Available online at www.sciencedirect.com

ScienceDirect

Procedia Social and Behavioral Sciences 1 (2009) 2589–2594

Procedia
Social and Behavioral Sciences

WCES 2009

Discovering Computer-Assisted Learning Processes based on Objective Exam Score Transformations

Patrick Wessa*

K.U.Leuven Association, Lessius Dept. of Business Studies, Belgium

Received October 25, 2008; revised December 24, 2008; accepted January 05, 2009

Abstract

Based on the implementation of a newly developed e-learning technology which features Reproducible Computing (www.wessa.net and www.freestatistics.org), it is investigated how the relationships between learning outcomes (as measured by exam questions) and objectively measured, computer-assisted learning activities can be investigated. The main contribution of this paper is that objective exam score transformations can have a huge effect on our ability to build models and understand the effects of computer technology on learning outcomes. The paper includes a theoretical description of an easy-to-use methodology and an illustrated case in which the effect of exam score transformations is clearly demonstrated.

Keywords: reproducible computing; computer-assisted learning; exams; statistical models; technology; data mining .

1. Introduction

The inability of scientists and students to reproduce empirical research results – as published in papers (or other documents) – has received a great deal of attention within the academic community. Some of the most prominent arguments and observations are described in [1], [2], [3], [4], [5], [6], [7]. While several solutions were developed ([5], [7], [8]), none have been actually implemented in education because of technical and practical reasons [9]. In an effort to overcome these issues, a new Compendium Platform [10] was developed which allows us to create constructivist learning environments ([11], [12], [13], [14]) that effectively support students in non-rote learning of statistical concepts [15]. This solution is based on the so-called R Framework ([16], [17], based on the R language [18]) which supports reproducible computing and allows scientists/educators to monitor actual learning processes [9] and control the quality of the e-learning environment which extends to the actual statistical software [19].

E-mail address: patrick@wessa.net

1877-0428/$–see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.sbspro.2009.01.457

The contribution of this paper consists of a simple methodology that allows educational researchers to assess and estimate the effectiveness of computer-assisted learning activities (based on the Compendium Platform) in terms of the actual learning outcomes as measured by an objective multiple-choice exam. It is mathematically shown that objective exam score transformations have the potential to improve our research models. In addition, it is demonstrated that the predictive improvement that is induced by such transformations can be easily tested (by the use of a simple t-test). Finally, an empirical illustration of the usefulness of the proposed approach is provided based on measurements of an undergraduate statistics course where the Compendium Platform was implemented.

2. Methodological Approach

First, a classical regression approach is used to model exam scores as a linear function of $(K-1) \in \mathbb{N}_0$ exogenous variables of interest. Let $\vec{y}$ represent an $N \times 1$ vector for all $N \in \mathbb{N}$ students (with $N > K$), containing the weighted sum of G item scores (c.q. scores on individual exam questions): $\vec{y} \equiv \sum_{j=1}^{G} \omega_j \vec{y}_j$ with $\sum_{j=1}^{G} \omega_j \equiv 1$. In addition, define an $N \times K$ matrix X that represents all exogenous variables (including a one-valued column which represents the constant), and a $K \times 1$ parameter vector $\vec{b}$ that represents the weights of the linear combination of all columns in X that is used to describe $\vec{y}$. The complete model becomes

$$\vec{y} = X\vec{b} + \vec{e} \quad \textbf{(1)}$$

where $\vec{e} \leftarrow \text{iid } \mathrm{N}(\vec{0}, \sigma_e^2)$ represents the prediction error. The model parameters $\vec{b}$ can be estimated by Ordinary Least Squares (OLS) through the following estimator $\hat{\vec{b}} = (X'X)^{-1} X'\vec{y}$. The prediction of the first model is simply $\hat{\vec{y}} = X\hat{\vec{b}} = X(X'X)^{-1} X'\vec{y}$ which can be interpreted as the ``best" approximation of $\vec{y}$ given a linear combination of the columns in X, under the traditional OLS assumptions (Gauss-Markov theorem).

