Title
The Use of Statistical Evidence in Litigation: A Historical Look at the Use of P-Values, Issues, and Considerations.

Author
Guinta, Jeremy John

Publication Date
2018

UNIVERSITY OF CALIFORNIA
Los Angeles

The Use of Statistical Evidence in Litigation: A Historical Look at the Use of P-Values, Issues, and Considerations.

A thesis submitted in partial satisfaction of the requirements
for the degree
Master's in Applied Statistics

by

Jeremy John Guinta

2018

ABSTRACT OF THE THESIS

The Use of Statistical Evidence in Litigation: A Historical Look at the Use of P-Values, Issues, and Considerations.

by

Jeremy John Guinta
Master's in Applied Statistics
University of California, Los Angeles, 2018
Professor Chad J Hazlett

The p-value has a long and storied use in expert testimony and litigation. However, researchers and academics in science and academia are questioning the use and reliance of p-values for the determination of statistical significance of findings. It is a natural extension to explore the concerns of science and academia to how p-values are used in litigation. This paper explores the history behind the use of the p-value in expert testimony and litigation, the potential issues with triers of fact and experts relying on p-values to make decisions based on statistical significance. Using a gender pay equity model as an example, this paper will develop several simulated datasets and models to show how p-values can be influenced to show statistically significant results or non-statistically significant results using simple modifications to the model specifications or the amount of data being used. This paper will also explore potential methods for analyzing the power of the test and considerations for regression analysis. The analysis suggests that more consideration is necessary for the use of a p-value "bright line" for gender pay analyses, and that a simple "bright line" consideration for litigation could be easily manipulated.

The thesis of Jeremy John Guinta is approved.

Erin K Hartman

Vivian Lew

Chad J Hazlett, Committee Chair

University of California, Los Angeles

2018

This is dedicated to my wife Nicole. Without your patience, understanding, and devotion to me and our family, none of this would have been possible.

To my children, Charles, and Isabella, thank you for your patience and understanding. Daddy is now able to go outside and play.

To the team at work. Thank you! Without your commitment to client service while I was in class or otherwise unavailable due to school commitments, this would not have been possible.

To Alex Krebs. A special thanks for giving me a sounding board for ideas.

Table of Contents

Tables

Figures

I. Introduction

The use of statistics in litigation has a very long and established history. The first use of statistical evidence in litigation that used the p-value was performed by the U.S. Supreme Court in 1977. The U.S. Supreme Court performed a standard deviation calculation, and in the footnotes to two separate rulings (*Castañeda v. Partida* and *Hazelwood School District v. United States*) the Supreme Court discussed how "two to three standard deviations" were necessary for statistical significance. Both cases involved discrimination of protected classes. In *Castañeda*, the issue before the Court was racial discrimination involving potential Mexican-American jurors in Texas.[1] In *Hazelwood*, the issue before the Court was racial discrimination in hiring of African-American teachers.[2]

Even though the U.S. Supreme Court also indicated in their *Hazelwood* ruling that the calculation was not always required, lower courts interpreted these footnotes and the U.S. Supreme Court's procedure to be a requirement to establish evidence of racial discrimination. The 4th Circuit issued the following ruling in *Moultrie v. Martin* in 1982.

> When a litigant seeks to prove his point exclusively through the use of statistics, he is borrowing the principles of another discipline, mathematics…[He] cannot be selective in which principles are applied. He must employ a standard mathematical analysis. Any other requirement defies logic to the point of being unjust. Statisticians do not simply look at two statistics…and make a subjective conclusion that the statistics are significantly different. Rather, statisticians compare figures through an objective process known as hypothesis testing.[3]

The precedent was set and spread from there to other circuits. The use of statistical evidence gained widespread adoption in the courts, and the courts naturally relied on experts to assist the trier of fact with the statistics. This evolved to the use of "bright-line"[4] tests to determine if the statistical evidence was significant.[5] Currently, many experts in the various disciplines of

---

[1] https://supreme.justia.com/cases/federal/us/430/482/ (Last Accessed December 28, 2017)

[2] https://supreme.justia.com/cases/federal/us/433/299/ (Last Accessed December 28, 2017)

[3] Is Proof of Statistical Significance Relevant? By David H. Kaye (Penn State Law, 1986)

[4] A "bright line" test is an unambiguous criterion or guideline that is used especially in law. Please see https://definitions.uslegal.com/b/bright-line-rule/ (Last Accessed November 24, 2018)

[5] The U.S. Supreme Court implied that a p-value of 0.05 or less was needed for a prima facie (latin for "at first look", this term implies that there is enough information to establish a presumption unless rebutted https://www.law.cornell.edu/wex/prima_facie)

academia and science are questioning the use of a "bright line" test to determine statistical significance as it relates to scientific findings.[6] Furthermore, other experts testifiers are also questioning the use of "bright line" tests of statistical significance.[7] Based on these above concerns, this thesis is focused on describing the role of experts in litigation, how experts use statistical tests in litigation, and how triers of fact interpret those statistics.

II. Preliminaries

Before this thesis begins, I will set up the framework in which the analysis is conducted. Primarily, this thesis will analyze the impact of changes to the model on the p-value. Thus, statistical theory regarding inference and hypothesis testing is required. The following definitions are needed:

- Hypothesis Testing – A procedure of assessing if a parameter is consistent or not consistent with an existing statement made about the data.[8]
- Null Hypothesis (denoted as $\boldsymbol{H_o}$) – The no difference hypothesis to be tested.[9] The default state.
- Alternative Hypothesis (denoted as $\boldsymbol{H_a}$ or $\boldsymbol{H_1}$) – The hypothesis against which the null hypothesis is tested.[10]

(Last Accessed November 24, 2018) of discrimination case. It is not clear how the Court made the determination that 0.05 was the appropriate "bright line" test, but this "bright line" was original devised by R.A. Fisher, and is commonly accepted in statistics as a line of "statistical significance" when testing the null hypothesis against the alternative hypothesis.

[6] Correspondence to Dr. Sander Greenland, Department of Epidemiology, School of Public Health, University of California, 2017, American Journal of Epidemiology Vol 186, No. 6 DOI: 10.1093/aje/kwx259 argues that the use of "bright line" tests is ineffective and serves no good purpose for inference and that even the use of confidence intervals cannot solve for bias and many other problems associated with hypothesis testing; Daniel J. Benjamin Et. Al in Nature Human Behaviour, 2017 (DOI: 10.1038/s41562-017-0189-z) argues for moving the "bright line" to 0.005; and Valentin Amrhein Et. Al in Nature Human Behaviour, 2017 (DOI: 10.1038/s41562-017-0224-0) argues that 0.005 still has all of the same problems as 0.05 and that "bright line" tests should be discarded for use in significance testing.

[7] Statistical Significance and Statistical Error in Antitrust Analysis, By Phillip Johnson, Edward Leamer, and Jeffrey Leitzinger, Antitrust Law Journal, American Bar Association, 2017 Volume 81 Issue 2.

[8] Everitt, Brian. *The Cambridge Dictionary of Statistics.* Cambridge, UK New York: Cambridge University Press. 4th Edition, 2010

[9] Everitt, Brian. *The Cambridge Dictionary of Statistics.* Cambridge, UK New York: Cambridge University Press. 4th Edition, 2010

[10] Everitt, Brian. *The Cambridge Dictionary of Statistics.* Cambridge, UK New York: Cambridge University Press. 4th Edition, 2010

- Type I Error (False Positive) – If $\boldsymbol{H_o}$ is true, but the hypothesis test rejects $\boldsymbol{H_o}$.[11]

- Type II Error (False Negative) – If $\boldsymbol{H_a}$ is true, but the hypothesis test rejects $\boldsymbol{H_a}$.[12]

- $\boldsymbol{\alpha}$ (can be referenced as Alpha) – The probability of Type I error is called the significance level of the test.[13] This quantity can be defined as:

$$\boldsymbol{\alpha = P(\bar{X} \geq \bar{x}; H_o) = P\left(\frac{\bar{X} - \mu_o}{\sqrt{\sigma^2/n)}} \geq \frac{\bar{x} - \mu_o}{\sqrt{\sigma^2/n)}} ; H_o \right)}$$

  However, $\boldsymbol{\alpha}$ is typically a set value that used to determine a threshold for statistical significance. See footnote five.

