Scientia Agricola
http://dx.doi.org/10.1590/1678-992X-2016-0067

Longitudinal model for categorical data applied in an agriculture experiment about elephant grass

Vinícius Menarin[1]*, Idemauro Antonio Rodrigues de Lara[1], Sila Carneiro da Silva[2]

[1]University of São Paulo/ESALQ – Exact Sciences Dept. – C.P. 09 – 13418-900 – Piracicaba, SP – Brazil.
[2]University of São Paulo/ESALQ – Animal Science Dept.
*Corresponding author <vinicius.menarin@gmail.com>

Edited by: Marcin Kozak

Received February 22, 2016
Accepted September 07, 2016

ABSTRACT: Experiments where the response is a categorical variable are usually carried out in many fields such as agriculture. In addition, in some situations this response has three or more levels without an order between them characterizing a multinomial (nominal) response. Statistical models for scenarios where the observations of a nominal response can be considered independent have an extensive literature, such as the baseline-category logit models. However, situations where this assumption is violated (as in longitudinal studies) require specific models that take into consideration the dependence between observations. In this paper, a fairly new extension of the generalized estimating equations is applied to analyze an experiment carried out to investigate the type of vegetation observed in an elephant grass pasture, according to some management conditions over time. This extension uses local odds ratios to explain the dependence among the categories of the outcome over the repeated measurements. Two different structures were compared to describe this dependence, and the Wald test was used to select the significant variables. Further, we built confidence intervals for the predicted probabilities of occurrence of each category and assessed the results comparing observed/predicted values and using the diagnostic analysis. The results allowed to conclude that there are various significant effects for treatments and for time. The structure of local odds ratio also proved as a good way to describe the dependence between categorical responses over time.

Keywords: type of vegetation, longitudinal multinomial data, generalized estimating equations, local odds ratio

Introduction

Understanding forage plant growth and colonization abilities (horizontal and vertical distribution of plant organs and parts) is important to provide a sound basis for planning and idealizing grazing management practices that ensure pasture productivity and persistence. For tall-tufted tussock forming species like elephant grass (*Pennisetum purpureum* Schum. cv. Napier), horizontal distribution is especially important because it is related to efficiency of carbon (energy) uptake (Ryel et al., 1994), competitive ability and stability of plant population and pasture productivity (Pereira et al., 2015a,b).

In many research fields, the analysis of categorical data is present, including agriculture. A categorical variable has its measurement scale formed by a set of categories (Agresti, 2007), for example, the type of vegetation present on an area. This type of variable can be classified as binary (2 categories) or polytomous (3 or more) and as ordinal and nominal. An ordinal categorical variable means that its categories have a natural order, while the nominal category has no order between the levels. This paper focuses on the nominal case for polytomous variables and the generalized linear models framework is commonly used to analyze these data. However, in some cases, the studies are carried out in such a way that several measurements are taken from the same subject (or sample unit). When measurements are taken over time, these studies generate data sets known as longitudinal data and it is necessary to consider some correlation measure (Diggle et al., 2002).

Among the several available approaches in the literature (Diggle et al., 2002; Pinheiro and Bates, 2000; Verbeke and Molenberghs, 2000), the generalized estimating equations (GEE) consist in a useful methodology for longitudinal data to consider the dependence between the observations (Liang and Zeger, 1986). However, when the response variable is multinomial, the original GEE approach does not apply and Lipsitz et al. (1994) and Touloumis et al. (2013) have developed some extensions. The objective of this paper was to describe some aspects of the GEE methodology and show an application in agriculture, where, in general, this methodology is not usual. Confidence intervals are not commonly used in this type of model thus, as a contribution, in this paper they are obtained for each of the response categories. Finally, the appendix presents the R code used in the analysis.

Materials and Methods

The data used in this work are results from an experiment carried out in Piracicaba, São Paulo State, Brazil, on an elephant grass pasture (*Pennisetum purpureum* Schum. cv. Napier) grazed by dairy cows (Pereira et al., 2015a,b). It is a complete randomized block design with the treatments allocated according to a 2 × 2 factorial arrangement, where treatments are the combinations of two pre-grazing conditions (95 % and maximum (98 %)

canopy light interception during regrowth) and two post-grazing heights (35 and 45 cm). The experiment was carried out from Jan 2011 until Apr 2012, period classified into six seasons presented in Table 1.

The response analyzed in the study is the type of vegetation observed in the field, which can be tussocks, bare ground and weeds. Forty (40) points were observed in each one of the four paddocks in each block (Figure 1). Since there are always 40 points observed in each paddock, we can analyze the proportions of each type of vegetation under the total, characterizing a multinomial outcome with three levels. There are 40 × 16 = 640 points per season, but in the early spring, one of the paddocks was affected by climate conditions and thus the total number of observations was $N = 640 \times 6 \times 40 = 3800$. Furthermore, there might be a spatial correlation between the observations, but it is not the main focus of this paper and the spatial coordinates of the points were not available.

Logit models

For independent observations of a multinomial outcome, the baseline logit models (also known as baseline-category logit models) are a well-known theory in the literature. These models are an extension of logistic regression models used to model binary outcomes (Agresti, 2002; Agresti, 2007; Dobson, 2008). Denoting the categories by (k = 1, ..., K), the model pairs each category with a baseline one and is defined as

$$ln\left(\frac{\pi_k}{\pi_K}\right) = \eta_k = x^{'}\beta_k, k = 1,\ldots,K-1$$

Table 1 – Seasons when the experiment was carried out.

Year	Months	Season
2011	Jan – Mar	Summer 1
2011	Apr – June	Autumn
2011	July – Sept	Winter
2011	Oct – mid-Nov	Early spring
2011	mid-Nov – Dec	Late spring
2012	Jan – Apr	Summer 2

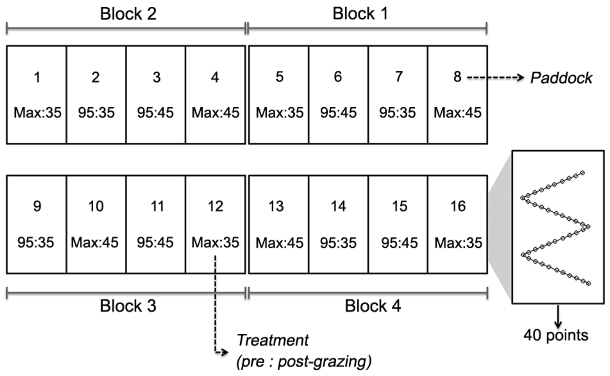

Figure 1 – Sketch of the experiment carried out.

where: η_k denotes the linear predictor, x is the design matrix and β is a vector of covariates, according to the generalized linear models (GLM) theory (Nelder and Wedderburn, 1972). The index k in the linear predictor indicates that each one of the (k – 1) ogits has separate parameters. Clearly, when k = 2, the model reduces to an ordinary logistic regression for binary outcomes. However, as this model is part of the GLM theory, it is suitable only when observations are independent from each other, requiring modifications to handle the temporal dependence.

