Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: http://www.tandfonline.com/loi/cjas20

A proportional odds transition model for ordinal responses with an application to pig behaviour

I. A. R. de Lara, J. P. Hinde, A. C. de Castro & I. J. O. da Silva

To cite this article: I. A. R. de Lara, J. P. Hinde, A. C. de Castro & I. J. O. da Silva (2016): A proportional odds transition model for ordinal responses with an application to pig behaviour, Journal of Applied Statistics

To link to this article: http://dx.doi.org/10.1080/02664763.2016.1191623

Published online: 02 Jun 2016.

JOURNAL OF APPLIED STATISTICS, 2016
http://dx.doi.org/10.1080/02664763.2016.1191623

Taylor & Francis
Taylor & Francis Group

A proportional odds transition model for ordinal responses with an application to pig behaviour

I. A. R. de Lara[a], J. P. Hinde[b], A. C. de Castro[c] and I. J. O. da Silva[c]

[a]Exact Sciences Department, Luiz de Queiroz College of Agriculture, University of São Paulo, Piracicaba, Brazil; [b]School of the Mathematics, Statistics and Applied Mathematics, National Univerisity of Ireland, Galway, Ireland; [c]Departament of Biosystems Engineering, Luiz de Queiroz College of Agriculture, University of São Paulo, Piracicaba, Brazil

ABSTRACT

Categorical data are quite common in many fields of science including in behaviour studies in animal science. In this article, the data concern the degree of lesions in pigs, related to the behaviour of these animals. The experimental design corresponded to two levels of environmental enrichment and four levels of genetic lineages in a completely randomized 2×4 factorial with data collected longitudinally over four time occasions. The transition models used for the data analysis are based on stochastic processes and Generalized Linear Models. In general, these are not used for analysis of longitudinal data but they are useful in many situations as in this study. We present some aspects of this class of models for the stationary case. The proportional odds transition model is used to construct the matrix of transition probabilities and a function was developed in the R system to fit this model. The likelihood ratio test was used to verify the assumption of odds ratio proportionality and to select the structure of the linear predictor. The methodology used allowed for the choice of a model that can be used to explain the relationship between the severity of lesions in pigs and the use of the environmental enrichment.

ARTICLE HISTORY
Received 7 October 2015
Accepted 16 May 2016

KEYWORDS
Longitudinal data; ordinal response; animal behaviour; transition probabilities

1. Introduction

The study of animal behaviour with a view to establishing better conditions for healthy development is an important area of research in animal husbandry. Indeed, Ítavo *et al.* [15] report that the study of animal behaviour is of great relevance for the rational management of animal production, including housing and diet. There is a strong interest in behaviour studies for various animals such as cattle, goats, pigs, etc, because both of a concern for animal welfare and for improved production and reproductive value. In these areas, it is common to have designed experiments that aim to identify the best management techniques and their relation to the behaviour of animals to improve overall performance. The development of models and statistical methods in this area is also a topic of considerable interest.

CONTACT I. A. R. de Lara idemauro@usp.br

The general approach is to make systematic observations of groups of animals and to record information relating to their behaviour, such as posture, aggressiveness, etc, under different conditions defined by one or more treatments. An intrinsic characteristic of the data measured in these studies is that they are on a nominal or ordinal scale (categorical data) and are generally polytomous, that is, there are more than two categories of possible response. According to Agresti [1] the development of methods for categorical data analysis began in 1960 and was initially stimulated by research related to social and biomedical studies, but the fact is that such data can occur in any scientific area. Paulino and Singer [25] point out that polytomous categorical data analysis is intrinsically multivariate, with the analysis of univariate discrete data (using Poisson, binomial, hypergeometric, negative binomial models) appearing as special cases. Moreover, for these particular cases, in cross-sectional studies where there is only one evaluation of each sample unit, one can use the techniques of Generalized Linear Models (GLM) and extensions, see [24].

On the other hand, an inherent characteristic of many animal behaviour studies is that they are longitudinal, so it is necessary to consider a possible correlation (dependence) between the observations made on the same animal. Two model classes commonly used for longitudinal data analysis are marginal models [18,35] and random effects models [11,23]. Lipsitz *et al.* [20] use a generalized estimating equations to extend marginal models to situations in which the response is multinomial. Hedeker [14] considers a multinomial regression model with the inclusion of random effects to allow for the time dependency over the repeated measurements and also to accommodate heterogeneity from excess zeros, a common feature in categorical data that can lead to overdispersion.

However, there are situations in longitudinal studies, particularly in animal research, where it is likely that the current state of an individual is influenced by the state of the individual on the previous occasion (or previous occasions). The interest focusses in studying the evolution of the individual's response category from one moment of time to another and assessing the possible effects of covariates. In this context, neither marginal models nor random effects models are able to capture directly the nature of these changes of response category over occasions, nor the effects involved in these changes. To meet this goal, we consider Markov transition models. These models are based on stochastic processes and different processes can be used to define transition models for different situations. Here, we consider a discrete time discrete state process (with a finite number of states) and a first-order Markov assumption with transition probabilities:

$$\begin{aligned}\pi_{ab}(t-1,t) &= P(Y_t = b \mid Y_{(t-1)} = a, Y_{(t-2)} = c, \ldots, Y_{(0)} = u) \\ &= P(Y_t = b \mid Y_{(t-1)} = a), \end{aligned} \tag{1}$$

where the states $a, b, c, \ldots, u \in S = \{1, 2, \ldots, k\}$, the finite set of states, and time $t \in \tau = \{0, 1, \ldots, T\}$, the set of observations times. Equation (1) says that an individual's state at time t, Y_t, depends only on the state at the immediately preceding occasion, $Y_{(t-1)}$, and not the complete history of the process. For any originating state a, these transition probabilities satisfy the condition $\sum_{b=1}^{k} \pi_{ab}(t-1,t) = 1$, that is at each occasion the process has to move to one of the k states, that is, the system is closed. To simplify the notation for this first-order process, we write $\pi_{ab}(t-1,t) = \pi_{ab}(t)$. In a more general modelling setting, these transition probabilities can also be made to depend upon explanatory variables, $\boldsymbol{x}$, and we will consider this extension in Section 2.

