Modelling Research Productivity Using a Generalization of the Ordered Logistic Regression Model

Delia North
Temesgen Zewotir
Michael Murray

Abstract
In South Africa, the Department of Education allocates funds to universities by means of a funding formula that focuses primarily on student throughput and academic staff-based research productivity. Accordingly, South African universities have developed their own strategies to help improve their student throughput and staff publication rates. In this paper we are concerned with identifying potential factors that affect the publication rates of academic staff at the University of KwaZulu-Natal (UKZN). Some extensions of the ordered logistic regression model will be considered with the final objective being to produce a model that can assign a particular academic (with a given set of demographic variables) to one of four possible publication-based productivity classes.

Keywords: research productivity, logistic regression, generalised ordered logistic regression

Introduction
Because the research output generated by a publishing academic forms a very important component of any funding that a South African university receives from the Department of Education, the identification of specific factors that affect research output has become a very important point of focus at the

University of KwaZulu-Natal (UKZN). In an earlier paper, North, Zewotir, and Murray (2011) found that a person's type of qualification and the size of the school in which they reside both play a very important role in determining the level of research output that will be produced. Their model, however, relies heavily on the fitting of a logistic model to a publishing versus non-publishing response variable. In this paper we would like to relax this restriction by making it possible for each academic member of staff to be assigned to one of four possible publication-based productivity classes. In particular, an ordinal response variable Y will be assigned a value 0 if any member of staff has not been able to produce, through the medium of publication or post-graduate supervision, any productivity unit points during a given calendar year. This response variable Y will be given the value 1 if they have been able to generate up to 60 productivity units for the year. If they have been able to generate more than 60 but not more than 120 productivity units in a given year, this response variable will be given the value 2. If they have been able to generate more than 120 productivity units in a given year then this response variable will be given the value 3. When a single paper has multiple authors, the productivity unit count of 60 points associated with the paper is apportioned equally between the authors.

The above concept of a productivity unit count arises from a discussion amongst the various faculties at UKZN on how they should `fairly' apportion the productivity units that can be associated with a published piece of work. As has been justifiably pointed out by a reviewer of this paper, it could be argued that such a rule may bias the point allocation process in favour of disciplines where a joint collaboration between researchers is less necessary. This effect may be mitigated by using the `impact rating' associated with a particular journal to adjust the productivity unit count that has been allocated to a published piece of work.

Research Activity	Productivity unit count
Journal article(sole author)	60
Entire book	100
Chapter in book	15
Graduating MSc student	16
Graduating PhD student	60
Patent	80

Because all members of staff at UKZN are expected to undertake some form of research, no distinction has been made between those members of staff (usually lecturers and senior lecturers) who in another university may be asked only to perform a teaching role, and professors who are also required to add a significant amount of research to their output metric. To include a teaching versus research publication scenario in our analysis, one could consider fitting a zero-inflation model to our data. Such an analysis has been done in another paper (in press) where we have fitted a zero-inflated negative-binomial and zero-inflated Poisson model to our data and then compared the obtained fit with that of a hurdle model.

By introducing an ordered response variable, one is now able to focus on comparing a publishing scenario (Y>0) with a non-publishing (Y=0) scenario as well as a prolific publishing scenario (Y=3) with a non-prolific publishing (Y<3) scenario. For example, one may find that some of the covariates (to be introduced in the next section) may exert a very different effect on a publishing versus non-publishing scenario when compared with a prolific versus non-prolific publishing scenario. It is this aspect of the data that we were not able to capture effectively in the paper published earlier (North, Zewotir & Murray 2011)

