A short guide to the statistical analysis employed in the NRC Assessment of Doctoral Programs

Vijay Nair, Kerby Shedden
UM Department of Statistics

September 14, 2009

The National Research Council will soon release an "assessment of doctoral programs." Here we present an example that illustrates the main steps in the statistical analysis used to construct the NRC's overall program ratings.

Our understanding of the analysis approach employed by the NRC is mainly derived from the NRC's publication *A Guide to the Methodology of the National Research Council Assessment of the Doctorate Programs* (`http://www.nap.edu/catalog/12676.html`). We found the example beginning on page 19, and the flow-chart in figure A-1 of the appendix, to be especially useful for understanding their statistical analysis.

Data used in the formal assessment of doctoral programs

The NRC's overall program assessment will be a range of rankings for each program, based on three types of data:

- *Program variables*: This is the result of the data collection process, which involved questionnaires at the institutional, program, and faculty levels. Where practical, the questionnaire data was subjected to a data review conducted by the NRC.
- *Faculty importance measures*: One section of the faculty questionnaire asked faculty members to identify up to four of the most important program variables in their field.
- *Faculty program ratings*: Selected faculty members were asked to rate selected graduate programs in terms of overall quality on a six-point scale.

An example to illustrate the data analysis approach

As noted above, the NRC's methodology guide provides an example of the overall program assessment using the actual data from one unidentified program. What we show here complements the NRC guide, by illustrating how most of the calculations are carried out on a small, artificial dataset involving 5 program variables and 10 programs.

The raw data for all program in a field will resemble table 1. The program variables V1, V2, etc. would be attributes of program quality like the average number of publications per faculty, or program graduation rates. All variables are polarized so that greater numerical values correspond to greater program quality.

Next, the variables are standardized across all ten programs in the field. This means that the numerical values for each program variable are centered and scaled so that the mean

value becomes zero and the variance becomes one, as in table 2 (for presentation, the results are scaled by 100 and rounded to the nearest integer). Doing this allows each variable to contribute equally to the final results, regardless of its variability, or the scale on which it is defined.

At this point, it would become possible to define overall ratings using some weighting of the program variables. For example, if we weighted the five program variables as 0.2, 0.2, 0.4, 0.2, and 0, the overall rating, the resulting ranks would be as shown in table 3. The overall rating for program 1 would be calculated as follows:

$$62 = -70 \cdot 0.2 + 79 \cdot 0.2 + 124 \cdot 0.4 + 53 \cdot 0.2 + 31 \cdot 0.0. \tag{1}$$

Rather than making an arbitrary decision about weighting program variables, the NRC attempted to identify appropriate field-specific weights for each program variable. This was done by combining two approaches. One approach used the "faculty importance measures" to construct a set of "direct weights," and one approach used a set of "faculty program ratings" to construct a set of "regression-based" weights.

Direct weights: The NRC asked each faculty member to state which variables are considered most important in his or her field. Specifically, faculty were instructed to (i) identify up to four "most important" variables, then (ii) select two of the four that are the most important, then (iii) assign relative weights to the categories "faculty variables," "student variables," and "program variables." Scores of 2 were assigned to the two variables selected in step (ii), and scores of 1 were assigned to the two variables selected in (i) that were not selected in (ii). These scores were then multiplied by the relative weights for the category to which each variable belonged. This produced a set of direct weights derived from the response of a single faculty member. These weights were then averaged over all faculty members in a field to produce the final set of direct weights for a field.

Table 4 illustrates how the direct weights are constructed, using artificial data on five program variables with three program raters. For simplicity, only two variable categories are used, with variables V1-V3 belonging to one category and variables V4-V5 belonging to the other category. Also for simplicity, each rater selects only two "most important" variables, then selects one of these to receive the higher score of 2.

Regression-based weights: The NRC sampled 50 programs in each field (or 30 programs in smaller fields) for the faculty program rating survey. The programs were selected as a stratified random sample, to ensure equal representation by program size and geographic region. Faculty members in a field were asked to rate 15 programs on a scale of 1 to 6 in terms of overall quality. The assignment of programs to faculty members for rating was stratified, so that approximately equal representation of raters by rank (e.g. full professors) rated each program. The NRC aimed to get 50 reviews for each program, and had a process for requesting additional reviews if the desired number was not reached due to non-response.

