Die Bodenkultur: Journal of Land Management, Food and Environment
Volume 68, Issue 2, 81–87, 2017. DOI: 10.1515/boku-2017-0008
ISSN: 0006-5471 online, © De Gruyter, www.degruyter.com/view/j/boku

Research Article

Ordinal regression model for pea seed mass

Ein ordinales Regressionsmodell für das Samengewicht von Erbse

Agnieszka Klimek-Kopyra[1]*, Jacek Strojny[2], Tadeusz Zając[1], Anna Ślizowska[1], Jana Klimešova[3], Reinhard W. Neugschwandtner[4]

[1] Institute of Plant Production, University of Agriculture in Kraków, Aleja Mickiewicza 21, 31-120 Kraków, Poland
[2] Department of Mathematical Statistics, University of Agriculture in Kraków, Aleja Mickiewicza 21, 31-120 Kraków, Poland
[3] Department of Crop Science, Breeding and Plant Medicine, Faculty of Agronomy, Mendel University in Brno, Zemedelska 1, 613 00 Brno, Czech Republic
[4] Division of Agronomy, Department of Crop Sciences, University of Natural Resources and Life Sciences Vienna (BOKU), Konrad-Lorenz-Straße 24, 3430 Tulln, Austria
* Corresponding author: klimek.a@wp.pl

Received: 14 March 2017, received in revied form: 13 June 2017, accepted 4 July 2017

Summary

The development of seeds at various positions in the pod is asynchronous. Thus, the differences of seed dry mass production because of environmental conditions may depend on the cultivar type, type of inoculants and interrelations between seeds per pod, pods per plant or seeds per plant. Presently, a mathematical description of pea seed categorisation is missing. The aim of the study was the assessment of two groups of variables (quantitative and qualitative) for pea seed weight categorisation by ordinal regression model. Year, cultivar and inoculant constituted the first group (qualitative variables), whilst seeds per pod, the pods per plant and seeds per plant (quantitative variables) were entered as covariates in the ordinal regression model. According to the ordinal regression model variables, seeds per pod, pods per plant, seeds per plant, year and cultivar are meaningful predictors of the seed mass categories. However, the variable inoculant is marginally significant.

Keywords: regression model, seed categories, inoculant, legume, *Pisum sativum*

Zusammenfassung

Die Entwicklung der Samen an den verschiedenen Positionen in der Hülse verläuft asynchron. Die Unterschiede in der Produktion des Samengewichtes von Erbse aufgrund von Umweltbedingungen könnten von der Sorte, dem Inokulum oder den Wechselbeziehungen zwischen der Kornanzahl pro Hülse, der Hülsenanzahl oder der Kornanzahl abhängig sein. Derzeit gibt es kein mathematisches Modell für die Klassifizierung von Erbsensamen. Das Ziel dieser Arbeit war es, den Einfluss von zwei Gruppen von Variablen (quantitative und qualitative) auf die Bildung des Samengewichtes von Erbse mittels eines ordinalen Regressionsmodelles zu bewerten. Jahre, Sorten und Inokulen bildeten die erste Gruppe (qualitative Variablen), während die Kornanzahl pro Hülse, die Anzahl der Hülsen und die Anzahl der Samen (quantitative Variablen) als Kovariaten in das ordinale Modell einflossen. Gemäß dem ordinalen Regressionsmodell sind Kornanzahl pro Hülse, Anzahl der Hülsen, Anzahl der Samen, Jahr und Sorte aussagekräftige Prädiktoren für die Kategorien des Samengewichts, während die Variable Inokulum geringen Einfluss hat.