The Data

Our study was restricted to identifying the per annum based research output of staff who are permanently employed at UKZN and who occupy the positions of lecturer, senior lecturer, or professor (i.e. associate professor or professor) at UKZN. Each year, a per annum based productivity unit count was obtained by looking at the total number of books, chapters in books, articles in peer reviewed journals, and supervision of MSc and PhD theses that a given academic had produced. Based on this productivity unit count, the academic was then assigned a value for Y which was recorded (separately for each year) together with a set of academic and demographic covariates that we hope will help to further improve the prediction capabilities of the models that we will be developing. In particular, associated with each response variable Y there was an indicator variable denoting whether the academic was female or male; a set of indicator variables denoting whether the academic was a lecturer, senior lecturer, or a professor; a set of indicator variables denoting the faculty in which the staff member was located (Education, Engineering, Health, Humanities, Law, Management, Medicine or Science); a set of indicator variables denoting the racial group to which the academic belongs (African, Coloured, Indian or White); an indicator variable denoting whether the academic has a PhD qualification or not; a variable denoting the number of academics in the school in which the academic resides (size); and an age category index determined by assigning a value 0 to this random variable if the academic (in that particular year) is in their twenties, a value 1 if they are in their thirties, and so on.

Focusing on the issue of gender, higher education institutions in South Africa have in the past been dominated by men (Teferra & Altbachl 2004). As a consequence, one might have expected to find that men had achieved a higher-level of research output than women in our study period. The following table contains a breakdown (according to gender) of the number of academics who were able to reach a particular productivity unit based category over a given year. A chi-square test for independence produced a value of 52.97 (with 3 degrees of freedom) which suggests that a strong relationship between gender and research productivity may exist in our dataset.

Table 1: A per annum based breakdown of research productivity versus gender over the period 2004 to 2008

	Y=0	Y=1	Y=2	Y=3
Male	1916	905	324	359
Female	1410	615	190	122
Total	3326	1520	514	481

Focusing on the issue of race, the system of apartheid that existed in South Africa will almost certainly have contributed to a lower level of research output achieved by black academics in the earlier years of our study period (Strydom & Fourie 1999). Special funding initiatives for black researchers introduced in more recent times may however see an improvement in the level of research output generated by black academics at UKZN. The following table contains a breakdown (according to race) of the number of academics who were able to reach a particular productivity unit based category. A chi-square test for independence produced a value of 127.4 (with 9 degrees of freedom) which suggests that a strong relationship between race and research productivity may exist in our dataset.

Table 2: A per annum based breakdown of research productivity versus race over the period 2004 to 2008

	Y=0	Y=1	Y=2	Y=3
African	652	241	63	56
Indian	1134	431	120	112
Coloured	98	19	9	3
White	1442	829	322	310
Total	3326	1520	514	481

Focusing on the age of a particular academic, one may expect to find that younger members of staff are more research active because they are still trying to build up research profiles. On the other hand, older members of staff may have become increasingly more involved with performing administrative duties or they may have already attained the research status that they desire. As a consequence, one might expect to see their research output beginning to taper off as they approach retirement age.

Because the completion of a PhD degree indicates an ability to conduct publication- based research, one would expect this variable to play a key role in determining the level of research output that a given academic generates at a given university. It may however be argued that once an academic has obtained a PhD, they could then enter a comfort zone in which research output begins to drop off. This paper will be able to provide a method of assessing whether this is indeed the case and whether this effect is different for a publish versus non- publishing scenario as opposed to a prolific publishing versus non-prolific publishing scenario.

Faculties generally tend to identify key areas of strength in their make-up which are then generously funded through the granting of scholarships and post-doctoral fellowships. As a result one would expect faculty membership to become an important predictor of research output

(Bland & Schmitz 1986). The following table contains a breakdown of the per annum based research productivity record of staff according to the faculties in which they reside. For example , looking at each per annum based record of members of staff in the Faculty of Science over the period 2004 to 2008, 578 out of a total of 3326 per annum based records recorded no output (Y=0) over a given year, and 350 recorded up to 60 productivity units over a given year. A Chi-square test for statistical independence between the rows and columns of this table produced a value 302.41 (with 21 degrees of freedom) which indicates a strong relationship between faculty membership and publishing productivity. One of the objectives of this paper is to determine which faculties are underperforming. With this knowledge in hand, management can begin to introduce measures that will help to stimulate research in underperforming faculties.