The goal of the direct weights is not to provide an assessment of programs based on faculty

perception, but rather as an indirect means for quantifying the importance of the program variables. This was done using regression analysis to relate the program quality rating (the dependent variable in the regression) to the program variables (the independent variables in the regression). The fitted coefficients from this regression analysis were then standardized so that their absolute values summed to one to produce the "regression based weights."

Table 5 shows an artificial example of the raw program rating data. For simplicity, only 10 programs and 10 raters are used, with each rater rating only 5 of the 10 programs. If we couple the program rating data in table 5 with the program variable data in table 2, we can use regression analysis to approximately express the program ratings in terms of the program variables. Using this artificial data, the fitted relationship is

$$\text{Program rating} \approx 3.0 + 0.45 \cdot V1 + 0.21 \cdot V2 - 0.25 \cdot V3 - 0.05 \cdot V4 + 0.79 \cdot V5. \qquad (2)$$

The regression-based weights are derived from these coefficients by dividing through by the sum of absolute values. The resulting regression-based weights are shown in table 6.

This example illustrates the most important step in the construction of the regression-based weights. The actual regression analysis used by the NRC is somewhat more sophisticated than this as it uses a "model selection" approach to provide more stable results. This may be important, since the number of data points (ratings) was not much larger than the number of variables. The NRC's methodology report describes how this was done.

Creating combined weights

The direct weights and regression-based weights are combined to form a set of "combined weights." The combined weights could be used to construct ratings of overall program quality, in the same way as done with the arbitrary weights 0.2, 0.2, 0.4, 0.2, 0 above.

The actual approach used by the NRC to produce the combined weights from the direct weights and regression-based weights is complex. However the NRC methodology guide acknowledges that the final result of the procedure generally is highly similar to what would be obtained by taking a simple average of the direct weights and regression-based weights. The combined weights that would be obtained in our example using a simple average of the direct and regression-based weights are shown in table 7, and the resulting overall program ratings and rankings are shown in table 8.

The method for calculating these ratings and rankings is important for understanding how the NRC overall summary is calculated. However, values as shown in columns A and R of table 8 will not be released by the NRC. The overall rating provided by the NRC includes an additional layer of uncertainty analysis, as described next.

Uncertainty analysis

As noted above, there are several sources of uncertainty in the data collected by the NRC. The sources of uncertainty can be grouped into factors affecting the values of the program

variables, and factors affecting the computed variable weights. The NRC has adopted a simulation-based approach to incorporate these two types of uncertainty into the overall program assessment. The basic idea is that 500 times, a new set of data (including new versions of the program variables, faculty importance measures, and faculty performance ratings) is used for the complete analysis described above. This produces 500 sets of direct and regression-based weights, and overall program ratings and rankings.

The new data are constructed as follows.

New program variables: Program variables such as the number of publications per faculty member reflect research activity over a window of time. A simple statistical approach was used to estimate the variances of these variables over different time windows. Other variables, for which variance estimation based on the available data was difficult, were assigned nominal variance levels. Variables such as student health insurance, which do not tend to fluctuate over time, were given zero variance. The "new" program data is a perturbed version of the actual program data, with the variance of the perturbation based on the estimated level of variability for each program variable.

New variable importance and rating data: The faculty importance measures and faculty performance ratings are aggregated over all reporting faculty to produce one set of weights per field, either by simple averaging (in the case of the importance measures), or through a regression analysis (in the case of the performance ratings). In both cases, it is possible to simulate new results (i.e. new direct and regression-based weights) using a "random halves procedure," in which case half of the respondents are selected at random, and used to construct the weights as discussed above.

As noted earlier, uncertainty in the performance ratings principally results from the fact that a subset of programs was sampled for assessment, and a subset of faculty was assigned to each program. Thus for the regression-based weights, the "random halves" approach mimics the variance resulting from the random assignment of raters to programs. For the direct weights based on faculty importance measures, all faculty rated all program variables. Thus sampling and random assignment are not sources of uncertainty. The random halves procedure in this case might best be viewed as capturing any degree of non-repeatability that would result when asking the same people to rate the variables on two occasions, or when averaging over the somewhat different populations of faculty members that exist at two different points in time.

