Multinomial regressions to identify phenotypes in Obsessive-Compulsive Disorder

Guaraci L. Requena*
Institute of Mathematics and Statistics, University of Sao Paulo (IME-USP), Sao Paulo, Brazil - guaraci@ime.usp.br

Roseli G. Shavitt
Department and Institute of Psychiatry, University of Sao Paulo (IPq - HCFMUSP), Sao Paulo, Brazil

Peggy M.A. Richter
Sunnybrook Health Sciences Centre, Department of Psychiatry, University of Toronto, Toronto, Canada

Gwyneth Zai
Centre for Addiction and Mental Health, Department of Psychiatry, University of Toronto, Toronto, Canada

Carlos A. B. Pereira
Institute of Mathematics and Statistics, University of Sao Paulo (IME-USP), Sao Paulo, Brazil

Abstract

Obsessive-Compulsive Disorder (OCD) is a psychiatric disorder characterized by intrusive thoughts (obsessions) and repetitive behaviors (compulsions), which affects about 2% of the population worldwide. Research in this area is under continuous development, with many interesting results being found, including genetic findings. In order to measure OCD severity, there are many interviews. The most used in the field is the Yale-Brown Obsessive-Compulsive Scale (Y-BOCS) and, most recently, the Dimensional Yale-Brown Obsessive-Compulsive Scale (DY-BOCS) has also been used by some authors. The DY-BOCS measures OCD severity discriminating five different and homogeneous symptom dimensions, plus one heterogeneous dimension. These measures reveal the OCD phenotype, which is of interest for genetic studies. On the other hand, even the Y-BOCS being the most used instrument to study OCD, it does not offer this phenotypical distinction. Our main objectives are 1. To build a technique to compare these two instruments in order to rewrite the DY-BOCS data in the same format of an existing Y-BOCS data set; and 2. To propose multinomial logistic regressions to model and to predict the phenotype (i.e., the most severe symptom dimension among the five homogeneous dimensions of the DY-BOCS) in a subject who has only the Y-BOCS data. This is an original study that aims to develop a statistical procedure to extract dimensional severity ratings in OCD from the existing Y-BOCS severity scores for use in genetic and other neurobiological research.

Keywords: OCD; DY-BOCS; Y-BOCS; multinomial logistic regression.

1. Introduction

Mental health studies have been a source of applied statistical problems. In this work, we chose a problem from the Obsessive-Compulsive Disorder (OCD) psychopathology. OCD is a neuropsychiatric disorder characterized by intrusive, recurrent and persistent thoughts and/or repetitive behaviors that usually have the function of neutralizing the distress caused by the unwanted thoughts.Distress, interference and time spent with symptoms are indicators of severity associated with OCD. Shavitt et al. (2014) defines:

- **Obsessions:** are intrusive, unwanted thoughts that cause distress and/or anxiety. The person attempts to ignore or suppress these obsessions with another thought or action (i.e., a compulsion).
- **Compulsions:** are repetitive behaviors/rituals or mental acts that the person feels driven to perform in response to an obsession. Compulsions are meant to neutralize or reduce the person's discomfort or to prevent a dreaded event.

Ruscio et al. (2010) states that OCD is the fourth most common psychiatric illness, with a lifetime prevalence of 1 to 3% and the World Health Organization has identified OCD as a leading global cause of nonfatal illness. OCD is a chronic disorder and manifests regardless of sex, race, intelligence, marital status, socio-economic status, religion or nationality.