Generalized estimating equations using a local odds ratios approach

A very important extension of GLMs for longitudinal data, the generalized estimating equations (GEE) briefly consist of an approach where the regression and within-subject correlation are modelled separately, i.e., besides the regression parameters β, we need to specify a "working" correlation matrix $R_i(\alpha)$ consider the dependence between the repeated measurements, reducing the standard errors of the parameter estimates. The parameter estimation methods are based in quasi-likelihood methods, no more in full-likelihood methods, as in GLM (Liang and Zeger, 1986; Zeger and Liang, 1992).

Even though this original approach can be used for discrete and continuous outcomes, it does not apply for the multinomial case (only for the binary) and Lipsitz et al. (1994) extended it to handle this type of data. This extension seems to be a very useful tool, but it is implemented in statistical software only for the ordinal case. Thus, as the focus of this paper is the nominal scenario, no more details will be discussed here and the referred reference can be used.

Another extension of the original GEE approach that deals with nominal and ordinal responses, in some aspects very similar to Lipsitz et al. (1994), was proposed by Touloumis et al. (2013). This approach uses local odds ratios to describe the association between the response categories across the repeated measurements. It is implemented in the R package *multgee* and further details will be discussed later.

Introducing notes used by Lipsitz et al. (1994) and Touloumis et al. (2013), Y is the categorized response with K levels (k = 1, ..., K), $K > 2$. K defines indicator variables $Y_{itk} = I(Y_{it} = k)$ showing if the i-th subject presented category k at time t (t = 1, ..., T_i). These indicator variables can be converted into a (K – 1) × 1 vector of responses $Y_{it} = [Y_{it1,} Y_{it2,} ..., Y_{it(K-1)}]'$ and $Y_i = [Y_{i1,} Y_{i2,} ..., Y_{iTi}]'$. The marginal distribution of Y_{it} is multinomial

$$f\left(y_{it} \mid x_{it}, \beta\right) = \prod_{k=1}^{K} \pi_{itk}^{y_{itk}},$$

where: π_{itk} is the probability of occurrence of category k at time t and $x_i = \left[x_{i1}^{'},\ldots,x_{iTi}^{'}\right]$ denotes the $T_i(K-1) \times p$ matrix of covariate values for subject i. The probabilities of interest π_{itk} are also converted into a vector $\pi_{it} = [\pi_{it1}, \ldots \pi_{it(K-1)}]'$. The marginal expected vector, $E[Y_{it} | x_i] = \pi_{it}$, is modelled by

$$g\,(E\,[Y_{it}\,|\,x_i]) = g(\pi_{it}) = x_{it}'\beta,$$

where: the choice of link vector g is the baseline-category logit model for the multinomial case.

Briefly, the association structure is described by the vector of local odds ratios $\alpha = [\psi_{1121}, \ldots, \psi_{112(K-1)}, \ldots, \psi_{(T-1)1T1}, \ldots, \psi_{(T-1)(K-1)T(K-1)}]'$. These local odds ratios $\psi_{tkt'k'}$ are taken at the cutpoints (k, k') at the marginalized contingency table for the time pair (t, t'). A generalized version of the rows and columns (RC) model (Becker and Clogg, 1989; Goodman, 1985) is fitted and, under this model, the local odds ratios satisfy

$$ln\left(\psi_{tkt'k'}\right) = \gamma_{tt'}\left(\omega^{tt'}_{tk} - \omega^{tt'}_{t(k+1)}\right)\left(\omega^{tt'}_{t'k'} - \omega^{tt'}_{t'(k'+1)}\right),$$

where: the sets of parameters γ and ω are called intrinsic and score parameters, respectively. All these odds ratios may be presented as a matrix of dimensions $T(K-1) \times T(K-1)$:

$$\begin{bmatrix} \\ 0 & \Psi_{12} & \cdots & \Psi_{1T} \\ \Psi_{21} & 0 & \cdots & \Psi_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ \Psi_{T1} & \Psi_{T2} & \cdots & 0 \end{bmatrix}.$$

Each block has dimension $(K-1) \times (K-1)$ and each element (k, k') is the local odds ratio estimate $\hat{\Psi}_{tkt'k'}$. This matrix can assume two different types representing the two association structures available for the multinomial case (Touloumis et al., 2013): i) Time exchangeable: under this simpler structure the local odds ratios are simplified to $ln\,(\psi_{tkt'k'}) = \gamma(\omega_k - \omega_{k+1})\,(\omega_{k'} - \omega_{k'+1})$, which does not assume any time dependency and implies that the matrix blocks are equal. ii) RC structure allows different odds ratios between the time pairs, i.e., there is time dependency. In this case, we have $ln\left(\psi_{tkt'k'}\right) = \gamma_{tt'}\left(\omega^{tt'}_{k} - \omega^{tt'}_{k+1}\right)\left(\omega^{tt'}_{k'} - \omega^{tt'}_{k'+1}\right)$.

As described by Touloumis et al. (2013), cases when the intrinsic parameters have a large variability should be analyzed with the RC structure; otherwise, if this variability is small, one should choose the time exchangeable structure.

The significance of effects of covariates can be assessed by the Wald statistical test described in Touloumis (2013). Two nested models can be tested and the rejection of the null hypothesis of the Wald test suggests that the model with fewer parameters is more appropriate.

Overall, the procedure for parameters estimation is very similar to the original GEE approach from Liang and Zeger (1986), more details can be seen in Touloumis et al. (2013) and a description of the functions available in *multgee* package is available in Touloumis (2015).

Statistical model and procedures for selection

Here we present the methods used to select the statistical model. The procedure can be understood as two steps where the first is the choice of the association structure and the second one is the way that the covariates enter the linear predictor. This is done in this order because the parameter estimates depend on the selected association structure, according to Touloumis (2013).

The range of intrinsic parameter estimates under the RC structure can be assessed to select the association structure: if these estimates do not differ much, we select the time exchangeable structure; otherwise, we use the RC one.

Once the association structure is selected, we proceed to the second step and selection of the linear predictor. Here, as previously described, we can use the Wald test to compare several nested models, available in the *multgee* package. As the experiment analyzed here is a factorial in a complete randomized block design, we proposed the models presented in Table 2, where the pre and post-grazing factors were abbreviated as pre and post respectively. These models allow to test the effects of interactions and main effects of the covariates and the selected model will be presented later. Note: the first model considers all the interactions between pre, post and seasons.

