_j(\mathbf{x})$ is related to the *odds*, $O_j(\mathbf{x})$, of $Y \leq j$ given \mathbf{x}

$$O_j(\mathbf{x}) = P(Y \leq j | \mathbf{x}) / P(Y > j | \mathbf{x})$$

by the equation

$$LO_j(\mathbf{x}) = \log[O_j(\mathbf{x})].$$

Odds come from GAMBLING, which stimulated the development of Probability Theory in the 17th and 18th centuries (Bernoulli brothers, among others).

More generally, the *odds* of an event equals the ratio of the probability of the event to the probability of the event *not* occurring:

$$\text{odds of event} = \frac{\text{probability of event occurring}}{\text{probability of event not occurring}}.$$

Example

Suppose a bookmaker offers odds of 2 to 1 on England beating Mexico in the next World cup (assuming they are drawn against each other). This means:

If I place \$100 on this bet and England wins, then I get \$200 from the bookmaker. If England loses - then so do I!

Odds of 2 to 1 on England winning are equivalent to the probability of England winning being 2/3.

Odds of 2 to 1 *against* England beating Mexico are equivalent to the probability of England winning being 1/3.

The *odds ratio*, $OR_j(\mathbf{x}, \mathbf{x}^*)$, of $Y \leq j$ given \mathbf{x} to $Y \leq j$ given \mathbf{x}^* is

$$OR_j(\mathbf{x}, \mathbf{x}^*) = O_j(\mathbf{x}) / O_j(\mathbf{x}^*),$$

as the name suggests!

3.2 Proportional Odds (PO) Models

(Walker & Duncan, 1967; McCullagh, 1980; Bender & Benner, 2000)

It assumes

- cutpoints $\alpha_0, \alpha_1, \dots, \alpha_J$ satisfying the condition

$$-\infty = \alpha_0 < \dots < \alpha_{J-1} < \alpha_J = \infty, \quad (1)$$

- a linear form for the logit of the cumulative probabilities $\gamma_j(\mathbf{x}) (j = 1, \dots, J - 1)$:

$$\gamma_j(\mathbf{x}) = \frac{\exp(\alpha_j - \boldsymbol{\beta}^T \mathbf{x})}{1 + \exp(\alpha_j - \boldsymbol{\beta}^T \mathbf{x})}. \quad (2)$$

(2) is called the *proportional odds assumption*, since it is equivalent to

$$LO_j(\mathbf{x}) = \alpha_j - \boldsymbol{\beta}^T \mathbf{x}, \quad (3)$$

and hence, for any \mathbf{x} and \mathbf{x}^* ,

$$LO_j(\mathbf{x}) - LO_j(\mathbf{x}^*) = \boldsymbol{\beta}^T (\mathbf{x}^* - \mathbf{x}),$$

which is independent of the choice of category j : the odds of $Y \leq j$ given \mathbf{x} and of $Y \leq j$ given \mathbf{x}^* are *proportional*. In other words, the odds ratio

$$\begin{aligned} OR_j(\mathbf{x}, \mathbf{x}^*) &= O_j(\mathbf{x})/O_j(\mathbf{x}^*) \\ &= \exp\{\boldsymbol{\beta}^T (\mathbf{x}^* - \mathbf{x})\} \end{aligned}$$

is constant for all j .

Notes:

- (1) We sometimes call $\boldsymbol{\beta}$ the *regression coefficient* or *slope* of the PO model.
- (2) Since $\alpha_0 = -\infty$ and $\alpha_J = \infty$ are fixed, we call $\alpha_1, \dots, \alpha_{J-1}, \boldsymbol{\beta}$ the *parameters of the PO model*.
- (3) Since the slope $\boldsymbol{\beta}$ of the log-odds ratio, $LO_j(\mathbf{x})$, in equation (3) is independent of j , we say that the PO model satisfies the *common slopes* assumption.