Schlagworte: logistisches Regressionsmodell, Samengewicht Inokulum, Legumiose, *Pisum sativum*

1. Introduction

The total global area of pea cultivation amounts to 6.8 million ha of dry peas and 2.3 million ha of green peas (FAO, 2013). Pea yield is determined by genotype variety and agro-climatic conditions in Europe (Dore et al., 1998). In Central and Eastern Europe, pea yield is lower than that in Western Europe. The phenomenon of different seed mass is accepted by farmers because of the natural diversity of the environment. However, the high variability of seed mass becomes a serious problem for certified seed production. In seed production, the unified seed mass guarantees high yield value. Ney et al. (1993) claim that seed setting within pods is asynchronic, which results in high variation of seed mass. Amongst the environmental factors modifying seed mass, weather and soil conditions are very important. Huang and Erickson (2007) showed that inoculation of pea seeds triggers several changes in morphological features and improves seed yield, whereas Zając et al. (2013) did not observe any significant reaction of the plant to inoculation. Zając et al. (2013) proved that the single seed mass significantly depends on the seed position in a pod with the seeds located in the central part of the pod having a larger mass and that the number of seeds in the pod is influenced by the node position on the stem. According to Illipronti et al. (2000), the higher seed mass can be directly connected to the amount of assimilates transported down the stem at the seed maturity stage, rather than to the position of seeds in pods set in the upper part of the stem. In most seed crops, individual seed weight is commonly analysed as the product of the individual seed growth rate and the duration of seed formation (Munier-Jolain et al., 1998). Dacko et al. (2016) reported that the single pea-seed mass is affected mostly by the following predictors: 'K index', which describes the hydrothermal conditions during pea sowing-emergence and emergence-maturity periods; 'length of pod'; and 'cultivar', an optimal K index and longer pods are associated with a higher seed mass. Gorden et al. (2016) confirmed the above findings and proved, basing on the seed size model, that the seed size is strongly related to climate diversity. The authors noticed that further studies about how abiotic factors (such as soil conditions and climate change) can change the difference in size of seeds should be conducted.

Presently, a mathematical description of pea seed categorisation is missing. The aim of the present research was to assess by means of ordinal regression analysis if the qualitative variables (year, cultivar and inoculant) or quantitative variables (seeds per pod, pods per plant and seeds per plant) have an impact on the seed mass diversification.

2. Material and Methods

2.1 Experimental setup

The research was carried out in the Experimental Research Station in Modzurów, Silesian province, Poland (50°09'N 18°07'E, 274 m. a.s.l.). The experimental field soil was Haplic Phaeozem. A randomised complete block design with four replications was used in the field experiment. The following factors were considered: year – 2010 and 2011 (seeds were sown on 9th of April in 2010 and 4th of April in 2011); cultivar – edible 'Tarchalska' and used as a fodder 'Klif'; and inoculant – of powder and gel form (Nitragina – powder form inoculant, produced by the BIOFOOD company (Poland) and non-commercial, gel inoculant (IUNG) produced by the Polish Institute of Soil Science and Plant Cultivation). Each plot had an area of 8.4 m^2. The following pre-sowing fertiliser doses were applied: phosphorus, 48 kg ha^{-1} (P_2O_5); potassium, 72 kg ha^{-1} (K_2O). Ammonium nitrate was applied as a starting dose with 20 kg ha^{-1} N. One hundred and twenty seeds of the cultivar 'Tarchalska' and 100 seeds of cultivar 'Klif' were sown per square meter. Row distance was 15 cm. Before harvest, 15 mature plants per plot (7.67 pods on average from each plant) were randomly selected and collected in order to conduct the biometrical analysis. The crops were harvested on 6th of August 2010 and on 3rd of August 2011. In the biometrical analysis, we focused on seeds per pod, pods per plant and seeds per plant. All collected seeds were divided into five classes of weight (<200, 200–240, 240–270, 270–300 and >300 mg) in order to assess how qualitative factors (year, cultivar and inoculant) and quantitative variables (seed mass in pod, the seeds per plant, seeds per pod) influence seed mass.

2.2 Statistical analysis

The pea experiment data was evaluated with an ordinal regression model. The reason of using this model is the prediction of the ordinal outcome. The outcome variable was pea seed's weight with five ordinal levels: <200, 200–240, 240–270, 270–300 and >300 mg. Ordinary logistic regression models apply with binary dependent variables, but generalisations of logistic regression apply with mul-

ticategorical responses. The ordinal regression enables to build models aiming to evaluate the importance of predictor variables, where the dependent variable is ordinal in nature. However, continuous variables could be considered as ordinal because of threshold effects.