- $\boldsymbol{\beta}$ (can be referenced as Beta) – The probability of a Type II error.[14] This quantity can be defined as:

$$\boldsymbol{\beta = P(\bar{X} \leq \bar{x}; H_a) = P\left(\frac{\bar{X} - \mu_a}{\sqrt{\sigma^2/n)}} < \frac{\bar{x} - \mu_a}{\sqrt{\sigma^2/n)}} ; H_a \right)}$$

  For example, assume that $\mu_a = 16$ and $\mu_o = 8$ with a σ^2 of four, and that the test is comparing if μ_a is different than μ_o with an observed $\bar{X} = 12.5$. Alpha and Beta can be visualized as follows:[15]

[11] Hogg, Robert V, et al. *Probability and Statistical Inference*. Pearson, 2015, pg. 355. See also, Lehmann, Erich Leo., and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, 2008, pg. 57.

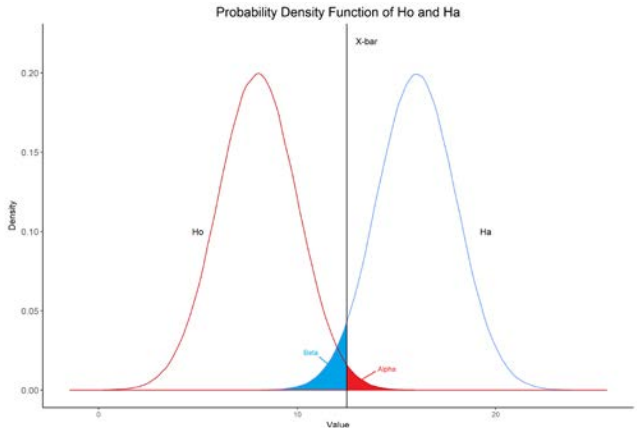
[12] Hogg, Robert V, et al. *Probability and Statistical Inference*. Pearson, 2015, pg. 355. See also, Lehmann, Erich Leo., and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, 20, pg. 57.

[13] Hogg, Robert V, et al. *Probability and Statistical Inference*. Pearson, 2015, pg. 355.

[14] Hogg, Robert V, et al. *Probability and Statistical Inference*. Pearson, 2015, pg. 355.

[15] Hogg, Robert V, et al. *Probability and Statistical Inference*. Pearson, 2015, pg. 356.

Preliminary Figure 1 – Comparison between Alpha and Beta.



- p-value – The probability of the observed data (or data showing a more extreme departure from the null hypothesis) when the null hypothesis is true.[16] This value is compared to α, or the level of significance.

$$\boldsymbol{p-value = P(\bar{X} \geq \bar{x}; \mu = \mu_o) = P\left(\frac{\bar{X} - \mu_o}{\sqrt{\sigma^2/n}} \geq \frac{\bar{x} - \mu_o}{\sqrt{\sigma^2/n}}\,; \mu = \mu_o\right)}$$

- Neyman-Pearson framework – This is the basis for the Neyman-Pearson lemma and is a departure from Fisher's logic on hypothesis testing. In Fisher's view, a hypothesis test consists only of a Null Hypothesis. The Null Hypothesis is tested based upon the data, and in this paradigm, there is not a determination of significance. There is only a reported p-value of the test. In the Neyman-Pearson framework there are two hypotheses that are considered: The Null Hypothesis and the Alternative Hypothesis. The Neyman-Pearson framework establishes the best critical region for which to judge tests as significant or not significant.[17]

[16] Everitt, Brian. *The Cambridge Dictionary of Statistics.* Cambridge, UK New York: Cambridge University Press. 4th Edition, 2010.

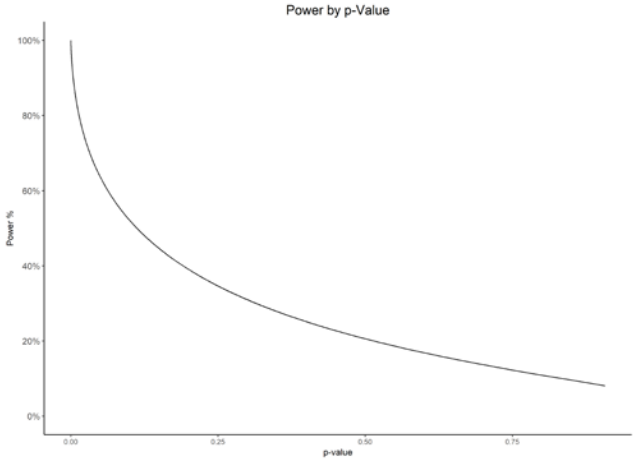
[17] Johannes Lenhard, *Models and Statistical Inference: The Controversy between Fisher and Neyman–Pearson,* Brit. J. Phil. Sci. 57 (2006), pgs. 69–91.

- Neyman-Pearson lemma – This is the best critical region of size $\boldsymbol{\alpha}$ for testing the simple null hypothesis $\boldsymbol{H_o: \theta = \theta_o}$ against the simple alternative hypothesis $\boldsymbol{H_a: \theta = \theta_a}$ that gives the greatest power among all critical regions of size $\boldsymbol{\alpha}$.[18]

- Power – The probability of rejecting the null hypothesis when it is false.[19] Power is calculated as $1 - \beta$.[20]

  The p-value and the Power are closely related. They have an inverse relationship such that as the p-value decreases, the Power increases. This can be visualized via the Power function of the test:[21]

Preliminary Figure 2 – Comparison between Power and p-Value



[18] Hogg, Robert V, et al. *Probability and Statistical Inference*. Pearson, 2015, pg. 400. The Neyman-Pearson lemma has three conditions for the joint PDF of $f(x;\theta)$ where θ_0 or θ_1 are two possible values of θ. If a positive constant k and subset C such that 1) $P[(X_1, X_2, X_3 \ldots X_n) \in C; \theta] = \alpha$; 2) $\frac{L(\theta_0)}{L(\theta_1)} \leq k\ for\ (x_1, x_2, x_3, \ldots x_n) \in C$; and 3) $\frac{L(\theta_0)}{L(\theta_1)} \geq k\ for\ (x_1, x_2, x_3, \ldots x_n) \in C'$, the C is the best critical region.

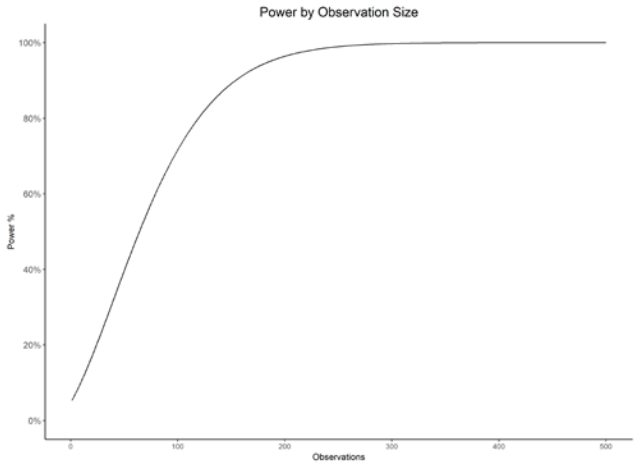
[19] Everitt, Brian. *The Cambridge Dictionary of Statistics.* Cambridge, UK New York: Cambridge University Press. 4th Edition, 2010.

[20] Hogg, Robert V, et al. *Probability and Statistical Inference*. Pearson, 2015, pg. 393.