These models provide an interesting approach to describe the process of moving from one response category to another at successive times and to assess the behaviour of the changes on each occasion. The set of transition probabilities can be written in matrix notation:

$$\boldsymbol{P}(t) = \begin{pmatrix} \pi_{11}(t) & \pi_{12}(t) & \dots & \pi_{1k}(t) \\ \pi_{21}(t) & \pi_{22}(t) & \dots & \pi_{2k}(t) \\ \vdots & \vdots & \dots & \vdots \\ \pi_{k1}(t) & \pi_{k2}(t) & \dots & \pi_{kk}(t) \end{pmatrix},$$

where the argument (t) indicates the dependence on time. For a longitudinal study, as we have here, there are in general T transition matrices with transition probabilities changing over time. These transition terms describe the evolution of the process over time, from one occasion to another. However, a common simplifying assumption is that the process is stationary, that is the transition probabilities are homogeneous over time, and the T transition matrices are stochastically equal, that is, $\boldsymbol{P}(t)$ is constant over time [33]. In this case we can use a common transition matrix $\boldsymbol{P}$. This is a strong assumption, but is often used in practice because it considerably reduces the number of model parameters and provides for a simpler presentation and interpretation of the results. So, for a stationary processes, there are transitions from one state to another but the transition probabilities, π_{ab}, do not depend on t.

For the estimation of transition probabilities, works such as [1,3,5,12,13,19] set this in the context of contingency tables and describe how standard techniques can be used for estimation and group comparison tests. The main advantages of this approach are its simplicity and computational ease. Thus for the case where the sample is homogeneous, that is, no explanatory variables, Anderson and Goodman [3] shows that maximum likelihood estimates of the transition probabilities at time t are $\hat{\pi}_{ab}(t) = n_{ab}(t)/n_{a.}(t-1)$, which are simply the row relative frequencies in the $(t-1)$ to t transition table, that is, the transition frequencies at time t, $n_{ab}(t)$, divided by the marginal totals for the originating states at occasion $(t-1)$, $n_{a.}(t-1)$. Additionally, when the process is stationary over time, the marginal transition frequencies, $\sum_t n_{ab}(t+1)$, are sufficient statistics for the estimation of the elements of the matrix $\boldsymbol{P}$. Thus, the T $(k \times k)$ contingency tables can be collapsed over time into a single table and $\hat{\pi}_{ab} = \sum_{t=1}^{T} n_{ab}(t) / \sum_{t=1}^{T} n_{a.}(t-1)$, the row relative frequencies in this collapsed table.

In practice, in many studies, the sample is rarely homogeneous and differences may be captured through various factors or covariates. For example, if the animals under study are of both sexes, we might expect different transition matrices for males and females and can stratify our data by gender and work with the two sets of contingency tables. Standard tests can used for comparison of the two groups. However, when we have many factors with multiple levels, the cross-table procedure may not be useful, because successive stratification can lead to sparse tables [1]. Also, any continuous covariates need to be categorized, again typically leading to sparse tables.

More generally, we can specify a Transition GLM, to describe the functional relationship of the transition probabilities to the set of available covariates. Here we will simply refer to such models as transition models. Some classical references for binary response data (just two states, i.e. $k=2$) are [4,9,11,12,17,19,23,36]. However, this categorization into dichotomous data often represents a simplification of reality. Ware *et al.* [33] consider

models with more than two response categories using a proportional odds model for the transition probabilities, including the possibility of non-stationarity, that is, allowing the transition matrix to depend on time. GLMs and extensions have some advantages over the cross-table procedure, since the characterization of the stochastic process is given by a conditional regression model, which can be fitted in any available software for GLM. This conditional model framework allows to consider the correlation (time-dependence) in an elegant analytical structure similar to other models for longitudinal data. With the use of these transition models, we can interpret the coefficients as the weight that each term has on the transition probabilities, as well as having various theoretical advantages for selecting significant covariates, testing chain order, assessing predictive ability, etc.

This work is aimed at presenting the extension of the proportional odds model to transition models applied to a longitudinal study with an ordinal response variable, the degree of severity of lesions in pigs, taking account of observed covariates. In contrast to the model presented by Ware *et al.* [33] we assume stationarity, that is the transition matrix is assumed to be homogeneous over time, $\boldsymbol{P}(t) = \boldsymbol{P}$. Likelihood ratio tests are used to assess the appropriateness of the proportional odds assumption and for model selection. A function written in the R software [26] is provided to organize data into a suitable form for fitting these stationary models.

2. Material and methods

2.1. Material

The data arise from part of a research project conducted by Castro [7], during the months of March to July 2014 at a commercial farm group ('Agroceres Pic Génétiporc, Brazil'), where male and female pigs are produced, along with semen of high genetic value. During the research, breeding males of pure and commercial genetic lines were exposed to two rearing conditions (with and without environmental enrichment) during their growth phase. The treatment structure was in a 2×4 factorial, corresponding to combinations of the two rearing conditions and four genetic lineages. The environmental enrichment, factor level E1, was where the housing pens were equipped at different times with suspended chains, a suspended 5 litre plastic container, and a loose 50 litre container. The objects were chosen because they are simple, low cost, easy to fit, and not harmful to the health of the animals. The absence of environmental enrichment, where no objects were supplied is denoted by factor level E2. The genetic lineage levels corresponded to: L1, a synthetic line of mixed breed animals from the races Pietrain, Duroc, Landrace and Large White; L2, a commercial product, for breeding purposes, resulting from crossing between two distinct lineages (Pietrain); L3, a genetic line coming from the Landrace breed of pigs; and L4, a genetic line coming from the Large White breed of pigs.

The animals used in the experiment were trained in childcare and maintained throughout the growth phase. Each group consisted of animals of the same sex, genetic line and size (uniform groups). In all 128 animals were used across the eight treatment combinations. Each treatment combination consisted of a pen with 16 animals, with the animal being considered as the experimental unit. For the purpose of recognition and measurement taking, the animals were identified with different colour and shape eartags. In the analysis, data from 124 animals are used, because for 4 animals there was missing data.