Table 3: A per annum based breakdown of research productivity versus faculty membership over the period 2004 to 2008

Faculty versus Research output	**Y=0**	**Y=1**	**Y=2**	**Y=3**
Science	578	390	133	151
Education	350	136	53	25
Engineering	192	121	26	25
Health	244	107	20	16
Humanities	758	402	163	178
Law	154	68	29	42
Management	433	101	32	20
Medicine	617	195	58	24
TOTAL	3326	1520	514	481

The above table indicates that a large proportion of academics at UKZN have produced no per annum based publication units whatsoever. Whereas one of the main objectives of our earlier study (North, Zewotir & Murray 2011) was to try and determine which demographic variables affect a publish versus non-publishing scenario, in this paper we are more interested in seeing if these demographic variables affect a publish (Y>0) versus non-publishing scenario (Y=0) in a different way to a prolific (Y=3) versus non-prolific (Y<3) publishing scenario.

A Generalised Ordered Logistic Regression

Because of the ordinal nature of our observed response variable, an obvious starting point is to fit an ordered logistic regression model (Fu 1998; Gould 2000) to our response variable *Y*. Such a model assumes that there exists an unobserved (but continuous valued) random variable Y^* that is linked to a set of exogenous variables *x* via a linear equation of the form

$$Y^* = x\beta + u$$

where the vector *x* contains the demographic variables that we have introduced above and *u* denotes an error term with a particular distribution function based specification which we will denote by *F(u).* This latent variable Y^* is then linked to a particular publication based productivity category in the following way:-

$$Y=j \text{ if and only if } c_j \leq Y^* < c_{j+1}$$

The points $\{c_0, c_1, c_2, c_3, c_4\}$ determine the so-called threshold (or cut-off) points for entry into a particular class with the model formulation then being completed by setting $c_0 = -\infty$ and $c_4 = +\infty$

Setting

$$F(u) = \frac{1}{1+e^{-u}} \qquad -\infty < u < \infty$$

produces an ordered logistic model that has

$$P(Y > j) = P(Y^* > c_{j+1})$$

$$= P\left(u > c_{j+1} - x\beta\right) = \frac{\exp(-c_{j+1}+x\beta)}{1+\exp(-c_{j+1}+x\beta)} = g(x, \beta)$$

Specifying a standard normal distribution for *F(u)* will produce an ordered probit model for our response variable *Y*.

Because the fitting of an ordered logistic model produces an odds ratio that can only vary in a proportional manner across each category of our response variable, viz.

$$\frac{P(Y>j)}{P(Y\leq j)} = e^{-c_{j+1} + x\beta} \quad (1)$$

such a modelling approach will not be able to help us determine, for example, whether having a PhD in a publishing versus non-publishing scenario (which would involve looking at the ratio $P(Y > 0)/\ P(Y \leq 0))$ plays a very different role to having a PhD in a prolific versus non-prolific publishing scenario which would require that we look at the ratio $P(Y > 3)/\ P(Y \leq 3)$.

One way of overcoming this problem would be to allow the regression parameter vector β itself to change in value as we move from one publication group to another. Known as a generalised ordered logistic model, when applied to our publication group based problem, this modelling approach then assumes that we have (setting $\alpha_j = -c_{j+1}$ in (1))

$$P(Y>j) = \frac{\exp(\alpha_j + x\beta_j)}{1+\exp(\alpha_j + x\beta_j)} = g(x, \beta_j)$$

with

$$P(Y = 0) = 1 - P(Y > 0) = \frac{1}{1+\exp(\alpha_1 + x\beta_1)}$$

And

$$P(Y = 3) = P(Y > 2) = \frac{\exp(\alpha_3 + x\beta_3)}{1+\exp(\alpha_3 + x\beta_3)}$$

respectively. From this model formulation we can then obtain the probabilities for entry into our other two publication classes

$$P(Y = j) = g(x, \beta_{j-1}) - g(x, \beta_j)$$

j=1,2

Results

As a starting point for our discussion, a logistic regression model was fitted to a binary response variable that was set equal to zero if the academic concerned had produced no research over a given year and was set equal to one otherwise. In terms of the response variable that we have introduced above, such a fitting procedure amounts to comparing an outcome from the group that we have labelled Y=0 with that from the combined group Y>0.

