The Relationship Between Teachers' Salaries and Student Performance Revealed with Modern Statistical Methods

Keith Kevelson
University of Massachusetts Dartmouth
Department of Electrical and Computer Engineering
MTH499: CSUMS Presentation Spring 2014

Introduction

- Considerable debate exists over the role of teachers' salaries in student performance.
- A rudimentary analysis comparing average teacher salaries in a given state with SERI(Science and Engineering Readiness Index) scores is conducted.
- Using a linear model, a small, but significant positive correlation is detected.

How to Analyze

- Mathematical relationships are inferred.
- This inference is called correlation.
- The type of correlation is named for the type of relationship inferred, so if a linear relationship is inferred, it is called a linear correlation.
- Many relationships must be analyzed, and many adjustments must be made.

Calculating Correlation

- Linear Correlation calculations are performed by using a method called “The Method of Least Squares.” To use this method to model data in one column(column B,) as a linear function of another column(column A,) do the following:

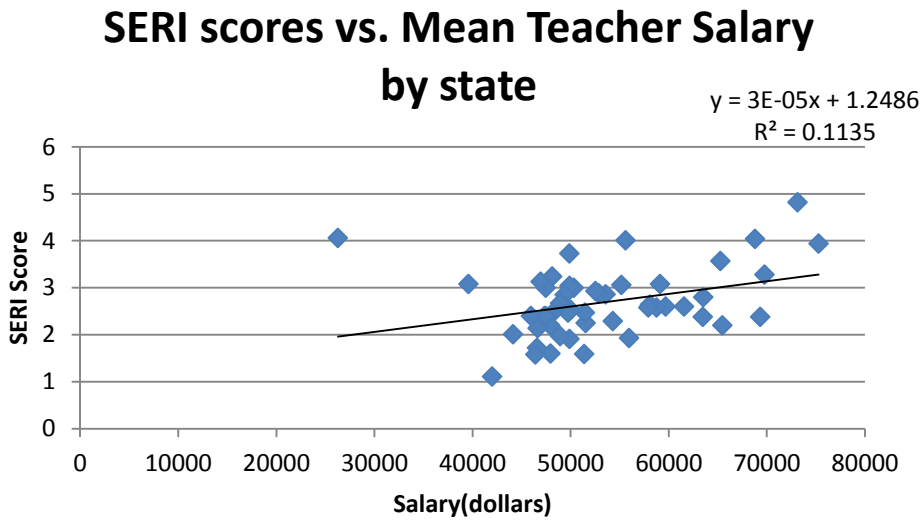
  1) Calculate the means of both columns
  2) Compute the sum of the squares in column A
  3) Compute the sum of each x-value multiplied by its corresponding y-value
  4) Calculate the slope of the line using the formula:

$$\left[\frac{\sum(xy) - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}\right]$$

The Data

- First, we'll look at teachers' salaries vs. SERI scores.(Stats subtract benefits and from2010.)

CA	69324	2.38	1.386294
NJ	68797	4.04	
Alaska	65468	2.2	
Maryland	65265	3.57	1.734601
PA	63521	2.8	1.516347
RI	63474	2.38	0.281851
Mich	61560	2.6	1.324925
DE	59679	2.6	-1.09861
IL	59113	3.08	1.516347
OR	58758	2.58	1.15268
OHIO	58092	2.64	0.944462
WY	57920	2.58	-1.45001
MN	26268	4.06	0.994623
NV	55957	1.93	-0.12014
NH	55599	4.01	0.489548
WS	55171	3.06	-0.08004
HW	54300	2.29	-0.40547



Analysis

- Slight correlation coefficient of 0.00003
- R-squared value is extremely low(0.11)
- Low R-squared value suggests state-by-state correlation is not that significant.
- Does this mean that there is no correlation under any circumstances? Probably not.

