Colloquium Biometricum 41
2011, 127–134

LOGISTIC MODEL IN ANALYSIS OF TWO– AND SIX–ROWED SPRING BARLEY DOUBLED HAPLOID LINES (*HORDEUM VULGARE* L.)

Ewa Bakinowska[1], Jan Bocianowski[1], Wojciech Rybiński[2]

[1]Department of Mathematical and Statistical Methods
Poznań University of Life Sciences
Wojska Polskiego 28, 60–637 Poznań, Poland
e–mail: ewabak@up.poznan.pl; jboc@up.poznan.pl
[2]Institute of Plant Genetics, Polish Academy of Science
Strzeszyńska 34, 60–479 Poznań, Poland
e–mail: wryb@igr.poznan.pl

Summary

The object of investigation constituted forty doubled haploid (DH) lines of spring barley (*Hordeum vulgare* L.). The DH lines were produced from seed of F_1 progeny obtained after crosses of two– and six–rowed cultivars. The control group (without use of mutagen) and the group with use of mutagen include twenty DH lines. In the both groups fourteen lines were two–rowed forms and six were six–rowed forms. Doubled haploid lines were analyzed with respect to the number of spikes. The considered trait is treated as discrete, so in the analysis the logistic model which belongs to the class of generalized linear models was applied.

Key Word and phrases: logistic model, spring barley, generalized linear model, doubled haploid lines

Subject classification AMS 2010: 62F10, 62J02, 62F12

1. Introduction

To analysis of experiments in which the observed trait is described by continuous random variable, the linear model usually is used. But sometimes, there are traits, which are naturally continuous in a character, but the results of their observations are statements about the membership of studied unit in definite category, as in the case of discrete random variables. Such traits include for example resistance to frost or tendency to big or small number of spikes of various variety of cereals. These tendencies are represented by continuous random variable which is hidden and only the results of classification of observed units (with respect to a symbolic ordinal scale) are analyzed in detail. The results of such classification are connected with some discrete random variable. In this case to use the standard methods based on linear model is not sufficient. Then to the statistical analysis of such experiments, the logistic model which to the class of generalized linear models belongs, is used (McCullagh and Nelder, 1989).

The relation of hidden continuous random variable with a discrete ordinal scale is determined by separation points of successive categories (borders of categories). In literature the borders are called thresholds (Misztal et al., 1989) or cutpoints (Miller et al., 1993).

For an experimenter who carries out an experiment, it may be interesting to estimate unknown cumulative probabilities and the possibility of the comparison of such probabilities for various treatments (compare Bakinowska and Kala, 2007).

In the literature we can find a lot of articles describing the application such generalized linear models in various fields of science: in medicine (Miller et al., 1993; Koch et al., 1989; Laframboise et al., 2007; Chen et al., 2009) or in economics (Cramer, 2003; Cramer and Ridder, 1988).

This paper is the illustration of application the logistic model in agriculture. Using the logistic model to analyze of such data allows to show how the same set of data can be described in other way (not by standard linear model). The aim of this paper is an application of logistic model to analysis of the number of spikes of various forms of spring barley.

2. Plant material

The doubled haploid (DH) lines were produced with use of *Hordeum bulbosum* method (Kasha and Kao, 1970) with additional mutagenic treatment of kernels obtained in results of crosses between two– and six–rowed polish spring barley cultivars: Maresi and Klimek. DH lines without mutagenic treatment constituted a control combination. For further analysis 40 DH lines were chosen: 20 control lines and 20 obtained after use of mutagen. In the both groups were this same number of two– and six–rowed lines. Phenotypic variability of lines were analysed on the base of field trial conducted with help of randomized blocks design in three replications. Kernels were sown in 1 m^2 plots with 15 cm between rows and 5 cm within the rows. After harvest of plants biometrical measurements were performed. Characteristics of analysed DH lines was earlier presented in detail by Rybiński et al. (2008). Among greater number of traits described by Rybiński et al. (2008), spike number per plant was chosen for achievement of goal of performed statistical analysis.