In the second model, the prediction of the first model is specified by a linear combination of the individual items (questions) that made up the total exam score. Let Y represent the $N \times G$ matrix that contains all G item scores, then it is possible to define the model

$$\hat{\vec{y}} = Y\vec{c} + \vec{a} \quad \textbf{(2)}$$

where $\vec{a} \leftarrow \text{iid } \mathrm{N}(\vec{0}, \sigma_a^2)$. Note that there is no constant term in this model. The OLS estimates for $\vec{c}$ are $\hat{\vec{c}} = (Y'Y)^{-1} Y'\hat{\vec{y}} = (Y'Y)^{-1} Y'X(X'X)^{-1} X'\vec{y}$, and the prediction for $\hat{\vec{y}}$ is obviously $\hat{\hat{\vec{y}}} = Y\hat{\vec{c}} = Y(Y'Y)^{-1} Y'\hat{\vec{y}} = Y(Y'Y)^{-1} Y'X(X'X)^{-1} X'\vec{y}$. The estimated parameters $\hat{\vec{c}}$ can be interpreted as the ``optimal" weights that could be applied to all examination items in order to obtain an alternative total exam score that approximates $\vec{y}$ and which is known to be predictable by X.

The third model simply combines model 1 and 2 by relating $\hat{\hat{\vec{y}}}$ to X in the regression model

$$\hat{\hat{\vec{y}}} = X\vec{f} + \vec{u} \quad \textbf{(3)}$$

where $\vec{u} \leftarrow \text{iid } \mathrm{N}(\vec{0}, \sigma_u^2)$.

The estimator for $\vec{f}$ becomes $\hat{\vec{f}} = (X'X)^{-1} X'\hat{\hat{\vec{y}}} = (X'X)^{-1} X'Y(Y'Y)^{-1} Y'X(X'X)^{-1} X'\vec{y}$. The third model is likely to yield different results from model 1 unless the estimated parameters (of model 2) are (nearly) equal to the original weights $\hat{\vec{c}} = (\hat{c}_1, \hat{c}_2, \hat{c}_3, ..., \hat{c}_G)' \simeq (\hat{\omega}_1, \hat{\omega}_2, \hat{\omega}_3, ..., \hat{\omega}_G)'$. If the weights are different then the third model can be extended by including $\vec{y}$ as an explanatory variable. This leads to an extended model $\hat{\hat{\vec{y}}} = X\vec{f} + \vec{y}g + \vec{u}$ which can be rewritten by substituting the estimated elements:

$$\begin{aligned}
&Y\hat{\vec{c}} = X\vec{f} + \vec{y}g + \vec{u} \\
&\tfrac{1}{g}Y\hat{\vec{c}} - \vec{y} = \tfrac{1}{g}X\vec{f} + \tfrac{1}{g}\vec{u} \\
&\tfrac{1}{g}Y(Y'Y)^{-1} Y'X(X'X)^{-1} X'\vec{y} - \vec{y} = \tfrac{1}{g}X\vec{f} + \tfrac{1}{g}\vec{u} \\
&\left(Y(Y'Y)^{-1} Y'X(X'X)^{-1} X' - gI_N\right)\vec{y} = X\vec{f} + \vec{u}
\end{aligned}$$

From this it can be concluded that the extended version of the third model is equal to the first model with a transformed endogenous variable. A special case arises in the limit when $g \to 0$ and $Y(Y'Y)^{-1} Y'X(X'X)^{-1} X' \to I_N$ because it leads to the first model with $\vec{f} = \vec{b}$ and $\vec{u} = \vec{e}$.

This can be easily shown by considering that in the extended model

$$\hat{\hat{\vec{y}}} = X\vec{f} + \vec{y}g + \vec{u} \qquad \textbf{(4)}$$

the estimator of any individual parameter $f_i \in \vec{f}$ (where $i = 1, 2, ..., K$) can be easily obtained as follows:
$\hat{f}_i = (\vec{x}_i'\vec{x}_i)^{-1}\vec{x}_i'\hat{\hat{\vec{y}}}$
$+ (\vec{x}_i'\vec{x}_i)^{-1}\vec{x}_i'\vec{u}$
$+ \sum_{j=1}^{i-1}(\vec{x}_i'\vec{x}_i)^{-1}\vec{x}_i'\vec{x}_j\hat{f}_j$
$+ \sum_{j=i+1}^{K}(\vec{x}_i'\vec{x}_i)^{-1}\vec{x}_i'\vec{x}_j\hat{f}_j$
$+ (\vec{x}_i'\vec{x}_i)^{-1}\vec{x}_i'\vec{y}\hat{g}$.

