Optimal Data Analysis
Vol. 8 (April 14, 2019), 81-83

2155-0182/10/$3.00

Regression *vs*. Novometric Analysis Predicting Income Based on Education

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

This study compares linear regression *vs*. novometric models of the association of education and income for a sample of 32 observations.[1] Regression analysis identified a relatively strong effect (R^2=56.4), but only 25% of point predictions fell within a 20% band of actual income. Novometric analysis identified a strong effect (ESS=81.7%) which was stable in jackknife validity analysis: the model correctly classified 91.7% of observations earning income less than $12,405, and 90.0% of those earning greater income. For people with an income which is less than the optimal threshold, and for those earning greater income, factors other than the number of years of education influenced earned income.

The Pearson correlation between education and income is r=0.751, p<0.0001, R^2=56.4. Figure 1 presents a scatterplot of the data as well as a plot of the linear regression model relating education (horizontal axis) and income (vertical axis): here F(1,30)=38.8, p<0.0001, R^2=56.4.

Figure 1: Scatterplot of Income by Education, Showing Linear Regression Model[1]

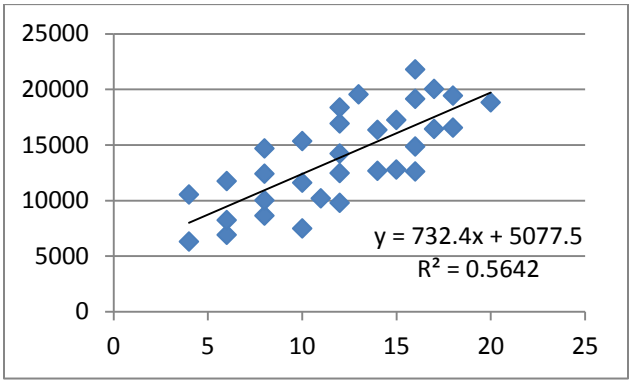

Table 1 summarizes the income point predictions made by the regression model for each observation: findings are sorted by years of education and then by income. Negative percent error is the percent by which the observation's actual income is *underestimated* by the model, and positive percent error is percent by which actual income is *overestimated*. Predicted values indicated in red fall within -10% to +10% of the actual income value—that is, they lie within a *20% band* of actual income. Only 8 of 32 (25%) point predictions made lie within a 20% band of actual income. The range of inaccuracy in predicted income—computed as maximum overestimation minus maximum underestimation for a given number of years of education—*exceeded* 50% for observations having 4, 6, 8, 10, 12, and 16 years of education.

2155-0182/10/$3.00

Table 1: Regression Point-Prediction Results

Years of Education	Predicted Income	Actual Income	Percent Error
4	8,007	6,281	-27.5
		10,516	23.9
6	9,472	6,898	-37.3
		8,212	-15.3
		11,744	19.3
8	10,937	8,618	-26.9
		10,011	-9.2
		12,405	11.8
		14,664	25.4
10	12,402	7,472	-66.0
		11,598	-6.9
		15,336	19.1
11	13,134	10,186	-28.9
12	13,866	9,771	-41.9
		12,444	-11.4
		14,213	2.4
		16,908	18.0
		18,347	24.4
13	14,599	19,546	25.3
14	15,331	12,660	-21.1
		16,326	6.1
15	16,064	12,772	-25.8
		17,218	6.7
16	16,796	12,599	-33.3
		14,852	-13.1
		19,138	12.2
		21,779	22.9
17	17,528	16,428	-6.7
		20,018	12.4
18	18,261	16,526	-10.5
		19,414	5.9
20	19,725	18,822	4.8

Moving to novometric analysis, Table 2 gives the descendant family of optimal models which emerged when treating income as being an ordered class or "dependent" variable, and education as an ordered attribute or "independent variable." Models were constrained to have the same classification accuracy in training and leave-one-out jackknife analysis: *p* in Table 2 is for LOO results (no other statistically viable models which explicitly maximized ESS existed for this application).[2-11] The optimal threshold value for all models in Table 2 was 11.5 years of education. Thus the structure of all optimal models in Table 2 was: IF education≤11.5 years then PREDICT income≤Cutpoint; IF education >11.5 years then PREDICT income>Cutpoint.

Table 2: Descendant Family of Optimal Models

	≤11.5 Yrs		>11.5 Yrs			
Cutpoint	N	Sens	N	Sens	ESS	*p*≤
10,186	8	87.5	24	75.0	62.5	0.0032
10,516	9	88.9	23	78.3	67.2	0.0009
11,598	10	90.0	22	81.8	71.8	0.0003
11,744	11	90.9	21	85.7	76.6	0.00005
12,405	12	91.7	20	90.0	81.7	0.000007
12,772	16	68.8	16	87.5	56.2	0.0016

The fifth model in the descendant family had the greatest ESS (81.7 is a strong effect[12]) and is thus the globally-optimal[11] (GO) model in this application: IF education≤11.5 years then PREDICT income≤$12,405; IF education>11.5 years then PREDICT income>$12,405. Table 3 presents the confusion matrix for the GO model.

Table 3: GO Model Classification Performance

Actual	Predicted Income		
Income	≤$12,405	>$12,405	Sensitivity
≤$12,405	11	1	91.7
>$12,405	2	18	90.0

Thus, novometric analysis finds that the crucial cutpoint affecting income is 12 years of education—ordinarily the period of time taken to obtain a high school diploma. Nine of ten people having fewer than 12 years of education earned $12,405 or less, and nine of ten people with at least 12 years of education earned more than $12,405. However, factors other than the

2155-0182/10/$3.00

number of years of education are responsible for additional differences in earned income. In this regard it is interesting that the correlation of years of education and income is r=0.39, R^2= 14.9, p<0.20 for the observations with fewer than a dozen years of education, and is r=0.43, R^2=18.2, p<0.07 for the observations having at least a dozen years of education.

References

[1]Lewis-Beck MS (1980). *Applied regression: An introduction*. Beverly Hills, CA: Sage.

[2]Yarnold PR (2017). What is optimal data analysis? *Optimal Data Analysis*, *6*, 26-42.

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

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

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

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

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

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

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

[10]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.

[11]Yarnold PR, Soltysik RC (2016). *Maximizing Predictive Accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286

[12]Yarnold PR, Soltysik RC (2005). *Optimal data analysis: Guidebook with software for Windows*. Washington, D.C.: APA Books.

Author Notes

No conflict of interest was reported.