Statistical Analysis to identify discriminant progression factors of the Parkinson's disease

Ana Rita Sousa
Instituto Superior Técnico, TULisbon
Lisboa, Portugal
anaritasousa@ist.utl.pt

ABSTRACT

The Parkinson's disease is a neurodegenerative disease and, after Alzheimer's disease, it is the most common disease among the world population, specially the elderly. This disease affects the motor system (tremor, rigidity, slowness of movements, postural instability), however no precise cause about its origin is known at the moment. Nevertheless it is known that two groups of patients can be identified: one first group showing fast progression of the disease and a second group where the progression is slow.

In this dissertation we intend to find the main features that enable the discrimination between these groups, using a Logistic Regression Model. There are a huge number of variables, so a previous variables selection, using a Principal Components Analysis method, is done.

Another important question brought by this study is about the way some variables, associated with surveys evaluation, are transformed from a Likert Scale to numerical. Sometimes, the objective characteristic depends on the type of the questions as well as on the individuals who answer them. This work also describes a mathematical relation between the characteristic under study and the ability of the individual, using a Rasch Model.

Keywords

Parkinson Disease, Progression, Rasch Models, Logistic Regression, Likert Scale, Principal Components Analysis.

1. INTRODUCTION

Parkinson's disease is a degenerative disease of the central nervous system with high prevalence. It affects 45 to 300 people in 100.000 inhabitants all over the world, and 2% of the worldwide population over 65 years old [3]. In Portugal, it is estimated that the number of cases of this disease is around 20,000. As it is known, the western world population is ageing,and since the appearance of the disease occurs between 55 and 65 years old, it is expected an increase in a near future.

The disease is named after the English doctor James Parkinson, who described it in 1817. Although some genetic and environmental factors have been identified in some restrict groups, so far there has not been identified as the cause of the disease any other factor that could explain the start and the evolution of the disease.

Considering the ageing of the population and consequent disease's prevalence increase, it is more and more important to identify the causes of the Parkinson's disease as well as getting further in the investigation of the disease's mechanism in order to find more adequate therapeutic solutions and develop new medications.

However, it is now possible to observe patients in which the symptoms of the disease evolve faster and others in which these symptoms evolve slower, but it is not yet possible to know what causes these different types of evolution.

Prompted by common interest, the Probabilities and Statistics Group of the CEMAT Instituto Superior Técnico together with Neurological Unit of Clinical Investigation of the Instituto de Medicina Molecular started a collaboration, in which the main objective was to use methods and statistical techniques to help identifying epidemiological and clinical factors that allow us to distinguish between two groups of progression (slow and fast) of the disease.

The data used in this study came from a clinical evaluation of each patient by a doctor and a demographic questionnaire answered by the patient or his caregiver.

All patients included in the study were diagnosed with Parkinson's disease, according to the United Kingdom Brain Bank Criteria (UKBBC), for over 10 years, and signed the informed consent. Patients with atypical forms of parkinsonism were excluded.

Since the number of variables is quite large, we used different methods to make a selection of them. In a preliminary analysis variables were identified with redundancies, omitted data and low variability. And then principal component analysis was applied to subgroups of variables.

In the next sections, we describe the methods used, and the results and conclusions we got.

2. METHODS

In this section we present the methods used along the study.

2.1 Likert Scale

The Likert Scale is one of the most widely used approaches in survey research [8]. A Likert Scale is uni-dimensional, ordered and it is constituted by a group of sentences (items) about which respondents specify their level of agreement or disagreement. The range of the scale can vary from "totally disagree" (level 1) to "totally agree" (level 5, 7 or 11).

Usually the attitude of the respondent is measured by the sum or the mean of the levels of response given for each item of the scale, which mathematically is not always the most correct. So, we tried to find a new method that would provide a more accurate codification of the respondent attitude. This technique is developed in the next section.

2.2 Rasch Model

The scales in general and the attitude scales in particular, have their limitations about its metric capacities, and consequently it is important to simulate results to verify the consistency of the measures and of measuring levels' analysis. The most common way to treat the Likert Scale simulations are the Item Response Theory (IRT) models. IRT is kind of a paradigm for the design, analysis, and scoring of tests, questionnaires, and similar instruments measuring abilities, attitudes, or other variables and it is based on the application of related mathematical models to testing data. Rasch model is a particular simplified case of the IRT model [7].