If we substitute $\hat{\hat{\vec{y}}}$ by $Y\hat{\vec{c}} = \sum_{j=1}^{G}\vec{y}_j\hat{c}_j$ and $\vec{y}$ by $\sum_{j=1}^{G}\omega_j\vec{y}_j$ then we obtain
$\hat{f}_i = (\vec{x}_i'\vec{x}_i)^{-1}\vec{x}_i'\sum_{j=1}^{G}\vec{y}_j(\hat{c}_j + \omega_j\hat{g})$
$+ (\vec{x}_i'\vec{x}_i)^{-1}\vec{x}_i'\vec{u} + \sum_{j=1}^{i-1}(\vec{x}_i'\vec{x}_i)^{-1}\vec{x}_i'\vec{x}_j\hat{f}_j$
$+ \sum_{j=i+1}^{K}(\vec{x}_i'\vec{x}_i)^{-1}\vec{x}_i'\vec{x}_j\hat{f}_j$
from which it is obvious that $\hat{f}_i \simeq \hat{b}_i$ when $\hat{c}_j \simeq \omega_j$ and $\hat{g} \simeq 0$. If $\hat{c}_j \simeq \omega_j$ then $Y\hat{\vec{c}} \simeq \vec{y}$ from which it follows that $Y(Y'Y)^{-1}Y'X(X'X)^{-1}X' \simeq I_N$.

This result is important because it is now easy to test if it is necessary to apply the transformation to the endogenous variable. The null hypothesis is simply $H_0 : g = 0$ versus $H_1 : g \neq 0$ which can be tested with the conventional t-test. In other words, if the null hypothesis is rejected then the transformation is necessary and the estimated parameters $\hat{\vec{c}}$ and $\hat{\vec{f}}$ are interpretable.

3. Data

The Compendium Platform was implemented in an undergraduate statistics course at the Lessius Business School in Belgium. The underlying technology allowed students to engage in a series of reflective review activities and at the same time it allowed us to accurately measure the actual computer-assisted learning activities. The main sections of the statistics course were built around a series of research-based workshops that require students to reflect and communicate about a variety of statistical problems, at various levels of difficulty. The workshops have been carefully designed and cannot be solved without additional information that is provided within the Virtual Learning Environment or by the tutor. Each workshop involved about 9 hours of work per student, per week. In addition, students were required to perform detailed peer reviews of about 5-7 submissions from other students which was facilitated by the Compendium Platform that allows anyone to reproduce and reuse statistical computations without the need to download or install anything on the client machine. This feedback-oriented process is similar to the peer review procedure of an article that is submitted to a scientific journal. The process of (anonymous) assessment by peers is an intrinsic part of scientific endeavor, and may help students in nurturing their scientific attitudes (through peer review experiences) and non-rote learning (through construction of knowledge).

A group of 240 undergraduate business students participated in the course and completed a total of 1907 workshops which were subjected to peer review. Every submission was assessed with respect to 3-6 criteria. For every graded criterion students had the ability to provide verbal feedback to the other student. As a consequence, a total of 41960 grades and 34438 verbal feedback communications were received by students. Based on the Compendium Technology, a large number of objective measurements were collected about the following computer-assisted learning activities (for each student): Gcount (the average number of students that collaborated during the workshop assignments), Abcount (average number of computations in the collaborating group), Abuniques (average number of unique/distinct computations in the collaborating group), nnzfg (number of meaningful feedback messages that were submitted by the student), nnzfr (number of meaningful feedback messages that were received by the student), mflg (median feedback length of submitted messages), mflr (median feedback length of received messages), Bcount (number of computations), aflg (average feedback length of submitted messages), aflr (average feedback length of received messages), iqrflg (inter-quartile range of submitted feedback messages), iqrflr (inter-quartile range of received feedback messages), WSMedian (median workshop score based on peer assessment).

4. Empirical evidence

It is now possible to illustrate the approach that was outlined in section 2 based on the data of section 3. This illustration relates to a subset of the original student population: only female students from the Preparatory programme were selected. The reason for this is because these students are a (more or less) homogeneous group [19] where self-reported data about learning activities are consistent with actual, objective measurements that were made based on the newly developed e-learning technology [19].