The response variable of interest is a score measuring lesions on the front of the animal. According to Turner *et al.* [30] lesions on the front are the result of reciprocal fights or fights in which the animal is challenged but avoids fighting. Counting the number of skin lesions (lesion scores) has been used to give an indicator of aggressive behaviour among animals and to investigate the development of aggression over time. Here, four evaluations were made, the first after housing the animals in the pens and other three between the changes of enrichment objects. The methodology used by Castro [7] was adapted from the work by Brown *et al.* [6], Melotti *et al.* [22], Tönepohla *et al.* [29], and Turner *et al.* [31].

2.2. Methods

2.2.1. The transition model

Consider a longitudinal study in which $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})'$ is the $(n_i \times 1)$ vector of response variables for the ith individual, where on occasion t there is also an associated $(p \times 1)$ vector of covariates, $\boldsymbol{x}_{it} = (x_{it1}, \ldots, x_{itp})'$. According to Diggle *et al.* [11], a Markov transition model specifies a GLM for the conditional distribution of Y_{it} given the previous responses and the set of covariates. For a general order q (integer) transition model we let $\boldsymbol{h}_{it} = (y_{i(t-1)}, y_{i(t-2)}, \ldots, y_{i,(t-q)})$ be the $(q \times 1)$ vector of the previous responses, that is, the q-step history for individual i at time t. Then, conditionally, the random variable $Y_{it} \mid \boldsymbol{h}_{it}$ is assumed to have a distribution that belongs to a canonical exponential family, that is:

$$f(y_{it} \mid \boldsymbol{h}_{it}) = \exp\left\{\frac{1}{\phi}\left[y_{it}\theta_{it} - b(\theta_{it})\right] + c(y_{it}, \phi)\right\}, \tag{2}$$

where ϕ is a dispersion parameter (supposed known), θ_{it} represents the canonical parameter, which can depend on the history $\boldsymbol{h}_{it}$ and covariates $\boldsymbol{x}_{it}$, and $b(\theta_{it})$ and $c(y_{it}, \phi)$ are functions depending on the specific distribution of the random variable Y_{it}. This is simply the specification of a GLM with the inclusion of an index t, for the repeated observations. It can be shown, as usual, that the conditional mean and variance are:

$$\mu_{it}^C = \mathrm{E}(Y_{it} \mid \boldsymbol{h}_{it}) = b'(\theta_{it}) \quad \text{and} \quad v_{it}^C = \mathrm{Var}(Y_{it} \mid \boldsymbol{h}_{it}) = \phi b''(\theta_{it}).$$

We suppose that the conditional mean and variance satisfy the following equations:

$$g(\mu_{it}^C) = \eta_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \sum_{r=1}^{s} f_r^*(\boldsymbol{h}_{it}; \boldsymbol{\alpha}) \quad \text{and} \quad v_{it}^C = \phi(\mu_{it}^C), \tag{3}$$

where $g(\mu_{it}^C)$ and $v(\mu_{it}^C)$, respectively, represent the link function and the variance function and f_r^* are functions that define the structure of the transition model in the linear predictor. Here f_r^* denotes a function of the history (previous responses) and can be interpreted as additional covariates in the linear predictor. To illustrate, if $s = q = 2$, a function of the history could be $\sum_{r=1}^{2} f_r^*(\boldsymbol{h}_{it}; \boldsymbol{\alpha}) = \alpha_1 f_1^*(y_{i(t-1)}) + \alpha_2 f_2^*(y_{i(t-2)}) = \alpha_1 y_{i(t-1)} + \alpha_2 y_{i(t-2)}$. In this example, the linear predictor of a second-order transition model has the set of covariates $(x_{it1}, \ldots, x_{itp}, y_{i(t-1)}, y_{i(t-2)})'$, with associated parameter vector $(\beta_1, \ldots, \beta_p, \alpha_1, \alpha_2)$.

Therefore, the parameters of primary interest are represented by the vector $\boldsymbol{\delta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})$, in which $\boldsymbol{\beta}$, of dimension $p \times 1$, is associated with the covariates, and $\boldsymbol{\alpha}$ is associated with the history (the previous responses) and has dimension that depends on both the order q

and the specific form of the functions f_r^*. In particular, the dimension of $\boldsymbol{\delta}$ may be greater than $(p+q)\times 1$, if we have more complex functions of the history or include interactions between covariates, between previous responses, or both.

The suggestion in [33] is to fit a model for each occasion as in a cross-sectional study. The $\boldsymbol{\delta}$ parameters are specified for each occasion and are estimated from the separate maximizations of the likelihood functions. This allows the separate modelling of the T possible transitions without any assumptions of stationarity.

For the stationary case, only one model is fitted using a sum of individual contributions to the likelihood function [4,11]. Assuming a Markov transition model of order q, the conditional distribution of $Y_{it} \mid \boldsymbol{h}_{it}$ is expressed by:

$$f(y_{it} \mid \boldsymbol{h}_{it}) = f(y_{it} \mid y_{i(t-1)}, y_{i(t-2)}, \ldots, y_{i(t-q)}),$$

so that the ith individual contribution to the likelihood is given by:

$$f(y_{i1}, y_{i2}, \ldots, y_{iq}) \prod_{t=q+1}^{n_i} f(y_{it} \mid y_{i(t-1)}, y_{i(t-2)}, \ldots, y_{i(t-q)}).$$

Note that the GLM (2) defines only the conditional distribution, so the likelihood of first q observations $f(y_{i1}, y_{i2}, \ldots, y_{iq})$ is not determined directly, except for the case of the normal distribution [11]. Consequently, the full likelihood function cannot be specified. Hence, an alternative is to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ by maximizing the conditional likelihood function:

$$\prod_{i=1}^{N} f(y_{i(q+1)}, \ldots, y_{in_i} \mid y_{i1}, \ldots, y_{iq}) = \prod_{i=1}^{N} \prod_{t=q+1}^{n_i} f(y_{it} \mid \boldsymbol{h}_{it}). \tag{4}$$

