Limited dependent variable models and probabilistic prediction in informetrics

LIMITED DEPENDENT VARIABLE MODELS AND PROBABILISTIC PREDICTION IN INFORMETRICS

Nick Deschacht[a] & Tim C.E. Engels[b]

[a]Faculty of Economics and Business, KU Leuven, Campus Brussel, Warmoesberg 26, 1000 Brussel, Belgium. Nick.Deschacht@kuleuven.be

[b]Centre for Research & Development Monitoring (ECOOM), Faculty of Political and Social Sciences, University of Antwerp, Middelheimlaan 1, 2020 Antwerp, Belgium; Antwerp Maritime Academy, Noordkasteel-Oost 6, 2030 Antwerp, Belgium. Tim.Engels@uantwerpen.be

Published in: Ying Ding, Ronald Rousseau, Wolfram Dietmar (Editors). Measuring scholarly Impact: Methods and Practice. Springer, 2014, ISBN-13: 978-3319103761, p 193-214, DOI 10.1007/978-3-319-10377-8_9.

Abstract

This chapter explores the potential for informetric applications of limited dependent variable models, i.e. binary, ordinal and count data regression models. In bibliometrics and scientometrics such models can be used in the analysis of all kinds of categorical and count data, such as assessments scores, career transitions, citation counts, editorial decisions or funding decisions. The chapter reviews the use of these models in the informetrics literature and introduces the models, their underlying assumptions and their potential for predictive purposes. The main advantage of limited dependent variable models is that they allow us to identify the main explanatory variables in a multivariate framework and to estimate the size of their (marginal) effects. The models are illustrated using an example data set to analyze the determinants of citations. The chapters also shows how these models can be estimated using the statistical software Stata.

Keywords: regression; categorical; binary; logit; ordered; Poisson; negative binomial;

1. Introduction

A topic search in the Social Science Citation Index on November 13th 2013 identified over 700 journal articles in Library and Information Science (LIS) that use regression analysis. In the top 25 of source titles, we find Scientometrics (64 articles), Journal of the American Society for Information Science and Technology (46), Information Processing & Management (24), Journal of Informetrics (12) and Journal of Documentation (9). Until 2004 the annual number of LIS papers that implemented a regression model did not exceed 20; then, in the period 2005 to 2010 a gradual increase to about 50 papers per year is apparent. Since 2011 the annual number jumped to about 100, illustrating the rise of regression models in LIS. In the aforementioned journals, binary, ordinal, Poisson and negative binomial regression models are common because classification issues (e.g. authorship attribution, classification of

journals, user profiles) and count data (e.g. number of papers, patents or their citations) abound in LIS. In this chapter we specify these limited dependent variable models with a view of facilitating the implementation of such models in LIS research.

Limited dependent variable models are a group of regression models in which the range of possible values of the variable of interest is limited. In some cases the outcome variable is binary, such as when the interest is in whether a journal article was cited over a certain period (yes or no). The outcome variable can also take multiple discrete values as is often the case in peer review and assessments. When frequencies are counted for a certain event the outcome variable consists of count data, such as the number of patents in a given year or the number of books published by publishing houses. In these cases the choice of the regression model may follow directly from the research question. Often, however, the choice of the regression model will be subject to careful deliberation and more than one model may be appropriate. Running multiple models on the same dataset may be instructive and can sometimes serve as a robustness check of the results. We illustrate this throughout this chapter. In the conclusions we provide the reader with some advice regarding model choice.

The strength of regression models is that they allow to estimate the size of the 'effect' of an explanatory variable on the dependent variable (the word 'effect' may be misleading because it suggests causation while a regression analysis in itself does not exclude the possibility of inverse causation or spurious causation resulting from omitted variables). As opposed to association analysis a regression analysis allows the researcher to quantify the effect of changes in the independent variables on the dependent variable. Another advantage is that regression analysis easily allows distinguishing and isolating the effects of different explanatory variables. An interesting example in this regard is the multilevel logistic analysis of the Leiden ranking by Bornmann, Mutz, & Daniel (2013), which shows that only 5% of the variation between universities in terms of the percentage of their publications that belong to the 10% most cited in a certain field is explained by between university differences, whereas about 80% is explained by differences among countries. Regression models can also be used for prediction, although the quality of such predictions is obviously conditional on the quality of the model. For most models, methods or rules of thumb to evaluate the quality of the resulting predictions are available.

The chapter introduces the main limited dependent variable models and illustrates their use to analyze the determinants of citations using data on the 2,271 journal articles published between 2008 and 2011 in the journals Journal of Informetrics (JOI), Journal of the American Society for Information Science and Technology (JASIST), Research Evaluation (RE), Research Policy (RP), and Scientometrics (SM). The data used in this illustration are available through the publisher's website for interested readers to experiment with them on their own. The next section introduces the data set and the variables used in the illustration. Section 3 discusses the logit model for binary choice. The models for multiple responses and count data are discussed in sections 4 and 5. The final sections present some concluding remarks and practical guidance on how to estimate these models using the statistical software Stata (Long & Freese, 2006). We opted to illustrate the models in Stata because this program appears to be most commonly used in informetrics. However, all the models mentioned here may be run in R, and many in SPSS and other packages.

The aim of this chapter is primarily to demonstrate the possibilities of limited dependent variable models in LIS and on comparing their strengths and weaknesses in an applied setting. The theoretical description of the various models was kept brief for reasons of space. Readers looking for more elaborate treatments are referred to econometric (Greene, 2011; Wooldridge, 2012) or specialized textbooks (e.g. Agresti, 2002; Agresti, 2010; Hilbe, 2011).