The objective exam consisted of 18 multiple choice questions which have been summed to obtain the overall score. This implies that $\omega_j = 1/G$ for $j = 1, 2, ..., G$ and consequently $\vec{y} \equiv 1/G \sum_{j=1}^{G} \vec{y}_j$. The parameter estimates of Model 2 (not shown) indicate that questions 5, 6, 7, 9, 13, and 14 seemed to be more important (c.q. more closely related to the exogenous variables) than the other questions. For example, question 6 obtains a weight that is more than twice as big as in the original exam score. The Adjusted R-squared of Model 2 was extremely high (0.94) which means that the predicted outcome $\hat{\hat{\vec{y}}}$ is closely related to the original prediction $\hat{\vec{y}}$. Moreover, some questions had near-zero or negative parameters in Model 2. This implies that these questions reduce the explanatory power of Model 1 and should not be included if one wishes to obtain high predictability of exam scores based on exogenous variables that are related to the use of computer-related learning activities.

Model 1 relates the overall exam score $\vec{y}$ to a set of pre-specified exogenous variables of interest (see section 3) and yields results that are presented in Table 1 (columns 2 and 3). The Adjusted R-squared of Model 1 is rather low and no parameter is significantly different from zero. This is disappointing and the researcher cannot learn anything useful from this analysis. There is reason to assume that Model 1 suffers from multicollinearity and extraneous variables which inflate the standard errors of the parameters. On the other hand it is unclear how the model's parameters should be reduced in order to obtain a parsimonious model with good predictability.

In the approach that was outlined in section 2 it was suggested that an objective transformation of the endogenous variable (the exam scores) might be helpful: the results of the transformed and extended model are shown in Table 1 (columns 4 and 5).

Table 1. OLS results (models 1 and 4)

Coefficients	Estimate M1	P-val M1	Estimate M4	P-val M4
(Intercept)	-8.57E+000	0.5	-7.28E+000	0.1
Gcount	-1.78E-001	0.54	**-2.15E-001**	**0.04**
Abcount	5.17E-002	0.65	2.56E-002	0.52
Abuniques	2.49E-001	0.45	7.36E-002	0.52
nnzfg	4.34E-002	0.4	**3.07E-002**	**0.09**
nnzfr	2.25E-002	0.65	**4.06E-002**	**0.02**
mflg	9.30E-002	0.43	4.66E-002	0.26
mflr	3.86E-002	0.79	7.00E-002	0.17
Bcount	5.07E-002	0.7	2.75E-002	0.55
aflg	-3.31E-002	0.7	-3.26E-002	0.28
aflr	-3.94E-002	0.71	**-6.13E-002**	**0.1**
iqrflg	1.79E-002	0.78	-2.08E-003	0.92
iqrflr	6.25E-002	0.54	3.06E-002	0.39
WSMedian	2.93E-003	0.95	**3.13E-002**	**0.04**
Abcount:Abuniques	-2.07E-003	0.53	-8.87E-004	0.44
nnzfg:nnzfr	-1.78E-004	0.51	-1.51E-004	0.11
mflg:mflr	-1.14E-003	0.57	-6.57E-004	0.34
mflg:Bcount	-8.38E-004	0.56	1.00E-004	0.84
mflr:Bcount	7.07E-005	0.98	-1.79E-004	0.83
aflg:aflr	5.03E-004	0.48	**4.57E-004**	**0.07**
iqrflg:iqrflr	-2.90E-004	0.64	-1.42E-004	0.51
mflg:mflr:Bcount	5.20E-006	0.82	-2.98E-006	0.7
TX18	NA	NA	**4.10E-001**	**2.38E-007**

Adjusted R-squared M1: 0.1621 p-value: 0.1575

Adjusted R-squared M4: 0.8232 p-value: 1.619e-09

5. Discussion and Conclusions

The illustrated results are spectacular from a statistical point of view. The objective exam score transformation yields beneficial results in terms of the adjusted R-squared and the significance of parameters. From Table 1 it is clearly seen that the exam scores of female students of the preparatory programme can be predicted by feedback-related communications (inbound and outbound) which is assisted by the newly developed Reproducible Computing technology. In addition, the group size (the number of collaborating students that work on assignments) should be kept small and the role of prior knowledge (as approximated by the median workshop scores) is not to be underestimated. These conclusions can be easily drawn - without the need of sophisticated model specification or selection. The main point here is that the introduction of new e-learning technology alone is not sufficient to model the relationship between learning outcomes and computer-assisted learning activities. The reason for this is obvious: the e-learning technology ensures accuracy of the measured variables on the right hand side of the equation whereas the endogenous variable (exam score) may still be biased because of the subjective nature of the weights ω_j that are attributed to the individual exam questions.