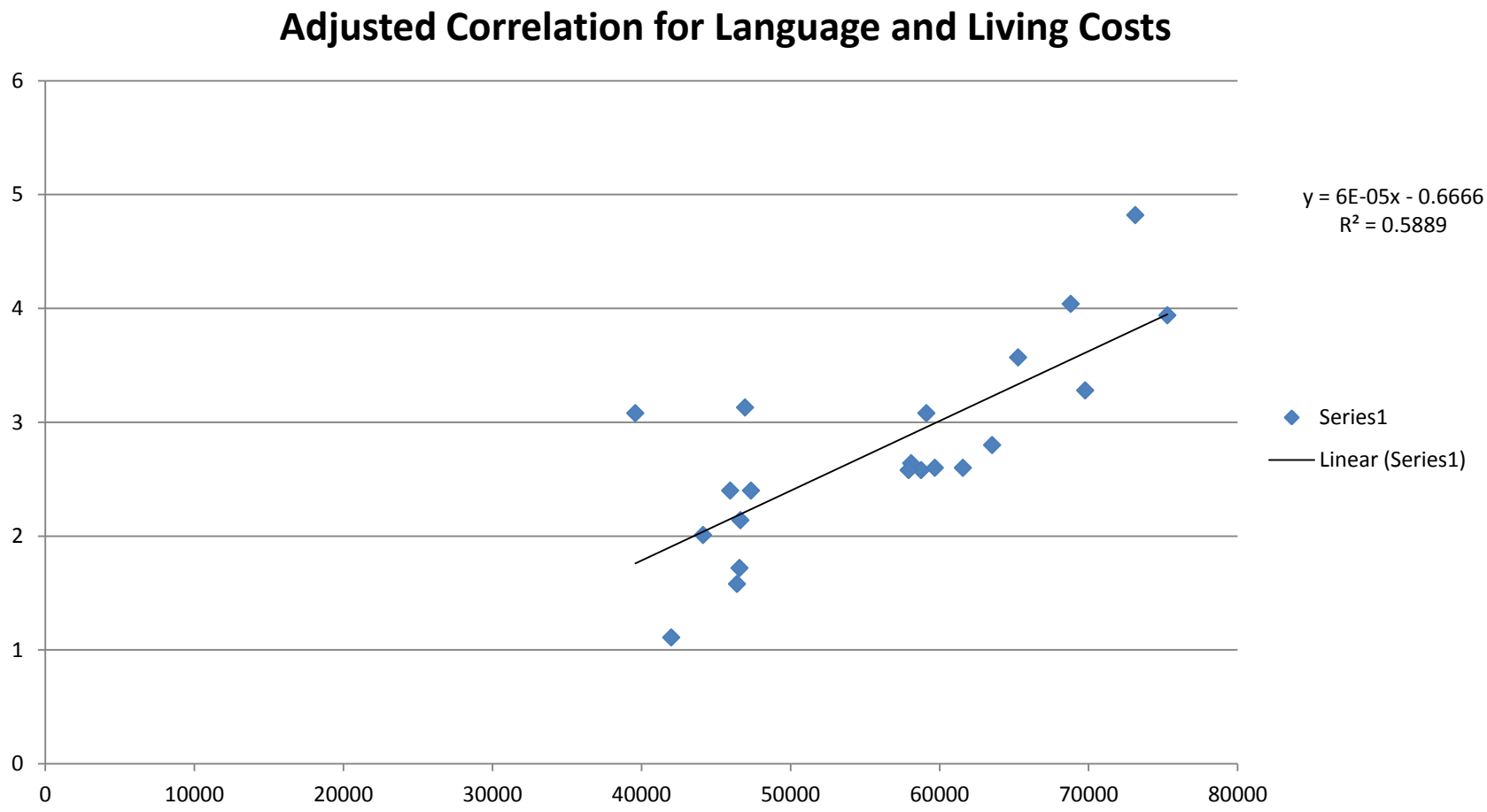
Other Factors

- If one compares states with low percentages(<10%) of English language-learners with other states with low percentages of English language-learners, a dramatic shift correlation is apparent.
- Some states and cities have dramatically higher costs of living than other states.

Adjusted Correlation

- Using Census Data, states that have over 90% of their students speaking English and average home prices between $250K and $350K are compared.
- Other means to measure multiple situations to be developed.

Huge Difference by Adjusting



Keith Kevelson—**A Study on Correlation of Teachers Salary and Student Performance**

Implications of this Correlation

- Within a narrow range of certain home prices and linguistic makeup of a student body, a state is more likely to receive a high SERI score if its teachers are paid more than average.
- By deduction, a state is less likely to receive a high SERI score if its teachers are paid less than average in these circumstances.
- R-squared value close to 0.6 indicates a much better fit than the fit in the previous correlation.