Sometimes, everyone has obsessive thoughts about some event, but a diagnosis of OCD implies that such activities consume at least 1 hour per day and interfere significantly with daily, family, or social functioning (American Psychiatric Association, 2013). To help psychiatrists to measure OCD severity, there are some instruments, such as:

- **Y-BOCS interview:** is the most used instrument to measure OCD severity. It is composed of: a list of 64 OCD symptoms which the subject answers 0 (absence), 1 (past) and 2 (present); a list of the 10 most severe symptoms; and a severity scale based on: time, interference, distress (0 is "none" and 4 is "extreme"), resistance (0 is "always resists" and 4 is "completely yields") and control (0 is "complete control" and 4 is "no control"). This severity scale is named Y-BOCS and ranges from 0 to 40 – it is the sum of these indicators cited above which are applied to obsessions and compulsions separately[1] (see Goodman et al., 1989).
- **DY-BOCS interview:** is an alternative instrument to measure OCD severity. It is based on the Y-BOCS and is composed of: a massive questionnaire about clinical and demographic variables; a list of 88 OCD symptoms, that are divided into six different OCD symptom dimensions:

  1. obsessions about harm due to aggression/injury/violence/natural disasters and related compulsions;
  2. obsessions concerning sexual/moral/religious obsessions and related compulsions;
  3. obsessions about symmetry/'just-right' perceptions, and compulsions to count or order/arrange;
  4. contamination obsessions and cleaning compulsions;
  5. obsessions and compulsions related to hoarding and
  6. miscellaneous obsessions and compulsions that relate to somatic concerns and superstitions, among other symptoms[2],

  which the subject also scores 0, 1 or 2; a list of 3 target symptoms (the 3 most severe symptoms); a severity scale for each of the six dimensions (ranging from 0 to 15) that takes into account time, distress and interference (these marginal OCD severity scores are named DY-BOCS)[3]; and a global OCD severity scale (ranging from 0 to 30) as well as the severity score Y-BOCS (for all details, see Rosario-Campos et al., 2006).

According to Matsunaga and Seedat (2007), the cross-cultural studies show that the OCD symptoms are not similar in different population and cultures, supporting the idea that biological and genetic factors can contribute to its etiology. In this sense, current research to explore genetic susceptibility factors in OCD has resulted in tentative to identification of a small number of genes. It is now broadly accepted that a major limitation to this work is the heterogeneous nature of this disorder, and that an approach taking into account different symptom dimensions may be more productive and more successful in identifying both common (shared) as well as dimension-specific vulnerability genes. However, as most existing genetic datasets did not collect dimensions severity ratings, some method to reliably extract dimensions ratings from the Y-BOCS is needed. Thus, this work aims to develop a statistical procedure to extract dimensional symptom severity ratings from the existing Y-BOCS data for use in OCD genetics and other neurobiological research.

The "OCD Spectrum Disorders Program" (PROTOC), located at the Department and Institute of Psychiatry, University of Sao Paulo, hosts a dataset built by the Brazilian OCD Research Consortium (Miguel et

[1] Usually, "Y-BOCS" is referred in the literature as a OCD severity scale (0 to 40). Through all this work, when we say "Y-BOCS", it means the Y-BOCS interview. For the OCD scale, we say "severity score Y-BOCS".

[2] The 1-5 dimensions are considered homogeneous and the 6 is considered heterogeneous.

[3] Through all this work, when we say "DY-BOCS", it means the DY-BOCS interview. For the dimensional OCD scale (0 to 15), we say "severity score DY-BOCS".

al., 2008) containing DY-BOCS information from 1001 patients. This is a collaborative work involving the University of Sao Paulo and the University of Toronto (which has a sample of 500 subjects with Y-BOCS data and DNA samples).

2. Methodology

The Y-BOCS is much more general than the DY-BOCS in the characterization of symptoms. For example, the Y-BOCS symptom #64 is "I have mental rituals (other than checking/counting)", while in the DY-BOCS we have five symptoms related with mental rituals: "I have mental rituals, other than checking, specifically related to" sexual or religious obsessions (#30); obsessions of symmetry, exactness, or just right perceptions (#41); contamination worries (#53); hoarding obsessions (#60); and somatic worries (#64). There are more examples, but, for most symptoms in the Y-BOCS, there is only one corresponding DY-BOCS symptom.

