STATISTICAL SIGNIFICANCE AND STATISTICAL ERROR IN ANTITRUST ANALYSIS

PHILLIP JOHNSON
EDWARD LEAMER
JEFFREY LEITZINGER*

Proof of antitrust impact and estimation of damages are central elements in antitrust cases. Generally, more is needed for these purposes than simple observational evidence regarding changes in price levels over time. This is because changes in economic conditions unrelated to the behavior at issue also may play a role in observed outcomes. For example, prices of consumer electronics have been falling for several decades because of technological progress. Against that backdrop, a successful price-fixing conspiracy may not lead to observable price increases but only slow their rate of decline. Therefore, proof of impact and estimation of damages often amounts to sorting out the effects on market outcomes of illegal behavior from the effects of other market supply and demand factors.

Regression analysis is a statistical technique widely employed by economists to identify the role played by one factor among those that simultaneously determine market outcomes. In this way, regression analysis is well suited to proof of impact and estimation of damages in antitrust cases. For that reason, regression models have become commonplace in antitrust litigation.[1]

In our experience, one aspect of regression results that often attracts specific attention in that environment is the statistical significance of the estimates. As is discussed below, some courts, participants in antitrust litigation, and commentators maintain that stringent levels of statistical significance should be a threshold requirement for the results of regression analysis to be used as evidence regarding impact and damages. They do so from two differ-

* Phillip Johnson, Econ One Research; Edward Leamer, Chauncey J. Medberry Chair in Management, Professor in Economics and Statistics, UCLA; Jeffrey Leitzinger, Econ One Research.

[1] Jonathan B. Baker & Daniel L. Rubinfeld, *Empirical Methods in Antitrust Litigation: Review and Critique*, 1 AM. L. & ECON. REV. 386, 387 (1999).

ent perspectives. First, it is argued that strict requirements on levels of statistical significance provide a necessary and appropriate limit on the frequency with which statistical results showing impact and damages are simply a statistical sampling accident. In a setting where the alleged anticompetitive behavior had no actual impact, the probability of getting a positive damage estimate is still 50 percent. More generally, the probability of a positive damage estimate approaches 50 percent, regardless of the true damages, as the damage estimate gets more and more imprecise. In this context, a statistical significance threshold provides a safeguard against false findings of damages. Second, proponents of statistical significance thresholds argue that the economics profession treats stringent levels of statistical significance as a necessary element for purposes of accepting regression-based results as valid, and that legal rules associated with expert evidence should require nothing less.

On the other hand, as is also described below, other scholars, antitrust practitioners, and courts maintain that a proper understanding of statistical significance argues against adoption of conventional statistical significance thresholds as an evidentiary requirement. Instead, according to this side of the debate, the inferences to be drawn from a regression result depend not only on its statistical significance but also on its interplay with other evidence in the case. Inflexible statistical significance requirements may unduly limit the information available to properly decide the case.

We count ourselves on this side of the argument. We think that the statistical standards need to fit the circumstances. We think it appropriate, for example, that a penalty in a criminal case requires a higher evidentiary standard than an antitrust damage award. In addition, the evidentiary standard should apply to the totality of the evidence, which means that the statistical regression-based evidence needs to be more conclusive if the rest of the evidence is weak, but less conclusive if the rest of the evidence is strong.

There is growing awareness within the economics and statistics professions that conventional significance thresholds have little real claim to act as standards to legitimize regression results, despite the widespread attention they receive.[2] Moreover, the evidentiary thresholds associated with proof of impact and estimation of damages in an antitrust case may differ from the confidence thresholds implicit in conventional significance measures. (We elaborate on the reasons for this potential misalignment below.) In that regard, a rule that requires statistical evidence of impact and damages to meet a stringent statisti-

---

[2] As detailed further below, the American Statistical Association (ASA) has recently taken a formal position critiquing the pervasive use of arbitrary thresholds, i.e., conventional statistical significance thresholds. Press Release, Am. Statistical Ass'n, American Statistical Association Releases Statement on Statistical Significance and P-values (Mar. 7, 2016) [hereinafter American Statistical Association Press Release].

cal significance threshold potentially could preclude regression results that nonetheless possess sufficient evidentiary weight to legally carry the day.

In recent years, courts have moved on various fronts to clearly articulate the nature of evidence necessary to sustain antitrust actions. As regression analysis plays an increasingly central role in those cases, courts soon will need to decide the proper role for statistical significance. The stakes for the conduct of antitrust litigation will be very high. Accordingly, the time is ripe to fully examine the underlying issues.

The purpose of this article is to contribute to that process. While the critics of conventional thresholds for statistical significance may offer a compelling argument about what not to do, they have little or nothing to say about what should be done instead. Faced with that lack of an alternative, it is therefore unsurprising that many practitioners have adopted the conventional approach. While we cannot offer a mechanical alternative to the traditional mechanical thresholds, what we do offer is a way of thinking about the choice of thresholds that embodies the non-statistical evidence as well as the evidentiary standard that may favor either the defense or the plaintiff.

Below, we discuss the intellectual foundations of statistical significance thresholds, alternative ways of viewing "significance," loss tradeoffs associated with inferential decision making, and the nature of evidentiary burdens (both implicit in conventional statistical significance levels and explicit in legal standards). Our recommendation is that regression evidence be viewed contextually, based both upon its economic significance and its statistical significance. Further, we recommend that such significance (in both respects) has to be evaluated against the backdrop of other evidence in the case that tends to make the specific implications of the regression results more or less plausible. More broadly, in deciding whether and how to use the regression results in antitrust matters, we urge an approach that explicitly recognizes not just the prospect of false positives (the focus of statistical significance) but also the consequent implications of any decision rule for false negatives. We conclude by offering an integrated Bayesian decision framework in which all of these elements can be incorporated.

I. THE ROLE OF REGRESSION ANALYSIS

In addition to establishing the presence of behavior that is illegal under the antitrust laws, the accuser, whether a government entity or a private antitrust plaintiff, faces two further requirements to obtain monetary recovery from the defendant(s). The accuser must establish impact—i.e., that there has been injury by the illegal behavior—and generally must also be able to provide a reasonable quantification of the damages that flow from that behavior. (In a government action, fines may be based on the benefits the accused received or

the damages caused.[3]) The focus in many antitrust cases, both for impact and damages, is on the extent to which the illegal behavior altered prices. As a matter of basic economics, prices are expected to reflect market supply and demand characteristics, one of which is the degree of competition operating on both sides of the market. Given the role played by competition in price formation, behavior that materially limits competition can be expected to impact prices.

However, that impact usually occurs against a market backdrop in which changes in other market characteristics also have affected prices. Those characteristics may include input costs, prices for substitutes and complements, factors that drive the willingness and propensity to pay on the part of buyers, the quality and extent of market information, and governmental rules and regulations. Thus, the impact of allegedly anticompetitive behavior may be obscured or incorrectly suggested by price movements tied to changes in other supply and demand factors. Even where price movements occurring in conjunction with challenged behavior are consistent with (and therefore supportive of) impact and damages, a plaintiff relying solely on those movements as proof of impact or damages typically faces a counter-argument from the defendant(s) that those movements were attributable (all or in part) to other market factors.[4] As a result, simple inspection of prices over time is often not sufficient for purposes of assessing impact or damages.