Let π_{itk} the probability of i-th point ($i = 1, \ldots, 640$) be classified in the k-th category in season t. The model is given by two logits

$$logit_{it1} = ln\left(\frac{\pi_{it1}}{\pi_{it3}}\right) logit_{it2} = ln\left(\frac{\pi_{it2}}{\pi_{it3}}\right).$$

The types of vegetation were ordered as tussock, weeds and bare ground in the model fitting ($k = 1, 2, 3$). Hence, the first logit relates tussocks and bare ground and the second logit is the relationship between weeds and bare ground. Furthermore, as an example for the first point ($i = 1$), Y_i is given by $Y_i = [(1,0,0)', (1,0,0)', (0,0,1)', (1,0,0)', (0,1,0)', (0,0,1)']'$ because the first point was observed as tussock, tussock, bare ground, tussock, weed and bare ground in each of the six seasons ($t = 1, 2, \ldots, 6$).

Confidence intervals for the predicted probabilities can be obtained as described in Agresti (2002), as this model is an extension of a baseline logit model. The difference is that here, we use the robust standard errors to build the intervals instead of the naive ones. The assessment of the goodness-of-fit of the model is done with plots of residuals and comparisons between observed and predicted proportions.

Table 2 – Statistical models proposed to the data analyzed.

Model	Linear predictor structure
1	Block + pre*post*season
2	Block + pre + post + season + pre × post + pre × season + post × season
3	Block + pre + post + season + pre × post
4	Block + pre + post + season + pre × post + post × season
5	Block + pre + post + season + pre × post + pre × season
6	Block + pre + post + season + pre × season
7	Block + pre + season + pre × season

Results and Discussion

Descriptive analysis

First, we present a description of the data. Figure 2 shows the proportions of each type of vegetation observed according to treatments and seasons. Tussocks are predominant over the other two types of vegetation, but when light interception is maximum, there are more places with bare ground and fewer with tussocks. Both the pre and post-grazing factors seem to influence the type of vegetation. Also, analyzing this graph, we can see a possible interaction between the pre-grazing factor and the seasons. For 95 % of light interception, the behavior of the trajectories of tussocks and bare ground is quite different for trajectories at maximum interception.

Statistical model

As the first step of the model building, the range of the estimates of intrinsic parameters (γ) varies from -0.0833 to 0.8438 indicating that the RC structure may be more appropriate in this case, as these estimates are not so close and cover positive and negative values.

Now we fit the models presented in Table 2 under the RC structure and compare them using the Wald test described above. According to Table 3, which displays the results of these comparisons, model 6 is selected to describe the data and considers the following effects: blocks, pre and post-grazing, season, and the interaction between pre-grazing and season already remarked in the descriptive analysis (Figure 1).

Therefore, the model can be written using dummy variables (first level as reference) as:

$$
\begin{aligned}
&ln\left(\frac{\pi_{itk}}{\pi_{it3}}\right)\\
&= \beta_{0k} + \underbrace{\beta_{1k}X_{11i} + \beta_{2k}X_{12i} + \beta_{3k}X_{11i}}_{Block} + \underbrace{\beta_{4k}X_{2i}}_{\text{pre−grazing}} + \underbrace{\beta_{5k}X_{3i}}_{\text{post−grazing}}\\
&+\underbrace{\beta_{6k}X_{41t} + \beta_{7k}X_{42t} + \beta_{8k}X_{43t} + \beta_{9k}X_{44t} + \beta_{10,k}X_{45t}}_{season}\\
&+\underbrace{\beta_{11,k}X_{2i}X_{41t} + \beta_{12,k}X_{2i}X_{42t} + \beta_{13,k}X_{2i}X_{43t} + \beta_{14,k}X_{2i}X_{44t} + \beta_{15,k}X_{2i}X_{45t}}_{\text{pre−grazing× season}}
\end{aligned}
$$

where: i = 1, ..., 640 is the measurement point in the field, t is the season and k = 1, 2. Table 4 displays the parameters estimates with the standard errors shown between parentheses, and significant parameters at 5 % of significance are followed by the * symbol.

Figure 3 presents a very satisfactory plot of observed and predicted values. In Figure 4, the ordinary residuals of the model are displayed confirming a good fit (there are no outliers or patterns in residuals, which have their means close to 0 – red line). Confidence intervals for the predicted probabilities are presented in Figure 5.

Table 3 – Results of comparisons in the Wald test between the nested models proposed.

Comparison	Degrees of freedom difference	*p*-value
Model 1 × model 2	10	0.7107
Model 2 × model 3	20	0.0157
Model 2 × model 4	10	0.0160
Model 2 × model 5	10	0.1623
Model 5 × model 6	2	0.1602
Model 6 × model 7	2	0.0212

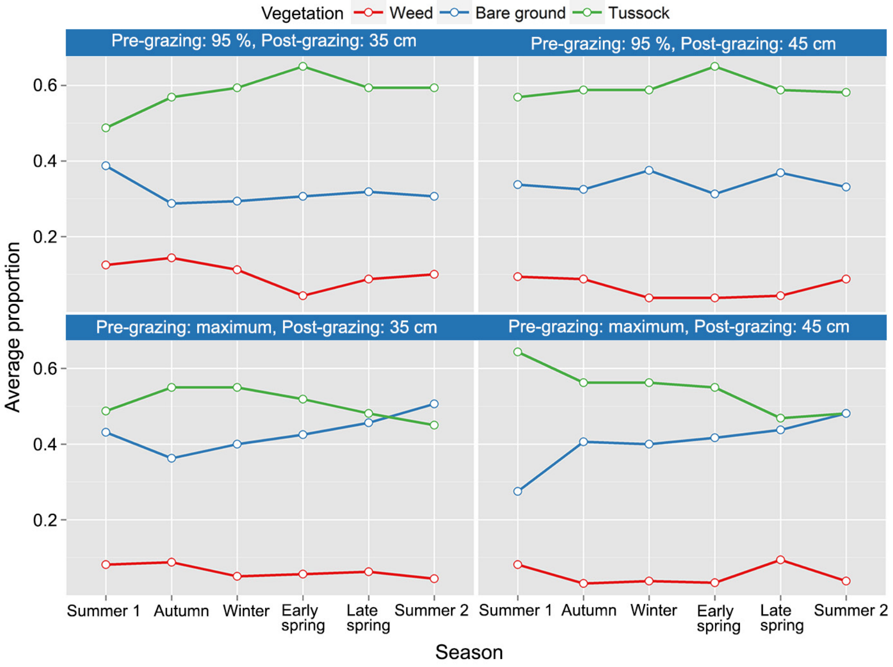

Figure 2 – Average proportions of each type of vegetation according to treatments and seasons.