3.2.1 Properties of PO Models

- PO models are the most widely used ordinal regression models due to their simplicity and the following properties:
- They are invariant under *reversal* and *collapsing* of categories.
 - invariance under *reversal* of categories (only the sign of the regression coefficient $\boldsymbol{\beta}$ changes).

Let $Y^* = J + 1 - Y$. Then Y^* is the *reversal* of Y :

$$\begin{array}{c|cccc} Y & 1 & 2 & \dots & J-1 & J \\ \hline Y^* & J & J-1 & \dots & 2 & 1 \end{array}$$

and

$$\begin{aligned} P(Y^* \leq j | \mathbf{x}) &= P(Y \geq J + 1 - j | \mathbf{x}) \\ &= P(Y > J - j | \mathbf{x}). \end{aligned}$$

Hence,

$$\begin{aligned}\text{logit}\{P(Y^* \leq j | \mathbf{x})\} &= \text{logit}\{P(Y > J - j | \mathbf{x})\} \\ &= -\text{logit}\{P(Y \leq J - j | \mathbf{x})\} \\ &= -\alpha_{J-j} + \boldsymbol{\beta}^T \mathbf{x}:\end{aligned}$$

if Y has a PO model with parameters $\alpha_1, \dots, \alpha_J, \boldsymbol{\beta}$, then its reversal Y^* also has a PO model, but with parameters $-\alpha_{J-1}, \dots, -\alpha_1, -\boldsymbol{\beta}$.

- invariance under *collapsing* of the ordered categories ($\boldsymbol{\beta}$ does not change when the response categories are collapsed or the category definitions are changed). We illustrate this for $J = 5$.

Collapse $\{1,2\}$ and $\{3,4\}$: the new response variable, Y^+ , is defined as follows

$$\begin{array}{c|ccc} Y & 1 & 2 & 3 & 4 & 5 \\ Y^+ & 1 & & 2 & & 3 \end{array}$$

Then

$$\begin{aligned}P(Y^+ = 1 | \mathbf{x}) &= P(Y \leq 2 | \mathbf{x}) \\ &= \alpha_2 - \boldsymbol{\beta}^T \mathbf{x},\end{aligned}$$

and

$$\begin{aligned}P(Y^+ \leq 2 | \mathbf{x}) &= P(Y \leq 4 | \mathbf{x}) \\ &= \alpha_4 - \boldsymbol{\beta}^T \mathbf{x}.\end{aligned}$$

Hence, Y^+ also has a PO model, but with $J^+ = 3$ values and parameters $\alpha_2, \alpha_4, \boldsymbol{\beta}$.

- The regression coefficients are easily interpreted:

Let

$$\mathbf{x}^{(k)} = (x_1, \dots, x_k + 1, \dots, x_p)^T,$$

that is, $\mathbf{x}^{(k)}$ is the vector \mathbf{x} with the k th component x_k of \mathbf{x} increased by one.

Then

$$LO_j(\mathbf{x}^{(k)}) = LO_j(\mathbf{x}) - \beta_k. \quad (4)$$

In other words, β_k is the change in $LO_j(\mathbf{x})$ corresponding to a unit increase in x_k .

By (4),

$$\begin{aligned}\beta_k &= LO_j(\mathbf{x}) - LO_j(\mathbf{x}^{(k)}) \\ &= \log[OR_j(\mathbf{x}, \mathbf{x}^{(k)})],\end{aligned} \quad (5)$$

and taking exponentials gives

$$\begin{aligned} e^{\beta_k} &= O_j(\mathbf{x})/O_j(\mathbf{x}^{(k)}) \\ &= OR_j(\mathbf{x}, \mathbf{x}^{(k)}) : \end{aligned} \tag{6}$$

e^{β_k} is the odds ratio corresponding to a unit increase in x_k , and is called *the odds ratio of x_k* (which does not depend on j).