Note that, in the DY-BOCS data, we have, directly, the information about the severity of each one of the five dimensions, whereas in the Y-BOCS data we do not have such information directly. So, if we have the data from both instruments entered in the same format, we could extract OCD dimensions severity information from the Y-BOCS data like we do from the DY-BOCS data (via multinomial regression, see Subsection 2.2).

2.1 Rewriting the DY-BOCS data

It is not possible to say if a patient scores "2" at the Y-BOCS symptom #64, then they score "2" at #30 or #41 etc. But, it is possible to say if a patient scores "2" at DY-BOCS symptom #30, then they score "2" at Y-BOCS symptom #64. This is summarized in the table bellow:

Table 1: Rewriting the DY-BOCS data toward Y-BOCS data.

DY-BOCS #30	**DY-BOCS #41**		**Y-BOCS #64**
0	0	→	0
0	1	→	1
1	0	→	1
1	1	→	1
0	2	→	2
2	0	→	2
1	2	→	2
2	1	→	2
2	2	→	2

Doing that, we will have the two data sets written in the same format. Therefore, the information about the OCD, on each one of the five dimensions, is present in the rewritten DY-BOCS data set as much as in the Y-BOCS data set. Thus, the idea is to use each symptom as a covariate (as well as, the target symptoms and the severity score Y-BOCS) in a model with the severity scores DY-BOCS as response on each dimension. In other words, now we have in the DY-BOCS data set: a list of symptoms, a list of target symptoms, the severity score Y-BOCS and the severity scores DY-BOCS of each dimension; in the Y-BOCS data set we have the same list of symptoms, a list of target symptoms and the severity score Y-BOCS, but we do not have a score telling us how much severe is the OCD on each dimension, like we have in DY-BOCS.

2.2 Multinomial Logistic Regression

Probably, there is a dependence relationship among the severity scores DY-BOCS of each dimension. For each patient we have a random vector with the severity scores DY-BOCS, say, (Y_1,Y_2,Y_3,Y_4,Y_5), where Y_i is the severity score DY-BOCS for the ith dimension ($i = 1, ..., 5$). In this work, we do not consider the dependence among them. Through this subsection, we fix $i = 1$, but it is analogous for $i = 2, ..., 5$.

The logistic regression (Hosmer & Lemeshow, 1989) is the most commonly used statistical model for predicting binary responses. Generalizations of it model categorical responses with more than two categories (Agresti, 2013). So, we are considering that Y_i is multinomially distributed. As Y_i ranges from 0 to 15, it is like we have the categories "0", "1", ..., "15". There are two kinds of multinomial logistic regressions: with nominal response and with ordinal response. We trait Y_i as nominal, because Y_i is self-reported by the patient (intermediated by a psychiatrist), so we think there is a lot of "noise" in the response, i.e., maybe the patient who scores $Y_i =$"10" has the OCD more severe than a patient who scores $Y_i =$"9". Y_i is subjective.

Thus, we have sixteen categories for Y_i. In this case, when we fit a multinomial logistic model, with nominal response, we are interested in the vector of the probabilities of each possible value to the response, say,

$(\pi_0, \pi_1, \pi_2, ..., \pi_{15})$. After we fit the model, we estimate this vector, obtaining, say, $(\hat{\pi}_0, \hat{\pi}_1, \hat{\pi}_2, ..., \hat{\pi}_{15})$. We fix as baseline the category "0". So, the multinomial logit model (nominal) is

$$\log\left(\frac{\pi_j}{\pi_0}\right) = \alpha_j + \boldsymbol{\beta_j'}\boldsymbol{x}, j = 1, 2, ..., 15,$$

where α_j and $\boldsymbol{\beta_j'}$ are the parameters of the regression and $\boldsymbol{x}$ are the covariates.