Table 4: **Parameter estimates for the logistic model (Response variable: publish)**

Covariate	**Estimate**	**95% Confidence Interval**	**p-value**
School size	0.004	[-0.00,0.01]	0.085
Male	-0.124	[-0.26, 0.01]	0.065
Lecturer*	-0.773	[-0.93,-0.62]	0.000
Professor*	0.839	[0.67,1.00]	0.000
Education*	0.309	[0.06,0.55]	0.013
Engineering	-0.503	[-0.33,0.22]	0.706
Health	-0.016	[-0.30,0.27]	0.913
Humanities *	0.265	[0.09,0.44]	0.003
Law	0.158	[-0.18,0.50]	0.335
Management*	-0.849	[-1.09,-0.60]	0.000
Medicine*	-0.302	[-0.53,-0.07]	0.010
African*	-0.288	[-0.46,-0.12]	0.001
Coloured*	-0.555	[-0.99,-0.12]	0.013
Indian*	-0.239	[-0.38,-0.09]	0.001
Age Index*	-0.194	[-0.27,-0.12]	0.000
PhD*	1.023	[0.88,1.17]	0.000
Constant	-0.206	[-0.52,0.11]	0.205

Using a 5% level of significance, the above results suggest that being a professor, having a PhD, or residing in the Faculty of Education or Humanities all help to increase the probability of being in a publishing group (Y>0) when compared with the probability of being in a non-publishing group (Y=0). In contrast, being a lecturer, residing in the Faculty of Management or Medicine, or being older all helps to decrease the probability of being in a publishing group when compared with a non-publishing group. Race also seems to play an important role with Africans, Indians and

Coloured being more likely to end up in the non-publishing group as compared with their white counterparts.

Because the purpose of this paper is to develop a method that can distinguish the effect, for researchers who publish prolifically, of each covariate in *x* from those who do not publish as prolifically or who do not publish at all, an ordered logistic model was fitted with the following results being obtained. In particular it should be noted that the estimates we have obtained for α_j relate to the threshold values that determine the entry of *Y* into a particular class, viz. we will have *Y*=0 if Y^* <0.2984, *Y*=1 if $0.2984 \leq Y^* < 1.935$, etc.

Table 5: Parameter estimates for the ordered logistic model (Response variable: *Y*)

Covariate	Estimate	95% Confidence Interval	p-value
School size*	0.007	[0.00,0.01]	0.004
Male	-0.020	[-0.14, 0.10]	0.750
Lecturer*	-0.773	[-0.92,-0.62]	0.000
Professor*	1.046	[0.90,1.19]	0.000
Education*	0.390	[0.17,0.61]	0.001
Engineering	-0.189	[-0.43,0.06]	0.133
Health	-0.036	[-0.30,0.22]	0.785
Humanities*	0.344	[0.19,0.50]	0.000
Law	0.277	[-0.03.592]	0.075
Management*	-0.851	[-1.09,-0.61]	0.000
Medicine*	-0.409	[-0.62,-0.20]	0.000
African*	-0.300	[-0.46,-0.14]	0.000
Coloured*	-0.544	[-0.98,-0.11]	0.014
Indian*	-0.209	[-0.34,-0.08]	0.002
Age Index*	-0.248	[-0.32,-0.18]	0.000
PhD*	1.054	[0.92,1.19]	0.000
Cut-off values α_1	0.299	[0.01,0.59]	
α_2	1.935	[1.64,2.23]	
α_3	2.905	[2.60, 3.21]	

The model based chi-square value of 1642.17 (with 16 degrees of freedom) obtained for our data indicates that the covariates (marked with asterisks) all have significant effects on our response variable Y.