[21] Hogg, Robert V, et al. *Probability and Statistical Inference*. Pearson, 2015, pg. 393. See also, Daniel J. Benjamin Et. Al in Nature Human Behaviour, 2017 (DOI: 10.1038/s41562-017-0189-z) Supplementary Information found at https://static-content.springer.com/esm/art%3A10.1038%2Fs41562-017-0189-z/MediaObjects/41562_2017_189_MOESM1_ESM.pdf (Last Accessed 11/24/2018). In Benjamin, Power is plotted over the False Positive Rate by Prior Odds Ratios. The Prior Odds Ratio is defined as $\frac{1-\phi}{\phi} = \frac{P(H_a)}{P(H_o)}$ and the False Positive Rate defined as $\frac{\alpha\phi}{\alpha\phi+(1-\phi)(1-\beta)}$. This equation demonstrates that as Power $(1-\beta)$ goes to one, the False Positive Rate will go to zero.

The Power is also related to the number of observations. Holding other parameters constant but increasing the number of observations will increase the Power. This can also be visualized via the Power function of the test:[22]

Preliminary Figure 3 – Comparison between Power and Number of Observations



- Ordinary Least Squares Regression – A statistical method that computes the relationship between two or more quantitative variables, so that a response or outcome variable can be predicted from the other, or others.[23] In matrix form, ordinary least squares regression is denoted as follows:[24]

$$\underset{nx1}{\boldsymbol{Y}} = \underset{nxp}{\boldsymbol{X}} * \underset{px1}{\boldsymbol{\beta}} + \underset{n\,x1}{\boldsymbol{\varepsilon}}$$

- Inference in Regression – The process of drawing a conclusion about the parameters based on the model. For example, one form of regression inference takes the form of

$$\boldsymbol{H_o}: \boldsymbol{\beta_k} = \mathbf{0}$$
$$\boldsymbol{H_a}: \boldsymbol{\beta_k} \neq \mathbf{0}$$

[22] Hogg, Robert V, et al. *Probability and Statistical Inference*. Pearson, 2015, pgs. 393,394. The text uses a binomial distribution to determine the Power function of the test defined as $K(p) = \sum_{y=0}^{i} \binom{n}{y} p^y (1-p)^{n-y}, 0 < p \leq \frac{1}{2}\ and\ i < n.$ The equation shows that as n increases, the power at any value of p will also increase. See Figure 8.5-1. See also, S.P. *Bloomberg Power Analysis Using R*, January 7, 2014.

[23] Kutner, Michael H. *Applied Linear Statistical Models*, McGraw-Hill Irwin, 2005, pg. 2.

[24] Kutner, Michael H. *Applied Linear Statistical Models*, McGraw-Hill Irwin, 2005, pg. 222.

which is testing if the slope of the of the regression coefficients is different from zero. This hypothesis test uses the standard error to determine if there is a difference. The test statistic takes on a t-distribution that can be calculated as $\boldsymbol{t^* = \frac{b_k}{s\{b_k\}}}$.[25] Where $\boldsymbol{s\{b_k\}}$ is calculated as $\boldsymbol{s\{b_k\} = MSE * (X'X)^{-1}}$[26] and MSE is calculated as $\boldsymbol{\frac{SSE}{n-p}}$.[27]

III. The Role of Experts in Litigation

An expert is an individual qualified to assist the trier of fact (e.g., judge, jury).[28] Experts provide their expertise to the legal proceeding that the trier of fact cannot provide for themselves. Often this expertise is in areas in which the legal proceeding is not skilled, such as data analytics, econometrics, statistics, accounting, engineering, and various science disciplines. Experts are often used when the legal proceeding requires specific knowledge of an industry, or process, or specialized mathematical calculations.

Since there are typically two sides to the litigation, there are often at least two experts: one for the plaintiff and one for the defense. Often, when the issues are complex, many experts from both sides are used to provide their expertise to the trier of fact on various issues. Experts will review the evidence obtained during discovery,[29] issue expert reports, prepare rebuttal reports, and provide testimony to defend his or her position. Even though experts offer independent analysis, experts are often "opposing" each other. Experts on both sides of a litigation will be provided the same set of evidence, and the experts often reach entirely opposite conclusions.

---

[25] Kutner, Michael H. *Applied Linear Statistical Models*, McGraw-Hill Irwin, 2005, pg. 228.

[26] $\boldsymbol{b = (X'X)^{-1}(X'X)Y}$. See Kutner, Michael H. *Applied Linear Statistical Models*, McGraw-Hill Irwin, 2005, pgs. 223, 227.

[27] Kutner, Michael H. *Applied Linear Statistical Models*, McGraw-Hill Irwin, 2005, pg. 225. $SSE = \boldsymbol{Y'Y - b'X'Y}$.

[28] A trier of fact is the individual (or individuals) that is tasked with determining the facts in a legal proceeding. The trier of fact can be a judge, a jury, an arbitrator, or any other person that presides over a legal proceeding in an official capacity. See https://www.law.cornell.edu/wex/trier_of_fact (Last Accessed November 24, 2018).

[29] Discovery is a pre-trial procedure in which both sides of a legal proceeding can request and receive evidence from each other. See https://www.law.cornell.edu/wex/discovery (Last Accessed November 24, 2018).

Each expert's work product is reviewed by the "opposing" side, and each side can be heavily criticized and rebutted. Experts often testify that their work product and conclusions are accurate. This process can lead to the "truth" of the evidence. It also provides the trier of fact a wide array of domain-specific knowledge that otherwise would not be introduced to the legal proceeding. When this process does not lead to the "truth" (i.e., the "opposing" experts do not agree) the trier of fact must interpret the conclusions reached by the expert's analyses provided and determine which expert has conducted a more reliable analysis.

IV. How Statistical Evidence is Used in Gender Pay Discrimination Litigation

As shown in Section I, the root of statistical evidence and the p-value are based on discrimination analysis for disparate impact for protected classes.[30] This type of analysis was the first to use statistical evidence that relies on the p-value, and it is firmly rooted in econometric modeling and theory. Econometric modeling involves proposing a theory and then using the available data to test the theory with a model.[31] For this type of analysis, a multivariate linear regression model is commonly used.[32] For example, in a gender pay discrimination analysis, the plaintiff (a female) will theorize that due to her gender she received less pay than an equivalent male. The plaintiff's expert would propose a model that would control for job title and location. In short, the plaintiff is hypothesizing that all employees with the same job title at the same location should earn approximately the same amount of money. The model would take on the following form.[33]

$$\boldsymbol{Pay = \beta_o + \beta_1 Gender + \beta_2 Occupation + \beta_3 Location + \varepsilon}$$

[30] The Civil Rights Act of 1964 and subsequent amendments established specific classes of individuals that were protected from discrimination. See https://www.archives.gov/education/lessons/civil-rights-act (Last Accessed November 24, 2018)

[31] Econometric modeling relies upon the Neyman-Pearson framework to make determinations of statistical significance after hypothesizing and testing of a theory based on the Neyman-Pearson's lemma. In this particular analysis the simple hypothesis is that the slope of the Gender Coefficient (which represents Female) is either zero or non-zero. The Null hypothesis is that there is no difference in pay between males and females. The alternative hypothesis is that there is a difference. The Alpha level of 0.05 is pre-determined by existing case law and is assumed to provide the most powerful test (see Preliminaries).

[32] See Preliminaries. An ordinary least squares regression gives the best unbiased estimator. This process allows for the analyst to conduct inference on a certain variable (in this example Gender), while controlling for various other factors.

[33] The pay variable is typically transformed using a natural log transformation. This would change the interpretation of the Gender variable to be approximately the percentage difference in pay. See Hansen, Bruce. *Econometrics*. University of Wisconsin, 2018, pgs. 14-16.