There are symptoms belong on dimension 1, on dimension 2 etc., and symptoms on the dimension "miscellaneous". The idea is to get homogeneous a heterogeneous illness. Thus, we have features (symptoms) that characterize OCD only on a specific dimension. Then, for each dimension, we consider as covariates: 1. the symptoms belong on the respective dimension; 2. the symptoms belong on "miscellaneous"; 3. the target symptom dimensions and 4. the severity score Y-BOCS (considering time, distress and suffering, ranging from 0 to 24).

Therefore, fitting all models with the covariates above, for all dimensions, we can estimate π_j by

$$\hat{\pi}_j = \frac{\exp(\alpha_j + \boldsymbol{\beta_j'}\boldsymbol{x})}{1 + \sum_{h=1}^{15}\exp(\alpha_h + \boldsymbol{\beta_h'}\boldsymbol{x})}, j = 1, 2, ..., 15. \tag{1}$$

For a specific subject, and considering all dimensions, we have the matrix

$$\boldsymbol{\hat{\pi}} = \begin{bmatrix} \hat{\pi}_{1,0} & \hat{\pi}_{1,1} & \cdots & \hat{\pi}_{1,15} \\ \hat{\pi}_{2,0} & \hat{\pi}_{2,1} & \cdots & \hat{\pi}_{2,15} \\ \dots & \dots & \ddots & \dots \\ \hat{\pi}_{5,0} & \hat{\pi}_{5,1} & \cdots & \hat{\pi}_{5,15} \end{bmatrix}_{(5\times16)},$$

where now we are denoting by $\pi_{k,j}, (k = 1, 2, ..., 5)$ e $(j = 0, 1, ..., 15)$, the estimated probability of a subject to have severity score DY-BOCS equal to j at the dimension k (Equation 1). Note that

$$\hat{\pi}_{k,0} = 1 - \sum_{j=1}^{15}\hat{\pi}_{k,j}, k = 1, 2, 3, 4, 5.$$

3. Results

As our objective is to predict the phenotype (the most severe dimension)[4], for each line of the matrix $\boldsymbol{\hat{\pi}}$ above, we identify the highest probability and take, as the predicted severity score DY-BOCS, the respective column $(j = 0, 1, ..., 15)$.

It is more common to fit the model using a part of the data set and valid it using the remaining observations (called training set). It is called "leave-p-out", where p is the size of the data set used to fit the model. In this work we use the "leave-one-out" cross validation. It is a similar idea, but as we have 917 subjects (excluding missing observations), for each dimension, we fit 917 models leaving one subject out, and we use these fitted models to predict the response to this "left-out" subject.

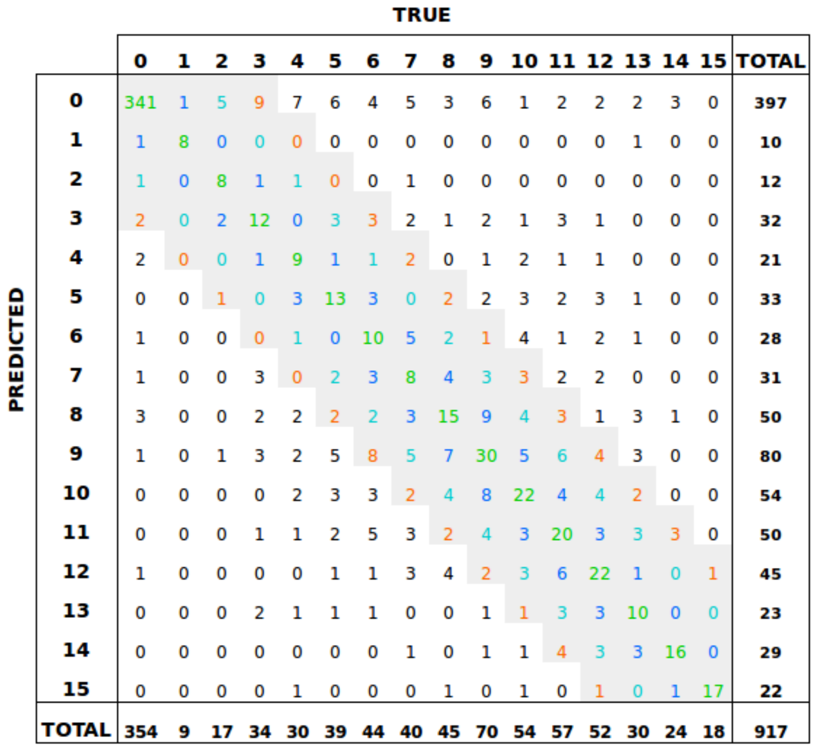
Note that the 917 fitted models are different and it can be seen as a defect of this procedure[5]. But use the usual 70% of the data set to fit it and the 30% remaining to valid it, one can ask: who does guarantee that this subsample used to valid the model is not a lucky subsample? In other words, who does guarantee that if we choose other subsamples, the fitted model is still good/bad? On the other hand, the leave-one-out use more observations and the validation is less susceptible to luck of chose subsamples.

