Global Journal of Management and Business Research: G
Interdisciplinary
Volume 15 Issue 4 Version 1.0 Year 2015
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 2249-4588 & Print ISSN: 0975-5853

Student Age as an Impact Factor for Student Evaluations of Instruction

By Katja Specht & Wolfgang Gohout

University of Applied Sciences Mittelhessen, Germany

Abstract- Student Evaluations of Instruction (SEI) are an important issue in countries like the USA, where the evaluation results can impact professional promotion chances and salary of faculty. According to Seldin [11], the percentage of American colleges using SEI grew from 29% in 1973 to 68% in 1983 and to 86% in 1993. Consequently, the adequacy of SEI has been examined extensively, and many statistical studies have been published. Non-instructional factors, which cannot be influenced by instructors, may bias the evaluation rating and should be identified and eliminated for a fair comparison. But in many cases, a mere linear regression of SEI on such potential factors is not adequate.

Keywords: evaluation, extrinsic impacts, generalized linear models, regression, student age.

GJMBR - G Classification : JEL Code : I29

Student Age as an Impact Factor for Student Evaluations of Instruction

Katja Specht [α] & Wolfgang Gohout [σ]

Abstract- Student Evaluations of Instruction (SEI) are an important issue in countries like the USA, where the evaluation results can impact professional promotion chances and salary of faculty. According to Seldin [11], the percentage of American colleges using SEI grew from 29% in 1973 to 68% in 1983 and to 86% in 1993. Consequently, the adequacy of SEI has been examined extensively, and many statistical studies have been published. Non-instructional factors, which cannot be influenced by instructors, may bias the evaluation rating and should be identified and eliminated for a fair comparison. But in many cases, a mere linear regression of SEI on such potential factors is not adequate. This paper proposes a proper approach to such situations, namely Generalized Linear Models (GLM). The estimation algorithm will be presented step-by-step so that it can be replicated with own data. Eventually, the estimated model will be used to eliminate the extrinsic impacts.

Keywords: evaluation, extrinsic impacts, generalized linear models, regression, student age.

I. Introduction

Student Evaluations of Instruction (SEI) are very widespread and common practice in countries, where the evaluation results are applied for professional promotion chances and salary of faculty. In countries like Germany, SEI are used as an instrument for the internal quality management and teaching improvement process. This instrument is also of growing interest in the accreditation process of study programs and universities.

The intrinsic impact factors of the evaluation ratings are the single items of the evaluation questionnaire, which are answered by the students. But many statistical investigations have shown that there are undesirable extrinsic factors, like class size or the quantitative exposition of the course, which are noninstructional by nature and, therefore, should be eliminated for a fair comparison of the evaluation ratings. Costin, Greenough and Menges [2] presented a review of empirical studies regarding student ratings. They concluded that SEI can provide reliable and valid information on the quality of courses and instruction but for further interpretation extrinsic factors should be taken into account. Already Heilman and Armentrout [6], Lovell and Haner [8], McDaniel and Feldhusen [9] and Hamilton [5] have shown that teachers of large classes may receive lower ratings. Hoefer, Yurkiewicz and Byrne [7] assessed significant differences between undergraduate and graduate SEI. For that matter, Brightman [1] states that it is unfair to compare a faculty member teaching a required core class with another faculty member teaching a senior–level elective course. Peterson, Berenson, Misa and Radosevich [10] have recommended to establish appropriate sets of norming reports in which possible semester factor effects are considered.

It is tempting to perform a linear regression of the evaluation ratings on the non–instructional factors by the least–squares principle and to use the estimated model for the compensation procedure. But, this will only be admissible, if the latent variable is normally distributed. This can be tested by using the residuals from the regression as a proxy for the latent variable. Frequently, a dependent variable, like evaluation ratings, is skewed to the right. This, in turn, usually prevents the residuals from being normal. At least, this occurs with our data.