In the sixties, the ideas of George Rasch, a Danish mathematician, had brought a new breath to psychometrics that was logged, and at the same time contested about its procedures. But the real evolution came with Andrich ideas, to apply a Rasch model to the measure analysis of Likert scale, and other similar scales[6].

The Rasch model has proposed new methods to the development and test analysis, in which the items and tests do not depend on the population sample and the individual ability is independent of the items. In fact, this model is a method to build scales in which the attitude of the individual is evaluated independently of the items that constitute the scale.

Therefore the main objective of the Rasch Model is to describe the mathematical relation between the latent characteristic of the individual and its performance in an item.

The following properties are associated with the Rasch Model:

- **unidimensionality,** the set of items measures one dimension only, a latent variable;
- **local independence of the items,** no item should contain information that can be used to respond to other item;
- **monotony,** the higher the ability of the individual the higher the probability to give a correct answer to the item;
- **sufficiency,** to estimate the difficulty of the item it is not necessary to know which individuals gave a correct answer and to estimate the individuals' ability it is not necessary to know which items were answered.

The formulation of the Rasch Model is based on the following equation:

$$P[X_{vi} = 1] = c_i + \frac{(1 - c_i)exp(\alpha_i(\theta_v - \sigma_i))}{1 + exp(\alpha_i(\theta_v - \sigma_i))} \qquad (1)$$

where $P[X_{vi} = 1]$ represents the probability that the individual v gives the correct answer to the item i, c_i is the randomness parameter (probability of individual, independent from its ability, answers correctly to the item i), α_i is the discrimination degree of the item i, θ_v is the ability of the individual v and σ_i is the difficulty parameter of the item i (corresponds to the median of the logistic distribution).

Originally the model was developed to be used to handle binary data, but later it was extended to polytomic variables.

The polytomous Rasch model must be used when there are more than two possible answers to the item, for example 0 (totally disagree) till m (totally agree). The values 0,...,m are called categories, according to Rasch Models terminology [4].

The polytomous Rasch model starts from equation (1)but considers that all items have the same discrimination power ($\alpha_i = 1$, $\forall i$) as well as the same randomness parameter ($c_i = 0$, $\forall i$), and generalizes it to every answer (x). So the probability that the individual v gives the correct answer to the item i depends on his ability and on the threshold between categories and it is given by:

$$P[X_{vi} = x] = \frac{exp(x\theta_v - \sum_{s=1}^{x} \tau_{is})}{\eta_{vi}}, x = 0, ..., m \qquad (2)$$

where $\eta_{vi} = \sum_{y=0}^{m} exp(y\theta_v - \sum_{s=1}^{y} \tau_{is})$ and we admit $\sum_{s=1}^{0} \tau_{is} = 0$ since there is no threshold to 0 category, and θ_v is the ability of the individual, τ_{ix} is the transition parameter or item i threshold that separates the answer in category *x-1* to category *x*.

So, in this model there are nm+1 parameters to estimate: θ_v and nm τ_{ix} (i=1,...n e x=1,...m).

2.3 Principal Component Analysis

The Principal Components Analysis (PCA) is the most used method to reduce the number of original variables [5]. It is a very convenient and useful method when the number of variables under study is large, a situation that is common in practice.

The PCA allows the transformation of the original set of variables, correlated with each other, into another set of variables that are liner combinations of the original variables. The new variables are not correlated, and we call it principal components. They are constructed in such a way that it is possible to choose a small number of principal components which explain a large percentage of the data variability.

The easiest way to obtain the principal components (PCs) of a data set is the following: if $\mathbf{X} = (X_1, ..., X_p)^t$, is the vector of p variables and $\mathbf{\Sigma}$ is the covariance matrix of $\mathbf{X}$, whith eigenvalues $\lambda_1 \geqslant ... \geqslant \lambda_p$, and eigenvectors γ_1,...,γ_p, the i-th principal component is given by $\gamma_i^t \mathbf{X}$.

Since $Var(X_i) = \lambda_i$, the total variance of the p original variables is $\sum_{i=1}^{p} Var(X_i) = \sum_{i=1}^{p} \lambda_i$. We use (3) to find the number k of PCs to save, where k is such that (3) corresponds to a high percentage (generally larger than 80%).