3. Description of method

Let us assume that in experiment there are *s* independent treatments, each represented by fixed number of units m_i. The studied units are classified to *k* separate categories. Let π_{ji} be the probability of the belonging of the studied unit to the *j*–th category corresponding to *i*–th treatment. The logistic model can be now written in the following form (see Miller et al. 1993, compare Bakinowska and Kala 2007):

$$\log\frac{\gamma_{ji}}{1-\gamma_{ji}}=\theta_j+\tau_i,\quad j=1,2,\ldots,k-1,\quad i=1,2,\ldots,s, \tag{3.1}$$

where θ_j is border (cutpoint) of *j*–th category, τ_i is the effect of *i*–th treatment (in result $\theta_j+\tau_i$ means the cutpoint of *j*–th category for *i*–th treatment), and γ_{ji} is the *j*–th cumulative probability corresponding to units of *i*–th treatment,

$$\gamma_{ji}=\pi_{1i}+\pi_{2i}+\ldots+\pi_{ji},\quad j=1,2,\ldots,k-1.$$

The results of classification of studied units are usually modeled with the use of multinomial distribution, which is determined by probabilities π_{ji}, $j=1,2,\ldots,k$, summing up to one, $\sum_{j=1}^{k}\pi_{ji}=1$, and the fixed number of units m_i. Our aim is to estimate the unknown cumulative probabilities in model (3.1) based on the experimantal data. The observed frequencies p_{ij}, which are natural estimators of unknown probabilities π_{ji}, $j=1,2,\ldots,k$, will be used in estimation by weighted least squares method.

4. Results and discussion

The earlier researches indicated on large difference between two– and six–rowed doubled haploid lines with respect to some traits: plant height, spike length, number of spikes per plant, number of grains per spike, grain weight per spike, grain number per plant, grain weight per plant, (Bocianowski and Rybiński, 2008; Rybiński et al., 2008). Therefore, the comparison between mutants and control genotypes for two– and six–rowed DH lines were conducted independently.

The studied trait was the number of spikes per plant. In the case of two–rowed form, the genotypes were divided (with respect to the number of spikes) into three separate ordered categories: less then 6, 6 or 7, more then 7 spikes. Similarly it was made with six–rowed forms. But the six–rowed forms are determined by smaller number of spikes, then the categories were established: less then 5, 5 or 6, more than 6 spikes.

To the analysis the model (3.1) will be used. The main aim it was to estimate the unknown cumulative probabilities γ_{ji}. From the model (3.1) we obtain:

$$\gamma_{ji} = \frac{\exp(\theta_j + \tau_i)}{1 + \exp(\theta_j + \tau_i)}, \quad j = 1,2 \quad i = 1,2. \tag{4.1}$$

We have *s*=2 treatments (mutants and control object) and *k*=3 separate categories, to which homogenous units are classified. The results of classifications were presented in Tables 1 and 2.

Table 1. Data concerning number of spikes for two–rowed barley doubled haploid lines

	Categories		
Treatments	<6	[6,7]	>7
Controls	15	15	12
Mutants	15	16	11

Table 2. Data concerning number of spikes for six–rowed barley doubled haploid lines

	Categories		
Treatments	<5	[5,6]	>6
Controls	9	4	5
Mutants	5	2	11

Let now τ_1 i τ_2 be the effects of control and mutant genotypes, respectively. The parameter θ_j in model (3.1) can be interpreted as the average, with respect to varieties, value of the *j*–th cutpoint, because effects τ_i sum to zero (compare McCullagh and Nelder 1989, p. 176):

$$\tau_1 + \tau_2 = 0, \tag{4.2}$$

what is equivalent to equality $\tau_1 = -\tau_2$. The equality (4.2) allows consider the model (3.1) with smaller number of parameters. So let τ be the effect of control genotypes, $-\tau$ effect of mutant genotypes, and θ_1, θ_2 cutpoints between categories. Then the model (3.1) in matrix form can be written as

$$\mathbf{C}^T \log(\mathbf{L}\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta}, \tag{4.3}$$

where

$$\boldsymbol{C^T} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}, \quad \mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\boldsymbol{\pi} = \begin{pmatrix} \boldsymbol{\pi}_1 \\ \boldsymbol{\pi}_2 \end{pmatrix} = \begin{pmatrix} \pi_{11} \\ \pi_{21} \\ \pi_{31} \\ \pi_{12} \\ \pi_{22} \\ \pi_{32} \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \tau \end{pmatrix}.$$