Therefore, our investigation focuses on a proper methodical approach of estimating a non–linear model. After a description of the data we shall present the Maximum–Likelihood (ML) estimation of a so–called Generalized Linear Model (GLM) step–by–step. The presentation is sufficiently detailed, so that the reader can, for instance, apply the procedure to own data with a matrix-based programming software like MATHLAB or GAUSS. We restrict our presentation to one non-instructional factor, namely 'student age' or, more precisely, the semester counter of the evaluated course. The proposed procedure can easily and obviously be extended to more non-instructive factors. Eventually, we shall show how to use the estimated model to correct actual and future evaluation ratings properly.

II. Data

We have collected $n = 140$ evaluation ratings z_i from seven-semester Bachelor programs from the Business Unit of a German University of Applied Sciences together with the semester counter (one to seven), to which the evaluated course regularly belongs. The evaluation ratings are means from a five–point Likert

Author α: Department of Business Administration and Engineering Technische Hochschule Mittelhessen (THM) Wilhelm-Leuschner-Str. 13 D-61169 Friedberg Germany. e-mail: katja.specht@wi.thm.de
Author σ: Department of Business Administration and Engineering Pforzheim University of Applied Sciences Tiefenbronner Str. 65 D75175 Pforzheim Germany. e-mail: wolfgang.gohout@hs-pforzheim.de

scale, where the choice 'one' is best and 'five' is worst. Unfortunately, the evaluation ratings are not normally distributed. More precisely, the standardized measure of skewness is 1.05 and the standardized measure of kurtosis is 4.51, indicating that the dependent variable is skewed to the right with a kurtosis much larger than that of the normal distribution. This results in non–normal residuals from a linear least-squares regression. And this prevents inferential conclusions of such a regression, like *t*–values and *p*–values. The usual methodology is no longer valid in this case.

Luckily, a Box Cox transformation of the evaluation ratings z_i can convert the ratings in (approximately) normally distributed values y_i:

$$y_i := g(z_i) := \frac{z_i^\lambda - 1}{\lambda} \sim N(\mu, \sigma^2).$$

The value of λ, which minimizes the absolute skewness of the transformed variables can be calculated numerically and is about 0.45 for our data. If we apply the rounded value 0.5, then we receive a standardized measure of skew-ness of about 0.03 and a standardized measure of kurtosis of about 2.85. The hypothesis of normality for the transformed variables y_i cannot be rejected by any test. The skew-ness–kurtosis test of D'Agostino, Belanger, and D'Agostino Jr. [3] yields a *p*–value above 90%. The transforming function *g* is called 'link function'.

The normal distribution belongs to the so–called 'exponential family'. This admits the estimation of a GLM, which will be specified in the next section.

III. Methodology and Exemplary Results

a) GLM estimation

The most general form of a regression model explains a variable by the sum of its (conditional) expected value and of some noise:

$$y_i = E(y_i|X) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

X denotes the design matrix. In our example, it consists of a first column of ones, representing the constant, and a second column with the semester counts. Further columns may be appended for additional non–instructional factors. The latent variables ε_i are independent and identically (iid) distributed, representing the noise.

In a GLM, the dependent variable must belong to the exponential family and its expected value, given the design matrix *X*, may be a non–linear function *h* of the linear predictor *X*β:

$$\mu = E(y_i|X) = h(X\beta)$$

In our example, the column vector β consists of two unknown parameters, β_0 and β_1, and *h* is the inverse of the link function *g* and is called 'response function'.