For the dimension 1, the results of the leave-one-out cross validation is summarized at the table below:

[4] We think it is possible to consider more than one dimension being a phenotype. For example, if a patient has the severity scores DY-BOCS's vector equals to $(15, 15, 0, 0, 0)$, then aggressive obsessions/compulsions is as severe as religious/sexual obsessions/compulsions, implying that both could be consider as phenotypes; or the phenotype could be consider as aggressive and religious/sexual obsessions/compulsions together.

[5] The final model is considered that one which is fitted using all observations.

Table 2: Results of the leave-one-out cross validation for dimension 1.



		TRUE																
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**TOTAL**
PREDICTED	**0**	341	1	5	9	7	6	4	5	3	6	1	2	2	2	3	0	**397**
	1	1	8	0	0	0	0	0	0	0	0	0	0	0	1	0	0	**10**
	2	1	0	8	1	1	0	0	1	0	0	0	0	0	0	0	0	**12**
	3	2	0	2	12	0	3	3	2	1	2	1	3	1	0	0	0	**32**
	4	2	0	0	1	9	1	1	2	0	1	2	1	1	0	0	0	**21**
	5	0	0	1	0	3	13	3	0	2	2	3	2	3	1	0	0	**33**
	6	1	0	0	0	1	0	10	5	2	1	4	1	2	1	0	0	**28**
	7	1	0	0	3	0	2	3	8	4	3	3	2	2	0	0	0	**31**
	8	3	0	0	2	2	2	2	3	15	9	4	3	1	3	1	0	**50**
	9	1	0	1	3	2	5	8	5	7	30	5	6	4	3	0	0	**80**
	10	0	0	0	0	2	3	3	2	4	8	22	4	4	2	0	0	**54**
	11	0	0	0	1	1	2	5	3	2	4	3	20	3	3	3	0	**50**
	12	1	0	0	0	0	1	1	3	4	2	3	6	22	1	0	1	**45**
	13	0	0	0	2	1	1	1	0	0	1	1	3	3	10	0	0	**23**
	14	0	0	0	0	0	0	0	1	0	1	1	4	3	3	16	0	**29**
	15	0	0	0	0	1	0	0	0	1	0	1	0	1	0	1	17	**22**
	TOTAL	**354**	**9**	**17**	**34**	**30**	**39**	**44**	**40**	**45**	**70**	**54**	**57**	**52**	**30**	**24**	**18**	**917**

The table below shows us the goodness of fit for all dimensions. Let Y_k be the observed severity score DY-BOCS for the subject k on a specific dimension. Now, let $\hat{Y}_k$ be the respective predicted value. If the model predicts, for example, $\hat{Y}_k = 10$ but the observed value is $Y_k = 9$, is it a good prediction? And if $\hat{Y}_k = 11$ and $Y_k = 9$? In the Table 3, we show the cumulated "deviations" $|Y_k - \hat{Y}_k| \leq 0, 1, 2, 3$ and 4.