Let now $\mathbf{p}$ be the vector of observed frequencies, corresponding to the probability vector $\boldsymbol{\pi}$. Then the estimator of unknown parameters $\boldsymbol{\beta}$ in model (4.3) obtained by weighted least squares method has the form:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{S}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{S}^{-1} \mathbf{C}^T \log(\mathbf{Lp}), \tag{4.4}$$

where

$$\mathbf{S} = (\mathbf{C}^T (\mathbf{Lp})^{-\delta} \mathbf{L}) \mathbf{V} (\mathbf{C}^T (\mathbf{Lp})^{-\delta} \mathbf{L})^T ,$$

and **V** is block diagonal matrix, where each block on main diagonal is the estimate of covariance matrix in multinomial distribution, $\mathbf{V}_i = \frac{1}{m_i}(\mathbf{p}_i^{\delta} - \mathbf{p}_i \mathbf{p}_i^{\ T})$, i=1, 2.

Using formula (4.4) for our data we obtain following estimates for two–rowed DH lines:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \hat{\tau} \end{pmatrix} = \begin{pmatrix} -0.586 \\ 0.975 \\ -0.030 \end{pmatrix}$$

and for six–rowed doubled haploid lines:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \hat{\tau} \end{pmatrix} = \begin{pmatrix} -0.444 \\ 0.176 \\ 0.599 \end{pmatrix}.$$

In consequence basing on (4.1) the searched probabilities can be calculated from

$$\hat{\gamma}_{ji} = \frac{\exp\left(\hat{\theta}_j + \hat{\tau}_i\right)}{1 + \exp\left(\hat{\theta}_j + \hat{\tau}_i\right)}, \quad j = 1, 2, \quad i = 1, 2.$$

These probabilities were presented in Table 3 (for two–rowed DH lines) and in Table 4 (for six–rowed doubled haploid lines). However, in figure 1 relation between estimated cumulative probabilities $\hat{\gamma}_{ji}$ obtained using the model (3.1) and cumulative probabilities $\tilde{\gamma}_{ji} = p_{1i} + ... + p_{ji}$ as the natural estimators of γ_{ji} (where p_{ji} are the observed frequencies) for number of spikes of plant was presented.

Table 3. Estimates of cumulative probabilities for two–rowed barley doubled haploid lines

	Two–rowed barley doubled haploid lines	
	$\hat{\gamma}_{1i}$	$\hat{\gamma}_{2i}$
Controls	0.351	0.720
Mutants	0.364	0.732

Table 4. Estimates of cumulative probabilities for six–rowed barley doubled haploid lines

	Six–rowed barley doubled haploid lines	
	$\hat{\gamma}_{1i}$	$\hat{\gamma}_{2i}$
Controls	0.538	0.685
Mutants	0.206	0.396