If $f_r^*(\boldsymbol{h}_{it}; \boldsymbol{\alpha}) = \alpha_r f_r^*(h_{it})$, then $g(\mu_{it}^C) = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \sum_{r=1}^{s} \alpha_r f_r^*(h_{it})$ is a linear function of both parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and to maximize the conditional likelihood function (4), it is possible to proceed as for the estimation of parameters in a GLM for independent data, by regressing Y_{it}, $t = q+1, \ldots, n_i$, against the $(p+s)$ covariates $(\boldsymbol{x}_{it}, f_1^*(h_{it}), \ldots, f_s^*(h_{it}))$. In general, for a random variable following model (2), the algorithm proposed by Nelder and Wedderburn [24] for fitting GLMs provides maximum likelihood estimates for the parameters $\boldsymbol{\delta}$ in the linear predictor $\boldsymbol{\eta}$ using iteratively weighted least squares. The *difficulty* for the stationary case is computational, as it is necessary to prepare the data in a stacked form and to create additional vectors of previous responses. While it is simple to stack the data, creating the vector of previous responses with missing data (compatible with the order of the chain and consistent with the idea expressed by the likelihood function (4)), needs more care. This is illustrated by the R-code presented in the Appendix and for more details see [11, 23].

In order to check the extent of historical dependence we need to test the Markov order of the process, which for this discrete time and discrete state space setting is a Markov chain,. The typical null hypothesis of interest is:

$$H_0: \text{ the chain is of order } q-1;$$

against the alternative that the chain is of order q. For the model in which there are explanatory variables can use the likelihood ratio test. We write $\ell(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \boldsymbol{y})_{(q-1)}$ and $\ell(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \boldsymbol{y})_{(q)}$ as

the logarithms of the maximized likelihood functions for order $(q-1)$ and q, respectively. The likelihood ratio test statistic is:

$$\lambda = 2(\ell(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \boldsymbol{y})_{(q)} - \ell(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \boldsymbol{y})_{(q-1)}), \tag{5}$$

and under H_0 we would expect $\lambda \sim \chi^2_v$ with degrees of freedom $v = \dim(\boldsymbol{\delta}_q) - \dim(\boldsymbol{\delta}_{(q-1)})$, provided that the same number of observations are used for fitting both models. Therefore, for the lowest order model, the first q observations must be omitted. The stationarity of the process can also be checked by an appropriate likelihood ratio test, using the transition probability estimates and the criterion in [3].

2.2.2. Modification for ordinal data

When the response variable is the ordinal, that is, taking values in ordered set $S = \{1, 2, 3, \ldots, k\}$, as in [21] the proportional odds model can be used to describe the dependence of this variable on one or more covariates. This model provides estimates of (ordered) cumulative probabilities and can be viewed as a multivariate extension of the GLM.

Now the response of the ith individual on the tth occasion becomes a $(k \times 1)$ vector, $\boldsymbol{y}_{it} = (y_{i1t}, y_{i2t}, \ldots, y_{ikt})'$, where $\{y_{ijt}\}$ represent a set of index variables for the response categories, with $y_{ijt} = 1$ if the ith individual is in the jth category at the time t, otherwise $y_{ijt} = 0$. The first-order Markov chain is characterized by adding the response category at the preceding time as an additional covariate in the regression model. In this context, $\boldsymbol{x}_{it} = (x_{it1}, x_{it2}, \ldots, x_{itp}, x_{it(p+1)})'$ is the vector of $(p+1)$ covariates associated with the ith individual at the tth transition, and $x_{it(p+1)}$ is taken as the value of the response at the previous time point. The proportional odds transition model is:

$$\eta = \log\left(\frac{\gamma_{ab}(t)(\boldsymbol{x})}{1 - \gamma_{ab}(t)(\boldsymbol{x})}\right) = \lambda_{ab(t)} + \boldsymbol{\delta}_t' \boldsymbol{x},$$

in which $\lambda_{ab(t)}$ is an intercept (it is not a parameter of practical interest), and the vector $\boldsymbol{x}$ represents the set of the explanatory variable values, which can also vary over time and $\boldsymbol{\delta}_t' = (\beta_{1t}, \ldots, \beta_{pt}, \alpha_t)$ is the vector of the unknown parameters.

Thus, transition probabilities can be estimated as marginal models. The transition cumulative probabilities are specified by:

$$\gamma_{ab}(t)(\boldsymbol{x}) = \frac{\exp(\lambda_{ab(t)} + \boldsymbol{\delta}_t' \boldsymbol{x})}{1 + \exp(\lambda_{ab(t)} + \boldsymbol{\delta}_t' \boldsymbol{x})} \quad b = 1, 2, \ldots, k-1, \tag{6}$$

where:

$$\gamma_{ab}(t)(\boldsymbol{x}) = \mathrm{P}(Y_t \leq b \mid Y_{(t-1)} = a)(\boldsymbol{x}) = \pi_{a1}(t)(\boldsymbol{x}) + \cdots + \pi_{ab}(t)(\boldsymbol{x}).$$

In order to calculate the individual probabilities for each transition we just difference the cumulative distribution function (6):

$$\pi_{aj}(t)(\boldsymbol{x}) = \gamma_{aj}(t)(\boldsymbol{x}) - \gamma_{a(j-1)}(t)(\boldsymbol{x}) \quad a, j \in S,$$

allowing the construction of the transition probabilities matrices.

A necessary condition for the use of the proportional odds model is that the proportional odds ratios assumption holds. In practical terms, this condition is equivalent to

a near linear increase of the odds ratios over the ordered categories. If this assumption is not valid the parameter estimates, associated with the covariates, change according to the response category level. The inclusion of indices j in Equation (7) is used to define the hypotheses relating to this assumption. Formally, we can test this condition using a likelihood ratio test. The functional form of the model to be tested is

$$\gamma_{ab}(t)(\boldsymbol{x}) = \frac{\exp(\lambda_{ab(t)} + \boldsymbol{\delta}'_{jt}\boldsymbol{x})}{1 + \exp(\lambda_{ab(t)} + \boldsymbol{\delta}'_{jt}\boldsymbol{x})} \quad j = 1, 2, \ldots, k-1, \tag{7}$$