Regression analysis is a widely used and accepted statistical tool for identifying the relationship between a market outcome and other market factors thought, at least potentially, to have some causal relationship with that outcome. In performing regression analysis, one embeds data for the market outcome (of interest in the matter at hand) and other likely causal factors in a model specification and then uses statistical methods to identify the relationships between the outcome and those other factors. The regression produces estimated coefficients linking changes in each factor to changes in the market outcome. Given the presence of certain fairly general statistical properties within the underlying data, these coefficients have statistically attractive characteristics as estimates of the impact of each of those factors on the market outcome.[5]

---

[3] 18 U.S.C. § 3571(c)(1)–(2), (d); 15 U.S.C. § 1; *see* William H. Page, *Impact: Injury and Causation*, *in* ABA SECTION OF ANTITRUST LAW, PROVING ANTITRUST DAMAGES: LEGAL AND ECONOMIC ISSUES 17–18 (2d ed. 2010).

[4] For additional resources on the application of econometric techniques in antitrust, see, e.g., ABA Section of Antitrust Law, Econ. Comm., Selected Readings in Antitrust Economics: Applied Econometrics (Apr. 2014).

[5] In technical parlance, they provide the best linear unbiased estimate. JEFFREY M. WOOLDRIDGE, INTRODUCTORY ECONOMETRICS: A MODERN APPROACH 101–02 (5th ed. 2013).

Turning specifically to the antitrust context, if one designs a regression model to explain price levels and also includes in the model variables representing other supply and demand factors, along with a variable that in some fashion (for instance, by time period) captures the illegal behavior alleged in the case, the coefficient associated with that behavior variable then provides an estimate of the impact of the alleged illegal behavior on prices, holding constant the effects of other market factors. Obviously, such an estimate has direct relevance to the issues of antitrust impact and damages. Similarly, regression analysis can be used to estimate the impact of alleged illegal behavior on other market outcomes, such as wages, output, or product/service offerings.

This ability to distinguish (at least statistically) the effects of illegal behavior from other market factors is why regression analysis is so often brought into the antitrust courtroom. Indeed, the ABA noted almost ten years ago that "[e]conometric and statistical analysis of data have come to play an important role in antitrust analysis."[6] As noted by Daniel Rubinfeld, "[J]udicial interest in using statistical methods also has been growing rapidly. Courts are finding, to a greater and greater degree, that reliable statistical evidence can be invaluable in deciding questions of impact, harm, and damages in a range of cases, including antitrust."[7]

Regression analysis can be an especially useful analytical tool in class action antitrust cases, where common methods of proof are important.[8] In particular, a single regression model can provide evidence that is common to class members. Moreover, a regression model also can be designed to analyze the results of illegal behavior by location, by product, or even by customer. Not surprisingly, then, when it comes to impact and estimates of damages, "class certification cases have relied on statistical analyses, including econometrics."[9]

---

[6] ABA SECTION OF ANTITRUST LAW, ECONOMETRICS 116 (Lawrence Wu ed., 1st ed. 2005); *see also* ABA SECTION OF ANTITRUST LAW, ECONOMETRICS 9 (Lawrence Wu ed., 2d ed. 2014) [hereinafter ABA, ECONOMETRICS SECOND ED.].

[7] Daniel L. Rubinfeld, *Market Definition with Differentiated Products: The Post/Nabisco Cereal Merger*, 68 ANTITRUST L.J. 163, 164 (2000).

[8] Current legal interpretations of the Federal Rules of Civil Procedure requirement that "questions of law or fact common to class members predominate over any questions affecting only individual members" focus attention on the availability and reliability of a method of proof of damages that does not require special treatment of individual members of the class and that is based on analysis of facts and data of the case, not a presumption of impact solely from theory. FED. R. CIV. P. 23(b)(3). *See, e.g.*, *In re* Hydrogen Peroxide Antitrust Litig., 552 F.3d 305, 310 (3d Cir. 2008).

[9] ABA, ECONOMETRICS SECOND ED., *supra* note 6, at 195.

II. THE STATISTICAL SIGNIFICANCE ISSUE

The existence of unexplained variation (which is inescapable as a practical matter)[10] means that coefficients estimated in a regression model are subject to statistical uncertainty. In effect, the estimates are drawn randomly from a distribution of potential estimates centered on the true value of the coefficients. Therefore, it is possible, purely as a statistical matter, to have an estimated coefficient that indicates a relationship between the variable representing the challenged conduct and prices where none exists in fact.

To see this graphically, Figure 1 illustrates a hypothetical probability distribution for potential coefficient estimates when the true coefficient is zero. This distribution of estimates around zero reflects that stochastic variability that would occur if the experiment were repeated over and over—for example, repeated random samples of size 50. The distribution is concentrated close to zero or spread widely apart depending on the quality of the experiment being studied—for example, as the sample size increases the distribution becomes more concentrated around zero.

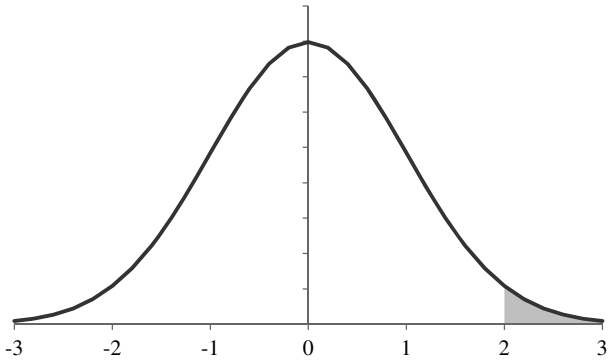

FIGURE 1:
A PROBABILITY DISTRIBUTION FOR A COEFFICIENT ESTIMATE WITH NO ACTUAL RELATIONSHIP

As is shown in Figure 1, notwithstanding the absence of an underlying relationship, there is a 50 percent chance of an estimated positive coefficient. This high probability of a finding of impact (or overcharge) when there was none (a false positive, also known as a Type I error) leads one to consider ways to

---

[10] Models must necessarily leave out various elements of reality to aid understanding. (If they didn't they would be reality itself.) The objective is that the elements omitted from the model do not affect substantially the relationships of interest captured by the model. As noted by George Box, "All models are wrong, some are useful." GEORGE E.P. BOX & NORMAN R. DRAPER, EMPIRICAL MODEL-BUILDING AND RESPONSE SURFACES 424 (1987).

control the probability of that outcome. Referring to Figure 1, if instead of deciding in favor of the hypothesis of impact if the estimate is positive, one required an estimate in excess of 2, the probability of a false positive would be far smaller, equal to the shaded area in the figure which is 2.2 percent of the total area under the curve. Under these circumstances, for example, a positive estimate of 1.4, which is smaller than 2, could be described as "not statistically significant at the 2.2 percent level." To that estimate of 1.4 one can attach a p-value, which is the just statistical significance found by acting as if the threshold were 1.4, not 2.0. The p-value attaching to an estimate of 1.4 is the probability to the right of 1.4 in Figure 1, which is 8.1 percent. Since the p-value is equal to 8.1 percent, the estimate of 1.4 could be described as statistically significant at the 10 percent level, but not the 5 percent or the 1 percent level, referring to the three conventional significance levels.[11]

There is no question that a p-value, properly understood, conveys important information regarding the accuracy of an estimate in relation to the hypothesis of no effect. It is also true that the three conventional levels for statistical significance (1%, 5%, and 10%) can facilitate a conversation about the statistical accuracy of an estimate, allowing the convenient asterisk notation: one, two, or three asterisks for a coefficient statistically significant at the 10%, 5%, or 1% level, respectively. But reporting results is not the same as using the conventional significance levels to make actual decisions. Still, over many years of conventional reporting (both in legal settings and in professional journals) regarding whether or not regression estimates achieve these significance levels, there is a strong impression, if not a reality, that economists view those significance levels as minimum thresholds for valid statistical proof, irrespective of the particular context.