The plaintiff would then perform a hypothesis test on the coefficient of the Gender dummy variable of the following form.[34]

$$H_o: \beta_{Gender} = 0$$
$$H_a: \beta_{Gender} != 0$$

If the p-value on the coefficient for the Gender variable is significant (using the U.S. Supreme Court's standard of 0.05) then the plaintiff would reject the null hypothesis and conclude that pay is different based on gender (while controlling for occupation and location).[35] If this model was the "true" model or at least the model accepted by the legal proceeding, the coefficient on Gender would also be used to calculate the damages. The coefficient would be interpreted as the average shortfall between male and female pay.[36]

The defense would counter that the plaintiff's model is not properly specified and that other key variables are needed to properly model pay. The defense may propose a model of the following form.

$$Pay = \beta_o + \beta_1 Gender + \beta_2 Education + \beta_3 Tenure + \beta_4 Occupation + \beta_5 Location + \varepsilon$$

The defense is testing the theory that all employees with the same education, same amount of tenure, the same occupation, and at the same location should earn approximately the same amount of money. The hypothesis test and conclusions would follow the same form.

$$H_o: \beta_{Gender} = 0$$
$$H_a: \beta_{Gender} != 0$$

[34] It would also be argued that the coefficient would also have to be negative (i.e., that female received less pay than male) for the model and hypothesis test to make sense.

[35] It should be noted that the proper interpretation of a hypothesis test for when the p-value reaches a pre-determined threshold is to reject the Null. A rejection of the Null Hypothesis should not be taken as an acceptance of the Alternative Hypothesis. (See Hogg, Robert V, et al. *Probability and Statistical Inference*. Pearson, 2015, pg. 360.) However, in practical application, this is often viewed as an acceptance of the Alternative Hypothesis.

[36] The plaintiff would also have to show that the regression model as a whole was statistically significant using a p-value level of 0.05 on an F-Test. See Kutner, Michael H. *Applied Linear Statistical Models*, McGraw-Hill Irwin, 2005, pg. 226.

If the p-value on the coefficient for the Gender variable is not significant based on the new model, then the defense would argue that there is not enough evidence to reject the null hypothesis. The overall conclusion is that there is no discrimination towards females when controlling for differences by education, tenure, occupation, and location. In this context, the battle of the two regressions is not necessarily which has more statistical significance, but which model:

1. Conforms with the law regarding what variables can be controlled for (e.g., that location is a bona fide factor that should be considered), and
2. Which model has the appropriate interpretation based on the other evidence presented during the legal proceeding.

For example, location would likely have to be used by the company to make pay decisions for the variable to be included in the model. However, courts can and do interpret the significance of the hypothesis test, and courts have ruled solely on the significance of the hypothesis test when determining the appropriate model.[37]

V. The Role of Daubert and the Gatekeeping Role of the Judge to the Trier of Fact to Scientific Testimony[38]

Daubert v. Merrell Dow Pharmaceuticals, Inc.[39] was a landmark case that changed the landscape for how expert testimony was used in a legal proceeding. Up until *Daubert*, scientific and statistical evidence was taken for granted by the courts, and experts were given wide latitude in their opinions.

Daubert was a civil litigation in which plaintiffs claimed a certain drug caused birth defects when taken during pregnancy. Plaintiffs relied on expert testimony based on a study that compared the

[37] *Segar v. Smith*, 738 F.2d 1249 (D.C. Cir. 1984). See https://law.justia.com/cases/federal/appellate-courts/F2/738/1249/135384/ (Last Accessed November 24, 2018)

[38] See generally David H. Kaye, *The Dynamics of Daubert: Methodology, Conclusions, and Fit in Statistical and Econometric Studies*, 87 *Va. L. Rev.* 1933 (2001).

[39] *Daubert v. Merrell Dow Pharmaceuticals, Inc. 509 U.S. 579 (1993*). https://supreme.justia.com/cases/federal/us/509/579/ (Last Accessed November 24, 2018)

molecular structure of the drug to other molecules known to cause birth defects, a re-analysis of existing epidemiological studies on the drug at issue, and a study based on the drug's impact on animals. During the trial-phase the presiding judge excluded all of plaintiffs' experts' testimony, conclusions, and granted summary judgment to the defense because the judge ruled that the methodology used by the plaintiff's expert was novel and did not have a proven methodology. This case made its way to the U.S. Supreme Court. The U.S. Supreme Court ruled that the lower court had erred in excluding the plaintiffs' experts' testimony and remanded the case back to the lower court.

The U.S. Supreme Court issued the following guidance that significantly changed the landscape of expert testimony and how scientific and statistical evidence is accepted by the courts.

1. Judges are required to be a gatekeeper for expert evidence before it is submitted to the trier of fact.
2. Scientific testimony or evidence must be relevant and reliable.
3. Scientific testimony or evidence must be based on proven methodology. There are four pillars that testimony or evidence from an expert must be based on.

As a practical matter, the U.S. Supreme Court's decision forced triers of fact to carefully consider experts' testimony. Before any testimony or scientific evidence is given to the trier of fact, the trier of fact in the legal proceeding must rule whether the expert's testimony is relevant, reliable, and based on a proven methodology.[40]

Furthermore, the four pillars of an expert's scientific evidence or testimony can be described as follows.

[40] This process is now commonly referred to as a "Daubert Challenge." In many civil litigation legal proceedings this challenge is common, and it requires the expert to defend their expertise, the relevancy of their testimony, and how reliable their methodology is. See https://definitions.uslegal.com/d/daubert-challenge/ (Last Accessed November 24, 2018).

1. The theory or methodology cannot be created just for the legal proceeding. The court should not be the gatekeeper of new or novel scientific techniques that have not been tested.
2. The theory or methodology has been peer-reviewed or published.
3. The error of the theory or methodology must be known. If the methodology or theory has not been peer-reviewed, the risk of error is unknown. The court is not in the position to accept novel techniques that have not been fully studied by the scientific community as the potential for error is great, and the court has no known way of estimating that error.
4. The theory or methodology must be generally accepted and known by the scientific community.

These pillars have wide latitude for the courts to interpret. For example, regression analysis is widely used, peer-reviewed technique used in numerous publications. In summary, regression analysis is not novel or new. However, if an expert used regression analysis is a new or novel way or applied regression techniques to data in a new or novel way, that expert may not be able to simply rely on the fact that regression analysis has been used before. More likely, they would have to demonstrate why regression analysis is not new or novel in the context of their analysis, and that the science community would accept their use of regression analysis in the stated context.

The standard as set forth by the original *Daubert* ruling has evolved over time. As new cases were tried in the post-*Daubert* world, courts naturally came to differing interpretations of *Daubert*. Later Supreme Court rulings, most notably *Kumho Tire Co. Ltd v. Carmichael*[41] and *General*

[41] *Kumho Tire Co. Ltd. v. Carmichael, 526 U.S. 137 (1999).* https://supreme.justia.com/cases/federal/us/526/137/ (Last Accessed November 24, 2018).

Electric v. Joiner[42] further defined how courts should treat and interpret expert evidence. In general, experts and the evidence that they opine on must:

1. Be based on sufficient facts and data AND
2. Be the product of reliable principles and methods AND
3. The expert must have reliably applied the sufficient facts and data based on reliable principles and methods to the facts of the case.[43]

Kumho stated that *Daubert* applied to technical testimony as well as scientific testimony. Furthermore, *Kumho*, indicated that the four pillars of expert testimony set forth in *Daubert* were not mandatory, but that the trier-of-fact must determine if the expert "employs in the courtroom the same level of intellectual rigor that characterizes the practice of an expert in the relevant field."[44] In *Joiner* the Supreme Court indicated that district courts are able to review both the conclusions and methodology of an expert's testimony.[45] The Ninth Circuit Court of Appeals, in the *Daubert* litigation, summarized the trier of facts role as it relates to expert testimony as follows:

> Our responsibility, then, unless we badly misread the Supreme Court's opinion, is to resolve disputes among respected, well-credentialed scientists about matters squarely within their expertise, in areas where there is no scientific consensus as to what is and what is not 'good science,' and occasionally to reject such expert testimony because it was not 'derived by the scientific method.' Mindful of our position in the hierarchy of the federal judiciary, we take a deep breath and proceed with this heady task.[46]

[42] *General Electric. Co. v. Joiner 522 U.S. 136 (*1997). https://supreme.justia.com/cases/federal/us/522/136/ (Last Accessed November 24, 2018).