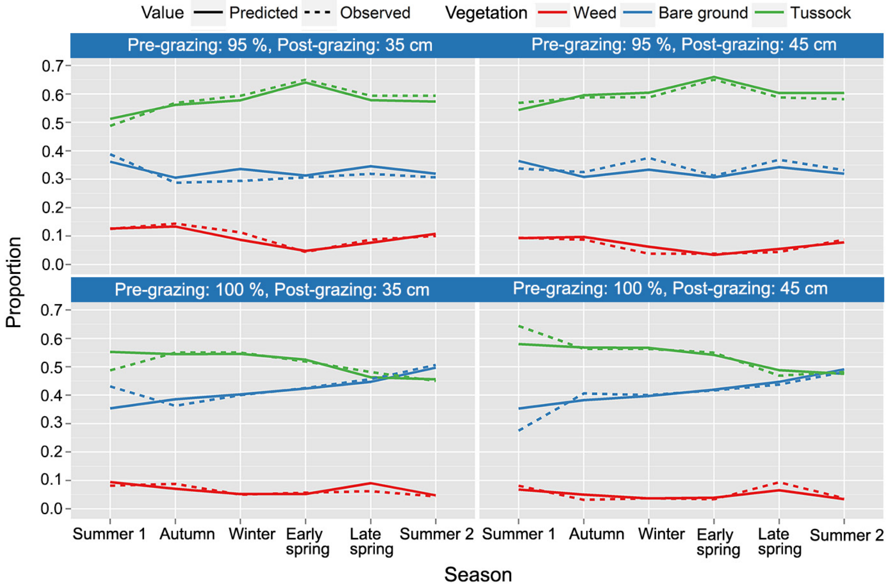

Figure 3 – Comparison between predicted values by the fitted GEE model and observed data.

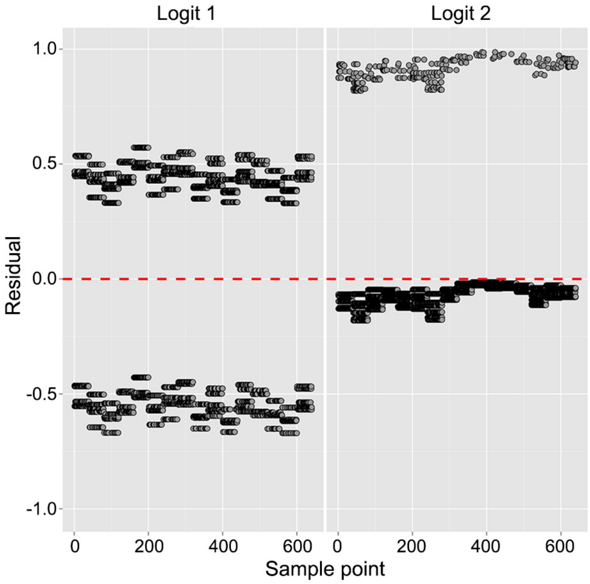

Figure 4 – Residuals of the fitted GEE model with their means represented by the red line.

Table 4 – Estimates for the regression parameters of the fitted GEE model using local odds ratios to describe the association structure.

Parameter		Logit 1 (k = 1)		Logit 2 (k = 2)	
		Tussocks vs bare ground		Weeds vs bare ground	
Intercept	β_{0k}	0.265	(0.137)	-0.764*	(0.220)
Block 2	β_{1k}	0.174	(0.092)	0.137	(0.176)
Block 3	β_{2k}	0.050	(0.089)	-1.195*	(0.233)
Block 4	β_{3k}	0.118	(0.096)	-0.442*	(0.191)
Pre (maximum)	β_{4k}	0.095	(0.170)	-0.282	(0.294)
Post (45 cm)	β_{5k}	0.052	(0.066)	-0.329*	(0.142)
Autumn	β_{6k}	0.259	(0.174)	0.220	(0.275)
Winter	β_{7k}	0.193	(0.170)	-0.314	(0.301)
Early spring	β_{8k}	0.366*	(0.176)	-0.851*	(0.349)
Late spring	β_{9k}	0.165	(0.181)	-0.469	(0.301)
Summer 2	$\beta_{10,k}$	0.235	(0.167)	-0.042	(0.272)
Pre : autumn	$\beta_{11,k}$	-0.361	(0.246)	-0.601	(0.428)
Pre : winter	$\beta_{12,k}$	-0.335	(0.240)	-0.422	(0.460)
Pre : early spring	$\beta_{13,k}$	-0.597*	(0.252)	0.075	(0.516)
Pre : late spring	$\beta_{14,k}$	-0.576*	(0.246)	0.194	(0.444)
Pre : summer 2	$\beta_{15,k}$	-0.766*	(0.238)	-0.990*	(0.468)

*Significant parameters at 5 % level.

Interpretations about the parameter estimates in baseline logit models are usually done by odds ratios. Among several interpretation that can be done (between blocks, seasons, pre and post-grazing), here the focus is only on the pre and post-grazing conditions, as they are of practical interest and are the main covariates. Table 5 summarizes the odds ratios and respective confidence intervals (CI) comparing places with maximum light interception versus 95 % (pre-grazing). Further, because of the interaction between these factors with seasons, there are different odds ratios for each season.

As an illustration, according to Table 5, the estimated odds of tussock is around 60 % of the odds of bare ground, for places where there is maximum light interception against those where there is 95 % in the early spring season:

$$\frac{exp\left(\hat{\beta}_{01}+\hat{\beta}_{41}+\hat{\beta}_{81}+\hat{\beta}_{13,1}\right)}{exp\left(\hat{\beta}_{01}+\hat{\beta}_{81}\right)}=exp\left(\hat{\beta}_{41}+\hat{\beta}_{13,1}\right)=exp\left(0.095-0.597\right)=0.61$$

Also, the estimated odds of tussock is 1.85 times greater than the odds of weeds, for places where there

Table 5 – Odds ratios estimates by the GEE model comparing pre-grazing levels.

Season	Pre-grazing (maximum × 95 %)					
	Tussock × bare ground		Weed × bare ground		Tussock × weed	
	Estimate	CI (95 %)	Estimate	CI (95 %)	Estimate	CI (95 %)
Summer 1	1.10	(0.79; 1.53)	0.75	(0.42; 1.34)	1.46	(0.84; 2.53)
Autumn	0.77	(0.55; 1.07)	0.41	(0.22; 0.76)	1.85	(1.03; 3.34)
Winter	0.79	(0.57; 1.09)	0.49	(0.24; 1.00)	1.59	(0.80; 3.18)
Early spring	0.61	(0.43; 0.85)	0.81	(0.36; 1.84)	0.74	(0.33; 1.66)
Late spring	0.62	(0.45; 0.86)	0.92	(0.48; 1.73)	0.68	(0.36; 1.27)
Summer 2	0.51	(0.37; 0.71)	0.28	(0.14; 0.56)	1.83	(0.92; 3.63)

CI = confidence intervals.