Generally, linear regression models do not work very well in prediction of ordinal response variables. Because linear regression models are drawing up to model the outcome variable measured on an interval scale the simplification of linear regression assumption may result in not accurately reflecting the relationship in data. A way to model ordinal variables is to use a generalisation of linear regression – generalised linear model to access cumulative probabilities for outcome variable categories. With this approach, the finding is a predicted probability of being in the corresponding category. The basic form of a generalised linear model is

$$link(\gamma_j) = \theta_j - [\beta_1 x_1 + \beta_2 x_2 + \mathrm{L} + \beta_k x_k] \quad (1)$$

where

γ_j – the cumulative probability for the j^{th} category of the dependent variable

$x_1 \ldots x_k$ – the predictor variables

k – the number of predictor variables

$\beta_1 \ldots \beta_k$ – the regression coefficients

Instead of predicting the cumulative probabilities, the ordinal regression model provides a function of those values. This function is called link function. Generally, an ordinal regression model consists of three major components:

- Location component – it is the part of equation (1) that includes predictor variables and related coefficients $[\beta_1 x_1 + \ldots + \beta_k x_k]$. It is used to access probabilities of membership of each observation in the categories considered.
- Scale component – it is a modification of the model (1) to account for differences in variability of the independent variables. The model comprising a scale component takes the following form:

$$link(\gamma_j) = \frac{\theta_j - [\beta_1 x_1 + \beta_2 x_2 + \mathrm{K} + \beta_k x_k]}{\exp(\tau_1 z_1 + \tau_2 z_2 + \mathrm{K} + \tau_k z_k)} \quad (2)$$

where

$z_1 \ldots z_k$ – the independent variables chosen from the set of predictor variables

$\tau_1 \ldots \tau_k$ – the coefficients for the scale component.

- Link function – it is the transformation of cumulative probabilities used in the estimated ordinal regression model. Link function should take the form that best fits the research question and the structure of data. Consideration of a link function is helpful to examine the distribution of values of the dependent variable. Because lower categories of the outcome variable were more probable, the negative *log–log* form of the link function was used:

$$-log(-log(\gamma)) \quad (3)$$

The process of selecting predictors for the location component takes both theoretical and empirical sides into consideration. Because the discrimination of potential predictor variables that are expected to influence seed weight was the aim of the study, the analysis started off by including all the variables involved and considered as might be influential. Stepwise procedure was used: variables not helpful in the model were removed from the analysis and the model was re-estimated. Qualitative variables (year, cultivar and inoculant) were entered as the factors in the model, and, on the other hand, quantitative variables (seeds per pod, pods per plant and seeds per plant) were entered as covariates in the ordinal regression model. The Wald test was used to determine whether a certain predictor variable is significant or not. It rejects the null hypothesis of the corresponding coefficient being zero.

A general appraisement of the goodness of fit of the final model enables pseudo-*R*-square coefficients. These indicators are based on the likelihood ratio. There are three pseudo-*R*-square measures of determination used to evaluate the ordinal regression model. Cox and Snell's indicator is a generalisation of the usual *R*-square designed to evaluate estimates obtained by the maximum likelihood method (Cox and Snell, 1989).

The total set of observations used to construct the regression model for the pea seed mass amounted to n = 5,527 (year – 2010: 2,505, 2011: 3,022; cultivar – Klif: 3,112, Tarchalska: 2,415; inoculant – Control: 1,765, IUNG: 1,981, Nitragina: 1,781).

3. Results and Discussion

3.1 Weather conditions

Weather conditions differed in two years of vegetation (Table 1). In 2010, intensive precipitation was observed in

Table 1. Weather conditions in the two experimental years
Tabelle 1. Wetterbedingungen in den zwei Versuchsjahren

	Years	April	May	June	July	August
Temperature (°C)	2010	7.5	11.7	16.7	20.4	18.5
	2011	9.7	13.2	17.4	17.3	18.9
	Long-term mean	8.1	13.7	16.5	18.2	17.9
Rainfall (mm)	2010	66.5	193.2	103.5	208.5	95.1
	2011	29.2	71.5	99.5	167.5	73.2
	Long-term mean	50.2	65.3	80.0	74.9	78.5

the seedling stage (May) and during the flowering phase, which contributed to the growth and development retardation of vegetative and generative stages. This reduced plant density, number of seeds and pods and consequently resulted in the low yield. The July 2010 was characterised by high temperature conditions, which contributed to the decrease in the yield parameters. In 2011, the precipitation and average temperature were optimal for the plant growth.