[43] Federal Rules of Evidence 702. https://www.law.cornell.edu/rules/fre/rule_702 (Last Accessed November 24, 2018).

[44] *Kumho Tire Co. Ltd. v. Carmichael, 526 U.S. 137 (1999),* at 152.

[45] *Joiner* 522 U.S. 136, 146. In summary the Supreme Court ruled that the 11th Circuit Court of Appeals erred in their ruling that stated that the district court could only review the methodology of an expert's opinions.

[46] *Daubert v. Merrell-Dow Pharms*., 43 F.3d 1311, 1316 (9th Cir. 1995).

VI. Summary of Recent Criticisms on the use of "Bright line" Benchmarks in Statistics

Recently, statisticians, academics, and scientists have commented on the reproducibility crisis in academia and science with regards to the use of the p-value for determining if a result is a "statistically significant." Some of these discussions can be summarized as follows.[47]

1. A singular Bright Line threshold leads to publication bias and reduces the reproducibility of results.
2. The power of the test at 0.05 conveys that the Null Hypothesis still has 3:1 odds of occurring (based on prior odds of 1:10).[48]
3. Using the p-value threshold of 0.05 can lead to an unacceptably high false positive rate.[49]
4. Hypothesis tests that do not list confidence intervals can be misleading.
5. Discussion of the term "significant" can bias the reviewer, and the use of the p-value with a "bright line" should not be considered with context of the methodology and biases of the analysis.

The discussions rightfully point out that adjusting for these concerns would not necessarily solve p-value hacking, omitted variables, or publication bias.

---

[47] Correspondence to Dr. Sander Greenland, Department of Epidemiology, School of Public Health, University of California, 2017, American Journal of Epidemiology Vol 186, No. 6 DOI: 10.1093/aje/kwx259; Daniel J. Benjamin Et. Al in Nature Human Behaviour, 2017 (DOI: 10.1038/s41562-017-0189-z); and Valentin Amrhein Et. Al in Nature Human Behaviour, 2017 (DOI: 10.1038/s41562-017-0224-0).

[48] Daniel J. Benjamin Et. Al in Nature Human Behaviour, 2017 (DOI: 10.1038/s41562-017-0189-z); and Valentin Amrhein Et. Al in Nature Human Behaviour, 2017 (DOI: 10.1038/s41562-017-0224-0). See Equation 1, Figure 1, and Figure 2. Equation 1 specifies that the ratio of the Alternative Hypothesis over the Null Hypothesis represents the strength of evidence for the Alternative Hypothesis relative to the Null Hypothesis. This ratio can be written as the Bayes Factor times the Prior Odds. The Bayes Factor for the Likelihood Ratio Bound is determined as $\frac{0.5}{e^{-0.5*1.96^2}} = 3.4$ Using this result and Equation 1 arrives at a result of $3.4 * \frac{1}{10} = 0.34$ or 3:1 odds in favor of the Null. Furthermore, other ratios using the Bayes Factor from the Local H_a bound derived as $BF = \frac{-1}{(e\text{^}1 * 0.05 * ln(0.05))} = 2.4$ arrives at a result of $2.4 * \frac{1}{10} = 0.24$ or 4:1 odds in favor of the Null.

[49] Daniel J. Benjamin Et. Al in Nature Human Behaviour, 2017 (DOI: 10.1038/s41562-017-0189-z); and Valentin Amrhein Et. Al in Nature Human Behaviour, 2017 (DOI: 10.1038/s41562-017-0224-0). The False Positive Rate is defined as $\frac{\alpha\phi}{(\alpha\phi+(1-\beta)(1-\phi))}$ where ϕ is the prior probability. Assuming a prior probability of $\frac{1}{10}$, a Power of 0.80, and an Alpha of 0.05, the False Positive rate would be $\frac{0.05*\left(\frac{10}{11}\right)}{0.05*\left(\frac{10}{11}\right)+(0.80)\left(1-\left(\frac{10}{11}\right)\right)} = 0.38.$ See Figure 2 in Benjamin for a plot of various Power levels over the False Positive Rate.

There are also theoretical reasons why the use of "bright line" tests should be of concern for academia, scientists, and experts. First, a p-value "bright line" of 0.05 assumes a false positive rate of 5%. This "bright line" leads to high false negative rate which may or may not be acceptable. Second, any test that uses a simple "bright line" threshold implicitly does not consider other evidence that may be available. Third, "bright lines" do not consider the power of the test and a balancing (if appropriate) of the false positive and false negative rates. Fourth, researcher degrees of freedom[50] can be manipulated to produce results that may or may not be accurate. Finally, p-values are easily influenced by the sample size of study. The results of a hypothesis tests can be influenced if many observations are considered. In other words, small potential impractical differences may show statistical significance because the model included a large number of observations.

VII. Why the Concerns of Science and Academia may not be Relevant to Litigation

First, legal proceedings are designed to be adversarial with many checks and balances on how evidence is produced and opined on. When an expert presents his or her work another expert reviews and critiques that work, offers up an opposing view, and points out any potential issues with the analysis. This adversarial process is repeated many times throughout the legal proceeding.

Second, civil courts use a "preponderance of the evidence"[51] standard for civil litigation. False negatives are when the alternative hypothesis is rejected when the null hypothesis is false.[52]

[50]Generally, "researcher degrees of freedom" is the ability for the researcher to influence the statistical significance or lack of statistical significance by changing the way the data was processed, interpreted, and/or modeled. https://simplystatistics.org/2013/07/31/the-researcher-degrees-of-freedom-recipe-tradeoff-in-data-analysis/ (Last Accessed November 24, 2018)

[51] The preponderance of evidence is typically a civil court standard to prove a case. A plaintiff must show that the fact presented are more likely than not true. https://legaldictionary.net/preponderance-of-evidence/ (Last Accessed November 24, 2018).

[52] Significant Statistics: The Unwitting Policy Making of Mathematically Ignorant Judges, Michael I. Meyerson & William Meyerson, Pepperdine Law Review Vol. 37: 771, 2010, pg. 829, "A judicial Type II error, failing to reject an erroneous null hypothesis of "no effect" or "no discrimination," is actually a legal acceptance of that false premise. When such a Type II error occurs, a truly-harmed plaintiff is denied relief. Nonetheless, courts should be somewhat cautious about finding liability."

Adjusting the "bright line" to a less stringent level to reduce false negatives based on other evidence may not be appropriate under the "preponderance of the evidence" standard.

Last, judges are gatekeepers to scientific evidence being introduced to the trier of fact, and scientific evidence must be based on sound, proven methodologies. Thus, issues such as p-value hacking,[53] misspecified models,[54] and omitted variable bias[55] should be uncovered during the adversarial process before the trier of fact considers an expert's analysis and testimony.

VIII. Why the Concerns of Science and Academia could be Relevant to Litigation

First, the p-value of 0.05 is still used to make a ruling of significance, and a "bright line" line test does not factor the continuous scale of a p-value. There is no valid reason that a statistical test that has a p-value of 0.051 is not "significant" but a statistical test with a p-value of 0.049 is "significant." Second, with a "bright line" test, experts do not necessarily need to factor in false positives, false negative rate, the power of the test, and whether one or two-sided tests are appropriate. Even though the other expert would be free to point out any such inconsistency or issue with the expert's work, a "bright line" has been established, and these other considerations may not have weight in the eyes of the trier of fact. Possibly worse, the trier of fact may not understand these issues and be "fooled" into considering a simple "bright line" standard. Third, the null hypothesis with a threshold of 0.05 may be considered to get too much weight. For example, in a pay disparate impact analysis, the null hypothesis can be interpreted as no pay disparity occurred. This initial starting point may or may not be appropriate based on other evidence. Last, as previously discussed the 0.05 threshold may not fit with the "preponderance of evidence" standard for civil litigation.[56]

[53] The process of analyzing and re-analyzing the same set of data in different ways to reach a target result.