In order to help with the development of an appropriate interpretation for some of the results given in Table 2, note that if one considers taking a logarithm of the odds ratio that appears in (1), then one obtains the following result, viz.

$$log\left(\frac{P(Y>j)}{P(Y\leq j)}\right) = \alpha_j + x\beta \qquad => \qquad P(Y > j) = e^{\alpha_j + x\beta} \; P(Y \leq j)$$

Thus a single unit increase in the value of the *k'*th component of *x* (keeping all the other components in *x* unchanged) will change the odds of observing $\{Y > j\}$ versus $\{Y \leq j\}$ by a multiplicative factor of $e^{\alpha_j + \beta_k}$ where β_k denotes the estimate that one has obtained for the *k'*th component of β. Because the estimates obtained for $\{\alpha_j$; j=1…3} are all positive valued, the above formula suggests that for any positive valued estimate for β_k, a single unit increase in the explanatory variable associated with β_k will make it more likely for that respondent to be placed in a higher category of Y. Similarly, for any negative valued estimate that we obtain for β_k a single unit increase in the value associated with the explanatory variable associated with β_k will make it less likely for that respondent to be placed in a higher category of Y.

With this explanation in hand it follows that any increase in the size of the school will make it more likely for that respondent to be placed in a higher category of publication. Being a professor, having a PhD, or being in the Faculty of Education or Humanities will also help to increase the probability of ending up in a higher category of publication. Being a lecturer or residing in the Faculty of Management or Medicine will reduce one's chances of being in a higher category of publication. Similarly, being of an older age or being of African, Coloured, or Indian origin also seems to reduce one's chances of ending up in a higher category of publication.

Having fitted an ordered logistic model, a test procedure (Brant, 1990) was run to see whether the fitting of an ordered logistic model is appropriate for the data that we observed. Brant's (1990) test procedure produced a significant chi-square value of 128.37 indicating that a parallel lines assumption is no longer appropriate for the evidence that we see in our data. As a consequence, a generalised ordered logistic model was fitted with the following results being obtained.

Table 6: Parameter estimates for the Generalised Ordered Logistic model

Group (Y)	Covariate	Estimate	95% Confidence Interval
Y>0 vs Y=0	School size*	0.005	[0.00,0.01]
	Male	-0.121	[-0.25, 0.01]
	Lecturer*	-0.758	[-0.92, -0.60]
	Professor*	0.852	[0.69,1.01]
	Education*	0.286	[0.04,0.53]
	Engineering	-0.041	[-0.31,0.23]
	Health	-0.010	[-0.29,0.28
	Humanities*	0.227	[0.05,0.40]
	Law	0.138	[-0.19,0.47]
	Management*	-0.857	[-1.11,-0.61]
	Medicine*	-0.346	[-0.58,-0.12]
	African*	-0.287	[-0.46,-0.12]
	Coloured*	-0.564	[-1.00,-0.13]
	Indian*	-0.220	[-0.36,-0.08]
	Age Index*	-0.193	[-0.27,-0.12]
	PhD*	1.020	[0.87,1.16]
	Constant	-0.222	[-0.54,0.09]
Y>1 vs Y≤ 1	School size*	0.011	[0.00,0.02]
	Male	0.144	[-0.02, 0.31]
	Lecturer*	-0.983	[-1.23, -0.74]
	Professor*	1.227	[1.03,1.42]
	Education*	0.684	[0.37,1.00]
	Engineering*	-0.478	[-0.84,-0.12]
	Health	-0.214	[-0.63,0.20]
	Humanities*	0.534	[0.33,0.74]
	Law*	0.465	[0.07,0.86]
	Management*	-0.5766	[-0.91,-0.24]
	Medicine*	-0.531	[-0.85,-0.21]
	African*	-0.258	[-0.49,-0.03]
	Coloured	-0.031	[-0.65,059]
	Indian	-0.185	[-0.38,0.01]
	Age Index*	-0.328	[-0.43,-0.23]