with the hypotheses to be tested: $\mathrm{H}_0 : \boldsymbol{\delta}_{jt} = \boldsymbol{\delta}_t \; \forall j$ against $\mathrm{H}_1 : \boldsymbol{\delta}_{jt} \neq \boldsymbol{\delta}_{lt}$ for some $j \neq l, j, l = 1, 2, \ldots, k-1$. The likelihood ratio test statistic is:

$$\Lambda = -2 \log \left[\frac{\mathrm{L}_{H_0}}{\mathrm{L}_{H_1}} \right] \tag{8}$$

in which L_{H_0} denotes the logarithm of the likelihood function under the hypothesis H_0, the proportional odds ratio model, and L_{H_1} represents the logarithm of the likelihood function under the H_1, the general cumulative logits model. If $\Lambda < \chi^2_{(m),\alpha)}$, the hypothesis L_{H_0} is not rejected at level α, indicating that the proportional odds model is adequate. If the test is significant, a more general cumulative logits model is required, with an increase in the number of parameters. Further details about cumulative logits and proportional odds models can be seen in [1,25,32]. These models can be fitted using the computational procedures in the R packages `VGAM` [34], `drm` [16], `mlogit` [10] and `ordinal` [8]. For fitting in SAS [27] the procedures `Logistic` [2] and `CATMOD` [28] are available and Molenberghs and Verbeke [23] provided a macro to fit the set of transitions in the stationary case.

2.2.3. Specific models

Here, for the purpose of analysis the lesions are classified into three categories: 1 corresponding to the absence of lesions; 2 corresponding to a moderate degree of lesions, and 3 corresponding to serious lesions. Moreover, with just four time occasions there are only three transitions of order one and so it is not sensible to consider higher order chains. In addition for simplicity, we also assume stationarity (a homogeneous process over time). Some of the models that are considered for the linear predictor (Equation (6)) are:

- Model 1: all main effects and interaction between lineage and enrichment

$$\begin{aligned}\eta_{klts} = \lambda_{ab} + [&\beta_l \,\text{lineage}_l + \beta_e \,\text{enrichment}_e + \beta_s \,\text{previous response}_s \\ &+ \beta_{le} \,\text{lineage} * \text{enrichment}_{le}].\end{aligned} \tag{9}$$

- Model 2: all main effects and no interaction

$$\eta_{klts} = \lambda_{ab} + [\beta_l \,\text{lineage}_l + \beta_e \,\text{enrichment}_e + \beta_s \,\text{previous response}_s]. \tag{10}$$

- Model 3: no lineage effect

$$\eta_{lts} = \lambda_{ab} + [\beta_e \,\text{enrichment}_e + \beta_s \,\text{previous response}_s]. \tag{11}$$

- Model 4: previous response only

$$\eta_{ts} = \lambda_{ab} + [\beta_s \,\text{previous response}_s]. \tag{12}$$

- Model 5: interaction between enrichment and previous response

$$\eta_{lts} = \lambda_{ab} + [\beta_e \text{ enrichment}_e + \beta_s \text{ previous response}_s + \beta_{es} \text{ enrichment} * \text{previous response}_{es}] \quad (13)$$

with $l = 1,2,3,4$; $e = 1,2$; and $s = 1,2,3$. To choose between the various models we again used likelihood ratio tests (at a 0.05 significance level), also taking into account the principle of parsimony, that is choosing the simplest model among competing adequate models. In this work, the fitting of transition models was done with the packages `VGAM` [34] and `ordinal` [8] available in `R`, version 3.2. To achieve this, a function was developed for fitting of stationary transition models, which enables responses from previous occasions to be embedded as additional covariates, similar to the `Dropout` macro proposed by Molenberghs and Verbeke [23]. This function (available in the appendix) is divided into two parts. The first part is responsible for reading the data set, which is in 'wide' format, that is, one column for each covariate and response (as in a cross-sectional study), and then it is changed to the 'long' format (stacked), as is normally used in longitudinal studies and, in particular, for stationary transition models. The second part of the function is responsible for creating a vector (or vectors for chains of higher order than one) of previous responses. In this part, n is the number of new vectors that we need create and t is the number of times that the response variable was observed.

The test of the proportional odds model, Equation (8), was implemented in the `VGAM` package [34] because it allows this assessment for a set of covariates. In the `ordinal` package this assessment can also be made, but only one covariate at a time. Once the model is selected by likelihood ratio tests, we construct the matrix of transition probabilities as in Equation (6), that is, using the coefficients of the parameters from the selected model. For this, the `ordinal` package [8] has a *predict* function which provides the estimated probabilities given the effects in the model.

3. Results and discussion

Initially we present an exploratory analysis of the data using contingency tables. Table 1 describes the total 372 transitions (124 animals by 3 transitions) of the degree of lesions on the front of the animals. Thus, there are 98 transitions from the condition 'absence of lesion' to others states, 169 transitions from the condition 'moderate degree' and 105

Table 1. Total transitions and estimates of transition probabilities for the degree of lesions in the pig behaviour study.

Previous response ($t-1$)	Future response (t)			Total
	1	1	3	
1	49 (0.50000)	42 (0.42857)	07 (0.07143)	98
2	58 (0.34320)	91 (0.53846)	20 (0.11834)	169
3	24 (0.22857)	48 (0.45714)	33 (0.31429)	105
Total				372

transitions from 'serious degree of lesion'. A simple chi-square test for homogeneity of the rows of Table 1 rejects the null hypothesis ($p < 0.001$), showing that the rows of Table 1 are not homogeneous, that is, the transition probabilities to the states 1, 2 and 3, depend on the previous response. Of course, this simple test disregards the environmental enrichment conditions and different genetic lineages that are present in this study.

Figure 1 illustrates the frequency of animals in states 1 (absence of lesion), 2 (moderate degree), and 3 (serious degree), on each occasion of the study period, for two groups of 62 animals with and without environmental enrichment. Thus, the points at time 0 represent the initial condition, and subsequently their transitions. There is clear evidence of an environmental enrichment effect.