Table 3: Goodness of fit for all models (all dimensions).

$\lvert Y_k - \hat{Y}_k\rvert \leq$	0	1	2	3	4
Dimension 1	61%	70%	76%	**82%**	87%
Dimension 2	67%	74%	78%	**83%**	87%
Dimension 3	44%	59%	71%	**80%**	86%
Dimension 4	58%	69%	77%	**86%**	90%
Dimension 5	71%	76%	83%	**88%**	92%

As we have 917 patients in the DY-BOCS data set, one can ask: for $r = 0, 1, ..., 15$, what is the $q_r \in [0, 1]$ such that the model is considered good if

$$\frac{\#\{k; |Y_k - \hat{Y}_k| \leq r\}}{917} \geq q_r?$$

If we consider the models are "good" when $r = 3$ and $q_r = 0.8$, for all dimensions, then our solution is considered "good" (see the highlighted column at the Table 3).

5. Conclusions

There are other manners to solve the multinomial regression problems. We can look at that as several binary logistic models (see Mello and Pereira, 2009). We also can look at that in a Bayesian perspective (some considerations can be found in Agresti and Hitchcock, 2005). It is important to say that we did not use a technique to select covariates because all are important by a clinic point of view, but it is possible chose different covariates for different j's at the Equation 1.

This work built a technique to extract the information about OCD dimensions aiming to allow the advances in genetic and neurobiological research in OCD. The idea, besides allow these advances in this area, is to solve this using several binary logistic models with a Bayesian perspective, letting the clinic experience builds the priors for the regression parameters (or for the vector of probabilities).
We are also interested in try other kinds of validations. As the objective is to predict the phenotype, maybe it makes sense to look how much good the model is only for the extremes cases. An idea is to recategorize the response as "none", "mild", "moderate", "severe" and "extreme", and consider as phenotype the extremes cases.

References

[1] Shavitt, R. G., et al. (2014). "Phenomenology of OCD: Lessons from a large multicenter study and implications for ICD-11." Journal of psychiatric research 57: 141-148.

[2] Ruscio, A. M., et al. (2010). "The epidemiology of obsessive-compulsive disorder in the National Comorbidity Survey Replication." Molecular psychiatry 15.1: 53-63.

[3] American Psychiatric Association (2013). "Diagnostic and statistical manual of mental disorders, (DSM-5)".

[4] Goodman, W. K., et al. (1989). "The Yale-Brown obsessive compulsive scale: I. Development, use, and reliability." Archives of general psychiatry 46.11: 1006-1011.

[5] Rosario-Campos, M. C., et al. (2006). "The Dimensional Yale-Brown Obsessive-Compulsive Scale (DY-BOCS): an instrument for assessing obsessive-compulsive symptom dimensions." Molecular psychiatry 11.5: 495-504.

[6] Matsunaga, H., and Seedat, S. (2007). "Obsessive-compulsive spectrum disorders: cross-national and ethnic issues." CNS spectrums 12.05: 392-400.

[7] Miguel, E. C., et al. (2008). "The Brazilian Research Consortium on Obsessive-Compulsive Spectrum Disorders: recruitment, assessment instruments, methods for the development of multicenter collaborative studies and preliminary results." Revista Brasileira de Psiquiatria 30.3: 185-196.

[8] Hosmer, D. W., and Lemeshow, S. (1989). "Applied logistic regression." New York: Johns Wiley & Sons.

[9] Agresti, A. (2013). "Categorical data analysis." John Wiley & Sons.

[10] Mello, J. and Pereira, C. A. B. (2009). "Model of credit loss and its use in spread decisions." Conference paper: Edimburg Conference on Credit Scoring.
URL: `http://www.business-school.ed.ac.uk/crc/conferences/conference-archive?a=45902`

[11] Agresti, A., & Hitchcock, D. B. (2005). "Bayesian inference for categorical data analysis. Statistical Methods and Applications". 14(3), 297-330.