	PhD*	1.137	[0.92,1.34]
	Constant*	-2.264	[-2.69,-1.84]
Y=3 vs Y≤ 2	School size*	0.011	[0.00,0.02]
	Male*	0.397	[0.16, 0.63]
	Lecturer*	-1.163	[-1.57, -0.76]
	Professor*	1.510	[1.22,1.79]
	Education	0.349	[-0.13,0.83]
	Engineering*	-0.553	[-1.03,-0.07]
	Health	-0.263	[-0.84,0.32]
	Humanities*	0.604	[0.35,0.86]
	Law*	0.591	[0.09,1.09]
	Management*	-0.767	[-1,27,-0.26]
	Medicine*	-1.050	[-1.54,-0.56]
	African	-0.097	[-0.42,0,22]
	Coloured	-0.282	[-1.46, 0.90]
	Indian	-0.050	[-0.30,0.20]
	Age Index*	-0.430	[-0.57,-0.21]
	PhD*	1.111	[0.78,1.44]
	Constant*	-3.371	[-3.97,-2.78]

The first section of the table represents results that one would obtain if one ran a binary type logistic regression where the dependent variable has been recoded so that it compares the outcomes from the non-publishing group 0 with those from the publishing groups 1+2+3. The second panel of estimates in the table would result from running a binary type logistic regression where the dependent variable has now been recoded for outcomes from group 0+1 versus those from group 2+ 3, and the third panel would result from running a binary type logistic regression where the dependent variable has been recoded for group 0+1+2 versus group 3.

Thus, when it comes to interpreting the results that we obtained for a given panel, one needs to keep in mind that each panel compares the outcome-based categories that are greater than some value with those that are less than this same value. For example, the negative valued estimate that we obtained for being a lecturer in the top panel indicates that being a lecturer reduces one’s odds of being able to publish something (i.e. being in groups 1+2+3) as compared with someone who is not a lecturer. In contrast, being a professor increases one’s odds of being able to publish something (i.e. being in groups 1+2+3) as compared with someone who is not a professor.

One important criticism with the use of this type model is that it can easily overfit the data. To overcome this problem one could consider performing a series of Wald type tests on each of the variables that appear in Table 6 to see whether their coefficients differ across the three panels that are given in the table. If they do not differ, the constraints can then be added (in a sequential manner) to the model until we eventually arrive at a final model to which no additional parallel line assumptions can be added. Known as a partial proportional odds model (Williams 2006), such a procedure produced the following results.

Table 7: **Parameter estimates for the partial proportional odds model**

Group (Y)	Covariate	Estimate	95% Confidence Interval	p-value
Y>0 vs Y=0	School size	0.0040	[-0.0008,0.00987]	0.107
	Male	-0.1184	[-0.2479, 0.0112]	0.073
	Lecturer*	-0.7476	[-0.9041, -0.5911]	0.000
	Professor*	0.8386	[0.6766,1.0005]	0.000
	Education*	0.3018	[0.0631,0.5404]	0.013
	Engineering	-0.0200	[-0.2902,0.2502]	0.885
	Health	-0.0466	[-0.3075,0.2143]	0.726
	Humanities*	0.2591	[0.0844,0.4297]	0.003
	Law	0.2684	[-0.0332,0.5701]	0.081
	Management*	-0.7906	[-1.0263,-0.5549]	0.000
	Medicine*	-0.2989	[-0.5221,-0.0578]	0.009
	African*	-0.2696	[-0.4297,-0.1094]	0.001
	Coloured*	-0.4963	[-0.9283,-0.0641]	0.024
	Indian*	-0.1951	[-0.3286,-0.0617]	0.004
	Age Index*	-0.1862	[-0.2589,-0.1134]	0.000
	PhD*	1.0457	[0.9092,1.1822]	0.000
	Constant	-0.2575	[-0.5648,0.0497]	0.100
Y>1 vs Y≤ 2	School size*	0.0121	[0.0063,0.0179]	0.000
	Male	0.1391	[-0.2829, 0.3065]	0.103
	Lecturer*	-1.0130	[-1.2508, -0.7753]	0.000
	Professor*	1.2373	[1.0430,1.4315]	0.000
	Education*	0.6595	[0.3489,0.9700]	0.000
	Engineering	-0.4886	[-0.8489,-0.1363]	0.007
	Health	-0.0466	[-0.3075,0.2143]	0.726
	Humanities*	0.4930	[0.3007,0.6852]	0.000
	Law	0.2684	[-0.0332,0.5701]	0.081
	Management*	-0.7906	[-1.0263,-0.5549]	0.000