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \tag{3}$$

This same procedure can be applied to a sample matrix $\mathbf{S}$. For this case we have, respectively, $\hat{\lambda_1} \geqslant ... \geqslant \hat{\lambda_p}$, $\hat{\gamma_1}$,...,$\hat{\gamma_p}$, ${\hat{\gamma_i}}^t \mathbf{X}$ and

$$\frac{\sum_{i=1}^{k} \hat{\lambda}_i}{\sum_{i=1}^{p} \hat{\lambda}_i} \tag{4}$$

PCA is frequently used to study correlation matrices, instead of covariance matrices. This is particularly advantageous when variables are measured on different scales or show different variability.

In the present study variables are in large number and are of different nature. A special difficulty is the presence of missing data. We found useful to use the function *PcaNA*, from R, package *rrcovna* [11] [12], to complete the data matrix. This increases the number of observations ready for principal components (R function *prcomp* [1] [10]) and makes the study more reliable and relevant.

2.4 Logistic Regression

The main objective of Logistic Regression is to find a function to describe the relation between the response (discrete) variable and numeric or categorical predictor variables (continuous/discrete).

Consider the response binary variable Y (assuming values 0 or 1, with probabilities π and $1 - \pi$, respectively), and $\mathbf{X}$= $X_1, ..., X_{p-1}$ the vector of *(p-1)* predictor variables. The Logistic Regression model is described by the following equation:

$$E[Y|\mathbf{X}] = \pi(\mathbf{X}) = \frac{exp(\beta_0 + \beta_1 X_1 + ... + \beta_{p-1} X_{p-1})}{1 + exp(\beta_0 + \beta_1 X_1 + ... + \beta_{p-1} X_{p-1})} \tag{5}$$

where β= $(\beta_0, \beta_1, ..., \beta_{p-1})^t$ is the vector of parameters.

Applying the *logit* transformation $\ln \frac{\pi}{1-\pi}$ to expression (5), leads to a new expression called the linear predictor:

$$\pi^* = \ln[\frac{\pi}{1-\pi}] = \beta_0 + \beta_1 X_1 + ... + \beta_{p-1} X_{p-1} \tag{6}$$

Using the maximum likelihood method to estimate the parameters, we have to maximize the likelihood function (7) that characterizes the data:

$$L(\beta|(x,y)) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \tag{7}$$

where, $\pi(x_i) = P(Y = 1|X = x_i)$ and $1 - \pi(x_i) = P(Y = 0|X = x_i)$.

The maximum likelihood estimates are denoted $\hat{\beta}$= $(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_{p-1})^t$.

3. RESULTS AND CONCLUSIONS

In this section we present the main results and ideas that can be found along the study.

First of all, since there are many variables, a variable selection was done. We based the selection in an exploratory data analysis and in principal components analysis.

After a careful data analysis (exploratory data analysis) we discharged the variables which were not statistically relevant (low variability and few observations), and from 204 we ended up with 170 variables.

After this first step, we identified the variables which represent items of a Likert Scale. They are the four groups of UPDRS scale (Non motor aspects of daily life (I), Motor aspects of daily life (II), Motor examination (III) and Motor complications (IV)) and the London scale. From the polytomous Rasch model implemented in R (function *PCM*, package *eRm* [9]), we obtained the individuals' ability and the difficulty item's parameters. Using the Rasch model parameters we transformed the items into new variables: UPDRSI, UPDRSII, UPDRSIII,UPDRSIV and London. With this procedure we arrived at 103 variables.

But, 103 is still a large number of variables compared to only 78 observations. So we decided to compute the Principal Components of some groups of variables. These groups were formed according to their meaning: ESC (group of variables that represent scales), NF (group of variables that represent non motor aspects), SINT (group of variables that represent symptoms of the disease), F (group of the rest of variables that represent motor aspects) and REST (the rest of the variables).

For each group we have retained the principal components that represent at least 60% of the data variability. This way

we had a huge reduction of the variables, and now we have 7 principal components: 2 for the group ESC (ESC1,ESC2)), 1 for the group NF (NF1), 2 for the group SINT (SINT1, SINT2), 1 for the group F (F1) and one for the REST group (REST1). After this process of reduction it was finally possible to apply a logistic regression, which turned out to be easy to interpret and statistically significant.

The response variable (Y) is the progression variable and it is defined as follows:

$$Y = \begin{cases} 0, & \text{the disease evolution is slow} \\ 1, & \text{the disease evolution is fast} \end{cases} \quad (8)$$

Using the function *lm* from R [2], we obtained the estimates of the logistics regression parameters for each principal component considered, shown in Table 1.