2. The data: Which articles get cited in informetrics?

Several studies have investigated intrinsic and extrinsic factors that influence the citation impact of papers. In the models in this chapter we include 12 explanatory variables – the first five of which are inspired by the literature review in Didegah & Thelwall (2013b) – to explain the number of citations (including self-citations) in the calendar year following publication. Our aim is to illustrate the applicability and the use of limited dependent variable models. The 12 variables included in the analysis are:

- The journal in which an article is published (8% of the articles in our sample were published in the JOI, 33% in SM, 21% in RP, 6% in RE and 32% in JASIST). The popularity of a journal tends to correlate positively with the impact of the articles that appear in it.
- The number of authors of the article (NumAut: min=1, max=11, avg=2.40; SD=1.34). We included this variable because collaborative articles tend to receive more citations.
- The number of countries mentioned in the address field of the article (NumCoun: min=1, max=9, avg=1.31, SD=0.60). International collaboration too tends to increase the number of citations.
- The number of cited references included in the article (NumRef: min=0, max=282, avg=40.20, SD=25.34). Papers with more references often attract more citations.
- The length in terms of number of pages of the article (NumPag: min=1, max=37, avg=13.33, SD=4.88). Longer papers can have more content, including more tables and/or figures, which in turn may translate into receiving more citations.
- The length of the article title in terms of number of characters (NumTitle: min=10, max=284, avg=87.48, SD=31.10). On the one hand shorter titles might be more to the point, on the other hand longer titles might occur more in article searches.
- Whether the article is the first in an issue or not (First: 8% are first articles). An article that is the first in an issue, is likely to attract more attention and may therefore receive more citations.
- Whether funding information is included in the acknowledgments of the article or not (Fund: 20% of the articles have funding information). Rigby (2013) reports a weak positive link between more funding information and impact of papers.
- The publication year of the article (PubYear: min=2008, max=2011, avg=2009.60, SD=1.10). Over time the number of citations tend to increase (e.g. because more source titles are added to the WoS), so we need to correct for publication year.
- The month in which the article appeared (PubMon: min=1, max=12, avg=6.63, SD=3.42). This measure is based on information in WoS on the date of the print publication and does not account for the fact that some journals may be late or that articles could be available for 'early view'. We included this variable because the number of citations received in the year following publication (dependent variable) is likely to be influenced by the timing of the publication of the articles.
- Whether the article deals with the h-index or not (H: 7% of the articles are about the h-index). Articles were classified as dealing with the h-index if 'h-ind*', 'h ind*' and/or 'Hirsch' occurred in their abstract. Among other things the h-bubble article by Rousseau, Garcia-Zorita, & Sanz-Casado (2013), which shows that h-index related articles inflated short term citations to a large extent, inspired us to include this variable.

- Whether the article deals with issues related to innovation and patenting or not (InnoPat: 18% of the articles are related to these topics). Articles were classified as related to innovation and patenting if 'innovation' and/or 'patent*' occurred in their title and/or abstract. As innovation is high on governments' agendas, we wondered whether researching innovation would also pay off in terms of number of citations.

An issue that should be kept in mind when estimating regression models is the degree of correlation between these explanatory variables. Too much correlation (multicollinearity) inflates the standard errors on the estimated coefficients so that the estimated effects become instable and sensitive to small variations in the data. As an indicator of the degree of collinearity one can calculate Variance Inflation Factors (VIF) for every explanatory variable.[1] In our dataset the maximal VIF was 2.0, which we consider to be tolerable since thresholds of 5 or more are common in the literature (Menard, 1995; O'Brien, 2007).

3. Binary regression

In bibliometrics and informetrics binary logistic models are often used for analyzing and/or predicting whether articles will be cited or not (Van Dalen & Henkens, 2005), whether patents are commercialized (Lee, 2008), used in military applications (Acosta, Coronado, Marín, & Prats, 2013) or will be infringed (Su, Chen, & Lee, 2012).These models are also used in studies of funding and editorial decisions (Fedderke, 2013), winning scientific prizes or awards (Heinze & Bauer, 2007; Rokach, Kalech, Blank, & Stern, 2011), career transitions and promotions (Jensen, Rouquier, & Croissant, 2009) and the use of public libraries by internet users (Vakkari, 2012). Many other outcomes that can be analyzed through binary regression can be thought of, e.g. whether a researcher belongs to the editorial board of a certain journal, is likely to collaborate or publish a book, will file a patent, will move to another institution, or will pass a certain threshold in terms of citations or h-index.

a) The binary logit model

If y_i is a binary variable that can take only the values 0 and 1, then the logit model writes the probability $P(y_i = 1)$ as a function of the explanatory variables:

$$P(y_i = 1|x_{1i}, x_{2i}, \dots, x_{ki}) = P_i = G(\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki})$$

where $G(z) = \frac{e^z}{1+e^z}$ is the logistic function. The range of the logistic function is between 0 and 1 which ensures that the predicted probabilities are limited to this same range. This is one of the reasons why logit models are more appropriate than OLS in the case of a binary dependent variable. OLS should also be avoided because it assumes that the error terms are normally distributed with constant variances, while neither of these conditions apply when the dependent variable is binary (for similar reasons OLS should be avoided in models of ordinal or count dependent variables). The interpretation of the coefficients is not straightforward in the logit model. This can be seen when we rewrite the model as

[1] If ${R^2}_k$ is the coefficient of determination of a linear regression model that predicts the explanatory variable X_k as a function of the other explanatory variables, then the Variance Inflation Factor $VIF_k = \frac{1}{1-{R^2}_k}$.

$$\ln\left(\frac{P_i}{1-P_i}\right) = \text{logit}(P_i) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}$$

where $\frac{P_i}{1-P_i}$ are the odds of $y_i = 1$ (e.g. if $P_i = 0.8$ then the odds are 4 to 1). From this equation it is clear that β_i is the change in the log-odds when x_i increases by one unit and the other variables are held constant. The exponentiated coefficients e^{β_i} can then be interpreted as the factor by which the odds increase when x_i increases by one unit (e^{β_i} is the odds ratio). However, effects in terms of odds cannot be interpreted unambiguously in terms of probabilities, because the change in probability when x_i increases by one unit depends on the level of x_i and on the values of the other explanatory variables. One way around this problem is to estimate the 'marginal effect at the means' of an explanatory variable x_i, which measures the change in the prediction function if x_i increases by one unit.[2] Although such marginal effects cannot substitute for the estimated coefficients or odds ratios (which remain correct even when the explanatory variables deviate from their means), the marginal effects are usually informative.

The coefficients of the logit model are estimated by maximizing the likelihood of the data with respect to the coefficients. Most statistical software packages carry out the necessary iterative numerical optimization and calculate corresponding standard errors, which allow for significance tests on the coefficients (which test whether the estimated effects could be attributable to sampling variability). A global test on all parameters in the model tests whether the likelihood of the observed data using the estimated coefficients is significantly greater than the likelihood of a model that has no independent variables. This test is referred to as the likelihood ratio test and uses a test statistic that has an approximate chi-square distribution under the null hypothesis that all parameters are zero.

b) Illustration

We now use the logit model to study the citation of journal articles in the field of informetrics. The dependent variable measures whether or not the article was cited in another published article during the calendar year following its publication. 66 percent of all articles in our sample were cited, whereas the remaining 34 percent were not. Table 1 presents the estimated coefficients in the model with standard errors and significance tests, and the corresponding odds ratios and marginal effects.