Returning now to the example, we illustrate in table 9 the application of the random-halves approach to the direct weights shown in table 4, by considering what would happen if we calculated the direct weights based on 2 of the 3 respondents. Note that here we only consider the effect of using a subsample of the respondents, we do not perturb the program variables as described above. In the actual analysis, these two perturbations of the data are done jointly. It is also important to note that in this illustrative example with only three respondents, the results will be more variable than in the actual study where there are many more respondents.

In table 10, we illustrate the random-halves approach applied to the regression-based weights by doing the regression analysis using 5 of the 10 raters. As above, this illustration uses the true program variables, rather than the perturbed program variables as in the actual NRC analysis.

The next step is to combine the random-halves direct weights with the random-halves regression-based weights. The NRC methodology guide states that these two sets of weights are statistically independent, which suggests that independent random halves and independent perturbations of the program variables were used in calculating the two sets of weights. The random-halves direct and regression-based weights were then paired and combined, as described above, to create random-halves combined weights. These combined weights are then used to form ratings for each program, which in turn determine rankings for each program. Since the random-halves procedure is repeated 500 times, we ultimately have 500 rankings for each program. At this point, the NRC does a statistical analysis to identify which program variables are not contributing significantly to the weights. These variables are then excluded and the weighting models are refit. Finally, the 500 refit weight vectors are used to define the overall program assessments. Specifically, the overall program assessment for a particular program will be the interval of the 500 ranks from the random halves analysis spanning from the 25^{th} percentile to the 75^{th} percentile.

Sources of uncertainty in the NRC analysis

Statistical uncertainty refers to any qualification of the results of a statistical analysis results due to limitations in the data. The goals of the NRC analysis methodology is to produce an interval of rankings for each program that accurately reflects all sources of uncertainty in the NRC data. Some tangible sources of uncertainty are described below.

1. The raw data (e.g. faculty publication, funding, and student graduate rates) are subject to reporting errors.

2. The faculty and program data describe a defined period of time, and may not reflect what happened either earlier or more recently in a particular program.

3. Only a subset of individuals was asked to rate each program in a field (with an average of 50 raters per program). Assignments of raters to programs was random, so if different random assignments had been made, somewhat different results would have been obtained.

4. The rating survey was separate from the primary faculty survey, and was subject to a different (potentially greater) degree of response selection bias.

5. Raters were not given an opportunity to indicate their knowledge level or confidence when rating programs. Thus, some raters may have had little or no basis for the ratings they provided, while others may have provided ratings based on detailed knowledge of the programs they rated.

	Program variables				
	V1	V2	V3	V4	V5
Program 1	34	84	19	77	107
Program 2	38	35	2	40	99
Program 3	49	80	4	99	49
Program 4	48	33	13	50	82
Program 5	98	42	6	97	80
Program 6	17	57	16	20	58
Program 7	92	25	18	78	166
Program 8	19	92	17	72	122
Program 9	64	99	2	66	119
Program 10	89	77	2	9	83

Table 1: Raw data for all ten programs in one field, where 5 program variables denoted V1, ..., V5 have been recorded.

	Program variables				
	V1	V2	V3	V4	V5
Program 1	-70	79	124	53	31
Program 2	-56	-101	-108	-68	7
Program 3	-19	65	-81	125	-139
Program 4	-23	-108	42	-35	-42
Program 5	145	-75	-53	119	-48
Program 6	-127	-20	83	-134	-113
Program 7	125	-137	111	56	203
Program 8	-120	109	97	37	75
Program 9	31	134	-108	17	66
Program 10	115	54	-108	-170	-40

Table 2: Standardized data corresponding to the raw data in table 1 (scaled by 100).

	Program variables						
	V1	V2	V3	V4	V5	Rating	Rank
Weight	0.2	0.2	0.4	0.2	0.0		
Program 1	-70	79	124	53	31	62	1
Program 2	-56	-101	-108	-68	7	-88	10
Program 3	-19	65	-81	125	-139	2	5
Program 4	-23	-108	42	-35	-42	-16	7
Program 5	145	-75	-53	119	-48	17	4
Program 6	-127	-20	83	-134	-113	-23	8
Program 7	125	-137	111	56	203	53	2
Program 8	-120	109	97	37	75	44	3
Program 9	31	134	-108	17	66	-7	6
Program 10	115	54	-108	-170	-40	-43	9

Table 3: Overall program rating and program ranks, based on variable weights 0.2, 0.2, 0.4, 0.2, and 0.