[54] This is when a model violates fundamental model assumptions, includes unnecessary variables, and/or the incorrect model function is used to model the data.

[55] This occurs when the model omits a key variable and mis-estimates the effects of one of the other factors as a result.

[56] Under this assumption, the argument would be for each side of the litigation to draw their own null and alternative hypotheses. For example, defendants would state that the null was that no pay disparity occurred, and plaintiffs would state the alternative was that some amount of pay disparity occurred. Then Type I and Type II errors could be balanced based on the selected hypotheses. See

IX. The Issue of Practical Significance and Other Evidence in Hand.

A. Practical Significance

An additional issue concerns the concept of practical significance. This can be as important as statistical significance in a legal proceeding. For example, suppose the average pay at a company is $50,000 a year. Now, suppose a regression model was conducted to test if any disparate impact occurred in female pay, and that the regression model did find that females were underpaid by $15,000 on average per year when compared to males. Consider that the p-value on the Gender coefficient was 0.07, and that the regression model p-value was 0.07. Are these results significant? Should the trier of fact ignore an average $15,000 pay difference between men and women because the p-value did not hit a specific threshold? Now, suppose in a different legal proceeding, but with the same average pay for the company of $50,000, a regression showed the difference between men and women's pay was $1.00 on average per year. The p-value on the Gender coefficient was 0.00001 and the p-value for the regression was 0.00001. In the first example, the result has a practical significance and should be further explored, even though the p-value did not reach a "bright line" threshold. In the second example, the $1.00 a year different is 0.01% of the average pay for the company, and although the result is statistically significant, there may be an argument that there is no practical significance of a pay difference of 0.01%.

B. Other Evidence in Hand

In litigation, evidence found via a regression analysis does not exist in a bubble. Often, there is other evidence available that may or may not suggest discrimination has occurred. Using a set threshold for the p-value ignores the weight of the other evidence present in the case.

For example, suppose there is email documentation indicating a conspiracy to pay female employees less than equivalent male employees. Under this context, suppose the Gender coefficient has a p-value of 0.055, and the value of the Gender coefficient is $-5,000 per year,

---

Significant Statistics: The Unwitting Policy Making of Mathematically Ignorant Judges, Michael I. Meyerson & William Meyerson, Pepperdine Law Review Vol. 37:771, 2010 pgs. 842, 843.

which implies that females were paid less than males. Although the Gender coefficient is not statistically significant under a 0.05 threshold, the regression analysis estimates that the pay difference between males and females could be considered substantial. It may not be reasonable to ignore statistical evidence because it did not reach a pre-determined threshold when there is other evidence that suggests a conspiracy was in place. Under the current interpretation that courts have adopted, the use of pre-determined significance levels would lead to the above example not being considered statistically significant, even though other evidence in hand suggests that discrimination may have occurred.

X. Introduction to the Civil Rights Act and Gender Discrimination Regulations

The Equal Pay Act (EPA) was passed in 1963 and it amended the Fair Labor Standards Act (FLSA).[57] Its purpose was to address pay inequality and to prohibit pay discrimination on the basis of gender. In 1964, the United States codified the Civil Rights Act of 1964[58] that expressly prohibited discrimination based upon gender, race, religion, or age. This act went further than the EPA, with the stated goals to end discrimination based upon gender, race, or religion. The Civil Rights Act further barred all forms of employment discrimination (pay, hire, termination, and promotions) based upon gender. The EPA and the Civil Rights Act provided key laws to make behavior that discriminated against certain groups of people illegal. Specifically, Title VII of the Civil Rights Act created the Equal Employment Opportunity Commission to enforce Title VII. This commission is an independent body charged with researching, investigating, and eliminating discrimination in the workplace. Since that time, this Act has set the standard for gender-based pay discrimination in the context of litigation and has been used to bring multi-million and multi-billion-dollar class action lawsuits against employers in the United States.

To bring a claim for gender discrimination under the EPA, an individual must show that:

[57] https://www.dol.gov/whd/flsa/ (Last Accessed November 24, 2018).

[58] https://www.eeoc.gov/laws/statutes/titlevii.cfm (Last Accessed November 24, 2018).

1. A man and a woman;
2. Work at the same place; and
3. Do substantially the same job (equal work) but receive unequal pay.

There are rebuttals that can be used to demonstrate that the unequal pay was a result of bona-fide factors that include: seniority, merit, and a measure of the quantity or quality of work. There is also an opportunity to demonstrate that the pay gap is a result of any other unknown factors other than gender. Often, this takes the form of differences based on location, cost of living, tenure, specific skills, education, or organizational structure of the company. Recently, state and local governments have been passing legislation to tighten the rules around what constitutes a bona-fide factor and limit the rebuttals to gender pay gaps that may exist.[59]

XI. Introduction to the American Community Survey Data[60] and Other Data Sources Used for the Analysis

A. Data Sources

American Community Survey[61]

The American Community Survey (ACS)[62] is a yearly national survey conducted by the US Census Bureau. This survey contains a wide variety of questions that are given to individual and household respondents. The data contains individual responses for 1% of the United States population for one-year. This data contains information related to individual's incomes, gender, race, education status, marital status, age, hours worked, occupation, and industry. There are key factors that were developed out of this data that are used to estimate an individual's income.

[59] California Labor Code 1197.5. Notably, California passed a more stringent equal pay act in 2016. The new act changed which factors could be considered when explaining an existing gender pay gap. Furthermore, it changed how companies must document bona-fide factors and how those factors relate to the specific work that employees are performing and being compensated for.

[60] https://www.census.gov/programs-surveys/acs/ (Last Accessed November 24, 2018).

[61] The American Community Survey data is useful as an example dataset. However, it should not be used to determine if there is gender pay disparity in the at large community. Gender pay disparity, at least in terms of the existing regulation, should only be analyzed while considering the specific company an employee works for. The American Community Survey does not provide that level of detail, and any results identified in this analysis should be construed to identify gender pay disparities in the population at large.

[62] https://www.census.gov/programs-surveys/acs/ (Last Accessed November 24, 2018).

Since this is survey data, the data are anonymized and certain variables (such as income) are top-coded to prevent identification of individual respondents.[63] The data is also broken into Public Use Microdata Areas (PUMA) in which survey responses are gathered and weighted by the population located within the PUMA. The PUMA boundaries are drawn for each census and can contain at least 100,000 people each. Each observation is accompanied by a weighting value.

PUMA Geocoding to County

Since the area that an individual resides may be an important consideration in an individual's wages, geographical data that links PUMA codes to geographical regions from the Missouri Census Data Center's Geographic Correspondence Engine are used to link the PUMA to actual location.[64]

B. Data Preparation

The data was gathered for 2016 from California, for a selected list of OCC codes.[65] The OCC codes describe specific job types along with areas of the economy or industry in which the job function was performed. Occupation classifications are based upon work performed and on skills required to perform the work.

The data is filtered to observations that indicated that the person worked exactly 40 hours a week, that they worked at least 48 weeks out of the year, were currently employed during the select year, and work in the private sector. The purpose of this filter is to identify observations that were most likely full time employees with a complete wage history for the entire year. Additionally, this filtered out certain types of survey responses, such as government employees.[66]

[63] Top-coding is the process of replacing extreme values in certain variables, such as income, to a preset high value. This has the effect of incorporating responses into a one set value but was done to prevent the identification of individual responses to the survey. The final dataset will be filtered to remove any responses with top coded values.