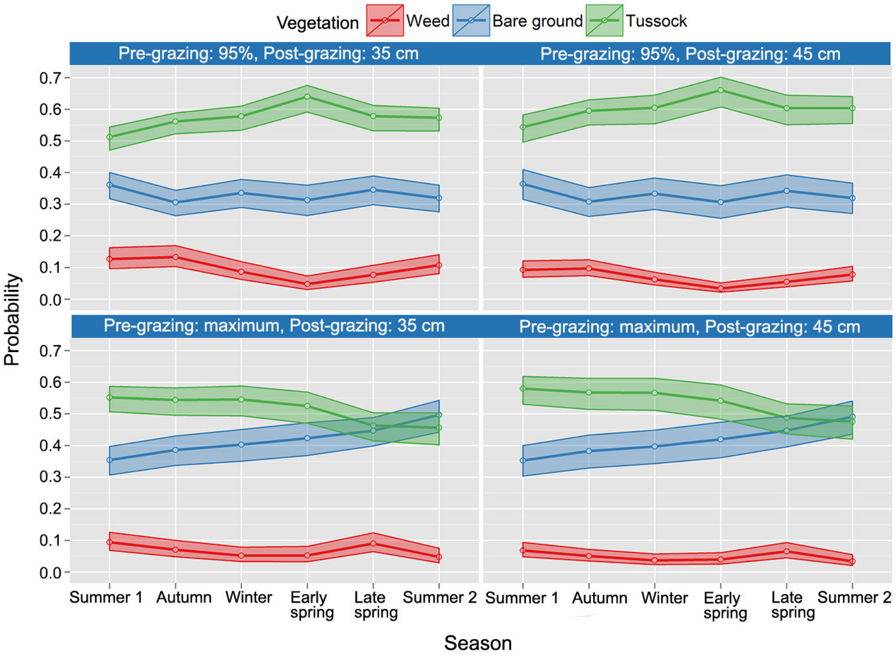

Figure 5 – Confidence intervals (95 %) for the predicted probabilities by the GEE model.

is maximum light interception against those with 95 %, in the autumn:

$$\frac{exp\left[\left(\hat{\beta}_{01}-\hat{\beta}_{02}\right)+\left(\hat{\beta}_{41}-\hat{\beta}_{42}\right)+\left(\hat{\beta}_{81}-\hat{\beta}_{82}\right)+\left(\hat{\beta}_{11,1}-\hat{\beta}_{11,2}\right)\right]}{exp\left[\left(\hat{\beta}_{01}-\hat{\beta}_{02}\right)+\left(\hat{\beta}_{81}-\hat{\beta}_{82}\right)\right]}$$

$$= exp\left[(0.095+0.282)+(-0.361+0.601)\right]=exp(0.617)=1.85$$

Similar interpretations are done for the rest of the table, remarking that confidence intervals containing the value 1 do not indicate significant differences.

Table 6 presents the estimated odds ratios estimates and respective 95 % confidence intervals comparing the levels of post-grazing (45 cm versus 35 cm). These results are simpler those in Table 5 because there is no interaction with seasons. As an example of interpretation, the estimated odds of weed are around 70 % of the odds of bare ground for places with height of 45 cm for 35 cm. Moreover, the estimated odds of tussocks is 1.46 times greater than the odds of weeds for places with height of 45 cm against 35 cm.

With less relevance to the model, we have the local odds ratios estimating related to the association structure (these parameters are treated as nuisance, i.e., of secondary interest), displayed in the following matrix:

$$\begin{array}{cc:cc:cc:cc:cc:cc}
0 & 0 & 1.20 & 0.79 & 1.42 & 0.66 & 1.00 & 1.00 & 1.01 & 0.99 & 2.89 & 0.43 \\
0 & 0 & 0.79 & 1.36 & 0.66 & 1.64 & 1.00 & 1.00 & 0.99 & 1.01 & 0.43 & 1.93 \\
\hdashline
1.20 & 0.79 & 0 & 0 & 3.13 & 0.42 & 2.24 & 0.57 & 2.60 & 0.30 & 2.78 & 0.32 \\
0.79 & 1.36 & 0 & 0 & 0.42 & 1.92 & 0.57 & 1.47 & 0.30 & 4.54 & 0.32 & 3.61 \\
\hdashline
1.42 & 0.66 & 3.13 & 0.42 & 0 & 0 & \mathbf{2.64} & \mathbf{0.31} & 1.13 & 0.76 & 1.00 & 1.00 \\
0.66 & 1.64 & 0.42 & 1.92 & 0 & 0 & \mathbf{0.31} & \mathbf{4.04} & 0.76 & 1.87 & 1.00 & 1.04 \\
\hdashline
1.00 & 1.00 & 2.24 & 0.57 & 2.64 & 0.31 & 0 & 0 & 1.83 & 0.54 & 3.87 & 0.29 \\
1.00 & 1.00 & 0.57 & 1.47 & 0.31 & 4.04 & 0 & 0 & 0.54 & 1.85 & 0.29 & 3.10 \\
\hdashline
1.01 & 0.99 & 2.60 & 0.30 & 1.13 & 0.76 & 1.83 & 0.54 & 0 & 0 & 1.10 & 0.88 \\
0.99 & 1.01 & 0.30 & 4.54 & 0.76 & 1.87 & 0.54 & 1.85 & 0 & 0 & 0.88 & 1.18 \\
\hdashline
2.89 & 0.43 & 2.78 & 0.32 & 1.00 & 1.00 & 3.87 & 0.29 & 1.10 & 0.88 & 0 & 0 \\
0.43 & 1.93 & 0.32 & 3.61 & 1.00 & 1.04 & 0.29 & 3.10 & 0.88 & 1.18 & 0 & 0
\end{array}$$

The bold highlighted block is interpreted as: i) the estimated odds of a point that was a tussock in the winter becomes a weed instead of a tussock in early spring is 2.6 times the corresponding odds for a point that was a weed in the winter; ii) the estimated odds of a point that was a tussock in the winter becomes bare ground instead of a weed in early spring is 30 % of the corresponding odds for a point that was a weed in the winter;

Table 6 – Odds ratios estimates by GEE model comparing post-grazing levels.