We explore this further in Table 2 where we have classified the transitions of the animal lesion condition by environmental enrichment. The estimate of the transition probability to the good state, that is, with no lesions, is, on average, 1.80 times greater for animals with the environmental enrichment. Focusing on the probability $\pi_{31} = P(Y_{ibt} = 1 \mid Y_{ia(t-1)} = 3)$, note that the odds of transition is almost 4 times higher for the animals with environmental enrichment. Similarly, the probability of leaving state 1 to state 3 is 2.19 times higher if the animal has no environmental enrichment. We could refine this exploration by further stratifying again by the genetic factor, but this begins to lead to rather sparse tables, compromising the data analysis.

A better approach is through modelling of these transitions using the longitudinal ordinal data models. Initially, we check the proportionality model. At this stage we included the effects of genetic lineage, environmental enrichment and the order one Markov covariate, previous response. The likelihood ratio test (Equation (8)) was not significant

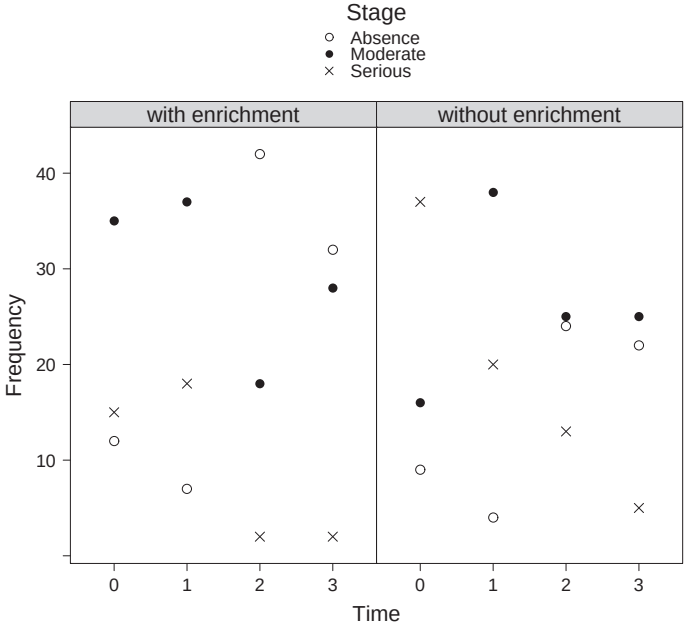

Figure 1. Observed frequency of lesions degree in pigs, at each time, conditioned on previous states of the animals, considering the effect of with and without environmental enrichment.

Table 2. Total transitions and estimates of probabilities for the degree of lesions in the pig behaviour, stratified by the presence or absence of environmental enrichment.

Previous response ($t-1$)	Future response (t) 1	2	3	Total
With environmental enrichment				
1	34 (0.55738)	24 (0.39344)	03 (0.04918)	61
2	31 (0.34444)	48 (0.53333)	11 (0.12222)	90
3	16 (0.45714)	11 (0.31429)	08 (0.22857)	35
				186
Without environmental enrichment				
1	15 (0.40541)	18 (0.48649)	04 (0.10811)	37
2	27 (0.34177)	43 (0.54430)	09 (0.11392)	79
3	08 (0.11429)	37 (0.52857)	25 (0.35714)	70
				186

Table 3. Transition models, null hypotheses, differences of degrees of freedom (d.f.), likelihood ratios statistic (L.R.) and selection of models.

Models	H_0	d.f.	L.R. statistic	p-Value	Selection
Model 1 vs. Model 2	$\beta_{le} = 0$	3	6.8876	0.07593	Model 2
Model 2 vs. Model 3	$\beta_{l} = 0$	3	4.2238	0.23830	Model 3
Model 3 vs. Model 4	$\beta_{e} = 0$	1	6.0589	0.01384	Model 3
Model 3 vs. Model 5	$\beta_{es} = 0$	2	8.2592	0.01609	Model 5

($p = 0.41641$), indicating that we could proceed with the proportional odds model and fit different transition models to explore the effects of the covariates. Various models were fitted to test the significance of the covariates and their interactions, including the particular models listed in Section 2.2.3. The results, which are shown in Table 3, support the selection of the transition model with a linear predictor as in Equation (13). Further, the effect of Markov covariate was also tested and found to be very significant (p-value $= < 0.001$). We also considered the possibility of a chain of order 2, but the likelihood ratio test, as in Equation (5), was not significant.

We consider the results for models 3 and 5. Here, the presentation of both is for instructional purposes. The first category was taken as the reference for all factors in these models. Table 4 shows the parameter estimates for model 3, which has no interaction and is therefore simpler. It is noted that both enrichment and previous response effects are significant.

Table 5 shows the parameter estimates for the selected model 5, which includes the interaction between previous response and enrichment, and allows us to see how much the previous response state may influence the effect of the enrichment covariate. When working with transition models it is always important to consider interactions of the previous response with other covariates. This role is played by the functions f_r^*, shown in

Table 4. Parameter estimates for the transition model 3 (Equation (11)).

Parameters	Estimates	Standard errors	*z* Value	*p*-Value
Intercepts				
λ_{a2}	0.2106	0.2116	0.995	
λ_{a3}	2.6458	0.2582	10.247	
Covariates				
Enrichment (E2)	0.5077	0.2070	2.453	0.0142
Previous response (2)	0.5623	0.2454	2.292	0.0219
Previous response (3)	1.3265	0.2881	4.604	< 0.001

Table 5. Parameter estimates for the transition model 5 (Equation (13)).

Parameters	Estimates	Standard errors	*z* Value	*p*-Value
Intercepts				
λ_{a2}	0.2522	0.2520	1.001	
λ_{a3}	2.7404	0.2937	9.330	
Covariates				
Enrichment (E2)	0.6341	0.4035	1.572	0.1160
Previous response (2)	0.8617	0.3232	2.667	0.0076
Previous response (3)	0.7408	0.4283	1.730	0.0837
Previous(2): enrichment(E2)	−0.6451	0.4987	−1.294	0.1957
Previous(3): enrichment(E2)	0.8144	0.5815	1.401	0.1613

Equation (3). Obviously, when selecting a model with interaction it makes no sense to interpret the main effects parameters.