[64] http://mcdc.missouri.edu/websas/geocorr12.html (Last Accessed November 24, 2018).

[65] OCC Codes are US Census occupation codes that group similar job into categories. Please see https://www.bls.gov/soc/ (Last Accessed November 24, 2018).

[66] Due to various transparency laws there is less likelihood of gender pay gap.

The data was also reviewed to assess the validity of the responses and adjustments were made to the data when the responses to the surveys do not appear to be accurate.[67] Additionally, survey responses in which the observations are for people currently enrolled in undergraduate college or high school were also removed.[68] The ACS data has nearly 300 variables and many of these variables are not relevant to this analysis. Variables related to gender, location, wages, occupation, age, and education were included. Also, for purposes of this analysis, the dependent variable, "wages," was simulated from the other data. This was performed so the "truth" of the model is pre-determined, and thus it is possible to test various hypothetical scenarios with the data.

XII. Introduction to Econometric Discrimination Analysis

As previously discussed, the gender pay gap analysis can be conducted by performing a regression and testing the statistical significance of the coefficient on the gender variable. The following model could be used to estimate the difference between men's and women's pay:[69]

$$\boldsymbol{Log(Pay) = \beta_o + \beta_1 Gender + \beta_2 Education + \beta_3 Tenure + \beta_4 Occupation + \beta_5 Location + \varepsilon}$$

This model will be referred to as "Model 1." This model is testing the percentage difference between men's and women's annual pay[70] after controlling for education, tenure, occupation, and location. The gender variable is the variable of interest, and the model is testing the impact of changes to the model on the p-value of the Gender coefficient. As the next section will discuss, the dependent variable (wages) will be simulated based on the data, and then I will make

[67] For example, when survey responses for the WAGP field (annual wages) are less than or equal to five dollars an hour (based on a 2080 work year), they are considered invalid and removed from the analysis.

[68] The purpose was to keep only individuals that are of working age, not in high school or college, and considered fully employed.

[69] The wage field can be highly skewed; therefore, it is common to perform a log transformation of the Wage field before conducted a regression analysis. This transformation changes the interpretation of the Gender variable from an absolute difference to an approximate percentage difference.

[70] Since the dependent variable, wages, has been log transformed, the coefficient on the Gender is analogous to the percent difference between men's and women's pay.

modifications to the model to evaluate the changes in the statistical significance (based on the p-value) of the Gender coefficient.

XIII. Areas of Analysis

A series of hypothetical analyses was conducted see how adjustments to the dependent variables included in the regression model can distort the underlying truth. I simulated the Wage field, so that I know the truth as to whether or not any gender pay difference exists based on the data and the model. However, I set up problems so that specific changes in how the analysis is conducted impact the p-value in various ways. The following hypotheticals were set up:

1. Gender pay differences occurred.
2. Gender pay differences occurred based on a specific known formula.
3. Gender pay differences did not occur.

For the above hypotheticals, I analyzed the following.

1. Analyze the effect on the hypothesis test based on different numbers of observations. The overall sample size can lead to no discrimination being found, even though discrimination actual did occur. More troubling is that discrimination can be found based only on sample size. A small difference in pay on a large enough dataset will show statistical significance. That pay difference may or may not be practically significance.
2. Analyze the power of the hypothesis test for the Gender coefficient based on the value of the p-value. The power of a hypothesis test is rarely analyzed or reviewed during the course of litigation. However, depending on the sample size, a hypothesis test can conclude statistical significance, but the hypothesis test is under-powered. This suggests that there is lower probability of properly rejecting the null hypothesis.
3. Analyze the effect of omitted variables from the model. Omitted variable bias occurs when the results of the hypothesis test would be different but for the removal of a variable or

variables. Although it is challenging to determine the proper model during litigation often due to the available data, understanding what the correct variables to include or exclude from the model is critical. Often, the exclusion of a single variable can show that discrimination occurred or did not occur.

I intend to show that under various scenarios how a "bright-line" threshold of the p-value can lead to biased or underpowered conclusions when compared to the truth of the data.

XIV. Simulated Data Preparation

I used the ACS data from 2016, and I filtered it to simulate a medium sized company. I selected the data from the following locations: San Francisco, San Diego, and Los Angeles[71] and filtered the data to specific occupations.[72] After filtering, my dataset contained 4,086 observations. I created my final analysis dataset with a simulated Wages variable by performing several analysis steps. First, I conducted an initial regression analysis using Model 1.

The results from this initial regression are below. They suggest that there is a 15% (Gender Female Coefficient) difference between Men and Women's pay, and that according to the p-value this result is statistically significant based on a significance level of 0.05.

Table 1: Initial Regression Based on American Community Survey Data

Coefficient	Estimate	Error	T-Stat	p-Value	Lower	Upper
(Intercept)	10.25	0.11	97.06	0.000	10.039	10.453
Gender Female	-0.15	0.02	-7.50	0.000	-0.184	-0.108
Highschool Or Equivalent	0.08	0.11	0.78	0.434	-0.126	0.293
Some College	0.28	0.10	2.75	0.006	0.080	0.477
Bachelors	0.54	0.10	5.38	0.000	0.342	0.735
Masters	0.75	0.10	7.46	0.000	0.556	0.953
Professional or Ph.D.	0.89	0.11	8.26	0.000	0.679	1.102
Computer Programmers (1010)	0.22	0.05	4.30	0.000	0.119	0.319
Computer System Analysts (1006)	0.26	0.05	5.16	0.000	0.159	0.354
Financial Managers (0120)	0.25	0.04	6.45	0.000	0.172	0.322

[71] The PUMA Geocoding provides very specific location information. For purposes of developing my dataset for analysis, I combined several PUMA codes into distinct location values. PUMA coded to "Alameda," "San Francisco," "San Mateo," "Santa Clara," and "Santa Cruz" were considered "San Francisco." PUMA coded to "San Diego" and "Orange" were considered "San Diego. PUMA coded to "Los Angeles" were considered "Los Angeles."

[72] I filtered the data to the following occupation descriptions based on OCC codes and descriptions: "Computer System Analysts (1006)," "Software Developers (1020)," "Information System Managers (0110)," "Misc. Engineers (1530)," "Accountants (0800)," "Financial Managers (0120)," "General and Operational Managers (0020)," "Marketing and Sales Managers (0050)," "Computer Programmers (1010)," and "Misc. Managers (0430)."

General and Operations Managers (0020)	0.15	0.05	3.17	0.002	0.055	0.235
Information Systems Managers (0110)	0.40	0.05	8.59	0.000	0.307	0.489
Marketing and Sales Managers (0050)	0.26	0.04	6.55	0.000	0.179	0.332
Misc. Engineers (1530)	0.30	0.04	6.88	0.000	0.213	0.383
Misc. Managers (0430)	0.25	0.03	8.50	0.000	0.191	0.306
Software Developers (1020)	0.43	0.03	13.14	0.000	0.368	0.497
San Diego	0.09	0.02	3.86	0.000	0.044	0.134
San Francisco	0.28	0.02	12.59	0.000	0.233	0.319
Tenure	0.01	0.00	16.46	0.000	0.011	0.014

The process for creating a zero-pay disparity model is simple. I performed another regression analysis but using only observations for "Males" and excluded the Gender variable. The model is of the following form:

$$\boldsymbol{Log(Pay) = \beta_o + \beta_1 Education + \beta_2 Tenure + \beta_3 Occupation + \beta_4 Location + \varepsilon}$$

This will be referred to Model 2. I used the results of this regression to create a new Wages field to build a dataset in which no gender pay difference can be found. First, I ran a regression on just the observations coded to "Males" in the data without the Female indicator variable. Then, I predicted the values of wages based on this regression, but only on observations coded as "Female." Last, I recoded the Wage variable to be equal to predicted values for when the observation was coded as "Female." I made no changes to the Wage variable when the observation was coded as "Male." I ran the same regression on this simulated dataset and now the Gender Female coefficient is showing a 0.00 estimate and no statistical significance.

