

## COMMENT

Joseph L. Gastwirth\*

**CITATION:** Joseph L. Gastwirth, Comment on the Age Discrimination Example, 42 *Jurimetrics J.* 333–340 (2002).

This is an interesting example<sup>1</sup> designed both to assist judges in understanding statistical evidence and evaluating its relevance or “fit” to the issues involved.<sup>2</sup> The concepts and techniques discussed (e.g., the odds ratio, Fisher’s exact test, and logistic regression) are quite important.<sup>3</sup> I am unsure, however, of the “fit” of the defendant’s logistic analysis in the legal context. This is especially true given the portion of *Texas Department of Community Affairs v. Burdine*<sup>4</sup> that states that

---

\*Joseph L. Gastwirth is Professor of Statistics and Economics, George Washington University, and currently Visiting Scientist at the Division of Cancer Epidemiology and Genetics, National Cancer Institute. He is the author of the textbook, *Statistical Reasoning in Law and Public Policy* (1988), and the editor of *Statistical Science in the Courtroom* (2000). He thanks Drs. Jay-Lubin, Marc Rosenblum, Binbing Yu, and Gang Zheng for helpful discussions about the legal and statistical issues and Dr. T. Reagan for providing the underlying data.

1. Robert Timothy Reagan, *Federal Judicial Center Statistical Examples Software Prototype: Age Discrimination Example*, 42 *JURIMETRICS J.* 281 (2002).

2. “Fit” plays an important role in the gatekeeping task the Court assigned trial judges in *Daubert v. Merrell Dow Pharms., Inc.*, 509 U.S. 579 (1993). For further discussion and references, see Marc Rosenblum, *On the Evolution of Analytical Proof, Statistics, and the Use of Experts in EEO Litigation*, in *STATISTICAL SCIENCE IN THE COURTROOM* 161 (Joseph L. Gastwirth ed., 2000), and Symposium, *At the Daubert Gate: Managing and Measuring Expertise in an Age of Science, Specialization and Speculation*, 57 *WASH. & LEE L. REV.* 901 (2000). But see D.H. Kaye, *The Dynamics of Daubert: Methodology, Conclusions, and Fit in Statistical and Econometric Studies*, 87 *VA. L. REV.* 1933, 1961 (2001) (arguing that “[a]s a logical matter, however, the fit requirement is superfluous”).

3. Reagan, *supra* note 1, at 284–85, 288.

4. 450 U.S. 248 (1981); see also *Reeves v. Sanderson Plumbing Prods.*, 530 U.S. 133 (2000). I am grateful to Dr. Marc Rosenblum for pointing out the importance of *Reeves*.

the plaintiff should have a full and fair opportunity to show that the defendant's explanation is a pretext. No discussion of the "pretext" stage is given; Section I demonstrates why this is a serious flaw in the current version. Section II identifies statistical issues that deserve more discussion, such as why the odds ratio is appropriate for layoff cases but probably not for typical hiring cases.<sup>5</sup> Section III offers some suggestions on how to improve the teaching utility of the prototype.

## I. THE AGE DISCRIMINATION EXAMPLE

The plaintiffs, whose specific ages are not given in a list, were terminated after a merger and reorganization.<sup>6</sup> The statistical data reported in Table 1<sup>7</sup> of the example are analyzed using Fisher's exact test, which considers all employees over 40-years-old as having the same probability of being terminated and tests whether this probability is the same as that of employees younger than 40. Because the test is a conditional one in that it uses the fact that 79 individuals were terminated, it is an appropriate procedure. However, it is less informative than a test that is directed at an increasing trend in firing due to age.<sup>8</sup>

**Table 1: The Number and Percentage of Employees Fired by Age Category**

Age Category	Under 40	40 to 50	50 to 60	60 and over
Retained	205	110	51	34
Fired	23	23	15	18
Total	228	133	66	52
% Fired	10.1	17.3	22.7	34.6

Table 1 groups the data by increasing age; the probability of being fired seems to rise with age. We test the null hypothesis of *no age effect* against a linearly increasing trend in the probability of being fired with age using the Cochran-Armitage (CA) test.<sup>9</sup> The resulting *p*-value of 0.0000074 is quite a bit

5. Integrating the software with the Federal Judicial Center's *Manual on Scientific Evidence* and appropriate statistical texts would provide greater understanding.

6. Reagan, *supra* note 1, at 282.

7. *Id.* at 284.