From the parameter coefficient estimates it is possible to construct the matrix of the fitted transition probabilities, as shown in Tables 6 and 7. In general, these probabilities were more favourable to animals with environmental enrichment. For example, from Table 6, given that an animal is in the good condition (state 1 = absence of lesions), with environmental enrichment, it has probability 0.55244 to continue in the same condition, whereas if it has no environmental enrichment, this probability falls to 0.42625. Whereas if the precondition of the animal is bad (state 3 = serious degree of lesions), it has probability 0.24677 to change to the good state, while without environmental enrichment this probability falls for 0.16470. This shows that there is dependence both on the previous state and the environmental factor, although there are no genetic lineage effects.

Now, in Table 7, given that an animal is in the good condition (state 1 = absence of lesions), with environmental enrichment, it has probability 0.56271 to continue in the same condition, whereas if it has no environmental enrichment, this probability falls to 0.40565. Whereas if the precondition of the animal is bad (state 3 = serious degree of lesions), it has probability 0.38021 to change to the good state, while without environmental enrichment

Table 6. Estimates of the transition probabilities by Model 3 (Equation (11)).

		Enrichment					
		With (E1)			Without (E2)		
Future response		1	2	3	1	2	3
Previous response	1	0.55245	0.38131	0.06625	0.42626	0.46830	0.10545
	2	0.41295	0.47634	0.11071	0.29745	0.53116	0.17139
	3	0.24677	0.54230	0.21092	0.16471	0.52776	0.30754

Table 7. Estimates for the transition probabilities for model 5 (Equation (13)).

		Enrichment					
		With (E1)			Without (E2)		
Future response		1	2	3	1	2	3
Previous response	1	0.56271	0.37665	0.06063	0.40565	0.48585	0.10848
	2	0.35215	0.51530	0.13254	0.35466	0.51404	0.13128
	3	0.38021	0.50054	0.11924	0.12596	0.50842	0.36561

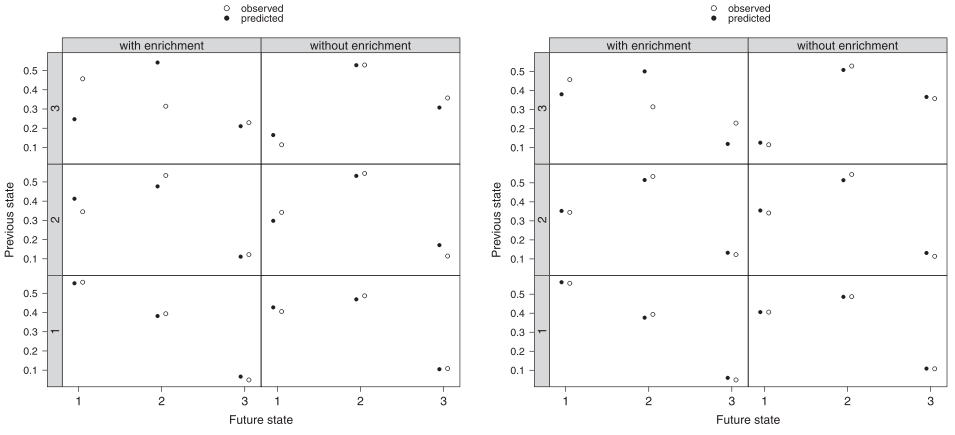

Figure 2. Observed and predicted probabilities from the first-order Markov transition models allowing for the effect of environmental enrichment. Left panel Model 3: without interaction and right panel Model 5: with interaction.

this probability falls to 0.12596. It is also observed that for previous response 2 (moderate degree), the transition probabilities do not differ between the treated and untreated groups.

Finally, to assess the predictive ability of the models, the observed and predicted probabilities are presented in Figure 2. Comparing these figures, we see that the model 5 gives more satisfactory results.

4. Conclusions

Transition models are a class of models that can be used in longitudinal studies when the dependent variable is categorical. The proportional odds model is a simpler alternative than the general model of cumulative logits, but it is necessary to check that proportionality applies for it to be valid, and this is not always the case. In this study, the initial motivation was answering a question of practical research interest: *Does the use of environmental enrichment modify the behavioural pattern of breeding animals?*. The methodology used was appropriate to the form of data and allowed the choice of a model that can be used to explain the relationship between the severity of lesions in pigs and the use, or not, of environmental enrichment. It was found that the use of environmental enrichment is beneficial and gives some degree of animal protection, in that if an animal has moderate or severe lesions then the probability that it will move to a better state is, in general, higher than for the animals receiving no enrichment.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding information

This work was supported by FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) funding agency, Brazil, grant number [2015/02628-2].

