MedCrave
Step into the World of Research

Biometrics & Biostatistics International Journal

Mini Review

Open Access

CrossMark

A Statistical mystery resolved

Abstract

During consulting work the regression analysis between the salaries and length of employment of a group of professional women gave an implausible, counter intuitive result. The resolution of this statistical mystery revealed a common, unrecognized misunderstanding of the nature and interpretation of regression.

Keywords: interpreting regression lines, sex-discrimination in government

Volume 8 Issue 3 - 2019

Othmar W Winkler
Georgetown University, USA

Correspondence: Othmar W Winkler, Professor emeritus, Georgetown University, USA, email winklero@georgetown.edu

Received: May 11, 2019 | **Published:** May 27, 2019

Background

Everybody dealing with data will, at one time or another, employ regression analysis. This very unusual case happened during the exploratory phase of the data of a sex-discrimination lawsuit.[1] The 32 librarians,16 male and 16 equally qualified female librarians of that government agency, appeared to be ideally suited to initiate discovery of the claimed discrimination in that professional workforce. Simple linear regressions of salary and length of employment, computed separately for the male and the female librarians, was a first approach to reveal the supposed existence and nature of sex-discrimination. Though expecting differences between these two regressions, the author was unprepared to make sense of the women's regression and incredulous. The fundamental insight gained by resolving this statistical puzzle should be of general interest.

Viewing the data

The relationship of 'Salary and Lengths of Service' for the male librarians, (Figure 1) was SALARY (M)=$16,900+$1,380∗ERDAEMPL, in other words, "the average starting salary of these 16 male librarians at ERDA was $16,900 with a yearly average salary increase of $1,380" which appeared to be reasonable. The relationship for the female librarians with comparable academic degrees and the same duties was: SALARY (W)=$26,500–1,020∗ERDAEMPL. The slope β=−$1,020 indicated that for each additional year of service, the salaries of those female librarians were reduced, on average, by $1,020 (Figure 2) while their entrance salaries were the highest when they were first employed. This just did not made sense, defying every experience with employment.

My first impression was that this obviously had to be an error in the data or some mix-up in the computer program. The person responsible for these data, however, swore that this represented the situation correctly, and that the computer program worked fine. This assurance was trustworthy because this 'discovery of facts' had been ordered by the District judge. I had hoped for errors in the data or in the computer program as the explanation of this implausible result. But no such easy explanation of that puzzling regression became available. Before continuing to the next chapter I like to invite the reader to stop reading and think of an explanation as a probable solution to the conundrum of Figure 2.

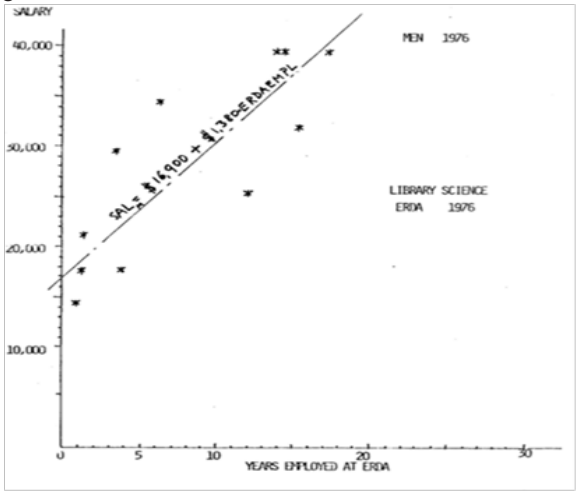

Figure 1 Male librarians, salary and length of employment.

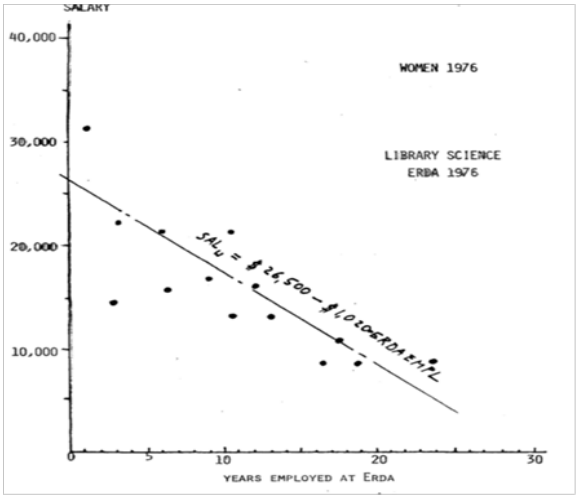

Figure 2 Female librarians, salary and length of employment.