8. Such a test is described in several statistics texts. See ALAN AGRESTI, *CATEGORICAL DATA ANALYSIS* 118-19 (1990); JOSEPH L. FLEISS, *STATISTICAL METHODS FOR RATES AND PROPORTIONS* 96-9 (1973); PETER SPRENT, *DATA DRIVEN STATISTICAL METHODS* 374-78 (1998). A trend test is suggested as an appropriate procedure for age cases in Joseph L. Gastwirth, *Statistical Evidence of Discrimination*, 160 J. ROYAL STAT. SOC'Y 289 (1997). The data from an age discrimination case that settled prior to trial are also discussed in SPRENT, *supra*, at 378.

9. See authorities cited *supra* note 8. The test essentially correlates the difference between the percentage of individuals in each group who were fired and the overall percentage with a trend of 1, 2, 3, and 4. These weights can be modified to account for other information. For instance, if there were written or oral comments that individuals should stop working at 50, then one could combine the 40 to 50 year olds with those under 40 using weights 1, 1, 2, and 3.

less than the value of 0.0002 obtained from Fisher's test. Clearly, this analysis strengthens the plaintiff's *prima facie* case.

However, the employer contended that age was not a factor when it considered the need for the employees' services, their performance evaluations, and the manager's assessment of their potential loyalty to the new management.<sup>10</sup> The defendant also asserted that senior employees were less likely to be loyal to management, using seniority as a surrogate or proxy for loyalty.<sup>11</sup> The first part of defendant's statistical evidence consists of two plots that clearly show that seniority and age are strongly related and that, in each age group, retained employees usually had less seniority than terminated employees.<sup>12</sup>

The final analysis offered by the defendant is a logistic regression relating the probability of being fired to years of service and age.<sup>13</sup> Figure 4 is a useful plot showing that the coefficient on seniority is bigger than that for age, indicating that although increased age still appears to be related to one's probability of termination, seniority plays a more significant role.<sup>14</sup> Formal statistical tests show that seniority is statistically significant at the commonly accepted 0.05 level, while age is not significant at that level. The legal analysis provided indicates that the data do not support an action for age discrimination under the *disparate treatment* theory and leaves open the appropriateness of the use of seniority as a proxy for loyalty in *disparate impact* cases. Although the prototype's analysis stops here, it is instructive to consider arguments that a plaintiff might raise to show that the reasons offered by the employer are pretextual. While the defendant claimed to consider past job evaluations, the company's future need for specific skills, and loyalty, the "explanation" only incorporated the rather "subjective" factor of loyalty. The company did not conduct an interview or use a previously validated loyalty test.<sup>15</sup> Rather, the logistic regression *only* incorporates one factor that is assessed by a proxy variable that is highly correlated with age ( $r = .77$ ). The plaintiff demonstrates the effect of this high correlation between variables age and seniority by submitting a logistic equation that incorporates an interaction term<sup>16</sup> along with age and seniority. The results given in Table 2 below indicate that although the model as a whole is highly predictive,<sup>17</sup> *none* of the individual predictors is significant, even though the interaction variable has the smallest *p*-value. This implies that it will be difficult to distinguish the effect of age from that of seniority, especially if they have a joint effect. This indicates that the factfinder

10. Reagan, *supra* note 1, at 291.

11. *Id.*

12. *Id.* at 291-92.

13. *Id.* at 293.

14. *Id.* at 294.

15. I do not know whether such a test exists. If one is not available, then the defendant should provide evidence of this fact to justify using the proxy of seniority, which is so clearly correlated with age.

16. This adds a new variable, age times seniority, that looks for a joint effect of age and seniority.

17. The overall test is at the 0.001 level.

needs to make sure that the "proxy" variable is not a "cover" for the legally protected characteristic (age).

**Table 2: Results of a Logistic Analysis  
with an Interaction Term**

Variable	Estimate	Standard Error	p-value
Constant	-2.54	.8364	.002
Age	.0058	.0202	.772
Seniority	.0038	.0970	.965
Interaction	.0013	.0016	.424

Plaintiffs might well question the use of seniority as a "proxy" for loyalty. Older employees are likely to be more loyal, in part because they have acquired firm-specific knowledge and also because their alternative job prospects may not be as bright as those of younger employees.<sup>18</sup> In particular, older employees do not turn over at the rate of younger ones, enabling the employer to save the cost of training new employees.<sup>19</sup> Thus, seniority, a factor usually positively related to pay or productivity, is now considered a substitute for a subjective assessment of "loyalty" to new management. The hypothetical indicates that the defendant claimed that the merger and reorganization inspired significant employee dissent; however, it presented no evidence indicating that more senior employees expressed greater unhappiness than junior ones.