A final remark about the above findings is related to an ethical dilemma. This paper does not imply – nor suggest – that educators should post transform the weights that are attributed to individual exam questions for the purpose of grading. Grading mechanisms should always be transparently communicated to students and comply with academic regulations and legislation. This is always true – even if the application of an objective exam score transformation would highlight the fact that some student groups are discriminated: in the data set under investigation several sub-populations were found that had optimal ω_j parameters that were significantly different from those that were found with the female students from the prep-programme. Even if we treat students equally, this does by no means imply that they respond to technological learning activities (or exam questions) equally.

This dilemma must be addressed in future development and research. Maybe it is possible to seamlessly integrate the examination/grading process into an e-learning environment in which the underlying technology addresses the needs of all students while still allowing educators to make fair comparisons?

References

1. J. de Leeuw, “Reproducible research: the bottom line,” in Department of Statistics Papers, 2001031101, Department of Statistics, UCLA, 2001
2. R. D. Peng, F. Dominici, and S. L. Zeger, “Reproducible epidemiologic research,” American Journal of Epidemiology, 2006
3. M. Schwab, N. Karrenbach, and J. Claerbout, “Making scientific computations reproducible,” Computing in Science & Engineering, vol. 2, no. 6, pp. 61–67, 2000
4. P. J. Green, “Diversities of gifts, but the same spirit,” The Statistician, pp. 423–438, 2003
5. R. Gentleman, “Applying reproducible research in scientific discovery,” BioSilico, 2005
6. R. Koenker and A. Zeileis, “Reproducible econometric research (a critical review of the state of the art),” in Research Report Series, no. 60, Department of Statistics and Mathematics Wirtschaftsuniversität Wien, 2007
7. D. L. Donoho and X. Huo, “Beamlab and reproducible research,” International Journal of Wavelets, Multiresolution and Information Processing, 2004
8. F. Leisch, “Sweave and beyond: Computations on text documents,” in Proceedings of the 3rd International Workshop on Distributed Statistical Computing, (Vienna, Austria), 2003
9. P. Wessa, “Learning statistics based on the compendium and reproducible computing,” in Proceedings of the World Congress on Engineering and Computer Science (International Conference on Education and Information Technology), UC Berkeley, San Francisco, USA, 2008
10. P.Wessa and E. van Stee, Statistical Computations Archive (online software at http://www.freestatistics.org). K.U.Leuven Association, Belgium, 2008
11. E. Von Glasersfeld, “Learning as a constructive activity,” in Problems of Representation in the Teaching and Learning of Mathematics, pp. 3–17, Hillsdale, NJ: Lawrence Erlbaum Associates, 1987
12. E. Smith, “Social constructivism, individual constructivism and the role of computers in mathematics education,” Journal of mathematical behavior, vol. 17, no. 4, 1999
13. P. Eggen and D. Kauchak, Educational Psychology: Windows on Classrooms. Upper Saddle River, NJ: Prentice Hall, 5th ed. ed., 2001
14. L. Moreno, C. Gonzalez, I. Castilla, E. Gonzalez, and J. Sigut, “Applying a constructivist and collaborative methodological approach in engineering education,” Computers & Education, vol. 49, pp. 891–915, 2007
15. P. Wessa, “How reproducible research leads to non-rote learning within a socially constructivist e-learning environment,” in Proceedings of the 7th European Conference on e-Learning, (Cyprus), 2008
16. P. Wessa, Free Statistics Software (online software at http://www.wessa.net). Office for Research Development and Education, 1.1.23-r2 ed., 2008
17. P. Wessa, “A framework for statistical software development, maintenance, and publishing within an open-access business model,” Computational Statistics, 2008
18. R Development Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0
19. P. Wessa, “Measurement and control of statistics learning processes based on constructivist feedback and reproducible computing,” in Proceedings of the 3rd International Conference on Virtual Learning, (Constanta, Romania), 2008.