Table 1: Estimates of the parameters, estimates of the standard errors and p-values

	Estimate	Stand. Error	p-Value
Intercept	2.7768	0.7922	0.0008
ESC1	-0.0014	0.0015	0.3437
ESC2	0.0085	0.0025	0.0009
NF1	-0.0028	0.0057	0.6195
SINT1	-0.0111	0.0908	0.9029
SINT2	-0.2096	0.1182	0.0807
F1	-0.0111	0.0048	0.0236
RESTO1	0.0035	0.0040	0.3813

As we can see the components with the lowest p-value are ESC2, SINT2 and F1, and so they are the most significant for the model under study. The component F1 is based on the original variables weight and height, but we had rejected them because some data are missing and they are more associated with molecular characteristics. So, in the Table 2, we can see the estimates of the model that considers only ESC2 and SINT2 as predictor variables.

Table 2: Estimates of the parameters, estimates of the standard errors and p-values

	Estimate	Stand. Error	p-Value
Intercept	0.7979	0.1040	4.98e-11
ESC2	0.0088	0.0022	0.0002
SINT2	-0.1844	0.1202	0.1293

This second model has a smaller degree of classification error estimate and lower over-fitting, so we consider it the chosen model.

In Table 3 we present the parameters' estimates, considering the original variables that are in the components ESC2 and SINT2.

Table 3: Parameters' estimates associated with the original variables

	Estimate
Intercept	0.798
UPDRS1	-0.001
UPDRS2	0.006
SEscale	-0.005
MMSE	-0.001
BDItotal	0.001
London	-0.001
FreezingEver	-0.086
MotorSleepBenefitEver	0.172
DysphagiaEver	-0.043
DysarthriaEver	0.086

So we can say that the symptoms referring to Motor Sleep Benefit and Dysarthria are indicators of a fast progression, and the symptoms referring to Freezing and Dysphagia are indicators of slow progression. On the other hand, the patients that are more independent in daily activities (SEscale) and are in a better mental state (MMSE) tend to have a slow progression of the disease, whereas the more depressive patients (BDI) tend to be in the fast progression group.

We expect that this study brings a new contribute to the understanding of Parkinson's disease, mainly in the distinction of the progression groups, and so we can anticipate precociously the group of progression the patient is going to be and then find the best treatment for him.

Another thing that came during this study was the use of the Rasch Models to model the qualitative scales, and how adequate this idea is, or if there is or there is not an easier method to understand and with better results.

4. REFERENCES

[1] R.A. Becker, J.M. Chambers, and A.R. Wilks. *The New S Language: A Programming Environment for Data Analysis and Graphics*. Wadsworth Brooks & Cole, Pacific Grove (EUA), 1 edition, 1988.

[2] J.M. Chambers. Linear models. In *Statistical Models in S*, chapter 4. Wadsworth & Brooks/Cole, 1992.

[3] M.C. de Rijk, L.J. Launer, M.M. Breteler, J.F. Dartigues, and M. Baldereschi. Prevalence of Parkinson's disease in Europe: a collaborative study of population based cohorts. *Neurology*, 54:21–23, 2000.

[4] G.H: Fischer and I.W. Molenaar. *Rasch Models: foundations, recent developments, and applications*. Springer-Verlag, New York, 1994.

[5] I.T. Jolliffe. *Principal Components Analysis*. Springer-Verlag, New York, 2 edition, 2002.

[6] J.A. Keats. Measurement in educational research. In *The International Encyclopedia of Educational Evaluation*, pages 237–244. Pergamon, 1990.

[7] H. Kelderman and C. Rijkes. Loglinear multidimensional irt models for polytomously scored items. *Psychometrika*, 59:149–176, 1994.

[8] R. Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 22:1–55, 1932.

[9] P. Mair and Hatzinger. Extended rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, 20:1–20, 2007.

[10] K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis*. Academic Press, London, 1 edition, 1979.

[11] S. Serneels and T. Verdonck. Principal component analysis for data containing outliers and missing elements. *Computational Statistics and Data Analisys*, 52:1712–1727, 2008.

[12] V. Todorov and P. Filzmoser. An object oriented framework for robust multivariate analysis. *Journal of Statistical Software*, 32:1–47, 2009.