[2] Next to the marginal effect at the means, other approaches are possible to calculate marginal effects. For a discussion and an example using bibliometric data, see Bornmann & Williams (2013).

Table 1. The binary logit model

	(1) Coefficients estimate/(SE)	(2) Odds ratio estimate/(SE)	(3) Marginal effects estimate/(SE)
JOI	0.851***	2.342***	0.175***
	(0.209)	(0.489)	(0.038)
RE	-0.706***	0.493***	-0.174***
	(0.212)	(0.104)	(0.052)
RP	0.510**	1.665**	0.112**
	(0.164)	(0.273)	(0.035)
SM	0.356**	1.428**	0.081**
	(0.128)	(0.183)	(0.029)
NumAut	-0.005	0.995	-0.001
	(0.036)	(0.036)	(0.008)
NumCoun	0.181*	1.198*	0.040*
	(0.086)	(0.103)	(0.019)
NumRef	0.007**	1.007**	0.002**
	(0.003)	(0.003)	(0.001)
NumPag	0.006	1.006	0.001
	(0.011)	(0.011)	(0.003)
NumTitle	-0.000	1.000	-0.000
	(0.002)	(0.002)	(0.000)
First	0.266	1.304	0.056
	(0.182)	(0.238)	(0.037)
Fund	-0.131	0.877	-0.029
	(0.124)	(0.109)	(0.028)
PubYear	0.048	1.049	0.011
	(0.043)	(0.045)	(0.010)
PubMon	-0.085***	0.918***	-0.019***
	(0.014)	(0.013)	(0.003)
H	0.953***	2.594***	0.176***
	(0.228)	(0.592)	(0.033)
InnoPat	-0.242	0.785	-0.055
	(0.139)	(0.109)	(0.032)
Constant	-96.290	0.000	
	(86.854)	(0.000)	
Chi^2	154.2	154.2	154.2
p	0.000	0.000	0.000
Pseudo-R^2	0.05	0.05	0.05
N	2271	2271	2271

* p.05; ** p.01; *** p.001.

The results indicate that – holding all the other explanatory variables in the model constant – articles published in the JOI, SM and RP have a significantly greater probability (than the reference category JASIST) of being cited, while that probability is lower for articles in RE. The publication month control variable has a negative effect, which was expected because the

probability of citation depends on the duration since publication. Other significant effects are found for international collaboration, the number of references listed in the article and for articles about the h-index.

The coefficients, odds ratios and marginal effects give an indication of the size of these effects. The estimated coefficient for JOI is .85, which implies that – ceteris paribus – the log-odds of JOI articles being cited are .85 greater than those of JASIST articles. The corresponding odds ratio is $e^{.85} = 2.34$, which implies that the odds of JOI articles being cited are 2.34 times greater than those of JASIST articles. The marginal effect provides an indication of the effect in terms of probabilities evaluated at the means of the explanatory variables: articles in the JOI are 18 percentage points more likely of being cited than articles in JASIST (remember that the overall unconditional probability of being cited is around 66 percent, so 18 percentage points is a substantial effect). Another sizeable marginal effect is that articles in RE are – holding all the other variables constant – 17 percentage points less likely of being cited than articles in JASIST. Articles about the h-index also increase their citation probability by 18 percentage points (compared to articles that do not write about the h-index). It appears that getting published in the JOI with an article on h-indices was a strategy worth considering for scholars in the field looking to improve their own h-index!

The model can now be used to make predictions by calculating predicted citation probabilities for articles with given values on the explanatory variables. Moreover, such predicted probabilities can also be calculated for the articles in our sample. This is a way to evaluate the predictive power of our model since we know whether these articles eventually were cited or not. If we use the common decision rule to predict citation when the predicted probability of an article is greater than .5, then the model makes correct predictions for 17 percent of the non-cited articles and 94 percent of the cited articles (table 2).

Table 2. Prediction table for the binary logit model

		Predicted category			
		Baseline		**Logit**	
		Not cited	Cited	Not cited	Cited
Observed category	Not cited	*0%*	*100%*	*17.1%*	*82.9%*
	Cited	*0%*	*100%*	*6.0%*	*94.0%*

In order to evaluate the quality of our model, these numbers should be compared to a baseline of correct predictions that would be made in absence of the explanatory variables. Since the overall proportion of cited articles is 66 percent, the best guess would then be to predict citation for any given article. In the non-cited category the proportion of correct predictions improves from 0 percent (baseline) to 17 percent in the logit model, while the proportion decreases from 100 percent to 94 percent among the cited articles. The sum of the proportions of correct predictions should be greater than 100 percent for a good model (Verbeek, 2008), which is the case in our example (the sum of the diagonal elements 17%+94% =111%).[3] The most common goodness-of-fit statistic for logit models is McFadden's R^2 (reported in Table 1 as 'Pseudo-R^2'), which is defined as the percent increase in the log-likelihood when moving from the baseline model with no explanatory variables to the full model.

[3] A related measure for model quality which is also based on the prediction table and which has the interesting property of ranging between 0 and 100%, is the Adjusted Count R^2 (see Long & Freese, 2006).

Note that this evaluation of the predictive power of our model relates to the internal validity of the model ('to what extent is the model capable of reproducing the sample data?'). Good internal validity does not imply that the model would perform equally well on new data. One way to assess external validity is to use only a subset (e.g. 90%) of the available observations to estimate the model (the training data) and to subsequently use the excluded observations (the test data) to evaluate the model's predictive capacity.

4. Ordinal regression

In the binary regression analysis of citations we lumped all articles that were cited (66% of the sample) together in one group. However, there may be important differences between articles that have just a few citations and those that have many citations. By not using this information the tests in the binary choice model have less power, which increases the risk of failing to demonstrate a true effect (a type II error). Ordered response or ordinal regression models are appropriate when the dependent variable is an ordinal scale.