The following ML estimation procedure is explained in more detail in Fahrmeir, Kneib, and Lang [4]. Let x_i be the *i*-th row of the design matrix *X*. Then we need the following symbols:

$$\eta_i := x_i'\beta, \quad \mu_i := h(\eta_i), \quad d_i := \frac{dh(\eta_i)}{d\eta_i}, \quad w_i := \frac{d_i^2}{\sigma^2}$$

$$y := (y_1, \ldots, y_n)', \quad \mu := (\mu_1, \ldots, \mu_n)'$$

Consequently, the following diagonal matrices depend on β:

$$D := \mathrm{diag}(d_1, \ldots, d_n), \quad W := \mathrm{diag}(w_1, \ldots, w_n)$$

The goal is to receive a solution of the non–linear equation system *s*(β) = 0, where *s*(β) is the functional vector of partial derivatives of the log–likelihood function:

$$s(\beta) = \frac{\partial l(\beta|y, X)}{\partial\beta} = X'D(y - \mu)/\sigma^2$$

Now, the ML estimator may be iteratively approximated by the following equations:

$$\hat{\beta}^{(k+1)} = (X'W^{(k)}X)^{-1}X'W^{(k)}\tilde{y}^{(k)}$$

$$\text{with} \quad \tilde{y}^{(k)} := X\hat{\beta}^{(k)} + D^{-1} \cdot (y - h(X\hat{\beta}^{(k)}))$$

We have started the iterations with the least–squares estimator $\hat{\beta}^{(0)} = (X'X)^{-1}X'y$.

In order to estimate σ^2, which depends on β, we first have to eliminate duplicate rows in X. We denote the reduced design matrix by $\widetilde{X}$. Note, that in our example it has just seven rows due to the seven semester counts. The y_i have to be averaged to $\overline{y}_j$, $j = 1, \dots 7$,within the seven groups of identical rows of X. Let n_j denote the number of observations in group j. Then, the variance can be estimated in each step of the iteration:

$$\hat{\sigma}^2(\hat{\beta}^{(k+1)}) = \frac{1}{7-p} \cdot \sum_{j=1}^{7} n_j \cdot \left(\bar{y}_j - h\left(\widetilde{x}_j' \hat{\beta}^{(k+1)}\right)\right)^2$$

Here, p is the number of columns of X, in our example: $p = 2$.

Table 1 shows the five iterations, which are needed for convergence in our example.

Table 1 : Iterations of the Fisher–Scoring algorithm

k	$\hat{\beta}_0^{(k)}$	$\hat{\beta}_1^{(k)}$	$s(\hat{\beta}_0^{(k)})$	$s(\hat{\beta}_1^{(k)})$
1	0.1519	−0.0307	−0.1772	−0.6619
2	−0.2679	−0.0419	−0.1482	−0.5554
3	−0.4696	−0.0502	−0.1323	−0.5017
4	−0.5006	−0.0523	−0.0166	−0.0651
5	−0.5012	−0.0524	0.0000	0.0000

Therefore, we receive the following estimated model:

$$\hat{y} = h(X\hat{\beta}) \qquad \text{with} \qquad \hat{\beta} = (-0.5012,\ -0.0524)'$$

The residuals from this model are clearly normal. Thus, they can be 'studentized' in order to eliminate outliers. In a first step, the residuals $\hat{\varepsilon}_i$ have to be 'standardized':

$$r_i := \frac{\hat{\varepsilon}_i}{\hat{\sigma} \cdot \sqrt{1 - h_{ii}}} \qquad \text{with} \quad h_{ii} = x_i'(X'X)^{-1}x_i$$

In a second step, the standardized residuals will be transformed into a Student distribution:

$$r_i^* := r_i \cdot \sqrt{\frac{n-p-1}{n-p-r_i^2}} \ \sim \ t_{n-p-1}$$

We choose to define an outlier as an observation with an absolute studentized residual above the percentage point of order 0.975. This yields a 5% probability of an error of first kind. In our example we have excluded ten observations leading to $n = 130$ observations, to which the whole procedure is applied again. This final estimation yields:

$$\hat{y} = h(X\hat{\beta}) \qquad \text{with} \qquad \hat{\beta} = (-0.5078,\ -0.0555)'$$

b) Model diagnostics

For model diagnostics, we can test the hypothesis $H_0 : C\beta = c$ by the asymptotically χ_r^2–distributed Wald statistic:

$$w = (C\hat{\beta} - c)'(CF(\hat{\beta})^{-1}C')^{-1}(C\hat{\beta} - c) \ \sim \ \chi_r^2$$

where r is the rank of C and $F(\hat{\beta}) = X'WX$ is the Fisher information matrix. In our example, the Wald statistic for $H_0 : \beta_0 = 0$ amounts to 34.32 with a p-value of almost zero. And the Wald statistic for $H_0 : \beta_1 = 0$ amounts to 6.79 with a p-value of 0.0091. Thus, both coefficients are highly significant.

