Optimal Data Analysis
Vol. 5 (October 23, 2016), 146-150

2155-0182/10/$3.00

Novometric Models of Smoking Habits of Male and Female Friends of American College Undergraduates: Gender, Smoking, and Ethnicity

Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

Novometric statistical analyses[1-15] were used to model smoking habits of one's male friends, and of one's female friends, for samples of 3,289 Anglo-American, 944 Mexican-American, and 733 Indian-American college undergraduates.[16] For both analyses the categorical attributes were ethnicity (a multicategorical attribute, dummy-coded using 1-3, respectively), and subject gender (0=female, 1=male) and smoking behavior (0=non-smoker, 1=smoker). The novometric findings are compared with results originally reported for this application obtained using disintegrated chi-square analysis.[13,17]

Data investigated here (Table 1) were originally analyzed using disintegrated chi-square analysis (legacy analyses presently used for this design are the log-linear model or logistic regression analysis[18-32]), and reported vis-à-vis partially documented[33] chi-square analysis.

For males, original analyses reported: "…data for all groups show that smokers tend to associate with smokers, and non-smokers to associate with non-smokers. This trend is somewhat more marked for males, with females tending to associate about equally with smokers and non-smokers (p. 168).

For females, original analyses reported: "…Significant (p<.01) differences were obtained for Anglo-American male and female smokers… Male smokers tend to associate with female smokers, and female smokers with other female smokers (PRY: notice that this is a degenerate result since not all of the class categories are used[1,34]). Non-smoking males and females associate with, and seem to be influenced by, the non-smoking peer-group to which they belong. Mexican-American females show a comparable difference (p<.01). Indian-American smokers, males and females, differed significantly in the frequency of association with female smokers. Females claimed an equal number of smoking and non-smoking female friends. Non-smoking males showed a significantly greater tendency (p<.01) than females to exclude smoking females from their circle of friends" (p. 168).

2155-0182/10/$3.00

Table 1: Study Data[16]

Friend's Gender	Subject Ethnicity	Subject Smokes	Subject Gender	% of Friends Smoke	N
Male	Anglo	Yes	Male	100	242
				50	118
				0	73
	Mexican	Yes	Male	100	98
				50	23
				0	13
	Indian	Yes	Male	100	160
				50	20
				0	35
	Anglo	Yes	Female	100	148
				50	42
				0	21
	Mexican	Yes	Female	100	36
				50	4
				0	1
	Indian	Yes	Female	100	41
				50	8
				0	8
	Anglo	No	Male	100	102
				50	156
				0	899
	Mexican	No	Male	100	61
				50	45
				0	178
	Indian	No	Male	100	30
				50	16
				0	78
	Anglo	No	Female	100	317
				50	269
				0	902
	Mexican	No	Female	100	228
				50	49
				0	208
	Indian	No	Female	100	149
				50	38
				0	150
Female	Anglo	Yes	Male	100	46
				50	60
				0	316
	Mexican	Yes	Male	100	10
				50	20
				0	103
	Indian	Yes	Male	100	18
				50	14
				0	177
	Anglo	Yes	Female	100	70
				50	64
				0	73
	Mexican	Yes	Female	100	13
				50	10
				0	20
	Indian	Yes	Female	100	21
				50	15
				0	22
	Anglo	No	Male	100	15
				50	25
				0	1083
	Mexican	No	Male	100	3
				50	5
				0	267
	Indian	No	Male	100	6
				50	2
				0	112
	Anglo	No	Female	100	31
				50	80
				0	1367
	Mexican	No	Female	100	12
				50	24
				0	449
	Indian	No	Female	100	35
				50	54
				0	254

Novometric Analysis: Males

Table 2 is a summary of the descendant family of optimal models predicting friends of males: all of the optimal models had identical performance in training (total sample) and jack-knife validity analysis.[1,2,34-36] Model six has the lowest D statistic and therefore it is the globally-optimal (GO) model presently.

Table 2: Descendant Family of Optimal Models Predicting Males' Friends

Step	ESS	Strata	Efficiency	D	Minimum Endpoint N
1	56.6	7	8.1	5.4	113
2	56.6	8	7.1	6.1	190
3	55.4	5	11.1	4.0	410
4	52.3	4	13.1	3.6	553
5	50.0	3	16.7	3.0	1071
6	48.5	2	24.2	2.1	2400

The two-strata GO model was: if ≤50% friends smoke, predict that the subject is female; otherwise predict the subject is male. Table 3 presents the confusion matrix for the GO model (moderate ESS=48.5, $p< 0.001$). As seen, the model accurately predicted 5 in 8 of the actual females (50% accuracy is expected by chance for each class category in two-category applications[1,34]), and 7 in 8 of the actual males.