PO is in the *Grouped Continuous* family of models, which assumes that Y is a discretisation of a continuous, unobserved latent variable, Z :

$$Y = j \text{ if } \alpha_{j-1} < Z \leq \alpha_j$$

Then

$$\gamma_j(\mathbf{x}) = F(\alpha_j - \boldsymbol{\beta}^T \mathbf{x}) \quad (j = 1, \dots, J - 1),$$

where $F(\cdot)$ is an appropriate cumulative distribution function (cdf). If Z exists, then F is the cdf of Z .

Note: Although PO is motivated by a latent variable Z , we can still use the PO model on data even if there is no obvious Z .

For PO,

$$F(x) = \frac{\exp(x)}{1 + \exp(x)},$$

the logistic cumulative distribution function.

Other common forms of F are

1. (Ordinal Probit model)

$$F(x) = \Phi(x),$$

where $\Phi(x)$ is the standard Normal cumulative distribution function.

2. (Complementary Log Log model)

$$F(x) = 1 - \exp\{-\exp(x)\}.$$

This model is also called the *discrete proportional hazards* model (McCullagh, 1980; Greenland, 1994).

3.3 Continuation Ratio (CR) Models

(Armstrong & Sloan, 1989; Bender & Benner, 2000)

For $j = 1, \dots, J$, let

$$\pi_j(\mathbf{x}) = P(Y = j|\mathbf{x})$$

and the (backward) *Continuation Ratio (CR)*

$$\begin{aligned}\delta_j(\mathbf{x}) &= P(Y = j | Y \leq j, \mathbf{x}) \\ &= \frac{\pi_j(\mathbf{x})}{\pi_1(\mathbf{x}) + \dots + \pi_j(\mathbf{x})}\end{aligned}$$

Then

$$\begin{aligned}\delta_1(\mathbf{x}) &= 1, \\ \delta_J(\mathbf{x}) &= \pi_J(\mathbf{x}).\end{aligned}\tag{7}$$

and for $j = 2, \dots, J$,

$$\pi_j(\mathbf{x}) = \frac{\delta_j(\mathbf{x})}{1 - \delta_j(\mathbf{x})} [\pi_1(\mathbf{x}) + \dots + \pi_{j-1}(\mathbf{x})].$$

The CR model is given by

$$\text{logit}\{\delta_j(\mathbf{x})\} = \alpha_j - \boldsymbol{\beta}^T \mathbf{x},\tag{8}$$

and the corresponding $\pi_j(\mathbf{x})$ is

$$\pi_j(\mathbf{x}) = \begin{cases} \frac{1}{\prod_{t=2}^J [1 + \exp(\alpha_t - \boldsymbol{\beta}^T \mathbf{x})]} & (j = 1), \\ \frac{\exp(\alpha_j - \boldsymbol{\beta}^T \mathbf{x})}{\prod_{t=j}^J [1 + \exp(\alpha_t - \boldsymbol{\beta}^T \mathbf{x})]} & (2 \leq j \leq J). \end{cases}\tag{9}$$

($\prod_{t=j}^J$ denotes the *product* from $t = j$ to J .)

$\pi_j(\mathbf{x})$ can also be expressed directly in terms of the δ_j s:

$$\pi_j(\mathbf{x}) = \begin{cases} \prod_{t=2}^J [1 - \delta_t(\mathbf{x})] & (j = 1), \\ \delta_j(\mathbf{x}) \prod_{t=j+1}^J [1 - \delta_t(\mathbf{x})] & (1 < j < J), \\ \delta_J(\mathbf{x}) & (j = J) \end{cases}\tag{10}$$

Notes:

- (1) Equation (10) holds for *any* model, since it makes no assumptions about $\delta_j(\mathbf{x})$.
- (2) Equation (9) follows from equation (10) using equation (8).
- (3) The slope $\boldsymbol{\beta}$ of $\text{logit}\{\delta_j(\mathbf{x})\}$ in equation (8) is the same for all j (the *common slopes* assumption).

(4) For the CR model, the α_j do not have to satisfy the relationship (1).