Assuming that an unusual amount of employee dissent arose from the merger, finding out when it began becomes important. For example, suppose the acquiring company had a reputation for laying off senior workers in previous mergers; suppose also that the president of the firm had circulated an e-mail to middle managers to proceed with the same strategy. Furthermore, the e-mail described senior employees as "old geezers." As often happens in the modern world, someone forwarded the e-mail to an employee in the firm being acquired, perhaps to warn the recipient of what the future might portend. If the dissension arose *after* these events, it seems that the distinct likelihood of age discrimination by the defendant *created* the dissent itself. Therefore, employee unhappiness is not an appropriate variable for the defendant to offer as an explanation, much less to use seniority as a "proxy" for it. Of course, there may well be alternative scenarios more favorable to the defendant. The main point is that the example<sup>20</sup> fails to discuss one of the three stages in *disparate treatment* cases and simply accepts the employer's claim about dissent, without showing that the degree of dissent was related to an employee's length of service. Indeed, 5 of the 11 employees with less

18. RICHARD A. POSNER, *AGING AND OLD AGE* 75 (1995).

19. In fact, in *Reeves v. Sanderson Plumbing Products*, 530 U.S. 133 (2000), the defendant had replaced the plaintiff three times because two of the replacements, all in their thirties, had left.

20. Reagan, *supra* note 1.

than one year of seniority were fired. This percentage of 44.5 exceeds that of any age group in our Table 1.

Plaintiffs might well point out that the defendant did not incorporate the job evaluations into the model. This is important because studies reviewed have shown that older employees perform similarly to younger ones.<sup>21</sup> This may be the result of selection; for example, employers may weed out poor performers so the abilities of remaining older workers are higher than average. For our purpose, the failure of a party to include a clearly relevant variable may render the analysis inadmissible under *Daubert*.<sup>22</sup> Indeed, in *Diehl v. Xerox Corp.*,<sup>23</sup> the trial judge did not credit a statistical analysis that failed to incorporate a major variable. In that case, plaintiff's expert did not include performance histories or a skills assessment study in a regression analysis whereas defendant's analysis using them along with the variable of seniority indicated that older workers were favored.<sup>24</sup> Thus, it seems that the logistic analysis may place too much weight on the variables age and seniority, even assuming that the defendant's explanation truthfully describes what its managers did. If in reality the employer *did not consider job evaluations* or favored more senior workers as in *Diehl*, then Reeves allows the fact-finder to conclude that the reasons offered were pretextual.<sup>25</sup>

The logistic model considers age as a continuous variable. This is true biologically, but the law treats all employees under the age of 40 as one group in that the Age Discrimination in Employment Act<sup>26</sup> (ADEA) only protects employees at least 40 years of age.<sup>27</sup> Thus, plaintiffs might argue that even if only one explanatory variable was appropriate, so the issue of omitted factors does not arise, the logistic model is so inappropriate that it should be deemed inadmissible. This may be too harsh a result, as the data clearly indicate that age or seniority is related to the firings. The defendant's regression should be accepted into evidence, but given less weight than a model reflecting the true legal status of employees under 40. In *Mangold v. California Public Utilities Commission*,<sup>28</sup> the defendant objected to the plaintiff's expert correlating the performance on a *subjective* promotion exam with age and asked for a comparison of test-takers over 40 with those under 40.<sup>29</sup> The opinion notes that the "favored" individuals need only be "substantially younger" than the plaintiff and so need not be under forty.<sup>30</sup> Thus, a logistic regression, incorporating the data on the relevant factors

21. POSNER, *supra* note 18, at 75 n.17.

22. See *Daubert v. Merrell Dow Pharms., Inc.*, 509 U.S. 579 (1993).

23. 933 F. Supp. 1157 (W.D.N.Y. 1996).

24. *Id.* at 1162, 1165.

25. See *Shannen v. Fireman's Fund Ins. Co.*, 156 F. Supp. 2d 279, 291 (S.D.N.Y. 2001) (citing cases allowing plaintiffs to establish pretext by demonstrating inconsistencies and contradictions in the defendant's reasons or factual errors in its statements).

26. 29 U.S.C. §§ 621-34 (1994 & Supp. V 1999).

27. 29 U.S.C. § 631.

28. 67 F.3d 1470 (9th Cir. 1995).

29. *Id.* at 1476.

30. *Id.*

that were available to the defendant, is likely to be more informative than a simple comparison of firing rates between those under and over forty. Additionally, checking of the fit of the data to the assumptions underlying the model is desirable.