Recent examples of applications of categorical or ordered logistic models in bibliometrics and informetrics include analyses and prediction of the factors that explain information seeking behavior of academic scientists (Niu & Hemminger, 2012), of the impact of international coauthorship on citation impact (Sin, 2011), of peer assessments of research groups (Engels, Goos, Dexters, & Spruyt, 2013) and of the popularity of new Twitter hashtags (Ma, Sun, & Cong, 2013). Other examples of outcomes that can be analyzed through ordered models include the outcomes of peer review of manuscripts submitted for publication (acceptance, minor review, major review, rejection), and the rank of professors (assistant professor, associate professor, full professor). In some cases, e.g. the published outcomes of a research project (academic papers only; patent only; academic papers plus patent; academic plus popularizing papers) the response categories may not be strictly ordered. In such cases a multinomial model can be used to analyze the data.

a) The ordered logit model

If y_i is an ordinal variable that can take only the values $j = 1, 2, \ldots, J$, then the cumulative probability is the probability that an observation i is in the j-th category or lower:

$$\gamma_{ij} = P(y_i \leq j)$$

The ordered logit model is then defined as

$$\text{logit}(\gamma_{ij}) = \alpha_j - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_k x_{ki}$$

The model has a different intercept α_j for each category j (the cutpoints), whereas the slope coefficients are assumed constant over the categories.[4] β_i is then the increase in the log-odds of being in a higher category when x_i increases by one unit while the other variables are held

[4] The negative signs before the coefficients are needed because cumulative probabilities were defined using a less-than or equal to symbol, while the coefficients should estimate the effect of explanatory variables on increasing levels of the dependent variable.

constant. As in the binary model the odds ratio e^{β_i} is the factor by which the odds of being in a higher category increase when x_i increases by one unit.

In the ordered logit model the slope coefficients are assumed equal at every categorical level. This 'proportional odds assumption' can be evaluated using the Brant test, which tests whether the slope coefficients are equal across separate binary models. In case the test does not support the assumption, an alternative model could be considered in which the coefficients are allowed to vary with the categorical levels (i.e. a multinomial logit model).

b) Illustration

We now use the ordered logit model to study the determinants of journal article citations. The dependent variable in the analysis is an ordinal variable with three categories: (1) no citations during the year following publication, (2) few citations (i.e. 1 or 2 citations), and (3) many citations (i.e. 3 or more). In our sample of 2,271 articles from the field of informetrics, 34 percent were not cited, 39 percent received one or two citations and the remaining 27 percent received three or more. Table 3 presents the estimated ordered logit model.

Table 3. The ordered logit model

	(1) Coefficients estimate/(SE)	(2) Odds ratio estimate/(SE)
JOI	0.841***	2.319***
	(0.166)	(0.385)
RE	-0.796***	0.451***
	(0.202)	(0.091)
RP	0.234	1.264
	(0.137)	(0.173)
SM	0.310**	1.363**
	(0.112)	(0.152)
NumAut	-0.012	0.989
	(0.032)	(0.031)
NumCoun	0.296***	1.344***
	(0.072)	(0.097)
NumRef	0.009***	1.009***
	(0.002)	(0.002)
NumPag	-0.006	0.994
	(0.010)	(0.010)
NumTitle	-0.001	0.999
	(0.001)	(0.001)
First	0.427**	1.533**
	(0.153)	(0.234)
Fund	-0.127	0.881
	(0.110)	(0.097)
PubYear	0.036	1.037
	(0.037)	(0.039)
PubMon	-0.088***	0.916***
	(0.012)	(0.011)
H	1.026***	2.789***
	(0.173)	(0.482)
InnoPat	-0.177	0.838
	(0.117)	(0.098)
cut1		
Constant	72.331	2.589e+31
	(74.884)	(1.939e+33)
cut2		
Constant	74.098	1.515e+32
	(74.885)	(1.135e+34)
Chi^2	222.2	222.2
p	0.000	0.000
Pseudo-R^2	0.04	0.04
N	2271	2271

* p.05; ** p.01; *** p.001.

The odds ratio for JOI is 2.3, which implies that – ceteris paribus – the odds of JOI articles being in a higher category are 2.3 times greater than those of JASIST articles. It is informative to compare these results with the ones from the binary model in table 3. Most of the coefficients have smaller p-values, which reflects the increased power by differentiating between articles with few and many citations. For example, the coefficient for international collaboration is now highly significant (p<.001 as opposed to p=0.035 in the binary model). The dummy variable indicating whether the article is the first article published in the journal issue is now significant (p=.005) while it was not in the binary model (p=.135). A further analysis shows the reasons for this finding: while first articles have a similar probability (than other articles) of not being cited, they have a much larger probability of having many citations. This indicates that the proportional odds assumption underlying the ordered logit model may be violated here (the effect on the odds of not being cited versus being cited is not the same as the odds of receiving a few citations versus many citations). There is one variable where an inverse scenario takes place: the indicator for articles published in RP is no longer significant in the ordered logit model. The reason is that in comparison with the JASIST reference category a large proportion of its articles are not cited (producing the effect in the binary analysis), while at the same time a slightly larger proportion of the RP articles have many citations (cancelling out the effect in the ordinal analysis).

The estimated ordered logit model can now be used to calculate predicted cumulative probabilities for every category. Because the cutpoints increase as the categorical level increases, the cumulative probabilities increase as well. Differences between adjacent cumulative probabilities yield predicted probabilities for each category. If we use as a decision rule to predict the category with the largest predicted probability, then the model makes correct predictions for 43 percent of the non-cited articles, 60 percent of the articles with few citations and 24 percent of the articles with many citations (table 4).

Table 4. Prediction table for the ordered logit model

		Predicted category					
		Baseline			**Ordered logit**		
		No citations	Few citations	Many citations	No citations	Few citations	Many citations
Observed category	No citations	*0%*	*100%*	*0%*	*43.4%*	*48.7%*	*7.9%*
	Few citations	*0%*	*100%*	*0%*	*29.2%*	*60.2%*	*10.6%*
	Many citations	*0%*	*100%*	*0%*	*16.5%*	*59.9%*	*23.6%*

A baseline model with no explanatory variables would predict a few citations for every article, because that is the category with the largest overall proportion (38%). The sum of the diagonal elements for the ordered logit model in table 4 (127.2%) is greater than that in the baseline model (always 100%), which is a minimum quality requirement for any model.