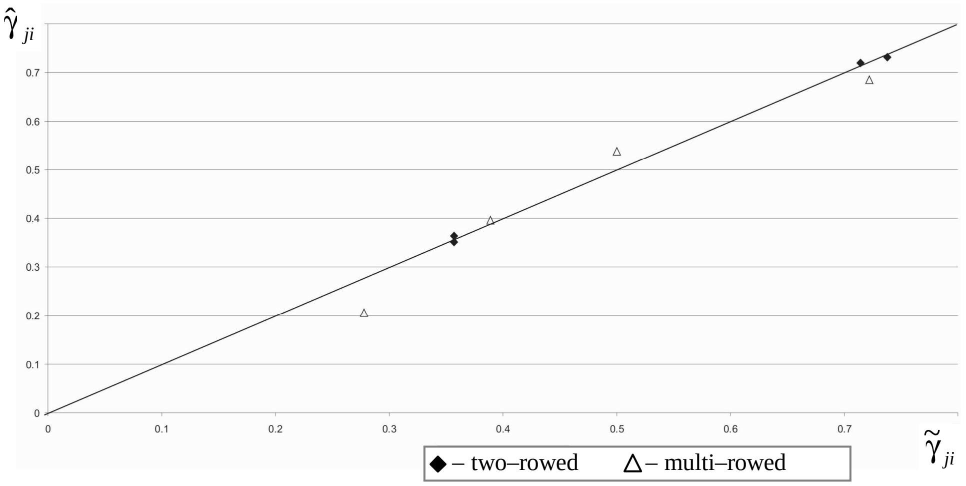

Fig. 1. Relation between estimated cumulative probabilities $\hat{\gamma}_{ji}$ obtained using the model (3.1), and cumulative probabilities $\tilde{\gamma}_{ji}$ for number of spikes per plant.

In the case of two–rowed control lines the probability, that one will observed less than 6 spikes per plant is 0.351, and less than 6 or less than 7 is 0.72. Let notice, that differences between control lines and mutants of doubled haploid lines, based on probability, for two–rowed are minimal, but for six–rowed significant.

5. Conclusions

Method described above seems to be good tool to analysis of such set of discrete data (Bocianowski et al., 2008). The main problem it was the comparison of various varieties with respect to number of spike. The problem was solved using the presented model to calculation the cumulative probabilities. This paper is only the illustration one way of application logistic models to analysis of agricultural experiments.

References

Bakinowska E., Kala R. (2007). An application of logistic models for comparison of varieties of seed pea with respect to lodging. *Biometrical Letters* 44(2), 143–154.

Bocianowski J., Bakinowska E., Rybiński W. (2008). Analysis of selected grasspea mutants by generalized linear model. *Colloquium Biometricum* 38, 161–171.

Bocianowski J., Rybiński W. (2008). Use of canonical variate analysis for the multivariate assessment of two– and six–rowed barley DH lines (*Hordeum vulgare* L.). *Annales Universitatis Mariae Curie–Skłodowska. Sectio E: Agricultura* LXIII (3), 53–61.

Chen C–C., Chuang C–L., Wu K–Y., Chan C–C. (2009). Sampling Strategies for Occupational Exposure Assessment under Generalized Linear Model. *Ann. Occup. Hyg.* 53(5), 509–521.

Cramer J. S. (2003). *Logit Models from Economics and Other Fields*. Cambridge University Press.

Cramer J. S., Ridder G. (1988). The Logit Model in Economics. *Statistica Neerlandica* 42(4), 297–314.

Kasha K. J., Kao K. N. (1970). High frequency haploid production in barley (*Hordeum vulgare* L.). *Nature* 225, 874–876.

Koch G. G., Carr G. J., Amara I. A., Stokes M. E., and Uryniak T. J. (1989). Categorical data analysis. In *Statistical Methodology in the Pharmaceutical Sciences*. D. A. Berry (ed), 391–475, New York, Marcel Dekker.

Laframboise T., Harrington D., Wier B. A. (2007). PLASQ: a generalized linear model–based procedure to determine allelic dosage in cancer cells from SNP array data. *Biostatistics* 8 (2), 323–336.

McCullagh P., Nelder J. A. (1989). *Generalized linear models*. 2nd ed. Chapman and Hall, London.

Miller M. E., Davis C. S., Landis J. R. (1993). The analysis of longitudinal polytomous data: generalized estimating equations and connections with weighted least squares. *Biometrics* 49, 1033–1044.

Misztal I., Gianola D., Foulley J. L. (1989). Computing aspects of a nonlinear method of sire evaluation for categorical data. *Journal of Dairy Science* 72, 1557–1568.

Rybiński W., Pankiewicz K., Bocianowski J., Rębarz M. (2008). Analysis of some traits on phenotypic and molecular level for two– and six–rowed doubled haploids of spring barley (*Hordeum vulgare* L.). *Biuletyn IHAR* 249, 141–155 (in Polish).