Post-grazing (45 cm × 35 cm)					
Tussock × bare ground		Weed × bare ground		Tussock × weed	
Estimate	CI (95 %)	Estimate	CI (95 %)	Estimate	CI (95 %)
1.05	(0.93; 1.20)	0.72	(0.54; 0.95)	1.46	(1.12; 1.91)

CI = confidence intervals.

iii) the estimated odds of a point that was a weed in the winter becomes bare ground instead of a weed in early spring is 4 times the corresponding odds for a point that was bare ground in the winter. We notice that many of the local odds ratios in the matrix are close to 1, which suggests that the association between the correlated measurements is not so strong as expected. Comparing the naive and robust standard errors of the parameter estimates (i.e., a model that does not consider the dependence compared to the proposed model here) the difference between them is quite small.

Lastly, as a practical brief of Table 5 and Table 6 (the main results of the model), if the objective is to stimulate the occurrence of tussocks instead of weeds, it is advisable to choose maximum light interception (only in the autumn) and/or 45 cm height (for any season). These management options either stimulate the occurrence of bare ground instead of weeds (for any season considering the post-grazing and in the autumn and the second summer for pre-grazing). If the objective is the occurrence of tussocks, instead of bare ground, the suggestion is to choose 95 % of light interception in the spring and the second summer.

Conclusions

This paper presented an application of a recent methodology for multinomial correlated responses. This methodology is an extension of the well-known generalized estimating equations (GEE) approach that consists in describing the dependence structure among correlated measurements in a way that makes sense for categorized (nominal) responses, using local odds ratios for that.

In the experiment, the purpose was to investigate how the management conditions affect the type of vegetation observed in the field. We conclude that both pre and post-grazing conditions affect the proportion of tussocks, weeds and places with bare ground, and there is also an effect of the season. The statistical model allows to choose a management procedure according to the type of vegetation of interest for any season. Further, with the association structure, we can understand how one place with some type of vegetation can change to another type in the other seasons. The confidence intervals obtained are certainly more correct than if we had used a model that did not consider the dependence between the repeated measurements.

Other possible approaches to analyze this experiment are in the context of random effects models (generalized linear mixed models) or transition models, but in these cases, the interpretations are different and cannot be compared to the results showed in this study. Briefly, in a generalized linear mixed model, the interpretations are made in a subject-specific level while in a model-like, as presented in this paper, are done for the population average. Other aspects that should be taken in account are the presence of missing values, if there are different numbers of observations per subject, equal/unequal time spacing. Therefore, as the results found here were satisfactory and answer the objectives of the study, these other approaches were not applied. Future studies can investigate them and be done to improve the residual analysis to check the goodness-of-fit obtained, whose tools for categorical data are still limited.

References

Agresti, A. 2002. Categorical Data Analysis. 2ed. John Wiley, Hoboken, NJ, USA.

Agresti, A. 2007. An Introduction to Categorical Data Analysis. 2ed. John Wiley, Hoboken, NJ, USA.

Diggle, P.J.; Heagerty, P.J.; Liang, K.Y.; Zeger, S.L. 2002. Analysis of Longitudinal Data. Oxford University Press, New York, NY, USA.

Dobson, A.J. 2008. An Introduction to Generalized Linear Models. 2ed. Chapman and Hall, New York, NY, USA.

Goodman, L.A. 1985. The analysis of cross-classified data having ordered and or unordered categories: association models, correlation models, and asymmetry models for contingency tables with or without missing entries. Annals of Statistics 13: 10-69.

Liang, K.Y.; Zeger, S.L. 1986. Longitudinal data analysis using generalized linear models. Biometrika 73: 13-22.

Lipsitz, S.R.; Kim, K.; Zhao, L.P. 1994. Analysis of repeated categorical data using generalized estimating equations. Statistics in Medicine 13: 1149-1163.

Nelder, J.A.; Wedderburn, R.W.M. 1972. Generalized linear models. Journal of the Royal Statistical Society Series A 135: 370-384.

Pereira, L.E.T.; Paiva, A.J.; Geremia, E.V.; Silva, S.C. 2015a. Grazing management and tussock distribution in elephant grass. Grass and Forage Science 70: 406-417.

Pereira, L.E.T.; Paiva, A.J.; Geremia, E.V.; Silva, S.C. 2015b. Regrowth patterns of elephant grass (*Pennisetum purpureum* Schum) subjected to strategies of intermittent stocking management. Grass and Forage Science 70: 195-204.

Pinheiro, J.C.; Bates, D.M. 2000. Mixed-Effects Models in S and S-PLUS. Springer, New York, NY, USA.

Ryel, R.J.; Beyschlag, W.; Caldwel, L.M.M. 1994. Light field heterogeneity among tussock grasses: theoretical considerations of light harvesting and seedling establishment in tussocks and uniform tiller distributions. Oecologia 98: 241-246.

Touloumis, A.; Agresti, A.; Kateri, M. 2013. GEE for multinomial responses using a local odds ratios parameterization. Biometrics 69: 633-640.

Verbeke, G.; Molenberghs, G. 2000. Linear Mixed Models for Longitudinal Data. Springer, New York, NY, USA.

Zeger, S.L.; Liang, K.Y. 1992. An overview of methods for the analysis of longitudinal data. Statistics in Medicine 11: 1825-1839.