The Brant test to evaluate the proportional odds assumption (equality of the slope coefficients over the categories) results in a test statistic value of $\chi^2 = 23.5$ ($p = 0.07$). If $p > 0.05$ then the evidence against the proportional odds assumption is not significant. It may be worth to keep in mind that significance tests are all about sample sizes, which in this case implies that even small differences in slope coefficients could result in a rejection of the null hypothesis if the sample is large (while large differences may be insignificant in small samples). Because our p-value is not much greater than the significance level we also estimated a multinomial

model, which consists of multiple binary logit models so that the slope coefficients are allowed to vary (the specification and estimates are not reported). For this model the sum of the diagonal elements in a prediction table (not shown) increases to 131.8 percent. However, this small increase in predictive power requires the estimation of much more parameters in the model, which increases the risk of overfitting. Although both the ordered and the multinomial models have their merits, the authors favor the ordinal model in this case because of its parsimony and the fact that the proportional odds assumption is not implausible. A multinomial model would be appropriate if the proportional odds assumption is clearly violated as well as in the case of a non-ordinal dependent variable.

5. Count data models

If the variable of interest measures the frequency of an event, then count data models may be appropriate to take advantage of the cardinal (rather than ordinal) nature of the data. The standard regression framework for analyzing count data is the Poisson model, but in most practical applications extensions of this model (the quasi-Poisson and negative binomial models) are needed to overcome violations of underlying assumptions (discussed below).

Abbasi, Altmann, & Hossain (2011) implement a Poisson model to identify the effects of co-authorship networks on performance of scholars; Niu & Hemminger (2012) complemented their logistic analysis of information seeking behavior with a Poisson regression. Negative binomial regression models have been applied to model the number of papers (Barjak & Robinson, 2007; Gantman, 2012) and in the study of citation counts, for example when comparing sets of papers (Bornmann & Daniel, 2008; Bornmann & Daniel, 2006) or the relative importance of authors and journals (Walters, 2006). Lee et al. (2007) pioneered the use of a zero-inflated negative binomial in informetrics in their analysis of citations of patents of the Korean Institute of Science and Technology (KIST). Zero-inflated models have two parts: A binary model to predict group membership and a count model for the data in the latter group (Hoekman, Frenken, & van Oort, 2009; Long & Freese, 2006). Recently, Chen (2012), Didegah & Thelwall (2013a) and Yoshikane (2013) implemented zero-inflated negative binomial models in their studies of, respectively, predictive effects of structural variation on citation counts, of citation impact in nanoscience, and of citations of Japanese patents. Zero-inflated models assume two sources and hence different underlying causes of zeros: Perfect zeros for which structural factors explain the observation of zeros (e.g. the number of academic papers per toddler) and zeros that occur in the count distribution (e.g. some academics may have no papers during a number of years). As illustrated by Didegah & Thelwall (2013b) hurdle models may provide a good alternative, at least in the case of citations, as receiving its first citation can be considered a real hurdle for a paper after which it becomes more likely to be cited again. In the section below we limit the explanation to the standard negative binomial regression; readers interested in truncated and other variations may consult (Hilbe, 2011).

a) The Poisson, the quasi-Poisson and the negative binomial regression models

If y_i is a count variable taking only non-negative integer values ($y_i = 0, 1, 2, ...$) and we assume that y_i conditional on the values of the explanatory variables has a Poisson distribution:

$$P(y_i = y|x_{1i}, x_{2i}, \dots, x_{ki}) = \frac{e^{-\mu_i}.{\mu_i}^y}{y!} \qquad y = 0, 1, 2, \dots$$

where μ_i is the expected value of the distribution. Note that the assumption refers to the *conditional* distribution of y_i and not to the unconditional distribution of y_i. Because the latter distribution also depends on the distribution of the explanatory variables, the distribution of the observed y_i is not a valid argument for preferring this model over another. The following example makes this clear: In a model with only one binary explanatory variable in which the conditional distribution is Poisson, the unconditional observed y_i would in many cases have a bimodal distribution (and so clearly not be Poisson).

The expected value μ_i is usually modelled by

$$\mu_i = E[y_i|x_{1i}, \dots, x_{ki}] = e^{\beta_0+\beta_1 x_{1i}+\dots+\beta_k x_{ki}}$$

For technical reasons, the log of the conditional mean of Poisson (and negative binomial) models is estimated, rather than the mean itself. The Poisson regression model can thus be defined as

$$P(y_i = y|x_{1i}, x_{2i}, \dots, x_{ki}) = \frac{e^{-(e^{\beta_0+\beta_1 x_{1i}+\dots+\beta_k x_{ki}})}.(e^{\beta_0+\beta_1 x_{1i}+\dots+\beta_k x_{ki}})^y}{y!} \qquad y = 0, 1, 2, \dots$$

of which the coefficients are usually estimated using maximum likelihood. How to interpret these coefficients becomes clear if we write the expected value as

$$\ln(E[y_i|x_{1i}, \dots, x_{ki}]) = \beta_0 + \beta_1 x_{1i}+\dots+\beta_k x_{ki}$$

which is a semi-log model familiar from linear regression. β_i is then the relative (percent) increase in μ_i when x_i increases by one unit while the other variables are held constant.[5]

A limitation of the Poisson regression model is that any Poisson distribution is completely determined by its mean and that the variance is assumed to equal that mean (the equidispersion assumption). This restriction is violated in many applications because the variance is often greater than the mean. In such cases there is overdispersion, by which we mean that the variance is greater than the variance implied by assuming a Poisson distribution. However, the maximum likelihood estimator is considered to produce consistent estimates for the coefficients regardless of the actual conditional distribution (Wooldridge, 1997). The procedure of using Poisson maximum likelihood estimation without assuming that the Poisson distribution is correct, is referred to as the quasi-Poisson model or the Poisson QMLE (quasi-maximum likelihood estimator). In the case of overdispersion the standard errors of the coefficients will be underestimated in the Poisson regression, thereby increasing the risk of making a type I error (incorrectly concluding that an effect is significant). The quasi-Poisson model adjusts the standard errors by estimating an additional parameter in the model (the quasi-Poisson assumes the variance to be a fixed multiple of the mean).[6] The Poisson and quasi-Poisson will always return the same estimates of the coefficients.