This leads to the statistical significance issue that is the focus of this article. What role should statistical significance play in the use of regression models for purposes of proving impact and damages in antitrust litigation? Does the presence of conventional statistical significance levels (10 percent or better) in academic literature mean that the courts should conclude that there is no impact when the statistical significance of the variable linking prices to the alleged illegal behavior does not reach those same levels? For example, should the courts refuse to allow a jury to even consider a regression-based damages

[11] It is important to recognize that statistical significance does not convey the everyday sense of significance which conveys importance or meaningfulness. Statistical significance means simply that, relative to the amount of statistical noise associated with the estimate, it falls a significant distance away from zero. A dataset that is well explained by a regression model (and therefore exhibits little statistical noise) might produce coefficient estimates that, in their practical implications, are really no different than zero—and therefore of little importance—but statistically significant nonetheless.

estimate that does not achieve at least a 10 percent statistical significance level?

III. ACADEMIC AND LEGAL CONSIDERATION OF CONVENTIONAL SIGNIFICANCE THRESHOLDS

As statistical analysis has become more commonplace in business and legal settings, and attention to statistical significance measures has grown, many have raised questions about the scientific legitimacy of those conventions.

A. Statisticians Recognize the Lack of Foundation for Conventional Significance Thresholds

The origin of the convention apparently began with Ronald Fisher's observation that a one in twenty occurrence was a rare event. Thus, the 0.05 threshold to make the rejection of a true null hypothesis is a rare occurrence. However, Fisher criticized the unthinking adoption of this threshold, noting that:

> [T]he calculation [at the 1% level] is absurdly academic, for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas. It should not be forgotten that the cases chosen for applying a test are manifestly a highly selected set, and that the conditions of selection cannot be specified even for a single worker; nor that in the argument used it would clearly be illegitimate for one to choose the actual level of significance indicated by a particular trial as though it were his lifelong habit to use just this level.[12]

In his econometrics textbook, Peter Kennedy calls hypothesis testing misleading,[13] observing that many believe that "hypothesis testing is overrated, overused, and practically useless as a means of illuminating what the data in some experiment are trying to tell us."[14] With regard to the widespread use of 5 percent statistical significance thresholds, Kennedy observes that, "[T]here is no good reason why 5% should be preferred to some other percentage. The father of statistics, R.A. Fisher, suggested it in an obscure 1923 paper, and it has been blindly followed ever since."[15] In another classic statistics textbook, Lehmann and Romano write,

> By habit, and because of the convenience of standardization in providing a common frame of reference, these values gradually became entrenched as

---

[12] Ronald A. Fisher, Statistical Methods and Scientific Inference 45 (3d ed. 1973).

[13] Peter Kennedy, A Guide to Econometrics 60–61 (6th ed. 2008) (citing Geoffrey R. Loftus, *A Picture Is Worth a Thousand p levels: On the Irrelevance of Hypothesis Testing in the Microcomputer Age*, 25 Behavior Research Methods, Instruments and Computers 250 (1993); Marks R. Nester, *An Applied Statistician's Creed*, 45 Applied Statistics 401 (1996)).

[14] Loftus, *supra* note 13, at 250 (quoted in Kennedy, *supra* note 13, at 61).

[15] Kennedy, *supra* note 13, at 60.

> the conventional levels to use. This is unfortunate, since the choice of significance level should also take into consideration the power that the test will achieve against the alternatives of interest. There is little point in carrying out an experiment which has only a small chance of detecting the effect being sought when it exists.[16]

Criticism continues. For example, a recent article in the *Journal of Empirical Finance* by Jae H. Kim and Philip Inyeob Ji chronicles criticisms that have been directed at significance testing over the years, such as:

> (i) arbitrary choice of the level of significance; (ii) little consideration of the power (or Type II error) of test; (iii) confusion between statistical and substantive importance (economic significance); and (iv) the practice of "sign econometrics" and "asterisk econometrics" with little attention paid to effect size. Despite these continuing criticisms, it appears that the practice of significance testing has not improved noticeably.[17]

Recently, *The Journal of the American Statistical Association*, one of the premier statistical journals, issued the following commentary on statistical significance and p-values:

> "The p-value was never intended to be a substitute for scientific reasoning," said Ron Wasserstein, the ASA's executive director. "Well-reasoned statistical arguments contain much more than the value of a single number and whether that number exceeds an arbitrary threshold. The ASA statement is intended to steer research into a 'post $p<0.05$ era.'" . . .
>
> The statement's six principles, many of which address misconceptions and misuse of the p-value, are the following:
>
> 1. P-values can indicate how incompatible the data are with a specified statistical model.
>
> 2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
>
> 3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
>
> 4. Proper inference requires full reporting and transparency.
>
> 5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
>
> 6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.[18]

---

[16] E.L. LEHMANN & JOSEPH P. ROMANO, TESTING STATISTICAL HYPOTHESIS 57 (3d ed. 2005); *see also* MORRIS H. DEGROOT & MARK J. SCHERVISH, PROBABILITY AND STATISTICS 617–20, § 9.9 (4th ed. 2012).

[17] Jae H. Kim & Philip Inyeob Ji, *Significance Testing in Empirical Finance: A Critical Review and Assessment*, 34 J. EMPIRICAL FIN. 1, 2 (2015).

[18] American Statistical Association Press Release, *supra* note 2.