	Raters									
	1			2			3			
	I	W	D	I	W	D	I	W	D	
V1	0	0.2	0.0	0	0.2	0.0	0	0.9	0.0	0.00
V2	0	0.2	0.0	2	0.2	0.4	1	0.9	0.9	0.28
V3	0	0.2	0.0	0	0.2	0.0	0	0.9	0.0	0.00
V4	2	0.8	1.6	1	0.8	0.8	2	0.1	0.2	0.55
V5	1	0.8	0.8	0	0.8	0.0	0	0.1	0.0	0.17

Table 4: Calculation of direct weights for 5 variables, based on 3 raters. Each rater selects his or her "most important" variables (column I), and assigns relative weights to the variable categories (column W). The direct weights for each of the three raters are in the columns denoted D, and the overall weights for the field are in the final column.

	Raters									
	1	2	3	4	5	6	7	8	9	10
Progam 1	3	3				3		3	3	
Progam 2	2	4				4		4		2
Progam 3	2					2	2	2		2
Progam 4		2	2	2		2			2	
Progam 5	3	3	3			5			3	
Progam 6			1	1	1		1			1
Progam 7	4				2		6	4		6
Progam 8		3	3	3	3				5	
Progam 9				4	4		4	2		4
Progam 10			4	6	2		4		4	

Table 5: Rating data for 10 raters and 10 programs. Each rater rates 5 programs and each program is rated by 5 raters (in the actual study, each rater rates 15 programs, and each program is rated by around 50 raters.

Program variables					
V1	V2	V3	V4	V5	
0.45	0.21	-0.25	-0.05	0.79	Regression coefficient
0.26	0.12	-0.14	-0.03	0.45	Regression-based weight

Table 6: Fitted regression coefficients and regression-based weights.

Program variables					
V1	V2	V3	V4	V5	
0.00	0.28	0.00	0.55	0.17	Direct
0.26	0.12	-0.14	-0.03	0.45	Regression-based
0.13	0.20	-0.07	0.26	0.31	Combined

Table 7: Direct weights, regression-based weights, and combined weights.

	Program variables						
	V1	V2	V3	V4	V5		
Weight	0.13	0.20	-0.07	0.26	0.31	A	R
Program 1	-70	79	124	53	31	21	5
Program 2	-56	-101	-108	-68	7	-35	8
Program 3	-19	65	-81	125	-139	5	6
Program 4	-23	-108	42	-35	-42	-49	9
Program 5	145	-75	-53	119	-48	23	4
Program 6	-127	-20	83	-134	-113	-96	10
Program 7	125	-137	111	56	203	58	2
Program 8	-120	109	97	37	75	32	3
Program 9	31	134	-108	17	66	63	1
Program 10	115	54	-108	-170	-40	-23	7

Table 8: Program ratings (column A) and rankings (column R) based on the combined weights.

	Program variables				
	V1	V2	V3	V4	V5
All	0.00	0.28	0.00	0.55	0.17
2,3	0.00	0.57	0.00	0.43	0.00
1,3	0.00	0.26	0.00	0.51	0.23
1,2	0.00	0.11	0.00	0.67	0.22

Table 9: Direct weights calculated using two out of the three respondents.

	Program variables				
	V1	V2	V3	V4	V5
All	0.26	0.12	-0.14	-0.03	0.45
3, 5, 7, 8, 10	0.26	0.05	-0.15	0.03	0.52
1, 6, 8, 9, 10	0.28	0.07	-0.05	0.05	0.55
2, 3, 5, 8, 9	-0.08	0.07	-0.35	0.06	0.44
2, 4, 5, 7, 10	0.27	0.12	-0.13	-0.05	0.43
1, 4, 7, 9, 10	0.34	0.22	0.07	-0.08	0.29
2, 3, 4, 5, 6	0.07	0.14	-0.43	0.05	0.31
3, 4, 6, 8, 9	0.23	0.10	-0.17	-0.09	0.40
1, 3, 4, 5, 6	0.27	0.22	-0.22	-0.00	0.29
1, 3, 7, 8, 9	0.27	0.16	-0.03	-0.07	0.47
3, 5, 6, 8, 10	0.15	-0.05	-0.23	0.14	0.43

Table 10: Regression-based weights calculated using all raters, then using 10 different random halves consisting of 5 out of the 10 raters.