The ML estimator $\hat{\beta}$ is (approximately) normally distibuted with covariance matrix $F(\hat{\beta})^{-1}$. Then, the transformed variable y may be estimated or predicted like this:

$$\hat{y} = E(y|X) = h(X\hat{\beta})$$

c) Back transformation

Eventually, we have to come back to the original evaluation ratings z_i. To this end, we apply a Taylor series approximation of the response function h, centered at $\mu = h(X\hat{\beta})$:

$$\begin{aligned} h(y) &= (1-y/2)^{-2} \\ \approx \tilde{h}(y) &:= (1-\mu/2)^{-2} + (1-\mu/2)^{-3}(y-\mu) + \\ &\frac{3}{4} \cdot (1-\mu/2)^{-4}(y-\mu)^2 \end{aligned}$$

The Taylor series approximation of the response function enables the conclusion for the evaluation ratings:

$$\begin{aligned} \hat{z} &= E(z|X) = E(h(y)|X) \approx E(\tilde{h}(y)|X) \\ &= (1-\mu/2)^{-2} + \frac{3}{4} \cdot (1-\mu/2)^{-4} \cdot V(y|X) \\ &= (1-h(X\hat{\beta})/2)^{-2} + \frac{3}{4} \cdot (1-h(X\hat{\beta})/2)^{-4} \cdot \hat{\sigma}^2(\hat{\beta}) \end{aligned}$$

Because the expectation values of odd powers in the Taylor series are zero, the approximation error (with some $\boldsymbol{\vartheta}$ ϵ [0, 1]) is limited to:

$$\left| \frac{h''''(\vartheta y + (1-\vartheta)\mu)}{4!} \cdot (y-\mu)^4 \right| < 0.0004$$

This may be imagined to be negligible.

Table 2 shows the estimated evaluation ratings in the last column for each group of identical co–variables:

Table 2 : GLM–estimated evaluation ratings

i	$\widetilde{X}_{i,1}$	$\widetilde{X}_{i,2}$	$\hat{y}_i$	$E(z_i\|X)$
1	1	1	0.6134	2.1661
2	1	2	0.5890	2.0853
3	1	3	0.5660	2.0136
4	1	4	0.5443	1.9495
5	1	5	0.5239	1.8921
6	1	6	0.5046	1.8402
7	1	7	0.4863	1.7933

It is clearly seen that the expected ratings in the last column are falling, and therefore getting better, with raising semester count in the third column. Thus, advanced students tend to be more patient with instructors.

d) Compensation

In the simple linear model $y = X\beta + \varepsilon$, the elimination of the impact of the 'extrinsic' factors in X is realized by the correction of the mean value of the dependent variable by the individual residual $\boldsymbol{\varepsilon_\bullet}$ of an actual or future observation $(y_\bullet, x'_\bullet)$:

$$y_\bullet = x'_\bullet\hat{\beta} + \hat{\varepsilon}_\bullet \quad \Rightarrow \quad y^*_\bullet = \bar{y} + \hat{\varepsilon}_\bullet = \bar{y} + y_\bullet - x'_\bullet\hat{\beta}$$

The analogous procedure in a GLM yields:

$$y_\bullet = h(x_\bullet'\hat{\beta}) + \hat{\varepsilon}_\bullet \;\Rightarrow\; y_\bullet^* = \bar{y} + y_\bullet - h(x_\bullet'\hat{\beta})$$

$$\Rightarrow\; z_\bullet^* = h(y_\bullet^*) = h(\bar{y} + g(z_\bullet) - h(x_\bullet'\hat{\beta}))$$

$$\text{with} \quad \bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} g(z_i)$$

Table 3 illustrates the consequences of these compensations for some randomly chosen ratings.