Appendix

```
Library (multgee)

# Step 1) Load data:
data = read.table('Experiment.txt',header=TRUE,dec=',',sep='\t')
data$season = factor(data$season, levels=c('Summer 1','Autumn','Winter','Early spring','Late spring','Summer 2'))
data$type = factor(data$type, levels=c('Weed','Bare ground','Tussock'))
data$pre = as.factor(data$pre) # pre-grazing
data$post = as.factor(data$post) # post-grazing
data$block = as.factor(data$block) # each one of the 4 blocks
str(data)

# Other variables:
# cod_type: codification of each type of vegetation (k = 1, 2, 3)
# cod_season: codification of each season (t = 1, ..., 6)
# point_experiment: index of each one of the 640 points (i = 1, ..., 640)

# Step 2) GEE model --> package MULTGEE:
# 2.1) Association structure choice based on intrinsic parameters:
newdata = with(data,data.frame(y=cod_type, id=point_experiment, repeated=cod_season))
intrinsic.pars(y=y, data=newdata, id=id, repeated=repeated, rscale='nominal')
# there is a considerable variability between the intrinsic parameters --> choose the RC structure

# 2.2) Linear predictor choice based on several nested models:
mod_gee1 = nomLORgee(cod_type ~ block + pre*post*cod_season, data=data, id=point_experiment, repeated=cod_season, LORstr = 'RC')

mod_gee2 = update(mod_gee1, formula = ~. - pre:post:cod_season)
waldts(mod_gee1,mod_gee2)

mod_gee3 = update(mod_gee2, formula = ~. - pre:cod_season - post:cod_season)
waldts(mod_gee3,mod_gee2)

mod_gee4 = update(mod_gee2, formula = ~. - pre:cod_season)
waldts(mod_gee4,mod_gee2)

mod_gee5 = update(mod_gee2, formula = ~. - post:cod_season)
waldts(mod_gee5,mod_gee2)

mod_gee6 = update(mod_gee5, formula = ~. - pre:post)
waldts(mod_gee5,mod_gee6)

mod_gee7 = update(mod_gee6, formula = ~. - post)
waldts(mod_gee6,mod_gee7)

mod_gee_final = mod_gee6

summary(mod_gee_final) # 16 parameters in each logit

# Comparison between naive and robust standard errors:
se_gee = cbind(as.data.frame(sqrt(diag(mod_gee_final$naive.variance))),as.data.frame(sqrt(diag(mod_gee_final$robust.variance))))
colnames(se_gee) = c('naive','robust')
se_gee$difference = se_gee[,2] - se_gee[,1]

# 2.3) Predicted values:
predicted = as.data.frame(mod_gee_final$fitted.values)
names(predicted) = c('p_tussock','p_weed','p_bare')

# 2.4) Residuals:
residuals = as.data.frame(mod_gee_final$residuals)
str(residuals)

names(residuals)[1] = 'logit1'
names(residuals)[2] = 'logit2'

mean_res1 = with(residuals, mean(logit1))
mean_res2 = with(residuals, mean(logit2))

# 2.5) Confidence intervals:
vcov_gee = mod_gee_final$robust.variance # variance-covariance matrix
vcov_1.1 = vcov_gee[1:16,1:16] # matrix of logit 1
vcov_2.1 = vcov_gee[17:32,17:32] # matrix of logit 2

# Fitting a "second" model setting tussock as reference (by doing this we can get the standard errors of estimates related to category 'bare ground'):
mod_gee_final2 = nomLORgee(cod_type2 ~ block + pre + post + cod_season + pre:cod_season, data=data, id=point_experiment, repeated=cod_season, LORstr = 'RC')
summary(mod_gee_final2)

vcov_gee2 = mod_gee_final2$robust.variance # variance-covariance matrix
vcov_1.2 = vcov_gee2[1:16,1:16] # matrix of logit 1, but of the "second" model
vcov_2.2 = vcov_gee2[17:32,17:32] # matrix of logit 2, but of the "second" model

# This file contains several vectors of values 0 and 1 arranged as a matrix (dimmension 16 × 95) that will be useful to obtain the confidence intervals:
coef_aux = read.table('Auxiliary_cofficients.txt',header=TRUE,dec=',',sep='\t')

betas_1.1 = as.matrix(coef(mod_gee_final)[1:16]) # parameters estimates of logit 1
betas_2.1 = as.matrix(coef(mod_gee_final)[17:32]) # parameters estimates of logit 2
betas_1.2 = as.matrix(coef(mod_gee_final2)[1:16]) # parameters estimates of logit 1 of "second" model (bare ground vs tussock)
betas_2.2 = as.matrix(coef(mod_gee_final2)[17:32]) # parameters estimates of logit 2 of "second" model (weed vs tussock)

crit = qnorm(1-0.05/2)
ic = c()
```

```
for (i in 2:96){
x = coef_aux[,i];x

l1.1 = t(x) %*% betas_1.1 # logit 1 value
l2.1 = t(x) %*% betas_2.1 # logit 2 value
l1.2 = t(x) %*% betas_1.2 # logit 1 value, model 2
l2.2 = t(x) %*% betas_2.2 # logit 2 value, model 2

se_l1.1 = sqrt(t(x) %*% vcov_1.1 %*% x) # logit 1 standard error
se_l2.1 = sqrt(t(x) %*% vcov_2.1 %*% x) # logit 2 standard error
se_l1.2 = sqrt(t(x) %*% vcov_1.2 %*% x) # logit 1 standard error, model 2
se_l2.2 = sqrt(t(x) %*% vcov_2.2 %*% x) # logit 2 standard error, model 2

ll_l1.1 = l1.1 - crit*se_l1.1 # lower limit of logit 1
ul_l1.1 = l1.1 + crit*se_l1.1 # upper limit of logit 1
ll_l2.1 = l2.1 - crit*se_l2.1 # lower limit of logit 2
ul_l2.1 = l2.1 + crit*se_l2.1 # upper limit of logit 2

ll_l1.2 = l1.2 - crit*se_l1.2 # lower limit of logit 1, model 2
ul_l1.2 = l1.2 + crit*se_l1.2 # upper limit of logit 1, model 2
ll_l2.2 = l2.2 - crit*se_l2.2 # lower limit of logit 1, model 2
ul_l2.2 = l2.2 + crit*se_l2.2 # upper limit of logit 1, model 2

ll_p_tuss = exp(ll_l1.1)/(1 + exp(ll_l1.1) + exp(ll_l2.1)) # lower limit of tussock probability
ul_p_tuss = exp(ul_l1.1)/(1 + exp(ul_l1.1) + exp(ul_l2.1)) # upper limit of tussock probability

ll_p_weed = exp(ll_l2.1)/(1 + exp(ll_l1.1) + exp(ll_l2.1)) # lower limit of weed probability
ul_p_weed = exp(ul_l2.1)/(1 + exp(ul_l1.1) + exp(ul_l2.1)) # upper limit of weed probability

ll_p_bare = exp(ll_l1.2)/(1 + exp(ll_l1.2) + exp(ll_l2.2)) # lower limit of bare ground probability
ul_p_bare = exp(ul_l1.2)/(1 + exp(ul_l1.2) + exp(ul_l2.2)) # upper limit of bare ground probability

ic = rbind(ic,cbind(ll_p_tuss,ul_p_tuss,ll_p_weed,ul_p_weed,ll_p_bare,ul_p_bare))}
ic = as.data.frame(ic)