3.2 Means of seed mass and quantitative variables

All three factors (year, cultivar and inoculant) had significant effects on seed yield parameters. Seed mass, seed per pod, pod per plant and seeds per plant were higher in 2011 than those in 2010. Fodder cultivar 'Klif' had more seeds per pod, pods per plant and seeds per plant, whilst 'Tarchalska' had a higher seed mass. The use of inoculants had no effect on the seed mass but significantly influenced the seeds per pod, pods per plant and seeds per plant with highest numbers IUNG application.

3.3 Model results

Tables 3 and 4 show the descriptive statistics of quantitative and qualitative variables used in the model. The process of the ordinal model construction involved specification of different arrangements of predictor variables based

Table 2. Means and analysis of variance of seed mass and quantitative variables (seeds per pod, pods per plant and seeds per plant) as affected by the qualitative variables (year, cultivar and inoculant)
Tabelle 2. Mittelwerte und Varianzanalyse des Samengewichts und der quantitativen Variablen (Samen pro Hülse, Hülsen pro Pflanze und Samen pro Pflanze) in Abhängigkeit der qualitativen Variablen (Jahr, Sorte und Inokulum)

	Seed mass (mg)	Seeds (per pod)	Pods (per plant)	Seeds (per plant)
Year				
2010	210^{b}	2.45^{b}	6.82^{b}	28.3^{b}
2011	270^{a}	2.88^{a}	7.18^{a}	33.5^{a}
Cultivar				
Tarchalska	270^{a}	2.55^{b}	6.50^{b}	27.4^{b}
Klif	220^{b}	2.76^{a}	7.42^{a}	33.9^{a}
Inoculant				
IUNG	250^{a}	2.87^{a}	7.30^{a}	33.2^{a}
Nitragina	240^{a}	2.69^{b}	6.86^{b}	30.6^{b}
Control	260^{a}	2.45^{c}	6.85^{b}	29.1^{c}

Significant differences are at $p < 0.001$. Different letters indicate significant differences between means.
No interactions of factors – year, cultivar and inoculant – are shown.

Table 3. Model fitting information
Tabelle 3. Informationen zur Modellanpassung

Model	–2 Log likelihood	Chi-square	df	Significance
Intercept only	10,564.1			
Final model	7,338.1	3,226.0	9	0.000

Link function: negative *log–log*

on their theoretical justification. The ultimate model was built on the premise of the significance of predictor variables in dependent variable categories description. Factors retained in the model were considered as important in explanation of seed mass. Additional variables that were initially considered got rejected. These factors, as not relevant for the seed mass modelling, are excluded from the selected model.

Model fitting information (Table 3) shows that using predictor variables significantly improves the model's ability of adequate seed weight predictions. McCullagh and Nelder (1989) indicate that the difference between –2 times the log-likelihood for the intercept-only model and for the final model can be interpreted as chi-square statistics. The significant chi-square statistic indicates that the final ordinal regression model gives better predictions than those based on marginal probabilities for the outcome variable categories.

Table 4, which contains the parameters' estimates, shows the final ordinal regression model. The parameters of the ordinal regression model were estimated by the iterative algorithm. Options set for the algorithm were link function (log–log), maximum number of iterations (100), maximum step-halving (5), parameter convergence (e^{-6}), delta (0 – no value was added to zero cell frequencies) and singularity tolerance (e^{-8}). The criterion of log-likelihood was not applied (the algorithm stops if the change in the log-likelihood is less than this criterion).