Table 3 : Some examples of proper corrections of evaluation ratings.

$z_\bullet$	$g(z_\bullet)$	$x_\bullet[2]$	$h(x_\bullet'\hat{\beta})$	$\bar{y}$	$y_\bullet^*$	$z_\bullet^*$
3.7	0.9602	1	0.6088	0.5395	0.8910	3.2521
1.5	0.3670	1	0.6088	0.5395	0.2977	1.3804
2.4	0.7090	2	0.5833	0.5395	0.6652	2.2452
1.4	0.3097	7	0.4768	0.5395	0.3724	1.5099

The arbitrary ratings $z_\bullet$ are corrected into the expected direction and yield the values in the last column. The ratings of early semesters are lowered, thus improved, and ratings of late semesters are raised, thus penalized.

e) *Semester dummies*

Now, we are going to model the impact of the categorical variable 'semester count' by semester dummies. This will drop the assumption of a monotonous influence in favour of more flexibility. We choose the first semester as the reference category. The dummy variables S_i, $i = 2, \ldots, 7$, are defined to be 'one', if the course is affiliated to semester i, and 'zero' otherwise. The related GLM reads:

$$y = h(X\beta) + \varepsilon = h(\beta_0 + \beta_1 \cdot S_2 + \cdots + \beta_6 \cdot S_7) + \varepsilon$$

with the $(n \times 7)$–dimensional design matrix

$$X = (1, S_2, \ldots, S_7)$$

The estimation procedure is the same as before. Six outliers can be identified in this model, leaving behind a sample number of $n = 134$ and the following vector of estimated coefficients:

$$\hat{\beta} = (-0.6484,\; 0.1065,\; 0.0545,\; -0.0216,$$
$$-0.3271,\; -0.3130,\; -0.1657)'$$

Again, the conclusion for the original ratings is performed by a Taylor series approximation of the response function. This yields the following expected evaluation rating values, dependent on the semester count:

Semester	1	2	3	4	5	6	7
$E(z_i\|X)$	2.0171	2.1669	2.0894	1.9907	1.7126	1.7224	1.8403

Evidently, with our data the evaluation ratings are 'raising' in the beginning and in the last three semesters. And they are 'falling back' in the middle part of the study program. But, remember that evaluation ratings are like 'grades' in our example, meaning that a 'high rating' is equivalent to a 'low grade'.

The residuals of this regression are clearly normal. The p–value of the skewness–kurtosis test is about 45%. The simultaneous significance of the dummy variables may be tested by the hypothesis $H_0 : C\beta = c$ with:

$$C = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad c = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

The asymptotically χ^2_6–distributed Wald statistic has a p–value less than 0.00001. Therefore, the simultaneous impact of the dummy variables is clearly significant.

Table 4 demonstrates the way of compensation for the non-instructional factor 'semester count' for seven exemplary evaluation ratings. Observed ratings $z_\bullet$ have to be reduced (i.e. improved) in the first four semesters and else raised (i.e. deteriorated). The corrected rates are listed in the last column.

Table 4 : Some examples of proper corrections of evaluation ratings with semester dummies.