---

[5] This interpretation is only approximately correct as it follows from differentiating $\ln(E[y_i|x_{1i}, \dots, x_{ki}])$ with respect to x_i. An exact interpretation is that the exponentiated coefficient e^{β_1} is the factor change in μ_i.

[6] The quasi-Poisson model assumes that $Var[y_i] = \varphi^2.E[y_i]$ where φ is an overdispersion parameter. An estimator for φ^2 is $\hat{\varphi}^2 = \frac{1}{n-k-1}\sum\frac{(y_i-\hat{y}_i)^2}{\hat{y}_i}$. Standard errors for the quasi-Poisson coefficients can then be obtained by multiplying those of the Poisson MLE by $\hat{\varphi}$.

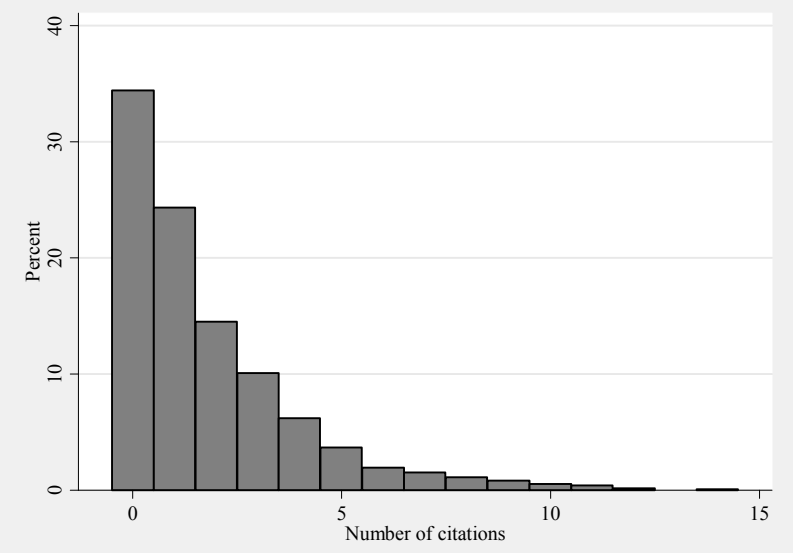
Another possibility in the case of overdispersion is to estimate a negative binomial regression model. This model also allows the conditional mean of y_i (μ_i) to differ from its variance ($\mu_i + \alpha.{\mu_i}^2$) by estimating an additional parameter (the dispersion parameter α). Since the negative binomial model assumes the variance to be a quadratic function of the mean, this model allows for far greater variances for large estimates of the mean than the quasi-Poisson.[7] The Poisson model can be regarded as a special case of this more general negative binomial model when α is zero. A significance test on α can thus be regarded as a test for the presence of overdispersion in the Poisson model. The probability mass function of the negative binomial distribution differs from the Poisson distribution so that the estimated coefficients – unlike those from the quasi-Poisson model – are not the same as in the Poisson model, although they tend to be similar.

A common goodness-of-fit statistic for count data regression models is a pseudo-R² that is calculated as the square of the correlation between the observed y_i and the values predicted by the model $\hat{y}_i$. This R² is indicative of the (internal) validity of the model when it is used for making predictions. An additional measure is the Akaike information criterion (AIC), which trades off goodness-of-fit with model complexity, by adding a penalty for the number of parameters estimated in the model.[8]

b) Illustration

We now apply the Poisson regression model to analyze journal article citations during the year following their publication. Graph 1 summarizes the distribution of the number of citations in our sample of 2,271 articles.

Graph 1. Frequency distribution of the number of citations



[7] An extension that could further improve the fit is the Generalized Negative Binomial Regression that models the overdispersion parameter (see the `gnbreg` command in Stata).

[8] $AIC = -2\ln(likelihood) + 2k$, where k is the number of parameters estimated in the model. So models with lower values for the AIC are to be preferred. Unlike R², the AIC is a relative measure and is only useful for comparing models on the same data.

The mean number of citations is 1.94 while the variance in the distribution is 7.38. This indicates that there may be overdispersion in a Poisson regression model, so alternatives should be considered. Table 5 presents the estimated coefficients of a Poisson, a quasi-Poisson and a negative binomial regression model.

Table 5. Count data models

	(1) Poisson estimate/(SE)	(2) Quasi-Poisson estimate/(SE)	(3) Negative binomial estimate/(SE)
JOI	0.470***	0.470***	0.549***
	(0.052)	(0.092)	(0.097)
RE	-0.716***	-0.716***	-0.690***
	(0.104)	(0.183)	(0.141)
RP	-0.079	-0.079	-0.046
	(0.053)	(0.094)	(0.090)
SM	0.094*	0.094	0.133
	(0.043)	(0.077)	(0.073)
NumAut	0.006	0.006	-0.000
	(0.012)	(0.022)	(0.020)
NumCoun	0.190***	0.190***	0.229***
	(0.022)	(0.039)	(0.044)
NumRef	0.007***	0.007***	0.007***
	(0.001)	(0.001)	(0.001)
NumPag	-0.009*	-0.009	-0.011
	(0.004)	(0.007)	(0.006)
NumTitle	-0.001	-0.001	-0.001
	(0.000)	(0.001)	(0.001)
First	0.148**	0.148	0.151
	(0.054)	(0.096)	(0.095)
Fund	-0.127**	-0.127	-0.122
	(0.045)	(0.079)	(0.072)
PubYear	0.009	0.009	0.023
	(0.014)	(0.025)	(0.024)
PubMon	-0.056***	-0.056***	-0.062***
	(0.004)	(0.008)	(0.008)
H	0.655***	0.655***	0.691***
	(0.049)	(0.087)	(0.097)
InnoPat	-0.072	-0.072	-0.099
	(0.047)	(0.083)	(0.080)
Constant	-16.849	-16.849	-46.448
	(28.634)	(50.618)	(48.594)
α			0.88
R-squared	0.09	0.09	0.08
AIC	9854	9854	8312
N	2271	2271	2271

* p.05; ** p.01; *** p.001.