The Cochran-Armitage trend test, however, can be extended to this situation by modeling the log-odds of being terminated as a function of an employee's age group. One would then use a trend test of the odds-ratios after incorporating the other legitimate factors.<sup>31</sup> Because neither party submitted the most objective information, that of the performance ability of employees, we cannot conduct a complete analysis. For exploratory purposes, we ran two regression models. In the first, which did not use seniority, the odds ratios, relative to the under forty group, were 1.864 for the 40-to-50 age group, 2.621 for the 50-to-60, and 4.719 for the sixty-plus group. When seniority was included, those odds ratios were reduced to 1.27, 1.537, and 1.57, respectively. A test for an increasing trend of these odds ratios is not statistically significant; however, there still appears to be an increasing trend with age. Thus, the ultimate decision may depend on the other circumstances surrounding the lay-off, the appropriateness of using seniority as a "proxy" for loyalty, or what the job evaluation data indicate. If the job evaluations did not correlate with seniority, they would likely diminish its importance in a complete logistic model.

## II. OTHER STATISTICAL ISSUES

While the odds ratio is the appropriate measure to examine termination data,<sup>32</sup> it may not be appropriate for hiring or promotion data where the ratio of the minority success rate to that of the majority group is often used. In *Bew v. City of Chicago*,<sup>33</sup> 98.595% of the blacks passed an exam compared to about 99.9952% of the whites.<sup>34</sup> The ratio of the success rates is 0.9864, far exceeding the simple "four-fifths" rule that has been used as a screening device.<sup>35</sup> Common sense suggests that with such a high percentage of minorities passing the exam and the ratio of the success rates being so high, the exam had a minimal impact on the prospects of a black applicant.<sup>36</sup> However, the odds-ratio, being symmetric in pass and fail rates, equals 28.87. Thus, blacks had only *one-twenty-fifth* the odds of passing as whites. But to conclude the data in *Bew* are stronger evidence of

---

31. This technique is commonly used in epidemiological studies. See Norman E. Breslow et al., *Multiplicative Models and Cohort Analysis*, 83 J. AM. STAT. ASS'N 1, 5 (1983).

32. See Joseph L. Gastwirth & Samuel W. Greenhouse, *Biostatistical Concepts and Methods in the Legal Setting*, 14 STAT. IN MED. 1641, 1642 (1995) (recalling that the odds ratio is the correct parameter specifying the distribution of the number of members of the protected class who are laid off).

33. 979 F. Supp. 693 (N.D. Ill. 1997).

34. *Id.* at 696 n.6.

35. *Id.*

36. The sample sizes were quite large and the usual test of significance yielded a difference of -5 standard deviations, well above the usual criteria of two to three. *Id.* at 696. Thus, at the summary judgment stage, it was reasonable for the judge to decide that there was a material issue of fact.

discrimination than the odds ratio of 2.56 in the FJC example is not sensible. The ratio of retention rates can be misleading in termination cases,<sup>37</sup> and judges should realize that several measures of the difference between two proportions can be used.<sup>38</sup>

It may be helpful to readers with a legal background to inform them of a basic difference in the assumptions underlying Fisher's exact test and the usual method for comparing two proportions: the exact test conditions on the number of employees fired. The numbers of fired employees in each of the age categories must add to the total. This implies that they are statistically dependent. In the typical case that concerns passing an entrance or promotion exam, the passing score is announced before the test so that the number of people passing in one group does not affect the number in another. Using Fisher's exact test in that situation typically entails a loss of power relative to recently developed methods.<sup>39</sup>

Although the prototype provides a good discussion of the logistic model and interpretation of the coefficients, some mention of the "goodness of fit" as well as its explanatory power would be useful. Goodness of fit is concerned with how well the model fits the data. For example, in Figure 1, it appears that the model underpredicts the probability of being fired for employees in the upper age range. For assessing the explanatory power of ordinary regression, one uses the proportion of variance explained ( $R^2$  or adjusted  $R^2$ ); however, the most appropriate measure for logistic and similar binary regression models is still an active research area.<sup>40</sup>

The inclusion of the chi-squared approximation to Fisher's exact test should be useful to judges, as it provides a relatively simple method to obtain a reliable result under certain conditions. Indeed, it would be helpful to mention still other approaches to analyzing the data. We have already provided one way and will simply observe that one could also have stratified the data in our Table 1 into seniority categories (e.g. by quintiles). The stratified version of the trend test

37. See Gastwirth & Greenhouse, *supra* note 32, at 1642.

38. The defendant in *Bew* prevailed by demonstrating that the test was job-related and the cut-off score was reasonable. 252 F.3d at 891. The measure of impact is important as it may play a role in evaluating the evidence validating a test. A test with a large impact will likely need a greater degree of predictive validity than one with a small impact.