names(ic) = c('p_tussock_ll','p_tussock_ul','p_weed_ll','p_weed_ul','p_bare_ll','p_bare_ul')

# 2.6) Odds ratios (comparing pre and post-grazing levels):
# Post-grazing (45 × 35 cm): --> no interaction, regardless of the season
# Tussock × bare ground (logit 1):
(ratio_est = exp(betas_1.1[6]))
(se = sqrt(vcov_1.1[6,6]))
(ratio_ll = exp(betas_1.1[6] - crit*se)) # lower limit
(ratio_ul = exp(betas_1.1[6] + crit*se)) # upper limit

# Weed × bare ground (logit 2):
(ratio_est = exp(betas_2.1[6]))
(se = sqrt(vcov_2.1[6,6]))
(ratio_ll = exp(betas_2.1[6] - crit*se)) # lower limit
(ratio_ul = exp(betas_2.1[6] + crit*se)) # upper limit

# Tussock × weed (logit 2 of model 2, multiplied by (-1)):
(ratio_est = exp((-1)*betas_2.2[6]))
(se = sqrt(vcov_2.2[6,6]))
(ratio_ll = exp((-1)*betas_2.2[6] - crit*se)) # lower limit
(ratio_ul = exp((-1)*betas_2.2[6] + crit*se)) # upper limit

# Pre-grazing (maximum × 95 %) --> depends on the season because of the interaction. A little bit more complicated because of the covariance that needs to be considered:
# Tussock × bare ground (logit 1):
# Summer 1:
(ratio_est = exp(betas_1.1[5]))
(se = sqrt(vcov_1.1[5,5]))
(ratio_ll = exp(betas_1.1[5] - crit*se))
(ratio_ul = exp(betas_1.1[5] + crit*se))

# Autumn:
(ratio_est = exp(betas_1.1[5] + betas_1.1[12]))
(se = sqrt(vcov_1.1[5,5] + vcov_1.1[12,12] + 2*vcov_1.1[5,12]))
(ratio_ll = exp(betas_1.1[5] + betas_1.1[12] - crit*se))
(ratio_ul = exp(betas_1.1[5] + betas_1.1[12] + crit*se))

# Winter:
(ratio_est = exp(betas_1.1[5] + betas_1.1[13]))
(se = sqrt(vcov_1.1[5,5] + vcov_1.1[13,13] + 2*vcov_1.1[5,13]))
(ratio_ll = exp(betas_1.1[5] + betas_1.1[13] - crit*se))
(ratio_ul = exp(betas_1.1[5] + betas_1.1[13] + crit*se))

# Early spring:
(ratio_est = exp(betas_1.1[5] + betas_1.1[14]))
(se = sqrt(vcov_1.1[5,5] + vcov_1.1[14,14] + 2*vcov_1.1[5,14]))
(ratio_ll = exp(betas_1.1[5] + betas_1.1[14] - crit*se))
(ratio_ul = exp(betas_1.1[5] + betas_1.1[14] + crit*se))

# Late spring:
(ratio_est = exp(betas_1.1[5] + betas_1.1[15]))
(se = sqrt(vcov_1.1[5,5] + vcov_1.1[15,15] + 2*vcov_1.1[5,15]))
(ratio_ll = exp(betas_1.1[5] + betas_1.1[15] - crit*se))
(ratio_ul = exp(betas_1.1[5] + betas_1.1[15] + crit*se))

# Summer 2:
(ratio_est = exp(betas_1.1[5] + betas_1.1[16]))
(se = sqrt(vcov_1.1[5,5] + vcov_1.1[16,16] + 2*vcov_1.1[5,16]))
(ratio_ll = exp(betas_1.1[5] + betas_1.1[16] - crit*se))
(ratio_ul = exp(betas_1.1[5] + betas_1.1[16] + crit*se))

# Weed × bare ground (logit 2):
# Summer 1:
(ratio_est = exp(betas_2.1[5]))
(se = sqrt(vcov_2.1[5,5]))
(ratio_ll = exp(betas_2.1[5] - crit*se))
(ratio_ul = exp(betas_2.1[5] + crit*se))

# Autumn:
(ratio_est = exp(betas_2.1[5] + betas_2.1[12]))
```

```
(se = sqrt(vcov_2.1[5,5] + vcov_2.1[12,12] + 2*vcov_2.1[5,12]))
(ratio_ll = exp(betas_2.1[5] + betas_2.1[12] - crit*se))
(ratio_ul = exp(betas_2.1[5] + betas_2.1[12] + crit*se))

# Winter:
(ratio_est = exp(betas_2.1[5] + betas_2.1[13]))
(se = sqrt(vcov_2.1[5,5] + vcov_2.1[13,13] + 2*vcov_2.1[5,13]))
(ratio_ll = exp(betas_2.1[5] + betas_2.1[13] - crit*se))
(ratio_ul = exp(betas_2.1[5] + betas_2.1[13] + crit*se))

# Early spring:
(ratio_est = exp(betas_2.1[5] + betas_2.1[14]))
(se = sqrt(vcov_2.1[5,5] + vcov_2.1[14,14] + 2*vcov_2.1[5,14]))
(ratio_ll = exp(betas_2.1[5] + betas_2.1[14] - crit*se))
(ratio_ul = exp(betas_2.1[5] + betas_2.1[14] + crit*se))

# Late spring:
(ratio_est = exp(betas_2.1[5] + betas_2.1[15]))
(se = sqrt(vcov_2.1[5,5] + vcov_2.1[15,15] + 2*vcov_2.1[5,15]))
(ratio_ll = exp(betas_2.1[5] + betas_2.1[15] - crit*se))
(ratio_ul = exp(betas_2.1[5] + betas_2.1[15] + crit*se))

# Summer 2:
(ratio_est = exp(betas_2.1[5] + betas_2.1[16]))
(se = sqrt(vcov_2.1[5,5] + vcov_2.1[16,16] + 2*vcov_2.1[5,16]))
(ratio_ll = exp(betas_2.1[5] + betas_2.1[16] - crit*se))
(ratio_ul = exp(betas_2.1[5] + betas_2.1[16] + crit*se))

# Tussock × weed (logit 2 of model 2, with inverted coefficient signs):
# Summer 1:
(ratio_est = exp((-1)*betas_2.2[5]))
(se = sqrt(vcov_2.2[5,5]))
(ratio_ll = exp((-1)*betas_2.2[5] - crit*se))
(ratio_ul = exp((-1)*betas_2.2[5] + crit*se))

# Autumn:
(ratio_est = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[12]))
(se = sqrt(vcov_2.2[5,5] + vcov_2.2[12,12] + 2*vcov_2.2[5,12]))
(ratio_ll = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[12] - crit*se))
(ratio_ul = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[12] + crit*se))

# Winter:
(ratio_est = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[13]))
(se = sqrt(vcov_2.2[5,5] + vcov_2.2[13,13] + 2*vcov_2.2[5,13]))
(ratio_ll = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[13] - crit*se))
(ratio_ul = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[13] + crit*se))

# Early spring:
(ratio_est = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[14]))
(se = sqrt(vcov_2.2[5,5] + vcov_2.2[14,14] + 2*vcov_2.2[5,14]))
(ratio_ll = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[14] - crit*se))
(ratio_ul = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[14] + crit*se))

# Late spring:
(ratio_est = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[15]))
(se = sqrt(vcov_2.2[5,5] + vcov_2.2[15,15] + 2*vcov_2.2[5,15]))
(ratio_ll = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[15] - crit*se))
(ratio_ul = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[15] + crit*se))

# Summer 2:
(ratio_est = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[16]))
(se = sqrt(vcov_2.2[5,5] + vcov_2.2[16,16] + 2*vcov_2.2[5,16]))
(ratio_ll = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[16] - crit*se))
(ratio_ul = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[16] + crit*se))
```