The overall assessment of the goodness of fit of the ordinal regression model was done according to the Nagelkerke (1991) pseudo-*R*-square for the reason that the coefficient attains values in the zero-to-one range. The estimate amounted value of 0.461, indicating the meaningful proportion of variance in the dependent variable associated with the predictor variables. From the theoretical standpoint, the threshold parameters are not as important as location estimates. The location parameters relate the predictor variables to the cumulative probabilities of dependent variable.

According to the statistical tests' results, the variables seeds per pod, pods per plant, seeds per plant year and cultivar are meaningful predictors for the outcome variable (pea seed mass categories). The variable inoculant is marginally significant. However, direct interpretation of the coefficients in the model in Table 4 is difficult because of the character of the link function. On the other hand, the sign of a particular parameter coefficient provides meaningful insight into the effect of a predictor variable in the model. The signs of the estimates in principle indicate the direction of a particular predictor effect. Positive signs of coefficients of seeds per pod and pods per plant suggest a positive relationship between these predictors and dependent variable. As such variable increases, so does the probability of being in one of the higher categories of the seed weight category. Conversely, negative estimates indicate inverse relationship. An increase in the number of seeds per plant variable predisposes a pea to be classified into the lower outcome variable categories.

Our results correspond with the observations of other authors. Halsted and Byron (1918) analysed the seed position in a pod of soybean. The author noticed that the number of seeds and the weight of seeds in a pod are variable. The seed mass mainly depends on the number of seeds and their position in a pod. Additionally, the seed mass is inversely proportional to seeds per pod. The author proved that the largest seed mass is gained in single seed pods. In double seed pods, an earlier developed seed is always the largest. However, in the triple seed pods, the largest seed mass is gained in the central part of the pod. Furthermore, node position on the main stem (Atta et al., 2004), cultivar and weather conditions (Zając et al., 2013; Dacko et al., 2016) influence the seed dry weight. Zając et al. (2013) and Dacko et al. (2016) supported the above study, claiming that cultivars and weather conditions have the largest impact on seed weight. Zając et al. (2013) proved that biometrical parameters of pods revealed considerable differences between the pea cultivars. The fodder pea cultivar 'Klif' was more productive because of the higher number

Table 4. Statistical evaluation of considered factors (ordinal regression model) for seed mass categories
Tabelle 4. Statistische Evaluierung der betrachteten Faktoren (ordinales Regressionsmodelles) hinsichtlich der Kategorien des Samengewichts

Model component	Variable level	Estimate	Standard error	Wald	df	Significance	95% Confidence interval	
							Lower bound	Upper bound
Threshold	[1] < 200 mg	–0.824	0.054	232.6	1	<0.001	–0.930	–0.718
	[2] 200–240 mg	–0.303	0.051	35.2	1	<0.001	–0.403	–0.203
	[3] 240–270 mg	0.232	0.052	19.6	1	<0.001	0.129	0.335
	[4] 270–300 mg	1.045	0.059	311.1	1	<0.001	0.929	1.161
Location	[Seeds per pod]	0.056	0.008	50.9	1	<0.001	0.041	0.072
	[Pods per plant]	0.048	0.010	22.2	1	<0.001	0.028	0.068
	[Seeds per plant]	–0.006	0.002	9.7	1	<0.002	–0.010	–0.002
	[Year = 2010]	–0.849	0.030	796.2	1	<0.001	–0.908	–0.790
	[Year = 2011]	0.000	.	.	0	.	.	.
	[Cultivar = Klif]	–0.727	0.028	691. 7	1	<0.001	–0.781	–0.673
	[Cultivar = Tarchalska]	0.000	.	.	0	.	.	.
	[IUNG = 1]	0.081	0.022	13.7	1	<0.001	0.038	0.125
	[Nitragina = 2]	0.038	0.023	2.9	1	0.088	–0.006	0.083
	[Control = 3]	0.000	.	.	0	.	.	.
Scale	[Cultivar = Klif]	–0.621	0.029	448.2	1	<0.001	–0.678	–0.563
	[Cultivar = Tarchalska]	0.000	.	.	0	.	.	.
	[Year = 2010]	–0.492	0.031	246.6	1	<0.001	–0.553	–0.430
	[Year = 2011]	0.000	.	.	0	.	.	.