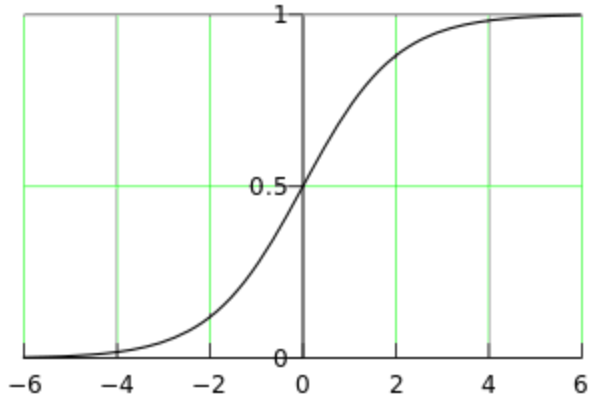
What About Other Correlations

- Logistic correlation might be more appropriate to analyze possible effects of teachers' salaries because intuition dictates that paying teachers $1,000,000 per year would not result in genius students.
- Academic performance could be measured in other ways, such as SAT scores.
- Adjustments can be performed by creating new coefficients

Logistic Regression

- A logistic function is expressed by:

$$[f(x) = \frac{1}{1 + e^{-x}}]$$



- The numerator is the maximum value that is approached over time.

How to Perform Logistic Regression

- Consider a logistic function:

$$[f(x) = 1/(1 + e^{-}b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots b_px_p)]$$

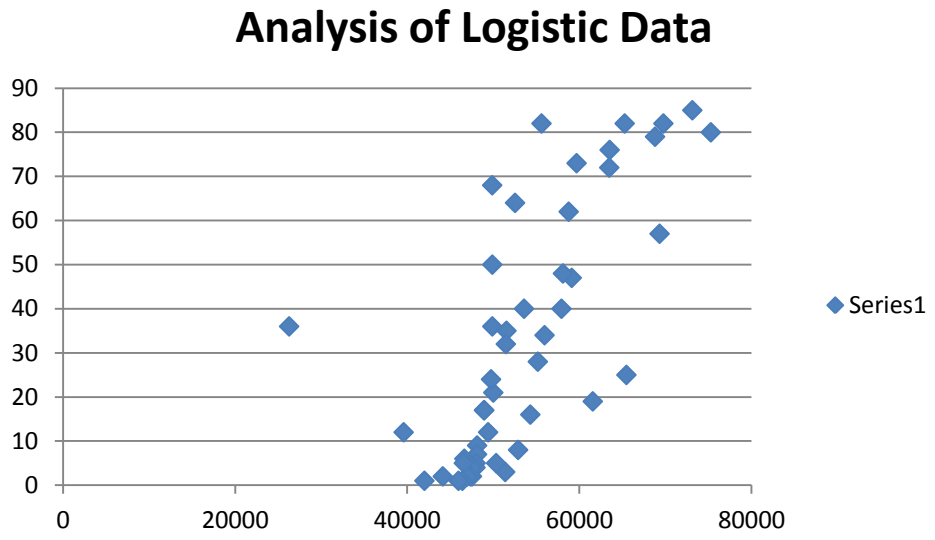
- Taking the natural log of both sides gives:

$$ln(\frac{p}{1-p}) = b_0 + b_1x_1 + b_2x_2 \ldots + b_px_p$$

- This is a linear function. This transformation allows us to analyze as many variables as we like, with as many patterns as we need.

Example Logistic Regression

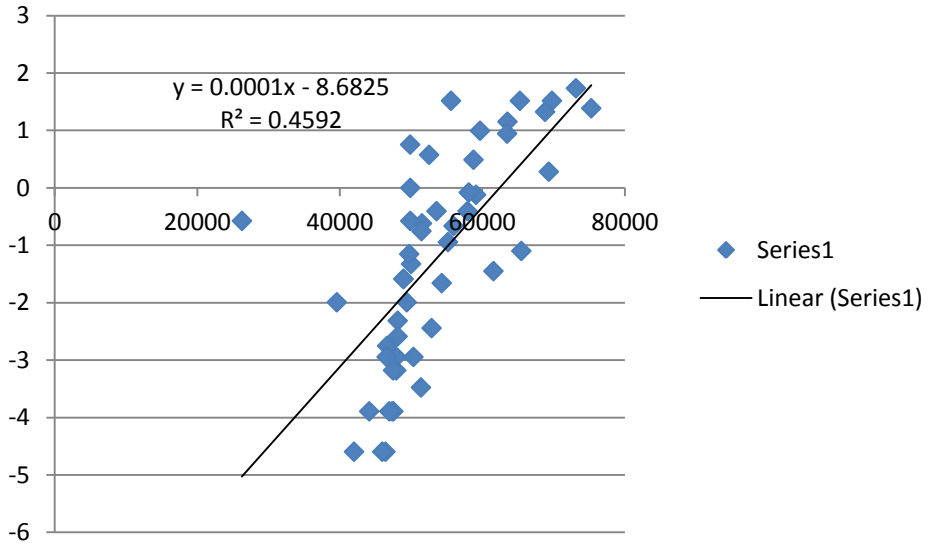
- Examine the relationship between salary and percent of students taking the SAT or ACT