(5) The *forward* CR

$$\begin{aligned}\delta_j(\mathbf{x}) &= P(Y = j | Y \geq j, \mathbf{x}) \\ &= \frac{\pi_j(\mathbf{x})}{\pi_j(\mathbf{x}) + \dots + \pi_J(\mathbf{x})}\end{aligned}$$

can also be defined, with corresponding changes in the expressions for $\pi_j(\mathbf{x})$.

Proof of equation (10)

There are two cases to consider: $1 \leq j < J$ and $j = J$.

$j = J$: This is just equation (7), which holds since $Y \leq J$ is always true, and hence

$$P(Y = J | Y \leq J, \mathbf{x}) = P(Y = J | \mathbf{x}),$$

as stated by equation (7).

$1 \leq j < J$: We start by showing that

$$1 - \delta_j(\mathbf{x}) = \frac{\sum_{t=1}^{j-1} \pi_t(\mathbf{x})}{\sum_{t=1}^j \pi_t(\mathbf{x})}. \quad (11)$$

The left-hand side equals

$$\begin{aligned}P(Y \neq j | Y \leq j, \mathbf{x}) &= P(Y < j | Y \leq j, \mathbf{x}) + P(Y > j | Y \leq j, \mathbf{x}) \\ &= \sum_{t=1}^{j-1} \pi_t(\mathbf{x}) / \sum_{t=1}^j \pi_t(\mathbf{x}),\end{aligned} \quad (12)$$

since the second probability in equation (12) is zero.

By (11), we get

$$\sum_{t=1}^j \pi_t(\mathbf{x}) = \prod_{t=j+1}^J [1 - \delta_t(\mathbf{x})], \quad (13)$$

since the right-hand side equals

$$\prod_{t=j+1}^J \left[\sum_{r=1}^{t-1} \pi_r(\mathbf{x}) / \sum_{r=1}^t \pi_r(\mathbf{x}) \right],$$

and all terms, except the numerator of the first factor, $\sum_{r=1}^j \pi_r(\mathbf{x})$, and the denominator of the last factor, $\sum_{r=1}^J \pi_r(\mathbf{x})$, cancel out. The former equals the left-hand side of equation (13), and the latter is one, since it is the sum of probabilities over all possible values.

3.3.1 Differences between the forward and backward CRs

- **Forward CRs** make sense when Y represents *discrete survival times*: since high-risk patients have short survival times, forward CRs compare high-risk patients with low-risk patients.
- **Backward CRs** are appropriate when Y represents a disease status given by ordered categories with higher values corresponding to more severe disease states. This is the case here with the CTS dataset.

3.4 Comparison of CR and PO Models

- In practice, there are *no major differences* between CR and PO results.
- PO is easier to interpret than CR (β_k is the log odds-ratio corresponding to a unit increase in x_k , as stated in (5)).
- PO has the nice property of *invariance* under *reversal* or *collapse* of categories; CR does not.
- CR can be fitted *more flexibly* than PO, using Binary Logistic Regression software. This allows the *common slopes assumption* (see Notes (3) of Sections 3.2 and 3.3) to be *relaxed* for some of the predictors.
- *Choice* of model depends on the investigator's preference for *cumulative probabilities* or *continuation ratios*.

3.5 Design (Dummy) Variables for Categorical Variables

In our dataset, these categorical variables are

- History Symptom Descriptors
- Clinical Signs

If $SIGN$ denotes one of the clinical signs, and has k categories, then we define k Design Variables $SIGN1, SIGN2, \dots, SIGNk-1, SIGNk$ as follows:

$SIGN$	$SIGN1$	$SIGN2$...	$SIGNk-1$	$SIGNk$
0 (<i>absent</i>)	0	0	...	0	0
1	1	0	...	0	0
2	0	1	...	0	0
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
$k-1$	0	0	...	1	0
k	0	0	...	0	1

SIGN's Design Variables are fitted instead of SIGN, since SIGN is *not* a quantitative variable. Similarly for History Symptom Descriptors.