B. LEGAL ATTITUDES TOWARDS CONVENTIONAL SIGNIFICANCE THRESHOLDS

The treatment of statistical significance in legal settings has been mixed. The *Reference Guide on Multiple Regression* published by the Federal Judicial Center comments on the reporting of a conventional level of statistical significance, observing that "[a]lthough the 5% criterion is typical, reporting of more stringent 1% significance tests or less stringent 10% tests can also provide useful information."[19] The court in *FTC v. Swedish Match North America*, for example, rejected an econometric analysis where the observed level of significance was 15 percent.[20] In *Sanner v. Board of Trade*[21] the court ruled that a demanding level of statistical significance is important to show that the expert's regression analysis is reliable.[22] In the *Photochromic Lens Antitrust Litigation*, the district court noted that "[t]he Magistrate Judge also found that [the expert's] use of a 50% statistical significance measure in his regressions, rather than the more rigorous 5% measure, rendered his models incapable of providing a reliable, working methodology through which [plaintiffs] could prove impact."[23]

However, the Supreme Court in *Matrixx* rejected the premise that "statistical significance is the only reliable indication of causation" because "[s]tatistically significant data are not always available" and the phenomenon being examined can be "subtle or rare" such that experts in the relevant field must rely on other tools.[24] In *High-Tech Employees*, the district court concluded:

> Defendants have not cited, nor has this Court found, any case holding that a regression model must reject a null hypothesis of zero effect at least at the 10% significance level in order to be admissible. In fact, there is authority holding otherwise. . . . See, e.g., Cook, 580 F. Supp. 2d at 1102, 1105 (rejecting argument that "statistical significance is a threshold requirement for establishing the admissibility of expert testimony involving the use of statistics" and holding that neither "the Tenth Circuit ([nor] any other court) has adopted a rule barring admission of any epidemiological study that was not statistically significant at the 95–percent confidence level."); Kadas, 255 F.3d at 362 (rejecting the idea that a study is inadmissible as a matter of law just because it is less statistically significant than the 5 % level).[25]

---

[19] Daniel L. Rubinfeld, *Reference Guide on Multiple Regression*, *in* REFERENCE MANUAL ON SCIENTIFIC EVIDENCE (THIRD) 303, 320 (2011).

[20] FTC v. Swedish Match, 131 F. Supp. 2d 151, 161 (D.D.C. 2000).

[21] Sanner v. Bd. of Trade, No. 89 C 8467, 2001 WL 1155277 (N.D. Ill. Sept. 28, 2001).

[22] ABA, ECONOMETRICS SECOND ED., *supra* note 6, at 35 n.39.

[23] *In re* Photochromic Lens Antitrust Litig., No. 8:10-CV-00984-T-27EA, 2014 WL 1338605, at *24 (M.D. Fla. Apr. 3, 2014).

[24] Matrixx Initiatives, Inc. v. Siracusano, 131 S. Ct. 1309, 1319 (2011) (citations omitted).

[25] *In re* High-Tech Employee Antitrust Litig., No. 11-CV-02509-LHK, 2014 WL 1351040, at *15 (N.D. Cal. Apr. 4, 2014) (internal footnote omitted). Two of the authors, Drs. Johnson and

As regression analysis assumes an expanding role in antitrust cases, the treatment by the courts of statistical significance, both for evaluating regression-based evidence and in discharging their gate-keeping role for the admissibility of regression evidence, is becoming increasingly important.

IV. PROBLEMS WITH ADOPTING CONVENTIONAL STATISTICAL SIGNIFICANCE THRESHOLDS AS EVIDENTIARY STANDARDS

A. THE NULL HYPOTHESIS OF NO EFFECT MAY HAVE LITTLE RELEVANCE

As explained above, a p-value is the probability of observing the estimated coefficient or larger under the assumption that there is no true relationship. In the parlance of statistics, that assumption is referred to as the null hypothesis. Since the coefficient can take on other values, there has to be some reason to single out the zero value for special treatment. One reason for special treatment is that the number zero may have "truth value," while all the other possible values of the coefficient may not. For example, cases may allege acts that either did not occur or did not cause damages, while the chance of damages exactly equal to any other precise figure—e.g., $8,786,529.52 (an Excel-driven random number), is vanishingly small. Zero damages has truth value; $8,786,529.52 does not.

A truth value is a probability that may be any number between zero (impossible) and one (certain). The scientific truth value of the null hypothesis should play a role in determining the level of statistical significance. If scientific opinion has the coefficient almost certainly equal to zero, it should take strong evidence to change our minds, whereas weaker evidence should be enough to alter our opinions if the scientific truth value of the zero hypothesis is small.

To express this differently, a regression analysis can serve two purposes: first, determining if the coefficient is different from zero (hypothesis testing); and second, if the first hurdle is passed, determining the value of the coefficient (estimation). A statistician's function in carrying out a hypothesis test is to report whether the data are statistically compatible with the hypothesis of no effect. A statistician's function in carrying out estimation is to report the best estimate and range of estimates that are compatible with the data—what is called a confidence interval. Below we explain how the optimal damage award can wisely use the results of both functions.

Leamer, were engaged by Class Plaintiffs in this litigation; Dr. Leamer testified on damages and common impact.

The relative importance of hypothesis testing versus estimation should depend on the scientific truth value of the null hypothesis. If the scientific truth value of the null hypothesis is great, emphasis should be put on hypothesis testing, and the data must play two roles: determining if the coefficient is different from zero (hypothesis testing) and, if zero is rejected, selecting a value for the coefficient (estimation). If the scientific truth value of the null hypothesis is small, the emphasis should be on estimation, with little energy expended on studying the null hypothesis of no effect.

Measures of statistical uncertainty including p-values can play a role in estimation as well as in hypothesis testing, and the presence of p-values in traditional reporting of regression results does not in any way imply that the null hypothesis of zero effect is especially relevant. A small p-value means that the statistically supportable estimates within a confidence interval are narrowly clustered around the best estimate, while a large p-value or small t-value means that the range is wide. This is where the conventional levels of p-values are undoubtedly helpful in creating standardized language for describing the width of the confidence intervals: limited, narrow, and very narrow, corresponding with p-values of 0.10, 0.05 and 0.01.

The antitrust litigation terms, impact and damages, correspond imperfectly with the two statistical activities: hypothesis testing and estimation. The evidence in litigation is both numerical and textual; the latter refers to historical documents and testimony. The textual information speaks primarily to impact: was there illegal behavior and did it likely have an impact? It is useful to suppose that the textual information can be summarized in one number: the scientific truth value of the hypothesis of no impact. This summary of the documents can be passed on to the statistician who uses it to conduct an analysis of the numerical data. If the documents are strong enough by themselves to establish impact, then the statistician uses the data only to estimate the damages, not to test the hypothesis of no impact. If the documents are weak, then the data must be strong enough both to reject the hypothesis of no impact and also to estimate the level of damages with adequate accuracy.

B. Economic Significance Is Not the Same as Statistical Significance

The word "significant" in everyday language is used to convey importance or consequence, but the word pair "statistically significant," properly used and understood, means measurable, not consequential. Thus, for example, a small and therefore inconsequential effect can be statistically significant, if the sample size is large enough to allow an accurate estimate of the small magnitude, and a large and consequential effect can be found statistically insignificant if the sample is too small to allow accurate estimation. The failure to distinguish carefully the difference between measurable and important has led to the ad-

monition by Stephen Ziliak and Deirdre McCloskey that "[i]nsignificant does not mean unimportant."[26] At least one academic journal has actually banned significance testing because it confuses the conversation.[27]

Daniel Rubinfeld explained the distinction between economic and statistical significance this way:

> Often, results that are practically significant are also statistically significant. However, it is possible with a large data set to find statistically significant coefficients that are practically insignificant. Similarly, it is also possible to obtain results that are practically significant but statistically insignificant. Suppose, for example, that a regression analysis suggests that prices are 7 percent higher in the period in which the alleged anticompetitive activity took place. If the data are such that only three or four years of data are available outside the period of alleged wrongful behavior, the 7 percent difference could be practically significant yet statistically insignificant.[28]

Professor Rubinfeld's point can be stated more broadly. All else equal, levels of statistical significance are closely related to the size of the underlying data set. Coefficient estimates will almost always be statistically significant when the data evidence is abundant.[29] In the practical environment of antitrust analysis, where historical data regarding long-running conspiracies can be difficult to resurrect and are sometimes plagued with missing information, the data limitations can be very important. Hence, it is not at all uncommon when conducting statistical analysis in antitrust matters to obtain results that have potentially important practical implications but, for lack of data, do not meet conventional thresholds for statistical significance. A rule rejecting the use of regression estimates that do not achieve conventional significance levels could well make the effectiveness of antitrust enforcement in any given case dependent upon the quantity and quality of historical data. Given the reality that the antitrust violators often control the relevant data, this situation would be problematic.