	Medicine*	-0.6096	[-0.9065,-0.3132]	0.000
	African*	-0.2696	[-0.4297,-0.1094]	0.001
	Coloured*	-0.4963	[-0.9283,-0.0641]	0.024
	Indian*	-0.1951	[-0.3286,-0.0617]	0.004
	Age Index*	-0.3363	[-0.4353,-0.2376]	0.000
	PhD*	1.0457	[0.9092,1.1822]	0.000
	Constant	-2.1682	[-2.5398, -1.7967]	0.000
Y=3 vs Y≤ 2	School size*	0.0139	[0.0065,0.0213]	0.000
	Male*	0.4029	[0.1707, 0.6351]	0.001
	Lecturer*	-1.2086	[-1.5988, -0.8184]	0.000
	Professor*	1.4999	[1.2279,1.7719]	0.000
	Education	0.3598	[-0.1104,0.8302]	0.134
	Engineering*	-0.5541	[-1.0229,-0.0852]	0.021
	Health	-0.0466	[-0.3075,0.2143]	0.726
	Humanities*	0.57421	[0.3351,0.8132]	0.000
	Law	0.2684	[-0.0332,0.5701]	0.081
	Management*	-0.7906	[-1.0263,-0.5549]	0.000
	Medicine*	-1.0958	[-1.5560,-0.6355]	0.000
	African*	-0.2696	[-0.4297,-0.1094]	0.001
	Coloured*	-0.4963	[-0.9283,-0.0641]	0.024
	Indian*	-0.1951	[-0.3286,-0.0617]	0.004
	Age Index*	-0.4534	[-0.5867,-0.3201]	0.000
	PhD*	1.0457	[0.9092,1.1822]	0.000
	Constant	-3.2706	[-3.7544,-2.7868]	0.000

Focusing on the two extremes of our publishing spectrum, namely those that do not publish at all (who have Y=0), and those that publish prolifically (who have Y=3), the following conclusions can be drawn from the estimates that appear in the above table. Because we are dealing with a non-randomised study, any conclusion relates to an association between the two variables rather than the conclusion that one variable is causing the other.

School Size: Because the estimate obtained for school size is statistically significant in the bottom panel (but not in the top panel), this result suggests that having a larger school size plays an important role in distinguishing a person who publishes prolifically (Y=3) from someone who does not publish prolifically (Y<3) but does not play a significant role in distinguishing a person who is able to publish (Y>0) from someone who is not able to publish at all (Y=0). Being positive valued, the result indicates that an increase in school size does seem help to improve the publication capabilities of academics in a given school but only at the high level end of the publishing spectrum.

Gender: Gender also seems to play a significant role but only when it comes to comparing a person who publishes prolifically (Y=3) with someone who does not publish prolifically (Y<3). The positive value obtained for this estimate indicates that males seem to perform better than females when only those who publish prolifically are considered.

Academic status: As one would expect, being a professor helps to improve one's publication capabilities. It could be argued, however, that one becomes a professor because one has a good publication record whereas one remains a lecturer because one has a poor publication record. These covariates therefore reflect rather than influence (i.e. cause) one's publication record. However, because we are dealing with an observational study without proper randomisation, it is important to emphasise that what we are talking about is really an association between the above two factors and not necessarily a causative relationship from one variable to another. Thus all we can conclude from this study is that having a PhD is positively associated with an increase in one's research productivity.