Example: Location of Wasting (*LocWast*)

LocWast	LocWast1	LocWast2	LocWast3
0	0	0	0
1	1	0	0
2	0	1	0
3	0	0	1

Notes:

- (1) Design variables are BINARY variables, taking the values 0, 1, which indicate the level of their variable. Thus, for example, $\text{LocWast3} = 1$ corresponds to $\text{LocWast} = 3$.
- (2) The coefficients of the design variables give the contribution made by each level of their variable. Thus, the coefficient of *LocWast1* gives the effect of *wasting in the Thenar Eminence* on the probabilities of the four diagnostic categories.

3.6 Classification Procedures

Let $d(\mathbf{x})$ be the classification function evaluated at the covariate vector \mathbf{x} :

$d(\mathbf{x})$ is the index of the group into which \mathbf{x} is classified.

Let $\hat{d}(\mathbf{x})$ denote the estimate of $d(\mathbf{x})$ obtained by substituting the maximum likelihood estimates for the corresponding parameters in the model.

3.6.1 Highest Probability (HP)

$$\hat{d}(\mathbf{x}) = \underset{1 \leq k \leq J}{\operatorname{argmax}} \hat{P}(Y = k | \mathbf{x}), \quad (14)$$

where $\hat{d}(\mathbf{x})$ is chosen to be the smallest index if several groups have the same maximum conditional probability.

(14): classify \mathbf{x} to the group k which *maximises* $\hat{P}(Y = k | \mathbf{x})$: **classify to the most probable group.**

Problems can occur if the estimated probabilities are close.

Example ($J = 4$)

k	<i>Carpal Tunnel Syndrome</i>			
	<i>NAD</i>	<i>Mild</i>	<i>Moderate</i>	<i>Severe</i>
	1	2	3	4
$\hat{P}(Y = k \mathbf{x})$	0.249	0.249	0.251	0.251

Suppose the above figures hold for one observation \mathbf{x} . Then

$$\hat{d}(\mathbf{x}) = 3 \text{ (Moderate CTS).}$$

But, how much confidence can we have in this classification? The estimated probabilities differ by at most 0.002.

3.6.2 Anderson and Philips (AP) for the PO model

(Anderson & Philips, 1981)

$$\hat{d}(\mathbf{x}) = j \text{ if } \hat{\alpha}_{j-1} < \hat{\boldsymbol{\beta}}^T \mathbf{x} \leq \hat{\alpha}_j \quad (j = 1, \dots, J).$$

This is a valid definition, since the α_j satisfy (1).

The AP method assumes that the continuous unobserved latent variable Z has *logistic density*: for $-\infty < z < +\infty$

$$f_Z(z) = \frac{\exp(z - \boldsymbol{\beta}^T \mathbf{x})}{\{1 + \exp(z - \boldsymbol{\beta}^T \mathbf{x})\}^2},$$

with

$$E(Z | \mathbf{x}) = \boldsymbol{\beta}^T \mathbf{x}.$$

4 General methodology for developing clinical prediction models for ordinal data

(Harrell *et al.*, 1996; Harrell *et al.*, 1998; Hosmer & Lemeshow, 1989 & 2001)

Hosmer & Lemeshow (1989 & 2001), the standard text for applied logistic regression modelling, give model building strategies and methods for logistic regression models;

Harrell *et al.* (1996) present a thorough discussion of modelling, and apply it to survival analysis;

Harrell *et al.* (1998) apply the general methods of Harrell *et al.* (1996) to modelling ordinal data.

4.1 Preliminary Steps

4.1.1 Define the response variable Y and predictor variable \mathbf{x} , after consultation with a medical expert.

4.1.2 Choose *interactions* carefully.

- These should correspond to biological relationships, and not indicate a general lack of fit of the model.
- Avoid too many interactions, since these will generate a large number of extra parameters.