In the Poisson model the estimated coefficient for the JOI is .47, which implies that – holding the other variables constant – the predicted number of citations of JOI articles is 47 percent higher than that of JASIST articles. Articles about the h-index have 66 percent more citations than other articles. Note that the coefficient estimates in the quasi-Poisson model are identical to those of the Poisson model. As could be expected the standard errors in the quasi-Poisson are substantially greater than those in the Poisson model. In fact they all are 77% greater because the overdispersion statistic was $\hat{\varphi}$ =1.77 (not in the table). This indicates that the Poisson distribution assumption was violated and that the Poisson model should not be used for inference. For example, it would be wrong to conclude that the effect of the variable 'First' is significant, since that result in the Poisson model is based on underestimated standard errors. Significance tests in the quasi-Poisson model show a positive effect of articles published in the JOI and a negative effect for articles published in RE (compared to articles in JASIST). We also find significant positive effects from international collaboration, the number of cited references, articles about the h-index and the publication month control variable. The results of the negative binomial model are very similar to those of the quasi-Poisson: the same effects are significant at the same significance levels and with very similar estimated effect sizes. For example, the predicted number of citations for an article on the h-index is 70 percent higher than for other articles, whereas the effect of one additional reference is .7 percent. Hence a 100 additional references had the same effect than switching to an h-index related topic, which illustrates the effect of the h-bubble (Rousseau, Garcia-Zorita, & Sanz-Casado, 2013). The estimated dispersion parameter α is .88 indicating overdispersion. A likelihood-ratio test that compares the negative binomial model with a model where α is zero (the Poisson model) confirms that the overdispersion parameter is significant ($\chi^2 = 1544, p < .001$) so that Poisson model is not reliable and the quasi-Poisson or the negative binomial should be preferred.

In order to evaluate the goodness-of-fit we calculated Pearson correlation coefficients between the observed number of citations and the predicted counts in the (quasi-)Poisson model (r=.300) and those in the negative binomial model (r=.290), resulting in values for pseudo-R^2 of .09 and .08 respectively. On the other hand, the Akaike information criterion (AIC) indicates a better fit in the negative binomial model, which has a smaller value for the AIC.

The effects that are found in the count data models are mostly the same effects that we found earlier in the categorical (binary and ordinal) models. This indicates that the main results of the analysis are robust to alterations in the model specification. Yet while the results were fairly robust, each approach did yield additional insights that might have been overlooked had only one approach been used. For example, the explanatory variable First, indicating whether an article is the first in a journal issue or not, did have a significant effect in the ordinal model, but not in the binary model nor in the count data models. With regard to the effect of the journals, one would draw similar conclusions from each of the models for JOI and for RE (the first yielding higher citation impact during the year following publication than papers in JASIST; the latter resulting in lower such citation impact). For RP and SM, however, a comparison of the results of the different models leads to a more nuanced idea as regards the citation impact of their papers in comparison with papers in JASIST.

6. Limited dependent variable models in Stata

The data used for the analyses presented above are available via the publisher's webpage. We now show how our results can be obtained using the statistical software Stata.

To estimate the binary logit model where the dummy variable 'D_cited' indicates the outcome of an article being cited and in which JASIST (the 2nd journal) is the reference category:

```
logit D_cited ib2.Journal NumAut NumCoun NumRef NumPag NumTitle
      i.First i.Fund PubYear PubMon i.H i.InnoPat
```

A prediction table, the adjusted count R^2, odds ratios and marginal effects at the means in the binary logit model were obtained by:

```
estat classification

fitstat

logit D_cited ib2.Journal NumAut NumCoun NumRef NumPag NumTitle
      i.First i.Fund PubYear PubMon i.H i.InnoPat, or

margins, dydx(*) atmeans
```

To reduce the code needed to estimate the other models, we first define a list of independent variables which we call 'indeps':

```
local indeps JOI SM RP RE NumAut NumCoun NumRef NumPag NumTitle First
      Fund PubYear PubMon H InnoPat
```

For the estimation of coefficients and odds ratios in the ordered logit model where the categorical variable 'citation_categories' contains the three outcome categories:

```
ologit citation_categories `indeps'

ologit citation_categories `indeps', or

brant, detail
```

The Poisson, quasi-Poisson and negative binomial models for the count variable 'citations' are obtained by:[9]

```
poisson citations `indeps'

glm citations `indeps', family(poisson) link(log) scale(x2)

nbreg citations `indeps'
```

[9] These models are part of a broader class of Generalized Linear Models (GLM). The quasi-Poisson model is estimated in Stata as a GLM in which the standard errors are adjusted ('scaled') using the Pearson chi-square ('x2') of the observed and predicted values in the model (i.e. the estimated overdispersion parameter $\hat{\varphi}$ that we discussed earlier).

7. Conclusion

Outcome variables that are categorical or frequency counts are common in informetrics. This chapter introduced and compared common limited dependent variable regression models that can be used to analyse such data. The use of linear models may often not be justified in informetrics, as the assumptions underlying them often do not apply in informetric datasets (Leydesdorff & Bensman, 2006). A practical issue for researchers is to decide which of the limited dependent variable models and their variations is most appropriate. In many cases the nature of the data will determine that choice (e.g. if the outcome variable is binary then there are no other options than to estimate a binary model). But sometimes the data will offer different options for modelling, as in the example of citations counts used throughout this chapter. In this case the researcher might strive to maximally exploit the information and variation in the data by avoiding to group observations into broader categories. However, there may be valid reasons for estimating categorical models in those cases too (e.g. if aggregated categories are considered more appropriate for a certain research question). In such cases it may be instructive to estimate and compare different models. Yoshikane (2013), for example, used linear, logistic as well as zero-inflated negative binomial models in his analysis of patent citation frequency; Niu & Hemminger (2012) ran a Poisson and two logistic models in their analysis of information seeking behaviour. Altering the specification is a way to check the robustness of the main results of a study and to detect interesting anomalies in the data.

Acknowledgements

The authors thank Fereshteh Didegah, Raf Guns, Edward Omey, and Ronald Rousseau for their suggestions during the writing of this chapter. We also thank the reviewers Richard Williams and Paul J Wilson for their feedback and excellent suggestions.

References

Abbasi, A., Altmann, J., & Hossain, L. (2011). Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics, 5,* 594-607.

Acosta, M., Coronado, D., Marín, R., & Prats, P. (2013). Factors affecting the diffusion of patented military technology in the field of weapons and ammunition. *Scientometrics, 94,* 1-22.

Agresti, A. (2002). *Categorical data analysis*. (2nd ed.) New York: Wiley.