Link function: negative *log–log*

of seeds and the higher mass of seeds in pods. The edible pea cultivar 'Tarchalska' produced a large amount of heavier seeds in the first two reproductive nodes. Application of IUNG inoculant resulted in a rapid increase in the number of pods and seed weight on the bottom part of reproductive shoots but the weight of seeds and pods decreased at consecutive nodes. Seed mass significantly depended on weather conditions. Plants were more productive in optimal weather conditions. Under high temperature conditions, as prevailing in July 2010, both seed position on the plants and time of pod set appeared to contribute considerably to the variance in seed mass. Seeds produced under high temperature conditions were lighter and smaller than those grown under normal temperature conditions.

4. Conclusion

The ordinal regression analysis showed that variables such as seeds per pod, pods per plant, seeds per plant, year and cultivar have an impact on the seed mass diversification and are meaningful predictors of the pea seed mass categories. However, an increase in seeds per plant predisposes a pea seed to be classified to the lower seed mass category. The inoculants were marginally significant for seed category.

The information gained from the study may help in shaping the seed production. Therefore, it can be useful to the seed industry, as each seed producer has the economic interest in gaining and producing large seeds of similar size.

References

Atta, S., Maltese, S. and R. Cousin (2004): Protein content and dry weight of seeds from various pea genotypes. Agronomy Journal 24, 257–266.

Cox, D.R. and E.J. Snell (1989): The Analysis of Binary Data. Chapman and Hall, London.

Dacko, M., Zając, T., Synowiec, A., Oleksy, A., Klimek-Kopyra, A. and B. Kulig (2016): New approach to determine biological and environmental factors influencing mass of single pea (*Pisum sativum* L.) seed in Silesia region in Poland using a CART model. European Journal of Agronomy 74, 29–37.

Dore, T., Meynard, J.M. and M. Sebillotte (1998): The role of grain number, nitrogen nutrition and stem number in limiting pea crop (*Pisum sativum*) yields under agricultural conditions. European Journal of Agronomy 8, 29–37.

FAO (2013): World Food and Agriculture. Statistical Yearbook. Food and Agriculture Organization of the United States, Rome.

Gorden, N.S., Winkler, K., Jahnke, M.R., Marshall, E., Horky, J., Hudelson, C. and J. Etterson (2016): Geographic patterns of seed mass are associated with climate factors, but relationships vary between species. American Journal of Botany 103, 60–72.

Halsted, S. and D. Byron (1918): Report of the department of botany. Annual Report 38, New Jersey State Agricultural Experimental Station, 369–424.

Huang, H. and R. Erickson (2007): Effect of seed treatment with *Rhizobium leguminosarum* on *phytium* damping-off, seedling height, root nodulation, root biomass, shoot biomass, and seed yield of pea and lentil. Journal of Phytopathology 155, 31–37.

Illipronti, J., Lommen, W., Langerak, J. and P. Struik (2000): Time of pod set and seed position on the plant contribute to variation in quality of seeds within soybean seed lots. Netherlands Journal of Agricultural Science 48, 165–180.

McCullagh, P. and J. Nelder (1989): Generalized linear models. Chapman and Hall, London.

Munier-Jolain, N.G., Munier-Jolain, N.M., Roche, R., Ney B. and C. Duthion (1998): Seed growth rate in grain legumes. I. Effect of photoassimilate availability on seed growth rate. Journal of Experimental Botany 49, 163–169.

Nagelkerke, N. (1991): A note on the general definition of the coefficient of determination. Biometrika 78, 691–692.

Ney, B., Duthion, C. and E. Fontaine (1993): Timing of reproductive abortion in relation to cell division, water content, and growth of pea seeds. Crop Science 33, 267–270.

Zając, T., Klimek-Kopyra, A., Oleksy, A. and A. Lenart (2013): Vertical distribution of pea (*Pisum sativum* L.) seed yield depending on the applied bacterial inoculants. Journal of Agricultural Science 5, 260–268.