Faculties: Academics who reside in the Faculties of Education or Humanities seem to be performing well whether one considers the publishing (Y>0) versus non-publishing (Y=0) scenario that is given in the top panel, or the prolific (Y=3) versus non-prolific publishing scenario that is given in bottom panel of the table. Academics in the Faculties of Management Sciences or Medicine however seem to be performing poorly from a publication point of view no matter which panel one looks at.

Race: The negative valued estimates obtained for the African, Indian, and Coloured racial groups seem to suggest that they do not, from an increased publication point of view, appear to perform as well as their white counterparts. The parallel lines constraint that we accepted for this model however indicates that these effects remain the same whether we compare someone who publishes with someone who does not (as in the top panel) or someone who publishes prolifically with someone who does not publish prolifically (as in the bottom panel) of the table.

Age: An increase in age seems to have a detrimental effect on research productivity with this effect being greatest in the bottom panel where we compare someone who publishes prolifically with someone who does not publish prolifically.

Qualification: As one would expect, having a PhD helps to improve one's publication capabilities. The parallel line assumption that was accepted for this covariate, however, does seem to indicate that this effect remains the same whether we compare someone who publishes with someone who does not (as in the top panel), or someone who publishes prolifically with someone who does not publish prolifically (as in the bottom panel) of the table. Such a result should not be unexpected because the possession of a PhD indicates an aptitude for doing research which in turn leads to the production of more papers. Having obtained this qualification, however, an academic may be tempted to `rest on their laurels' which in turn may lead to a reduction in research output. The acceptance of the parallel lines assumption, along with the positive value that we obtained for this estimate, however, seems to indicate that this is not the case.

Conclusions

We have been able to improve upon the publishing versus non-publishing scenario that we developed in an earlier paper (North, Zewotir & Murray 2011). By modelling the above publication process as a generalised ordinal logistic model, which we have done in this paper, one is able to separate out the effect of covariates based on whether one wants to consider a publishing versus non-publishing scenario or a prolific versus non-prolific publishing scenario.

References

Allison, P 1999. Comparing Logit and Probit Coefficients across Groups. *Sociological Methods and Research* 28,2: 186 – 208.

Bland, CJ & CC Schmitz 1986. Characteristics of the Successful Researcher and Implications for Faculty Development. *Journal of Medical Education* 61,1: 22 - 31.

Boshoff, N 2009. Shanghai Academic Ranking of World Universities (ARWU) and the 'big five' South African Research Universities. *South African Journal of Higher Education* 23,4: 635 - 655.

De Villiers AP & AGW Steyn 2009. Effect of Changes in State Funding of Higher Education on Higher Education Output in South Africa: 1986 - 2007. *South African Journal of Higher Education* 23,1: 43 - 68.

Fu, V 1998. Estimating Generalized Ordered Logit Models. *Stata Technical Bulletin* 44: 27 – 30; *Stata Technical Bulletin Reprints* 8: 160 - 164. College Station, TX: Stata Press.

Gould, W 2000. Interpreting Logistic Regression in all its Forms. *Stata Technical Bulletin* 53: 19 - 28.

North, D, T Zewotir & M Murray 2011. Demographic and Academic Factors Affecting Research Productivity at UKZN. *South African Journal of Higher Education* (forthcoming).

Strydom, AH & M Fourie 1999. Higher Education Research in South Africa: Achievements, Conditions and New Challenges. *Higher Education* 38,2: 155 - 167.

Teferra, D & PG Altbachl 2004. African Higher Education: Challenges for the 21st Century. *Higher Education* 47,1: 21 - 50.

Williams, R 2006. Generalized Ordered Logit/ Partial Proportional Odds Models for Ordinal Dependent Variables. *Stata Journal* 6,1: 58 – 82.

Williams, R 2009. Using Heterogenous Choice Models to Compare Logit and Probit Coefficients across Groups. *Sociological Methods and Research* 37,4: 531 - 559.

Delia North
northd@ukzn.ac.za
Temesgen Zewotir
zewotir@ukzn.ac.za

Michael Murray
murraym@ukzn.ac.za
School of Statistics and Actuarial Science
University of KwaZulu-Natal