---

[26] STEPHEN T. ZILIAK & DEIRDRE N. MCCLOSKEY, THE CULT OF STATISTICAL SIGNIFICANCE: HOW THE STANDARD ERROR COSTS US JOBS, JUSTICE, AND LIVES 43 (2008).;

[27] David Trafimow & Michael Marks, *Editorial*, 37 BASIC & APPLIED SOC. PSYCHOL. 1, 1 (2015). "The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it. However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. From now on, BASP is banning the NHSTP." *Id.* (citation omitted). Instead BASP journal requirements will focus on good descriptive statistics, including on the size of an effect and information on its frequency or distribution in the data. David Trafimow, *Editorial*, 36 BASIC & APPLIED SOC. PSYCHOL. 1, 1 (2014).

[28] Daniel L. Rubinfeld, *Quantitative Methods in Antitrust, in* 1 ABA SECTION OF ANTITRUST LAW, ISSUES IN COMPETITION LAW AND POLICY 723, 738–39 (W. Dale Collins ed., 2008).

[29] Richard Lempert, *The Significance of Statistical Significance*, 34 L. & SOC. INQUIRY 225, 234–45 (2009). Lempert describes an example of this phenomenon drawn from the pharmaceutical arena.

The interplay between data issues and levels of statistical significance can be especially important when regression analysis is used to analyze impact or damages for individual customers, say for instance in a class action antitrust case. As noted above, regression analysis is an important tool for isolating the effects of anticompetitive behavior from the effects of other market factors. But the amount of transactional information available for a given customer over time is often limited. Hence, the need to account for other market factors (particularly where there are many at play) may use up much of the available customer data, making statistical significance difficult to achieve.[30] As a result, were estimates that achieve conventional significance deemed legally necessary for purposes of proving impact (or damages), regression would frequently be unavailable as a means a proof at the individual customer level simply because of limited data.

One solution is just to turn this confusing situation over to the judge and jury. As Rubinfeld and Steiner put it:

> The real question is whether [the estimate] marks a measurable effect of the variable. . . . Rather than choose any particular significance level, a better procedure might be to state that [the estimated value is] the best estimate . . . and . . . the probability of getting this sample result when the true parameter [would indicate that there are no damages]. The court and the jury are then left with the problem of evaluating the importance of the statistical results, rather than leaving the decision entirely in the hands of the expert.[31]

But courts have properly asked whether it really makes sense to leave it to juries (or judges) to sort through all of this. Hence, there is an understandable desire to find clearly applicable standards like conventional significance thresholds. We recognize the need for practical guidance as an important challenge with regression evidence, but believe that strict adherence to conventional significance thresholds is not the answer. We offer our suggestions below.

C. The Type II Error Should Also Be Considered When Choosing the Significance Level

Even if one accepts null-hypothesis testing as relevant for impact and damage analysis in antitrust cases, there remains another clear problem with insisting upon conventional significance thresholds in regression analysis used for

---

[30] *See* ABA, ECONOMETRICS SECOND ED, *supra* note 6, at 359.

[31] Daniel L. Rubinfeld & Peter O. Steiner, *Quantitative Methods in Antitrust Litigation*, 46 L. & CONTEMP. PROBS. 69, 100 (1983). Zvi Griliches has noted that "[h]ere and subsequently, all statements about statistical 'significance' should not be taken literally. . . . Tests of significance are used here as a metric for discussing the relative fit of different versions of the model. In each case, the actual magnitude of the estimated coefficients is of more interest . . . ." ZVI GRILICHES, R&D AND PRODUCTIVITY: THE ECONOMETRIC EVIDENCE 89–90 n.2 (1998); *see also* ZILIAK & MCCLOSKEY, *supra* note 26, at 111–12.

those purposes. There are two kinds of errors that can be made when testing a null hypothesis, not just one. The null hypothesis can be wrongly rejected when it is in fact true (the Type I error addressed by statistical significance), and the null hypothesis can fail to be rejected when it is in fact false (referred to as Type II error). In the legal setting, Type I error would occur by awarding damages when there were none. Type II error occurs when antitrust violators do not pay damages.

Very occasionally, analysts discuss Type II error—accepting the null hypothesis when it is false, e.g., failing to make an award when there actually were damages. We believe Type II error (and its counterpart, the power of the test, which is the probability of rejecting a null hypothesis when it is false) warrants more attention, especially when the null hypothesis has relatively low truth value (i.e., it is highly unlikely there was no impact and damages). For instance, suppose one uses a regression model to determine the impact of a proven cartel on prices. The conventional significance approach would be to accept the regression results as evidence of impact only if the relevant coefficient(s) achieved conventional significance levels. But zero-effect is not necessarily the only relevant hypothesis in this setting. Another hypothesis could come from the many empirical studies of overcharges occurring in actual cartels.[32] Suppose that one draws from those studies the conclusion that the typical overcharge percentage associated with acknowledged cartels is 10 percent. Given that, we would argue that it makes more sense to ask statistically whether the regression results allow one to reject the null hypothesis that, like cartels typically, this cartel produced a 10 percent overcharge, in favor of the alternative hypothesis that the overcharge was zero. As an inferential matter, this approach has no less validity than the conventional hypothesis testing approach. Indeed, it is more in line with the presumption of liability that triggers the need to examine overcharges in the first place.

Once Type II error is explicitly recognized as something that merits attention, the fundamental problem arises: Interpreting the data more strictly to reduce the (conditional) probability of Type I error (the significance level) increases the (conditional) probability of Type II error, and vice versa. The solution for optimally trading off these two probabilities begins with the recognition that the significance level is the probability of Type I error (rejecting a true null hypothesis) conditional upon the truthfulness of the null hypothesis. The unconditional probability of Type I error is the significance level multiplied by the truth value of the null hypothesis. This means that if the null hypothesis has high truth value (e.g. a perceived likelihood of 90 percent or more), Type I error becomes much more likely and more caution—i.e., more

---

[32] John M. Connor, *Price-Fixing Overcharges: Revised 3d Edition* (2014), papers.ssrn.com/sol3/papers.cfm?abstract_id=2400780.

stringent levels of statistical significance—should be required before rejecting it. But if the null and alternative hypotheses have roughly similar truth values (50 percent), the prospect for Type II error increases, and less-demanding significance levels would be appropriate in deciding to accept regression results rejecting the null hypothesis. If the truth value of the null hypothesis is very small, the chance of rejecting a true null hypothesis is correspondingly small, and concern should shift toward the Type II error.