References

[1] A. Agresti, *Categorical Data Analysis*, 3rd ed., Wiley, New York, 2012.
[2] P. Allison, *Logistic Regression Using the SAS System: Theory and Application*, SAS Institute Inc., Cary, NC, 1999.
[3] T.W. Anderson and L.A. Goodman, *Statistical inference about Markov chains*, Ann. Math. Statist. 28 (1957), pp. 89–110.
[4] A. Azzalini, *Maximum likelihood estimation of order m for stationary stochastic processes*, Biometrika 70 (1983), pp. 381–387.
[5] Y.M.M. Bishop, S.E. Fienberg, and P.W. Holland, *Discrete Multivariate Analysis: Theory and Practice*, The MIT Press, Massachusetts, 1975.
[6] J.A. Brown, C. Dewey, C.F.M. Delange, I.B. Mandell, P.P. Purslow, J.A. Robinson, E.J. Squires, and T.M. Widowski, *Reliability of temperament tests on finishing pigs in group-housing and comparison to social tests*, Appl. Anim. Behav. Sci. 118 (2009), pp. 28–35.
[7] A.C. Castro, *Comportamento e desempenho sexual de suínos reprodutores em ambientes enriquecidos*, Unpublished doctoral diss., University of São Paulo, Piracicaba, Brazil, 2015.
[8] R.H. Christensen, *Analysis of Ordinal Data with Cumulative Link Models Estimation with the R-Package Ordinal*, 2011. Available at http://www.R-project.org.
[9] D.R. Cox, *The Analysis of Binary Data*, Methuen, London, 1970.
[10] Y. Croissant, *Multinomial Logit Model*, 2013. Available at http://www.r-project.org.
[11] P.J. Diggle, P.J. Heagerty, K.Y. Liang, and S.L. Zeger, *Analysis of Longitudinal Data*, Oxford University Press, New York, 2002.
[12] G.M. Fitzmaurice and N.M. Laird, *A likelihood-based method for analysing longitudinal binary responses*, Biometrika 80 (1993), pp. 141–151.
[13] L.A. Goodman, *Statistical methods for analysing processes of change*, AJS 68 (1962), pp. 57–78.
[14] D. Hedeker, *A mixed-effects multinomial logistic regression model*, Stat. Med. 22 (2003), pp. 1433–1446.
[15] L.C.V. Ítavo, R.M.B.O. Souza, J. Rmoli, C.C.B.F. Ítavo, and A.M. Dias, *Diurnal intake behavior of bovines in continuous or rotational grazing*, Arch. Zootec. 7 (2008), pp. 43–52.
[16] J. Jokinen, *Regression and Association Model for Repeated Categorical Data*, 2013. Available at http://www.helsinki.fi/ ∼ jtjokine/drm.
[17] E.L. Korn and A.S. Whittemore, *Methods for analysing panel studies of acute health effects of air pollution*, Biometrics 35 (1979), pp. 715–802.
[18] K.Y. Liang and S.L. Zeger, *Longitudinal data analysis using generalized linear models*, Biometrika 73 (1986), pp. 13–22.
[19] J.K. Lindsey, *Statistical Analysis of Stochastic Processes in Time*, Cambridge University Press, New York, 2004.
[20] S.R. Lipsitz, K. Kim, and L. Zhao, *Analysis of repeated categorical data using generalized estimating equations*, Stat. Med. 13 (1994), pp. 1149–1163.
[21] P. McCullagh, *Regression methods for ordinal data*, J. R. Stat. Soc. Ser. B 42 (1980), pp. 109–142.
[22] L. Melotti, M. Oostindjer, J.E. Bolhuis, S. Held, and M. Mendl, *Coping personality type and environmental enrichment affect aggression at weaning in pigs*, Appl. Anim. Behav. Sci. 113 (2011), pp. 144–153.
[23] G. Molenberghs and G. Verbeke, *Models for Discrete Longitudinal Data*, Springer-Verlag, New York, 2005.

[24] J.A. Nelder and R.W.M. Wedderburn, *Generalized linear models*, J. Roy. Statist. Soc. Ser. A 135 (1972), pp. 370–384.
[25] C.D. Paulino and J.M. Singer, *Análise de Dados Categorizados*, Edgard Blücher, São Paulo, 2006.
[26] R Development Core Team, *A Language and Environment for Statistical Computing 2.14.1*. Available at http://www.R-project.org.
[27] SAS Institute Inc., *User's Guide SAS/STAT, version 9.3*, Cary/NC, 2011.
[28] M.E. Stokes, C.S. Davis, and G.G. Koch, *Categorical Data Analysis Using the SAS System*, SAS Institute Inc., Cary/NC, 2000.
[29] B. Tönepohla, A.K. Appela, S. Welpa, B. Voß, U.K. von Borstela, and M. Gaulya, *Effect of marginal environmental and social enrichment during rearing on pigs' reactions to novelty, conspecifics and handling*, Appl. Anim. Behav. Sci. 140 (2012), pp. 137–145.
[30] S.P. Turner, M.J. Farbworth, I.M.S. White, S. Brotherstone, M. Mendl, P. Knap, P. Penny, and A.B. Lawrence, *The accumulation of skin lesions and their use as a predictor of individual aggressiveness in pigs*, Appl. Anim. Behav. Sci. 96 (2006), pp. 245–259.
[31] S.P. Turner, R. Roehe, R.B. D'Eath, S.H. Ison, M. Farish, M.C. Jack, N. Lundeheim, L. Rydhmer, and A.B. Lawrence, *Genetic validation of postmixing skin injuries in pigs as an indicator of aggressiveness and the relationship with injuries under more stable social conditions*, J. Anim. Sci. 87 (2009), pp. 3076–3082.
[32] G. Tutz, *Regression for Categorical Data*, Cambridge, New York, 2011.
[33] J.H. Ware, S. Lipsitz, and F.E. Speizer, *Issues in the analysis of repeated categorical outcomes*, Stat. Med. 7 (1988), pp. 95–107.
[34] T.W. Yee, *The VGAM package for categorical data analysis*, J. Stat. Softw. 32 (2010), pp. 1–34.
[35] S.L. Zeger and K.Y. Liang, *Longitudinal data analysis for discrete and continuous outcomes*, Biometrics 42 (1986), pp. 121–130.
[36] S.L. Zeger, K.Y. Liang, and S.G. Self, *The analysis of binary longitudinal data with time-independent covariates*, Biometrika 72 (1985), pp. 31–38.

Appendix

```
#####################################################################

### Data set preparation routine ###

#####################################################################

#First part: changing the data structure

data=read.csv("lesao_frentemodfsep.csv", head=TRUE,
    sep=";", dec=",")
data=na.omit(data)
attach(data)
data2=rbind(data,data,data, data)[,1:3]
id=rep(seq(1:nrow(data)),4)
time=c(rep('t1',nrow(data)),rep('t2',nrow(data)),
rep('t3',nrow(data)), rep('t4',nrow(data)))
response=c(resp1,resp2,resp3,resp4)
newdata=data.frame(id,time,data2,response)
detach(data)
attach(newdata)
newresp=response
newdata=data.frame(newdata,newresp)
newdata<-newdata[order(id),]
head(newdata)
```

```
#Second part: "dropout" function for create the
   previous response

desloca=function(data,response,n,t){
varind=matrix(0,nrow(data),n)
for(j in 1:n){
varind[,j]=c(rep(0,j),response[1:(nrow(data)-j)])
varind[seq(j,nrow(data),t),j]=NA
if(j>1)
varind[which(is.na(varind[,(j-1)])),j]=NA}
list('newdata'=data.frame(data,varind))}

detach(newdata)
drop=desloca(newdata,newdata[,"newresp"],1,4)
drop=data.frame(drop)
head(drop)
names(drop)
names(drop)=c("id","time","lineage","enrch","animal",
"resp","newresp","prev1")
drop=na.omit(drop)
head(drop)

#####################################################################
```