$z_\bullet$	$g(z_\bullet)$	$x_\bullet[2]$	$h(x'_\bullet\hat{\beta})$	$\bar{y}$	$y^*_\bullet$	$z^*_\bullet$
2.7	0.7828	1	0.5703	0.5389	0.7514	2.5659
2.5	0.7351	2	0.6191	0.5389	0.6549	2.2107
2.2	0.6516	3	0.5945	0.5389	0.5960	2.0291
2.1	0.6199	4	0.5611	0.5389	0.5976	2.0339
2.0	0.5858	5	0.4518	0.5389	0.6729	2.2711
1.8	0.5093	6	0.4561	0.5389	0.5921	2.0179
1.5	0.3097	7	0.5051	0.5389	0.4008	1.5640

IV. Conclusions

Evaluation ratings are an important instrument in quality management of teaching. Several non–instructional factors may bias the intended evaluation of the instructor. It is essential to assess the quantitative influence of those non–instructional factors in order to compensate the evaluation ratings for these extrinsic factors and achieve a fair comparison.

It is tempting to perform a linear least–squares regression of the evaluation ratings on the non-instructive factors. The estimated model could easily be used to eliminate the extrinsic impact. But, if the residuals from this regression are not normally distributed, the results will not be reliable. Another method of estimation has to be applied.

At least with our SEI data, the residuals from a linear least–squares regression on student's age are skewed and far from beeing normal. But a proper Box–Cox transformation of the evaluation ratings yields a normally distributed dependent variable. This, in turn, enables the maximum likelihood estimation of a GLM. This procedure is not quite common. Therefore, it is explained in detail in this paper.

Once we have estimated a valid model, we can use it to eliminate the impact of the considered co–variable. Due to the non–linear GLM approach, this task requires a Taylor series approximation of the response function, which can be fairly easily performed. In our example, the expected evaluation ratings are getting better with rising semester count. Students seem to get more indulgent with growing age.

Finally, we have conducted the GLM regression of the transformed evaluation ratings on semester dummy variables. Now we receive more flexible, non–monotonic impacts of the semester count on evaluation ratings. Especially with small data sets, this might be the better approach.

An important message of this paper should be to carefully inspect the assumptions of an applied method. In many cases, these assumptions may not be met by the data. In these cases a less familiar procedure may serve as an alternative.

References Références Referencias

1. H.J. Brightman: Mentoring faculty to improve teaching and student learning“, Decision Sciences Journal of Innovative Education, vol. 3, pp. 191-203 2005.
2. F. Costin, W.T. Greenough, and R.J. Menges: Student Ratings of College Teaching: Reliability, Validity, and Usefulness“, Review of Educational Research, vol. 41, pp. 511-534, 1972.
3. R.B. D'Agostino, A. Belanger, and R.B. D'Agostino Jr.: A suggestion for using powerful and informative tests of normality“, The American Statistician, vol 44(4), pp. 316-321, 1990.
4. L. Fahrmeir, T. Kneib, and S. Lang: ”Regression“, 2nd ed., Heidelberg: Springer, 2009.

5. L.C. Hamilton: Grades, class size, and faculty status predict teaching evaluations", Teaching Sociology, vol. 8, pp. 47-62, 1980.
6. J.D. Heilman, and W.D. Armentrout: The ratings of college teachers on ten traits by their students", Journal of Educational Psychology, vol. 27, pp. 197-216, 1936.
7. P. Hoefer, J. Yurkiewicz, and J.C. Byrne: The Association between Students' Evaluation of Teaching and Grades", Decision Sciences Journal of Innovative Education, vol. 10(3), pp. 447-459, 2012.
8. G.D. Lovell, and C.F. Haner: Forced-choice applied to college faculty rating", Educational and Psychological Measurement, vol. 15, pp. 291-304, 1955.
9. E.D. McDaniel, and J.F. Feldhusen: »Relationship between faculty ratings and indexes of service and scolarship", Annual Convention of the American Psychological Association, vol. 5, pp. 619-620, 1970.
10. R.L. Peterson, M.L Berenson, R.B. Misra, and D.J. Radosevich: An Evaluation of Factors Regarding Students' Assessment of Faculty in a Business School", Decision Sciences Journal of Innovative Education, vol. 6(2), pp. 375-402, 2008.
11. P. Seldin: The »use and abuse of student ratings of professors", The Chronicel of Higher Education, vol. 21, 1993.

This page is intentionally left blank