One other defect of the conventional approach (i.e., a fixed significance level) is that it leaves the conditional probability of Type I error completely independent of the strength of the data evidence. Any improvement in the quality or quantity of the evidence is devoted exclusively to reducing the conditional probability of Type II error. Thus for example, under the conventional approach a preliminary analysis (say, in the class certification phase of a case) using a small preliminary data set might have a 0.05 Type I error probability and a corresponding 0.50 Type II error probability. If later information improves and increases the size of the data set, the conventional Type I error probability would remain at 0.05 but the corresponding Type II error probability might be, for example, only 0.000005. A balanced approach requires the additional evidence to be used to reduce both probabilities, making the significance level a decreasing function of sample size.

D. The Relative Importance of Awards Too High and Awards Too Low Should Affect the Significance Level

The documents and the data together comprise the information on which a damage award is made. The way that information is ideally translated into an award depends on whether it is more important to avoid awards too high or awards too low, in other words, Type I and Type II errors. To discourage frivolous lawsuits, excessive awards should be avoided, but to discourage illegal acts, too small of an award should also be avoided.

The statistical uncertainty associated with a failure to pass a conventional significance test of a null hypothesis does not indicate an absence of damages. Rather, it reflects statistical uncertainty in the magnitude of the damage estimate, a state of evidence which could support a zero award in the hypothesis testing mode but an award—even one that is larger than the damage estimate—in the estimation mode. Damage awards larger than the best estimate should occur in this uncertain state when the documents and testimony make it virtually certain that impact from illegal acts has occurred and when it is more important to avoid awards that are too small than those that are too large. Zero damage awards should occur in this uncertain state when the documents and testimony leave impact uncertain and/or when it is more important to avoid awards that are too large than those that are too small. If these two kinds of errors are similarly consequential, and if the documents and testi-

mony make it virtually certain that impact from illegal acts has occurred, then the damage estimate should be the award, regardless of its statistical significance.

We illustrate below the different optimal choices of statistical significance when these two kinds of errors are equally consequential versus when an award being too high is ten times more consequential than an award too low.

E. POTENTIAL MISMATCHES IN THE EVIDENTIARY BURDEN

By setting statistical significance thresholds for regression-based evidence, one effectively establishes an evidentiary burden associated with that evidence. Insofar as the legal elements required in a private antitrust case, such as for impact and damages, have their own legally established evidentiary burdens, another potential difficulty with minimum statistical significance thresholds is that the implied burden for regression evidence could be in conflict with other existing legal burdens.

For instance, the evidentiary standard that must be met to prove impact is preponderance of the evidence. As described by J. Thomas Rosch, this "really just requires that the party bearing the burden of proof show that it is more probable than not that it has met the standard of proof it bears."[33] That is, evidence showing that more likely than not there was impact (i.e., some damages) would be sufficient legally to meet plaintiffs' evidentiary burdens.

Moreover, this burden refers to the combined evidentiary effect of documents, testimony, and statistical analysis viewed together. If this burden is already met by documents and testimony offered in the case, then a regression result will not be tasked with proving in isolation that the hypothesis of no impact is highly implausible. Yet, in effect, this is what a conventional statistical significance threshold would do. On the other hand, if the non-statistical evidence of impact is weak, more statistical support will be needed to satisfy the evidentiary requirement—i.e., greater levels of statistical significance will be needed.

Further, regression is used in antitrust matters in various kinds of cases (private class actions and governmental enforcement actions, for example) and at various stages of the case (class certification, motions, merits, etc.). And, as we understand it, the legal burdens faced by the parties in proving the

---

[33] J. Thomas Rosch, Comm'r, Fed. Trade Comm'n, Remarks Before the EU Competition Law and Policy Workshop, Observations on Evidentiary Issues in Antitrust Cases 4 (June 19, 2009). When it comes to proving the amount of damages, the Supreme Court has described the evidentiary standard as "show[ing] the extent of the damages as a matter of just and reasonable inference, although the results be only approximate." Story Parchment Co. v. Paterson Parchment Paper Co., 282 U.S. 555, 563 (1931).

requisite elements may vary by type of case, procedural stage, and issue.[34] There may well be settings within these combinations in which the evidentiary burden would align reasonably with conventional statistical significance thresholds. But it also seems clear, as outlined above, that this is not always the case. If so, then a conventional statistical significance threshold will be at odds with the applicable legal burden.

V. INTEGRATING REGRESSION EVIDENCE WITH OTHER EVIDENCE UNDER DIFFERENT DECISION RULES

We have discussed above the three factors (the strength of the non-statistical evidence, the strength of statistical evidence, and the relative importance of Type I and Type II errors) that should play a role in reaching decisions. In this Part, we present a stylized example to illustrate how these factors interact to determine an optimal damage award, and how that optimal award contrasts with the decisions based on the conventional significance criterion (i.e., award the estimated damage if it is statistically significant by conventional standards and zero otherwise). Our main argument is that the p-value standard should not be one of the conventional numbers (0.10, 0.05, or 0.01), but instead should be tuned to the circumstances. We recognize that this approach puts an extra burden on experts, judges, and juries to consider the circumstances and not just rely on conventional one-size-fits-all statistical standards. With that in mind, we offer suggestions below on how a court could reasonably and wisely deal with this burden and move the conversation between experts in a direction that is more meaningful to the court.

We take as inputs into the decision: (1) a hypothetical regression estimate indicating damages equal to a 10 percent overcharge and a corresponding p-value, (2) other case evidence summarized by a prior probability of positive damages, and (3) the relative social importance of awards too small versus awards too great. We show how the optimal award depends on these three factors.

We employ a Bayesian decision theoretic framework in which the strength of the regression results along with the truth value of the damages hypothesis (i.e., the prior probability of damages based on the other evidence) combine to determine a posterior probability distribution for the damage amount that summarizes all the evidence. Using this posterior probability distribution of damage amounts, we find the damage award that minimizes the expected loss as determined by the relative importance of Type I and Type II errors.

---

[34] The nature of damages may also vary, e.g., fines from enforcement actions or treble damages under private enforcement.

The first of the two decision rules we illustrate treats both types of error with equal weight (EBOE for equal balancing of error). Under this standard, the objective is to minimize the expected difference between the award and the true damages. That is, society accords equal weight to the error associated with the failure to award sufficient damages as to the error associated with overpayment. The second decision rule we illustrate below gives much more weight to Type I error, embodying a much higher level of concern with avoiding excessive damage penalties than under-penalizing firms that are guilty. For this illustration we assume that with this favor-low-over-high (FLOH) approach society (and courts) would be equally troubled by damages awards $1 in excess of true damages (Type I error) as awards $9 below actual damages (Type II error). This is a counterpart, effectively, to imposing a conventional 10 percent significance level as an evidentiary threshold.

As shown in the Appendix to this article, the optimal award under the EBOE standard is the median (50th percentile) of the damage probability distribution. Under the FLOH standard, the optimal award is the 10th percentile of the distribution (i.e., it is nine times more likely that true damages are greater than the award than that they are smaller than the award). To make a positive damage award requires more than 50 percent certainty of positive damages under the EBOE standards, but more than 90 percent certainty of positive damages under the FLOH standard.