[1]The class-action sex-discrimination lawsuit of the female professional employees as plaintiffs against the Energy Resource Development Agency of the Federal Government, ERDA was filed in 1976 as Chewning vs. ERDA,. It had been established Jan 1975 from the split earlier of the Federal Atomic Energy Commission. In October 1977, together with other Commissions and Departments, ERDA was merged into the Department of Energy, DOE-I was summoned to establish the facts of the women's claim that equally qualified professional women were paid less and been promoted less often than their male counterparts. To accomplish this, the U.S. District judge of the District of Columbia ordered the full, unrestricted access to the confidential employment records of all professional men and women employed by ERDA. It is important to note that misusing and tampering with those data would have been a Federal crime.

Resolving the paradox

Searching for reasons to explain this negative slope, implying that the female librarians at ERDA were paid $1,020 less for every additional year they were employed at ERDA, it then dawned on me that the error was in the interpretation, not in the data. The usual interpretation of the slope β as the change in Y corresponding to a one unit change in X implicitly assumes a longitudinal, dynamic situation. Yet both, the regressions of the male and of the female librarians, represented a 'static cross-section' of the situation at the time of this lawsuit. These data required a cross-sectional, static understanding of β, as if these 16 women, at the time of this lawsuit, were lined up for a group photo. Imagine the recent hires, employed the shortest time, were standing to the left of the group, with low values on the horizontal axis of the graph, but with the largest salaries, their Y-values. Those longer employed librarians, hired in previous years and decades, standing to their right, with larger X-values, happened to have the lower incomes, Y-values. The average difference between the sizes of the salaries of any two female librarians in this line-up, the one employed one year longer, hired a year earlier, happened to earn a lower income, smaller on average by $1020. β was the average salary difference between any two of these 16 female librarians whose length of employment differed by one year.[2] The key to resolve this puzzle was recognizing these data as a 'cross-section' at a given point in time, instead of wrongly interpret them to be the 'longitudinal development' of their salaries over the years.

How did this unbelievably dismal situation happen?

The employment histories of these women revealed that those who entered ERDA before the promulgation of 'Title VII' – anti sex-discrimination signed into law in 1969 and expanded in 1974 – were hired decades earlier at starting salaries that were substantially lower than the starting salaries of the male librarians hired at that time, and also lower than the salaries female librarians who had been hired after the anti-sex-discrimination laws had been enacted. A colleague opined that the earlier hired female librarians had fewer educational opportunities and less professional education available to them. This was used as justification for their lower salaries. All librarians had received raises of similar percentages but those of the older-tenured female librarians, due to their lower starting incomes, amounted to smaller pay increases. The existence of these discrepancies in 1975 was due to management's improper, selective implementation of the anti-sex-discrimination laws. ERDA's management had assumed that only the averages of the women's salaries of a department, not Individual cases of discrimination would be checked for compliance with anti-discrimination laws.

Conclusions

The statistical paradox, focused on female librarians, who appeared to be paid less the longer they were employed, originated in the failure to recognize the cross sectional, static nature of the data to correctly interpret the regression. This misinterpretation was subconsciously encouraged by the preceding wrong, and just as inappropriate, similar interpretation, of the regression of male librarians which was easily overlooked because their positive correlation of length of employment and income seemed to agree with common sense and general experience. In conclusion, the peculiar circumstances that lead to this lawsuit revealed the unnoticed, preferred custom to interpret regression lines longitudinally as a dynamic "average change in Y for a change or increase in X," even when the data are a cross section not warranting such an interpretation. This incorrect interpretation of the women's data became the puzzle of an obviously implausible employment situation. If it had not been for this very unusual employment situation such misinterpretations of the regression line, its error, would continue to remain unrecognized.

This class action lawsuit, by the way, had a happy ending. The women of that class-action lawsuit won a decisive victory against their agency proving convincingly, through statistics, the presence of substantial social and economic discrimination.

Acknowledgments

None.

Conflicts of interest

Author declares that there are no conflicts of interest.

[2]Considering a 'dynamic' situation in contrast to this 'static 'one, additional information about the evolution of each employee's yearly salary for the length of his/her employment would be needed. Then each one of these employment histories would be plotted as a line on a graph with 'salary' on the vertical axis, but 'calendar years' instead of 'length of employment in years' on the horizontal axis. Each librarian's length of employment would be represented as a line, 16 lines per gender. Due to the yearly raises of their salaries, those lines would be up sloping, from left to right. Those just having joined, say in 1975, would have a short, up sloping line, at the far right of the graph. Those having been employed e.g. for 30 years, who would have entered in 1947, would be represented by a long line, beginning at the far left, extending to the end of the graph at right. All of these 32 employees' lines would move up from left to right due to the yearly salary increases depicted as lines with positive slopes. The traces of the female librarian's lines would be on a lower level and flatter, due to their lower starting salaries and smaller increases. The slope β of the simple linear regression lines of the 16 female librarians' would have a positive β in such a hypothetical, 'dynamic, longitudinal' situation. Only in such a situation would the usual interpretation of β as "average increase in salary per one year" be correct and unproblematic.