2155-0182/10/$3.00

Table 3: Confusion Matrix: Male GO Model

		Predicted Friend		
		Female	Male	
Actual	Female	2,415	1,460	62.3%
Friend	Male	151	940	86.2%

Novometric Analysis: Females

Table 4 is a summary of the descendant family of optimal models predicting friends of females: all optimal models had identical performance in training (total sample) and jackknife validity analysis. Models one (sex is root variable, rating of percentage of friends who smoke is second attribute) and two (rating of percentage of friends who smoke is root variable, sex is second attribute) returned isomorphic classification performance. In applications in which such equivalent optimal models are identified—and in which the models cannot be distinguished on the basis of substantive theory or pragmatic considerations, primary (applied first) and secondary (applied next, if necessary) *a priori* selection heuristics—specified by the investigator—are used to select the final (GO) model.[1,34] Here the primary selection heuristic—select the model having the largest minimum endpoint sample size (used to inhibit overfitting and to maximize statistical power[1]), identifies model two as being the GO model.

Table 4: Descendant Family of Optimal Models Predicting Females' Friends

Step	ESS	Strata	Efficiency	D	Minimum Endpoint N
1	43.4	3	14.4	3.9	429
2	43.4	3	14.4	3.9	653
3	31.6	2	15.8	4.3	2,614

The three-strata GO model (Figure 1) indicates that 55% (5 in 9) of the 653 subjects who reported having some male friends who smoke were smokers. Of the remaining 4,243 subjects without male friends who smoke, 29% (3 in 10) of 2,058 male subjects, *vs.* 5% (1 in 20) of 2,185 female subjects, were smokers.

Figure 1: Friends of Females GO Model

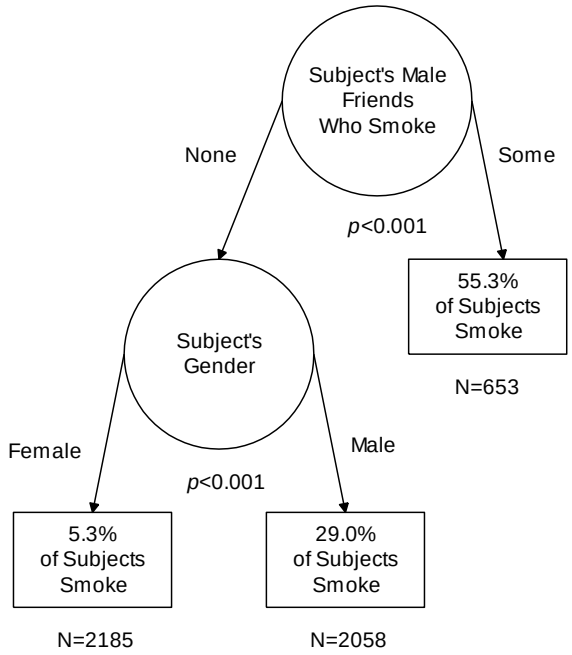

Table 5 is the confusion matrix for this model: 9 in 10 females and 6 in 10 males were accurately predicted, yielding relatively strong ESS=50.0.

Table 5: Confusion Matrix: Female GO Model

		Predicted Friend		
		Female	Male	
Actual	Female	3,465	410	89.4%
Friend	Male	430	661	60.6%

Not surprisingly, novometric results are straightforward compared to results obtained via disintegrated, partially documented chi-square.

References

[1]Yarnold PR, Soltysik RC (2016). *Maximizing predictive accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286

2155-0182/10/$3.00

[2]Yarnold PR (2016). Novometrics *vs*. ODA *vs*. One-Way ANOVA: Evaluating comparative effectiveness of sales training programs, and the importance of conducting LOO with small samples. *Optimal Data Analysis*, *5*, 131-132.

[3]Yarnold PR (2016). How many EO-CTA models exist in my sample and which is the best model? *Optimal Data Analysis*, *5*, 62-64.

[4]Yarnold PR, Linden A (2016). Novometric analysis with ordered class variables: The optimal alternative to linear regression analysis, *Optimal Data Analysis*, *5*, 65-73.

[5]Yarnold PR, Bennett CL (2016). Novometrics *vs*. correlation: Age and clinical measures of PCP survivors, *Optimal Data Analysis*, *5*, 74-78.

[6]Yarnold PR, Bennett CL (2016). Novometrics *vs*. multiple regression analysis: Age and clinical measures of PCP survivors, *Optimal Data Analysis*, *5*, 79-82.

[7]Yarnold PR (2016). Novometrics *vs*. regression analysis: Literacy, and age and income, of ambulatory geriatric patients. *Optimal Data Analysis*, *5*, 83-85.

[8]Yarnold PR (2016). Novometrics *vs*. regression analysis: Modeling patient satisfaction in the Emergency Room. *Optimal Data Analysis*, *5*, 86-93.

[9]Yarnold PR (2016). Matrix display of pairwise novometric associations for ordered variables. *Optimal Data Analysis*, *5*, 94-101.

[10]Yarnold PR (2016). Novometric theorem generalized to unrestricted class variables. *Optimal Data Analysis*, *5*, 102-103.