Tables 1 and 2 show the optimal damage awards (given a regression point estimate for damages of 10 percent) associated with varying degrees of strength both in the non-statistical evidence and in the regression result. Table 1 contains the optimal awards based on the EBOE standard and Table 2 uses the FLOH standard. The hypothetical strength of the non-statistical evidence varies across rows in these tables, beginning in the first row with weak non-statistical evidence which determines only a 10 percent chance of positive actual damages and ending in the last row with strong non-statistical evidence that makes it 90 percent certain that damages occurred. The hypothetical strength of the statistical evidence varies across the columns in these tables, beginning in the first column with weak statistical evidence yielding a p-value of 0.75 and ending in the last column with strong statistical evidence and a p-value of 0.01. In both tables the optimal award is higher when either the non-statistical evidence or the statistical evidence is stronger.

With the EBOE standard, the optimal award is at or close to the damage estimate of 10 either when the data evidence is strong enough to make the p-value 0.01 or lower, or when the non-statistical evidence is strong enough that it supports a truth value of the positive-impact hypothesis at 90 percent or more. A statistically significant damage estimate is sufficient but not necessary to make an award close to the damage estimate. For example, if the non-statistical evidence makes it highly likely (90 percent) that damages occurred,

then a p-value of only 0.5 supports a damage award of 9.3, not far from the damage estimate of 10.

TABLE 1:
OPTIMAL AWARD GIVEN DAMAGE ESTIMATE EQUAL TO 10 AND EBOE EVIDENTIARY STANDARD

		Strength of Statistical Evidence P-value of Damage Estimate					
		0.75	0.50	0.25	0.10	0.05	0.01
Strength of Non-Statistical Evidence (probability)	0.10	0.0	0.0	0.0	0.0	6.3	9.4
	0.25	0.0	0.0	2.4	7.6	8.8	9.8
	0.50	0.0	3.9	7.7	9.2	9.6	9.9
	0.75	5.0	8.0	9.3	9.7	9.9	10.0
	0.90	8.3	9.3	9.7	9.8	9.9	10.0

The results in Table 2 following the FLOH decision rule have the same basic features as the EBOE awards—larger damage awards are associated with increased statistical significance or stronger non-statistical evidence. Here, however, all the award levels are much reduced compared with the EBOE awards, so much so that even with strong statistical evidence (0.01 p-value) and strong non-statistical evidence (90 percent probability of impact), the optimal award is only about half the damage estimate of 10.

TABLE 2:
OPTIMAL AWARD GIVEN DAMAGE ESTIMATE EQUAL TO 10 AND FLOH EVIDENTIARY STANDARD

		Strength of Statistical Evidence P-value of Damage Estimate					
		0.75	0.50	0.25	0.10	0.05	0.01
Strength of Non-Statistical Evidence (probability)	0.10	0.0	0.0	0.0	0.0	0.0	0.0
	0.25	0.0	0.0	0.0	0.0	0.0	4.1
	0.50	0.0	0.0	0.0	0.0	2.4	4.8
	0.75	0.0	0.0	0.5	2.5	3.5	5.0
	0.90	0.0	0.0	0.8	2.0	3.9	5.1

By contrast, applying a conventional significance standard (making an award equal to the damage estimate when the p-value is at or below 0.05) would result in a table comparable to these, but having awards of 10 recorded in the last two columns, with zero awards everywhere else. Thus, unlike a conventional significance standard, the integrated approach results in positive

damage awards even when the statistical evidence is not that strong. In addition, when the statistical evidence is strong, the optimal award can be considerably below the damage estimate under the FLOH standard when it is important to avoid awards that exceed actual damages.

VI. CONCLUSION

There are many good reasons not to use conventional statistical significance thresholds for determining whether regression results should be accepted as evidence of harm in antitrust cases.[35] Conventional statistical significance thresholds lack logical connection with the good decision making in a legal context. They reflect a hypothesis testing context which often will not apply, given the evidentiary context for statistical analysis of impact and damages analysis. Conventional statistical significance thresholds embody a maximum tolerance for Type I error that may be far less than is called for by the relevant legal burdens. The stringent limits on Type I error may also lead to probabilities of Type II error that are much too high from society's standpoint. All these and other attacks on conventional significance levels boil down to the implications of the results in Tables 1 and 2 in the previous section: When all the elements of a decision are considered using a fully defined decision-theoretic framework, conventional significance levels sometimes lead to optimal decisions and sometimes do not.

This statement could amount to just another attack on the use of conventional statistical levels, a practice that has survived decades of attacks, probably for want of any alternative. But Table 1 and Table 2 could be the makings of an alternative, if this abstract example were turned into something practical. We think that is not so difficult. Suppose that the court advised the testifying experts how the legal setting translates into the relative importance of awards too small versus awards too large. Thus, the court effectively picks the table. If, for example, the court decides that errors favoring the plaintiff and those favoring defense are equally important, then the court could choose something like Table 1 to help organize the expert testimony. Although this table would not need to be put directly in front of the jury, experts for the plaintiffs and the defense could be expected to summarize the documents and testimony by, in effect, recommending one or two rows of the table and offering testimony to explain their recommendations. Experts could offer their statistical estimates and p-values, and translate them into recommended damage amounts suited to the circumstances as suggested by Table 1. This would shift the testimony from what, for a jury, is probably a mysterious conversation

[35] This is not to say, as the above examples indicate, that we believe regression results lacking conventional significance would always be sufficient to justify an award of damages in the amount shown by the regression, or any award at all for that matter.

about choice of p-values and significance levels into a more understandable discussion of the strength of the non-statistical evidence. Ultimately, it would be up to the jury to make its own determination of the strength of the non-statistical evidence when a damage award is decided, something the jury instructions could make clear.

Of course, much more needs to be done to make this work well. Fact finders cannot be expected to easily reduce the non-statistical evidence to probabilities of impact, to line them up graphically with legal decision rules, and solve for the optimal amounts. Careful attention is needed to developing straightforward guidelines and jury instructions that implement this approach in a practical, meaningful way. But the potential benefit is well worth the effort. Regression analysis has become a mainstay in antitrust litigation—an analytical tool capable of solving some of the most difficult issues posed in such cases. It is important that it is used correctly and to the fullest extent of its capabilities.

APPENDIX

A NUMERICAL EXAMPLE OF THE EBOE AND FLOH STANDARDS

We assume for the illustration here and in Tables 1 and 2 in the text, that the task is to determine whether there are damages and, if so, of what magnitude. We assume that the non-statistical evidence (documents, testimony, etc.) regarding damages can be summarized with a probability distribution with a mass $(1-\pi)$ at zero (i.e. no damages) and the remainder (π) uniformly distributed between zero and M, the maximum plausible damage amount. For the illustrations below, we assume M to be 20.0 and π to be either 0.10, 0.25, 0.50, 0.75 or 0.90, representing weak, uncertain, supportive, or compelling (non-statistical) evidence that damages are positive. We refer to π as the prior probability of damages derived from review of the non-statistical evidence.