4.2 Data Reduction

4.2.1 Methods ignoring Y

- Principal Component Analysis (PCA)
- Cluster Analysis

on \mathbf{x} alone

The main principal components (with eigenvalues at least one) are used as variables in the logistic regression.

4.2.2 Variable Selection

Many statistical packages offer automatic variable selection, whose dangers are well-known (Harrell et al., 1996 & 1998).

The main dangers are

- The actual significance level of the tests used is much higher than the nominal value, since multiple hypothesis testing is involved: too many variables will be selected.
- None of the test statistics used has got an exact χ^2 distribution. Hence, the usual Normal distribution theory results are at best approximate.

4.3 Verify the Model's Assumptions

The extent to which a model's assumptions hold will determine the usefulness of the model. If the model's assumptions don't hold, then any prediction based on the model will be uncertain.

Model checking usually involves plots of suitable functions of the data.

Our situation involves

- PO assumption
- CR assumption

4.3.1 Linearity of continuous predictors

Can be checked by

- adding squares of the predictors to the model, and seeing whether the model's fit is significantly improved,
- drawing smoothed plots of appropriate residuals.

4.3.2 Additivity of predictors (no interaction)

- test whether adding cross-product terms to the model significantly improves its fit.
- interactions could take the form of change of shape of the predictors' distributions (for example, linear age relationship for males, quadratic one for females).

4.3.3 Predictors' Distribution

No assumptions are made for logistic regression.

This check applies, for example, to parametric survival analysis models (Weibull or log-Normal survival distributions).

4.3.4 Influential Observations

These are mainly detected using residuals where appropriate.

4.4 Fit the Model

determined by earlier steps. Also, fit the full model for comparison purposes.

4.5 Compute measures of *predictive accuracy*

- calibration (extent of bias)
- discrimination (predictors' ability to separate patients with different responses)
- percentage correct classification

This is the only measure considered in these talks.

Examples

Calibration: If the average *predicted* mortality for a group of similar patients is 0.3, and the *actual* proportion dying is also 0.3, then the predictions are *well-calibrated*: the predicted values are close to the actual (observed, future data).

Discrimination: A weather forecaster has *high discrimination* if his predicted risks of rain for days when it actually rained are higher than for dry days.

4.6 Validate the model

4.6.1 Resubstitution

This involves testing the model on the dataset used to fit it, and is not recommended in general, since it tends to give over-optimistic results. It is, however, the easiest to carry out, and can be used for comparison with other, more sophisticated, methods.

4.6.2 Data splitting

This randomly splits the dataset into disjoint design (training) and test (validation) sets, and is an improvement on section 4.6.1.

However, it needs a large dataset to be effective.

4.6.3 Cross-validation

This repeatedly splits the dataset into design and test sets, and averages the results over all the test sets. Its advantages over data splitting are

- a much larger training sample
- reduced variability of the performance evaluation indicators

4.6.4 Bootstrap

(Efron & Tibshirani, 1993)

This is the best, if less well-known, method:

- it provides nearly unbiased estimates of predictive accuracy that are of relatively low variance
- fewer model fits are required than with cross-validation

The bootstrap is a data-based simulation method for statistical inference.

Bootstrap: from the phrase “to pull oneself up by one’s bootstraps”.

The Bootstrap Method

Let $\mathbf{z}_i = (\mathbf{x}_i^T, Y_i)^T$ be the i th observation in the dataset ($i = 1, \dots, n$), and $\mathbf{z} = (\mathbf{z}_1^T, \dots, \mathbf{z}_n^T)^T$. The \mathbf{z} ’s are assumed independent.

A *Bootstrap Sample* $\mathbf{z}^* = (\mathbf{z}_1^{*T}, \dots, \mathbf{z}_n^{*T})^T$ is obtained by randomly sampling n times, *with replacement*, from the original data points $\mathbf{z}_1, \dots, \mathbf{z}_n$.