Keith Kevelson—**A Study on Correlation of Teachers Salary and Student Performance**

Logistic(continued)

- Taking the graph of the trendline and the natural log of the data gives:



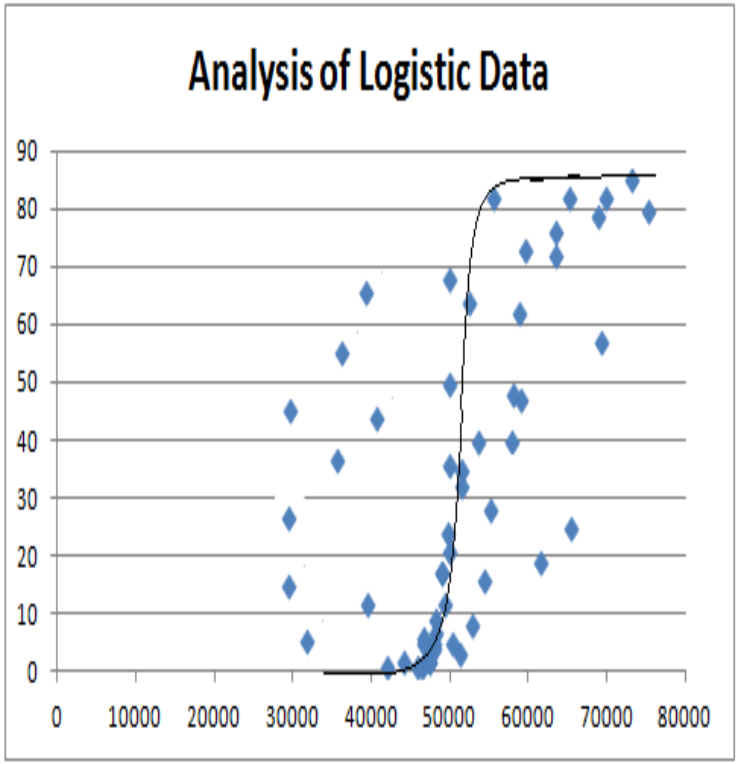
Logistic Trend from Previous Data

Regression Fit equation

- Note that max is 95 and not 100

$$P(s) = \frac{95}{1 + (94e^{(} - 0.00086s))}$$

Graph of best fit among data



Total Least Squares Method

- For faster algorithm that executes in linear time, T(n)=O(n), a novel bash shell-script was designed to gather data based on a previous algorithm(Guo and Renaut)
- Uses Golub and Van Loan Theorem

Snippet of Powershell Script

```
Function Get-SearchResults {
   param([string] $searchstring=$(throw "Please specify a search string."))

   $client = New-Object System.Net.WebClient
   $url="http://www.google.com/search?q={0}`&format=rss" -f $searchstring
   [xml]$results = $client.DownloadString($url)
   $channel = $results.rss.channel

  foreach ($item in $channel.item) {
     $result = New-Object PSObject
     $result | Add-Member NoteProperty Title -value $item.title
     $result | Add-Member NoteProperty Link -value $item.link
     $result | Add-Member NoteProperty Description -value $item.description
     $result | Add-Member NoteProperty PubDate -value $item.pubdate
     $sb = {
        $ie = New-Object -com internetexplorer.application
        $ie.navigate($this.link)
        $ie.visible = $true
     }
     $result | Add-Member ScriptMethod Open -value $sb
     $result
   }
}

Get-SearchResults "Teachers' Salaries by district"
```

Big Conclusions to Draw

- With the right variables, correlations can be drawn between any two characteristics, but that doesn't mean there's causation.
- Lack of correlation does not mean lack of causation.
- Apparent relationship between teachers' salaries and student performance appears to exist to a certain extent.

Future Work

- More shell-scripting and faster data-mining
- More research into compound independent variables. For example, student performance could have a much stronger correlation with:

$$\varsigma = \sqrt{(salary)^2 + (students' family income)^2}$$

Acknowledgements

- H. Guo and R. A. Renaut, **A regularized total least squares algorithm**, in Total Least Squares and Errors-in-Variables Modeling: Analysis, Algorithms and Applications, S. Van Huffel and P. Lemmerling, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002, pp. 57–66.

- Very Special Thanks to Dr. Akil Narayan