39. See Roger L. Berger and Dennis D. Boos, *P Values Maximized Over a Confidence Set for the Nuisance Parameter*, 89 J. AM. STAT. ASS'N 1012 (1994) for the analysis of a two-by-two table, and Boris Freidlin and Joseph L. Gastwirth, *Unconditional Versions of Several Tests Commonly Used in the Analysis of Contingency Tables*, 55 BIOMETRICS 264 (1999) for the extension to combination and trend tests.

40. See Edward L. Korn & Richard Simon, *Explained Residual Variation, Explained Risk and Goodness of Fit*, 45 AM. STATISTICIAN 201 (1991); J.G. Pigeon & Joseph F. Heyse, *An Improved Goodness of Fit Statistic for Probability Prediction Models*, 41 BIOMETRICAL J. 71 (1999); Efstathia Bura & Joseph L. Gastwirth, *The Binary Regression Quantile Plot: Assessing the Importance of Predictors in Binary Regression Visually*, 43 BIOMETRICAL J. 1 (2001). Basic texts that could be cited in teaching materials include P.W. HOSMER & STANLEY LEMESHOW, *APPLIED LOGISTICAL REGRESSION* (1989) and DAVID G. KLEINBAUM, *LOGISTIC REGRESSION: A SELF-LEARNING APPROACH* (1994). For more technical discussions, see PETER MCCULLAGH & JOHN A. NELDER, *GENERALIZED LINEAR MODELS* (1989) and ALAN AGRETI, *CATEGORICAL DATA ANALYSIS* 84-96 (1990).

could then be used to assess the role of age, assuming that seniority was a proper substitute for loyalty. It is important for the judiciary to appreciate that often there are several reasonable approaches that typically yield similar, but not identical, results.

### III. SUGGESTED CHANGES

Although this comment has raised several concerns about the appropriateness of the analysis in the context of the prototype, the problems can be remedied by modifying the explanation offered by the defendant. As we saw, seniority reduces the apparent age effect to a nonstatistically significant one using two alternative approaches based on logistic regression. The issues then became (1) the appropriateness of using seniority as a proxy for loyalty, and (2) the lack of the most objective data—job evaluations—in the model. If the FJC replaced seniority by the average of the last two years of job evaluations, then the logistic analyses would be proper. This is especially true here, as the former management made those evaluations and, consequently, no “bias” in them can be attributed to the new management. Then the “pretext” phase might concern whether the new management really used the evaluations or “adjusted” them in some manner. These are factual, rather than statistical, issues.

By modifying the scenario, the example could then introduce some of the basic assumptions underlying the various statistical procedures. This should be germane to other uses of two-by-two tables and logistic regression as legal evidence. Many epidemiological studies are submitted as evidence in toxic tort and environmental cases, so a basic knowledge of how the analysis relates to the way the data were collected should be useful to judges.

Finally, as the discussion of *Bew*<sup>41</sup> reminds us, statistical significance depends on the sample size as well as magnitude of the difference. Hopefully, other parts of the instructional materials will discuss this as well as the important issue of the power, or ability to detect a true difference, of a statistical test.<sup>42</sup>

41. See *supra* notes 33–38 and accompanying text.

42. See MICHAEL O. FINKELSTEIN & BRUCE LEVIN, *STATISTICS FOR LAWYERS* 186–188 (1990); 1 JOSEPH L. GASTWIRTH, *STATISTICAL REASONING IN LAW AND PUBLIC POLICY* 180–84, 257–58 (1988); David H. Kaye & David A. Freedman, *Reference Guide in Statistics*, in *REFERENCE MANUAL ON SCIENTIFIC EVIDENCE* 85–177 (2d ed. 2000). In this comment, we adopted the 0.05 level commonly used in the social sciences to determine when a result is statistically significant. In contrast to those studies where the researcher often can decide on the sample size in the discrimination setting, the numbers of employees in the various protected groups have been determined by the employer's reaction to economic circumstances. To be fair to both parties, the power of the test used to detect a meaningful disparity should be considered in setting the cut-off level for significance. See RICHARD A. POSNER, *FRONTIERS OF LEGAL THEORY* 373–74 (2001) (noting that because there is no special legal significance in the 0.05 level and because the 0.05 convention is rooted in considerations unrelated to litigation, statistical evidence not reaching it should not be excluded).