To estimate a statistic $s(\mathbf{z})$ and its standard error,

- (1) generate a large number B (typically, 200) of independent Bootstrap Samples $\mathbf{z}^{*1}, \dots, \mathbf{z}^{*B}$, each of size n ,
- (2) for $b = 1, \dots, B$, compute the *bootstrap replication* $s(\mathbf{z}^{*b})$ of s . Then
- (3) the *bootstrap estimate* of $s(\mathbf{z})$ is

$$\hat{s}_{boot} = \sum_{b=1}^B s(\mathbf{z}^{*b})/B,$$

- (4) the *bootstrap estimate* of standard error is the standard deviation of the bootstrap replications,

$$\hat{se}_{boot} = \left\{ \sum_{b=1}^B [s(\mathbf{z}^{*b}) - \hat{s}_{boot}]^2 / (B - 1) \right\}^{\frac{1}{2}}$$

5 Results

We only give them for the resubstitution method and the PO model.

5.1 Full Model

This involved the full set of 79 predictor variables, 13 of which were linearly dependent on other variables. The full model was thus based on 66 linearly independent variables.

% correct: 80.3%.

Doctor's Diagnosis	Predicted Diagnosis				Total
	NAD	Mild CTS	Mod CTS	Severe CTS	
NAD	694	83	0	0	777
Mild CTS	77	489	54	0	620
Mod CTS	0	82	192	18	292
Severe CTS	0	6	29	50	85
Total	771	660	275	68	1774

5.2 Forward Selection of Variables

This is the most economical and fastest variable selection method, since it builds up from zero variables. Eleven variables were selected by this method.

% correct: 80.1%.

Doctor's Diagnosis	Predicted Diagnosis				Total
	NAD	Mild CTS	Mod CTS	Severe CTS	
NAD	700	77	0	0	777
Mild CTS	80	485	55	0	620
Mod CTS	1	81	188	22	292
Severe CTS	0	6	31	48	85
Total	781	649	274	70	1774

5.3 Backward Elimination of Variables

This is a much slower and less economical method, since it starts with all 79 variables. It removed variables until 22 significant ones were left.

% correct: 79.5%.

Doctor's Diagnosis	Predicted Diagnosis				Total
	NAD	Mild CTS	Mod CTS	Severe CTS	
NAD	693	84	0	0	777
Mild CTS	82	492	46	0	620
Mod CTS	0	89	179	24	292
Severe CTS	0	6	32	47	85
Total	775	671	257	71	1774

5.4 PCA Selection of Variables

This was not successful: 35 factors were selected as having eigenvalues at least one. This does not constitute a variable reduction!

Probable Reason: PCA assumes Normally distributed, or at least *continuous* predictor variables, whereas the CTS dataset uses many *design variables (binary)*.

5.5 Conclusions

- No evidence was found of the need to include non-linear or interaction terms.
- Forward selection (FS) seems much better than backward elimination (BE):
 - half as many variables were selected with FS as with BE;
 - FS had a slightly higher % correct than BE, though probably not statistically significant;
 - FS had a much shorter execution time than BE.
- All three models validated to give very similar results (certainly, no statistical evidence of a difference between them in terms of % correct classification).
- Six variables in the full PO model, four in the FS, and eight in the BE were particularly important (in terms of the magnitude of their coefficients).

Two were important in all three models

- median motor latency at the wrist
- clinical weakness located elsewhere than the thenar or hypothenar eminences

Three were important in two of the three models

- the presence of tingling
- weakness relieved by other than shaking the affected hand
- wasting in the thenar eminence

Selection of variables using the magnitude of their coefficients is based on *Wald's Test*, an approximate result for logistic regression models: let

$$\begin{aligned}\hat{\beta}_k &= \text{maximum likelihood estimate of } \beta_k, \\ SE(\hat{\beta}_k) &= \text{Standard Error of } \hat{\beta}_k.\end{aligned}$$

Wald's test assumes that $\hat{\beta}_k$ is the observed value of a Normal variable with mean β_k . Under the null hypothesis $\beta_k = 0$ (x_k has no effect), $\hat{\beta}_k/SE(\hat{\beta}_k)$ has an approximate Standard Normal distribution, and its square is approximately distributed as a χ^2 with one degree of freedom.