We assume also that regression analysis has generated an estimate of damages $\hat{D}$ which is normally distributed with mean μ (the actual damage amount) and variance σ^2. In the discussion below, we address two related questions. How should the value of π affect the interpretation of the damage estimate provided by the regression analysis? What is a wisely chosen critical value for the statistic $\hat{D}/\sigma$ (a measure of statistical significance associated with $\hat{D}$) below which no damages should be awarded?

With *A* standing for the court's damage award[36] and μ standing for the true damages, we suppose for purposes of this illustration that the legal standard for proof of damages takes one of two different forms. We refer to the first form as EBOE (equal balancing of error). Under this standard, the objective is to minimize the absolute difference between the award A and the true damages, $|A-\mu|$. That is, society accords equal weight to the error associated with the failure to award sufficient damages and the error associated with overpayment. This balanced treatment affords no special emphasis on avoiding Type I error and is akin to a "more likely than not" standard. The second form of legal standard used below is FLOH (favor low over high). This puts a nine times higher penalty on the award of damages that are too high (Type I error) versus damages that are too low (Type II error) and is a counterpart, effectively, to imposing a conventional 10 percent significance level as an evidentiary threshold.[37]

---

[36] Certain types of cases may have automatic adjustments to damages, e.g., treble damages in antitrust or for willfulness in patent infringement cases. These adjustments apparently reflect a policy decision to discourage bad behavior and/or incentivize parties to seek redress. For this exposition, we generally disregard that there are adjustments that may increase an award above the finding of actual damages. Incorporating these policies into the analysis would be straightforward (i.e., adjusting both A and μ to reflect the rule) and would not change the implications of this illustration.

[37] Thus the social cost is reflective of $|A-\mu|$ if $A \leq \mu$ but $9*|A-\mu|$ if $A>\mu$.

The prior probability of damages π can be combined with the data evidence ($\hat{D}$) to form a cumulative (posterior) probability function for the true damages (μ). The expected social penalty under the EBOE standard is minimized when the award *A* is set equal to the median of that distribution. If, given the non-statistical evidence and the estimated damages ($\hat{D}$), the probability of $\mu>0$ is equal to or less than 0.5, then zero is the median and no damages should be awarded.

Under the FLOH legal standard, the expected social loss is minimized when the award is set at the 10th percentile of the posterior distribution. Then, if the evidence (both statistical and documentary) shows that the probability of $\mu>0$ is equal to or less than 0.9, which means zero is in the 10th percentile, no damages should be awarded. The higher FLOH hurdle affects the outcome in two ways. First, the FLOH standard increases the chance that the evidence is insufficient to warrant any award of damages. Second, even when a damage award is warranted, under the FLOH standard the optimal award will be smaller than under an EBOE (more likely than not) standard to reduce the likelihood of the more serious error of overpaying damages.

Figure 2 (below) shows how differences in the strength of the non-statistical evidence and the legal standard affect the outcome when the statistical evidence is weak. Specifically, we assume that the regression offers a damages estimate equal to 10.0 with a standard error of 20. The implied Z-statistic is ½ and the implied p-value is 0.62. This estimate is not statistically significant by conventional standards and, following the suggestions of some commentators, might be used as evidence against the presence of damages (or, at minimum, precluded as a basis for expert opinion regarding damages on the part of plaintiff's experts).

The implications of the non-statistical and statistical evidence viewed in combination are summarized by cumulative probability curves showing, for each possible damage award (the horizontal axis), the probability of actual damages below that amount. Those probabilities (necessarily) increase as one considers higher damage awards (moving from left to right). They reach 100 percent at M=20.0, the assumed maximum possible damage amount. This figure has two heavy horizontal lines, the higher one corresponding to the median (the 50 percent point in the probability distribution) and the lower one corresponding to the bottom 10th percentile in the distribution. These lines correspond with the two evidentiary burdens used in this analysis. The damage award that minimizes the expected social costs of error under each of the two evidentiary burdens is the amount (on the horizontal axis) corresponding to the intersection of the cumulative probability curve associated with the evidence and the bold horizontal line that represents that burden.

Under the EBOE standard (the bold horizontal line at 50 percent), the proper damage award is positive in all scenarios except when the non-statistical evidence is so weak that it provides a prior probability for damages of only 25 percent. This is notwithstanding the assumed statistical weakness of the regression result. If the non-statistical evidence had indicated a 50 percent chance of damages, then the median damage is 1.0, well below the estimated damages of 10.0 but not zero. On the other hand, if the non-statistical evidence is strong enough to make the prior probability of damages equal to 90 percent, then the best damage award is 9.0, only slightly smaller than the estimated damages of 10.0 derived from the statistical analysis of the weak data alone.

This shows that if the non-statistical evidence is strong, the role played by the statistical analysis is not really to test for the presence of damages but to quantify them, and the statistical accuracy of the estimate becomes less important. A second message here is that the statistical results, weak though they may be, nonetheless play an important role in the ultimate outcome. Hence, those results should not be excluded from the analysis or simply treated as equivalent to a finding of zero damages.

Under the FLOH standard (which corresponds to the bold horizontal line at 10 percent), there would only be a very small damage award even when the documentary evidence is strong enough by itself enough to meet the 10th percentile standard. The statistical uncertainty is too great to support a higher damage award in an evidentiary setting that strongly penalizes excess damages.

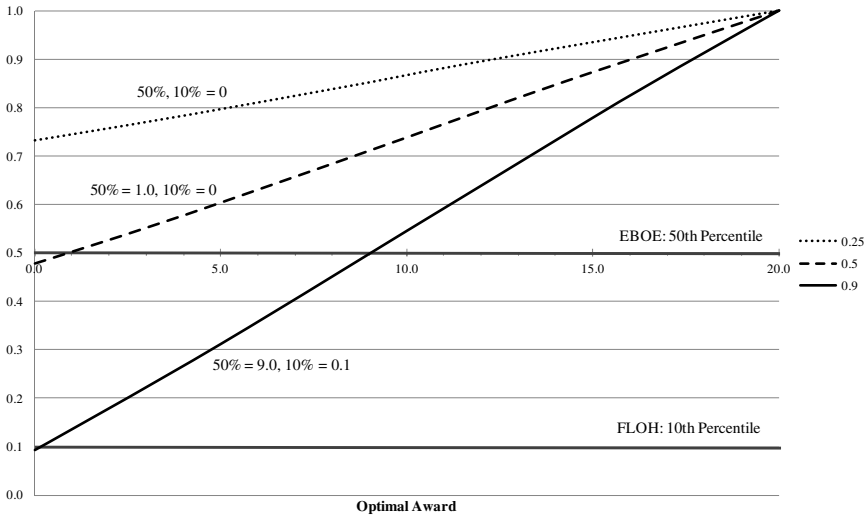

FIGURE 2: CUMULATIVE PROBABILITY OF DAMAGES

Images like Figure 2 underlie the two tables above which contain damage awards implied by the analysis described above, assuming in all cases the same statistically-based damage estimate equal to 10.0. The tables indicate how the damage award varies with the statistical significance of the damage estimate and with the strength of the non-statistical evidence as measured by the truth value of the hypothesis of impact.