Agresti, A. (2010). *Analysis of ordinal categorical data*. (2nd ed.) New York: Wiley.

Barjak, F. & Robinson, S. (2007). International collaboration, mobility, and team diversity in the life sciences: Impact on research performance. In D. Torres-Salinas & H. F. Moed (Eds.), *Proceedings of ISSI 2007* (pp. 63-73). Madrid: ISSI.

Bornmann, L. & Daniel, H.-D. (2008). Selecting manuscripts for a high-impact journal through peer review: A citation analysis of communications that were accepted by *Angewandte Chemie International Edition*, or rejected but published elsewhere. *Journal of the American Society for Information Science and Technology, 59,* 1841-1852.

Bornmann, L. & Daniel, H. D. (2006). Selecting scientific excellence through committee peer review - A citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. *Scientometrics, 68,* 427-440.

Bornmann, L., Mutz, R., & Daniel, H.-D. (2013). Multilevel-statistical reformulation of citation-based university rankings: The Leiden ranking 2011/2012. *Journal of the American Society for Information Science and Technology, 64,* 1649-1658.

Bornmann, L. & Williams, R. (2013). How to calculate the practical significance of citation impact differences? An empirical example from evaluative institutional bibliometrics using adjusted predictions and marginal effects. *Journal of Informetrics, 7,* 562-574.

Chen, C. (2012). Predictive effects of structural variation on citation counts. *Journal of the American Society for Information Science and Technology, 63,* 431-449.

Didegah, F. & Thelwall, M. (2013a). Determinants of research citation impact in nanoscience and nanotechnology. *Journal of the American Society for Information Science and Technology, 64,* 1055-1064.

Didegah, F. & Thelwall, M. (2013b). Which factors help authors produce the highest impact research? Collaboration, journal and document properties. *Journal of Informetrics, 7,* 861-873.

Engels, T. C. E., Goos, P., Dexters, N., & Spruyt, E. H. J. (2013). Group size, h-index and efficiency in publishing in top journals explain expert panel assessments of research group quality and productivity. *Research Evaluation, 22,* 224-236.

Fedderke, J. W. (2013). The objectivity of national research foundation peer review in South Africa assessed against bibliometric indexes. *Scientometrics, 97,* 177-206.

Gantman, E. R. (2012). Economic, linguistic, and political factors in the scientific productivity of countries. *Scientometrics, 93,* 967-985.

Greene, W. H. (2011). *Econometric analysis*. (7th ed. ed.) Upper Saddle River: Prentice Hall.

Heinze, T. & Bauer, G. (2007). Characterizing creative scientists in nano-S&T: Productivity, multidisciplinarity, and network brokerage in a longitudinal perspective. *Scientometrics, 70,* 811-830.

Hilbe, J. M. (2011). *Negative binomial regression*. (2nd ed.) Cambridge, UK: Cambridge University Press.

Hoekman, J., Frenken, K., & van Oort, F. (2009). The geography of collaborative knowledge production in Europe. *Annals of Regional Science, 43,* 721-738.

Jensen, P., Rouquier, J.-B., & Croissant, Y. (2009). Testing bibliometric indicators by their prediction of scientists promotions. *Scientometrics, 78,* 467-479.

Lee, Y. G. (2008). Patent licensability and life: A study of US patents registered by South Korean public research institutes. *Scientometrics, 75,* 463-471.

Lee, Y.-G., Lee, J.-D., Song, Y.-I., & Lee, S.-J. (2007). An in-depth empirical analysis of patent citation counts using zero-inflated count data model: The case of KIST. *Scientometrics, 70,* 27-39.

Leydesdorff, L. & Bensman, S. (2006). Classification and powerlaws: The logarithmic transformation. *Journal of the American Society for Information Science and Technology, 57,* 1470-1486.

Long, J. S. & Freese, J. (2006). *Regression models for categorical dependent variables using Stata*. College Station, TX: Stata Press.

Ma, Z., Sun, A., & Cong, G. (2013). On predicting the popularity of newly emerging hashtags in Twitter. *Journal of the American Society for Information Science and Technology, 64,* 1399-1410.

Menard, S. (1995). *Applied logistic regression analysis*. Thousand Oaks, CA: Sage.

Niu, X. & Hemminger, B. M. (2012). A study of factors that affect the information-seeking behavior of academic scientists. *Journal of the American Society for Information Science and Technology, 63,* 336-353.

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance unflation factors. *Quality & Quantity, 41,* 673-690.

Rigby, J. (2013). Looking for the impact of peer review: does count of funding acknowledgements really predict research impact? *Scientometrics, 94,* 57-73.

Rokach, L., Kalech, M., Blank, I., & Stern, R. (2011). Who is going to win the next Association for the Advancement of Artificial Intelligence fellowship award? Evaluating researchers by mining bibliographic data. *Journal of the American Society for Information Science and Technology, 62,* 2456-2470.

Rousseau, R., Garcia-Zorita, C., & Sanz-Casado, E. (2013). The h-bubble. *Journal of Informetrics, 7,* 294-300.

Sin, S.-C. J. (2011). International coauthorship and citation impact: A bibliometric study of six LIS journals, 1980-2008. *Journal of the American Society for Information Science and Technology, 62,* 1770-1783.

Su, H. N., Chen, C. M. L., & Lee, P. C. (2012). Patent litigation precaution method: Analyzing characteristics of US litigated and non-litigated patents from 1976 to 2010. *Scientometrics, 92,* 181-195.

Vakkari, P. (2012). Internet use increases the odds of using the public library. *Journal of Documentation, 68,* 618-638.

Van Dalen, H. P. & Henkens, K. (2005). Signals in science - On the importance of signaling in gaining attention in science. *Scientometrics, 64,* 209-233.

Verbeek, M. (2008). *A guide to modern econometrics*. New York: Wiley.

Walters, G. D. (2006). Predicting subsequent citations to articles published in twelve crime-psychology journals: Author impact versus journal impact. *Scientometrics, 69,* 499-510.

Wooldridge, J. M. (2012). *Introductory econometrics: A modern approach*. (5th ed. ed.) Andover: Cengage Learning.

Yoshikane, F. (2013). Multiple regression analysis of a patent's citation frequency and quantitative characteristics: The case of Japanese patents. *Scientometrics, 96,* 365-379.