Example

For the full model, the MEDIAN MOTOR LATENCY AT THE WRIST (MMLW) was selected for the following reason (output reproduced from SAS):

Parameter	DF	Estimate	Standard Error
MMLW	1	1.1950	0.1390
χ^2		Probability > χ^2	
73.8731		< 0.0001	

6 Open Problems/Further Work

- Examination of the various performance measures proposed by Harrell et al. (1998).
- Investigation of approaches for handling Non-Responses in Nerve Conduction Measurements. These include the Multilevel Models of Goldstein (1995).
- Checking of the CR and PO assumptions, which requires sophisticated plotting of suitably defined residuals; at present this is best done using S-Plus (see Harrell et al., 1998, and Bender & Benner, 2000).
- Flexible fitting of the CR model by any Binary Logistic Regression software using a clever trick (see Bender & Benner, 2000).

7 References

- Armstrong, B. G. & Sloan, M. (1989). Ordinal regression models for epidemiologic data. *American Journal of Epidemiology*, **129**, 191-204.
- Bender, R. & Benner, A. (2000). Calculating ordinal regression models in SAS and S-Plus. *Biometrical Journal*, **42**, 677-699.
- Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, London. **Reprinted 1998 by CRC Press.**
- Goldstein, H. (1995). *Multilevel Statistical Models*. Arnold, London.
- Greenland, S. (1994). Alternative models for ordinal logistic regression. *Statistics in Medicine*, **13**, 1665-1677.
- Harrell, F. E., Lee, K. L. & Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, **15**, 361-387.
- Harrell, F. E., Margolis, P. A., Gove, S., Mason, K. E., Mulholland, E. K., Lehmann, D., Muhe, L., Gatchalian, S., Eichenwald, H. F., & The WHO/ARI Young Infant Multicentre Study Group (1998). Development of a clinical prediction model for an ordinal outcome: The World Health Organization multicentre study of clinical signs and etiological agents of pneumonia, sepsis and meningitis in young infants. *Statistics in Medicine*, **17**, 909-944.
- Hosmer, D. W. & Lemeshow, S. (1989). *Applied Logistic Regression*. John Wiley & Sons, New York, USA.
- Hosmer, D. W. & Lemeshow, S. (2001). *Applied Logistic Regression, 2nd Edition*. John Wiley & Sons, New York, USA.
- McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Series B*, **42**, 109-142.
- Rosenbaum, R. B. & Ochoa, J. L. (1993). *Carpal Tunnel Syndrome and Other Disorders of the Median Nerve*. Butterworth-Heinemann, Stoneham, MA, USA.
- Rudolfer, S. M., Paliouras, G. & Peers, I. (1996). Diagnostic strategies for carpal tunnel syndrome. *Abstract Book, Sixth Biennial Conference of the European Society for Medical Decision Making (June 16-18, 1996, Torino, Italy)*, 65.
- Rudolfer, S. M., Paliouras, G. & Peers, I. (1999). A comparison of logistic regression to decision tree induction in the diagnosis of carpal tunnel syndrome. *Computers and Biomedical Research*, **32**, 391-414.

- Rudolfer, S. M. & Peers, I. (1999). Statistical modelling of the diagnosis of carpal tunnel syndrome. *Proceedings of the 14th International Workshop on Statistical Modelling, Graz, Austria, July 19-23, 1999* (Friedl, H., Berghold, A. & Kauermann, G. (Editors)), 638-642.
- Rudolfer, S. M., Watson, P. C. & Lesaffre, E. (1995). Are ordinal models useful for classification? A revised analysis. *Journal of Statistical Computing and Simulation*, **52**, 105-132.
- Walker, S. H. & Duncan, D. B. (1967). Estimation of the probability of an event as function of several independent variables. *Biometrika*, **54**, 167-179.