[11]Yarnold PR, Batra M (2016). Matrix display of pairwise novometric associations for mixed-metric variables. *Optimal Data Analysis*, *5*, 104-107.

[12]Yarnold PR (2016). Restricted *vs*. unrestricted optimal analysis: Smoking behavior of college undergraduates. *Optimal Data Analysis*, *5*, 124-128.

[13]Yarnold PR (2016). Parental smoking behavior, ethnicity, gender, and the cigarette smoking behavior of high school students. *Optimal Data Analysis*, *5*, 136-140.

[14]Yarnold PR (2016). Using gender of an imaginary rated smoker, and subject's gender, ethnicity, and smoking behavior to identify perceived differences in peer-group smoking standards of American high school students. *Optimal Data Analysis*, *5*, 141-143.

[15]Yarnold PR (2016). Assessing hold-out validity of models of smoking behavior developed for male Anglo-American college undergraduates applied to classify comparable Mexican-American and Indian-American samples. *Optimal Data Analysis*, *5*, 133-135.

[16]Zagona SV (1967). Psycho-social correlates of smoking behavior and attitudes for a sample of Anglo-American, Mexican-American, and Indian-American high school students. In: Zagona SV (Ed.), *Studies and issues in smoking behavior*. Tucson, AZ: University of Arizona Press (pp. 157-180).

[17]Yarnold PR (2016). CTA *vs*. disintegrated chi-square: Integrated *vs*. piecemeal analysis. *Optimal Data Analysis*, *5*, 118-120.

[18]Grimm LG, Yarnold PR (1995). *Reading and Understanding Multivariate Statistics*. Washington, DC: APA Books.

2155-0182/10/$3.00

[19]Grimm LG, Yarnold PR (2000). *Reading and Understanding More Multivariate Statistics*. Washington, DC: APA Books.

[20]Yarnold PR (2015). UniODA-based structural decomposition *vs*. log-linear model: Statics and dynamics of intergenerational class mobility. *Optimal Data Analysis, 4*, 179-181.

[21]Yarnold PR (2015). Modeling religious mobility by UniODA-based structural decomposition. *Optimal Data Analysis, 4*, 192-193.

[22]Yarnold PR (2015). UniODA-based structural decomposition *vs*. legacy linear models: Statics and dynamics of intergenerational occupational mobility. *Optimal Data Analysis, 4*, 194-196.

[23]Yarnold PR (2015). UniODA-based structural decomposition *vs*. legacy linear models: Statics and dynamics of intergenerational occupational mobility. *Optimal Data Analysis, 4*, 194-196.

[24]Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Win-dows*. Washington, DC, APA Books.

[25]Yarnold PR, Soltysik RC (2016). *Maximizing predictive accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286

[26]Yarnold PR, Soltysik RC (1991). Refining two-group multivariable classification models using univariate optimal discriminant analysis. *Decision Sciences, 22*, 1158-1164.

[27]Yarnold PR, Hart LA, Soltysik RC (1994). Optimizing the classification performance of logistic regression and Fisher's discriminant analyses. *Educational and Psychological Measurement, 54*, 73-85.

[28]Yarnold PR, Soltysik RC, McCormick WC, Burns R, Lin EHB, Bush T, Martin GJ (1995). Application of multivariable optimal discriminant analysis in general internal medicine. *Journal of General Internal Medicine, 10*, 601-606.

[29]Yarnold PR, Soltysik RC, Lefevre F, Martin GJ (1998). Predicting in-hospital mortality of patients receiving cardiopulmonary resuscitation: Unit-weighted MultiODA for binary data. *Statistics in Medicine, 17*, 2405-2414.

[30]Yarnold PR (2013). Univariate and multivariate analysis of categorical attributes with many response categories. *Optimal Data Analysis*, 2, 177-190.

[31]Yarnold PR (2014). UniODA *vs*. logistic regression analysis: Serum cholesterol and coronary heart disease and mortality among middle aged diabetic men. *Optimal Data Analysis*, *3*, 17-18.

[32]Yarnold PR (2015). UniODA *vs*. logistic regression and Fisher's linear discriminant analysis: Modeling 10-year population change. *Optimal Data Analysis, 4*, 139-145.

[33]Yarnold PR (2016). ODA *vs*. undocumented chi-square: Clarity *vs*. confusion. *Optimal Data Analysis, 5*, 121-123.

[34]Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*. Washington, DC, APA Books.

[35]Yarnold PR (2016). Using UniODA to determine the ESS of a CTA model in LOO analysis. *Optimal Data Analysis, 5*, 3-10.

[36]Yarnold PR (2016). Determining jackknife ESS for a CTA model with chaotic instability. *Optimal Data Analysis, 5*, 11-14.

Author Notes

The study analyzed de-individuated data and was exempt from Institutional Review Board review. No conflict of interest was reported.

Mail: Optimal Data Analysis, LLC
6348 N. Milwaukee Ave., #163
Chicago, IL 60646