Table 2: Regression Based on Modified American Community Survey Data to Remove Pay Differences

Coefficient	Estimate	Error	T-Stat	p-Value	Lower	Upper
(Intercept)	10.26	10.26	122.49	0.000	10.092	10.420
Gender Female	0.00	0.00	0.00	1.000	-0.030	0.030
Highschool Or Equivalent	0.01	0.01	0.12	0.903	-0.156	0.177
Some College	0.23	0.23	2.85	0.004	0.072	0.386
Bachelors	0.51	0.51	6.36	0.000	0.350	0.661
Masters	0.71	0.71	8.81	0.000	0.550	0.865
Professional or Ph.D.	0.85	0.85	9.93	0.000	0.682	1.018
Computer Programmers (1010)	0.19	0.19	4.76	0.000	0.113	0.272
Computer System Analysts (1006)	0.24	0.24	6.08	0.000	0.162	0.317
Financial Managers (0120)	0.31	0.31	10.06	0.000	0.246	0.365
General and Operations Managers (0020)	0.09	0.09	2.36	0.018	0.015	0.157
Information Systems Managers (0110)	0.42	0.42	11.34	0.000	0.345	0.489
Marketing and Sales Managers (0050)	0.33	0.33	10.59	0.000	0.267	0.388

Misc. Engineers (1530)	0.27	0.27	7.84	0.000	0.202	0.336
Misc. Managers (0430)	0.24	0.24	10.51	0.000	0.198	0.289
Software Developers (1020)	0.42	0.42	16.25	0.000	0.373	0.476
San Diego	0.10	0.10	5.69	0.000	0.068	0.139
San Francisco	0.30	0.30	17.56	0.000	0.271	0.339
Tenure	0.01	0.01	22.00	0.000	0.012	0.014

Although the regression analysis using Model 1 shows that there is no difference between male and female wages, there still exists a sizable pay difference between male and female wages. This is because the regression model is able to develop estimates that account for multiple factors present in the data. A simple view of average wages by job title and gender shows significant pay differences as shown in Table 3 below. Even though there are pay differences between male and female wages, under a plain interpretation of the law and court precedent, the results from Model 2 may not be considered discriminatory.[73]

Table 3: Average Wage Summary by Gender and Occupation

Occupation	Gender	Mean Wage	Observations
Accountants (0800)	Male	96,386	206
	Female	71,685	443
Computer Programmers (1010)	Male	106,144	138
	Female	90,122	24
Computer System Analysts (1006)	Male	98,587	94
	Female	86,609	72
Financial Managers (0120)	Male	126,522	153
	Female	89,048	190
General and Operations Managers (0020)	Male	83,770	115
	Female	73,716	93
Information Systems Managers (0110)	Male	129,134	142
	Female	118,648	62
Marketing and Sales Managers (0050)	Male	112,987	164
	Female	92,303	159
Misc. Engineers (1530)	Male	121,118	215
	Female	106,591	45
Misc. Managers (0430)	Male	114,735	590
	Female	91,811	433
Software Developers (1020)	Male	137,954	618
	Female	124,650	130

[73] This statement assumes that no other evidence exists to suggest discrimination toward females, and that the model was properly specified and properly modeled the data.

The purpose of this procedure is to start with a dataset that shows no pay differences between males and females and then introduce differences in a controlled manner. From here, I can adjust various parameters that can change in the real world and analyze the impact on the p-value.

XV. Hypothetical 1 – Random Pay Differences

Now suppose that the company had pay differences between male and female employees, but that the differences were not intentional. I will assume that the pay differences were simply a result of hiring individuals at different wage rates based on prior wages at a different company without consideration to current employee's wages in the same occupation.[74]

Using the same regression model as Model 1, Table 4 shows that there is a 5% pay difference between male and female wages, and Table 4 shows that the 5% difference is significant based on a p-value threshold of 0.05 as the statistical significant threshold.

Table 4: Regression Based on Modified American Community Survey Data with Pay Differences Assigned Randomly

Coefficient	Estimate	Error	T-Stat	p-Value	Lower	Upper
(Intercept)	10.20	0.09	118.39	0.000	10.033	10.371
Gender Female	-0.05	0.02	-2.93	0.003	-0.078	-0.015
Highschool Or Equivalent	-0.01	0.09	-0.12	0.907	-0.181	0.161
Some College	0.23	0.08	2.83	0.005	0.072	0.395
Bachelors	0.52	0.08	6.41	0.000	0.364	0.684
Masters	0.73	0.08	8.81	0.000	0.566	0.890
Professional Or Phd	0.88	0.09	9.96	0.000	0.704	1.050
Computer Programmers (1010)	0.21	0.04	5.11	0.000	0.131	0.295
Computer System Analysts (1006)	0.25	0.04	6.26	0.000	0.174	0.333
Financial Managers (0120)	0.32	0.03	10.26	0.000	0.259	0.382
General And Operations Managers (0020)	0.11	0.04	2.83	0.005	0.032	0.179
Information Systems Managers (0110)	0.43	0.04	11.48	0.000	0.360	0.508
Marketing And Sales Managers (0050)	0.34	0.03	10.66	0.000	0.277	0.402
Misc. Engineers (1530)	0.28	0.04	7.89	0.000	0.210	0.348
Misc. Managers (0430)	0.26	0.02	10.71	0.000	0.209	0.303
Software Developers (1020)	0.44	0.03	16.37	0.000	0.387	0.493
San Diego	0.11	0.02	5.67	0.000	0.070	0.143
San Francisco	0.31	0.02	17.55	0.000	0.279	0.349
Tenure	0.01	0.00	21.98	0.000	0.012	0.015

This regression uses 4,086 observations, which for a regression analysis is a substantial number of observations. However, the p-value is susceptible to sample size, and increasing the number

[74] For this analysis, I adjusted the original Wage variable created by Model 2, which shows no gender pay differences, by applying differing ranges of normally distributed random variables to observations coded as male and separately to females.

of observations that are analyzed can lead to a statistical significant result, even if the difference between the male and female wages are slight. This can be demonstrated by performing a simulation of hundreds of random draws of data, artificially expanding the number of observations and running a regression on the artificially expanded data. Even though there may or may not be a significant difference based on the sample data, simply enlarging the data can create a significant result. A proportional random sample of the data by occupation, with a minimum draw amount of 30 observations was drawn from the original 4,086 observations to produce a sample dataset of 706 observations. Table 5 shows the representative sample drawn from the original 4,086 observations.

Table 5: Proportional Random Sample Draw by Occupation

Occupation	Observations	% of Population	Sample Amount
Misc. Managers (0430)	1,023	0.25	255
Financial Managers (0120)	343	0.08	30
Information Systems Managers (0110)	204	0.05	30
Computer System Analysts (1006)	166	0.04	30
Computer Programmers (1010)	162	0.04	30
Marketing and Sales Managers (0050)	323	0.08	30
Misc. Engineers (1530)	260	0.06	30
Software Developers (1020)	748	0.18	134
General and Operations Managers (0020)	208	0.05	30
Accountants (0800)	649	0.16	103

The data is simulated via a random draw process of 100 times, and for each draw the data is expanded from 1 to 100 times the original sample size at each simulation loop. In total, 10,000 models with a range in observations from 702 to 70,200 were created. After the 10,000 models were created, the results are averaged by the number of copies generated during the process. As Figure 1 shows below, by increasing the number of observations a significant result can be achieved, even though the underlying sample may or may not have had a statistically significant result. This occurs because the number observations are found in the denominator of the standard error calculation for the regression coefficient. Therefore, as the number of observations increases the standard error on the coefficient decreases. This makes finding a significant result

using a “bright line” test more likely because the model and the number of observations can be adjusted to achieve the “bright line” threshold.[75]

Figure 1: Simlulated p-value and Standard Error over Expanded Observation Size

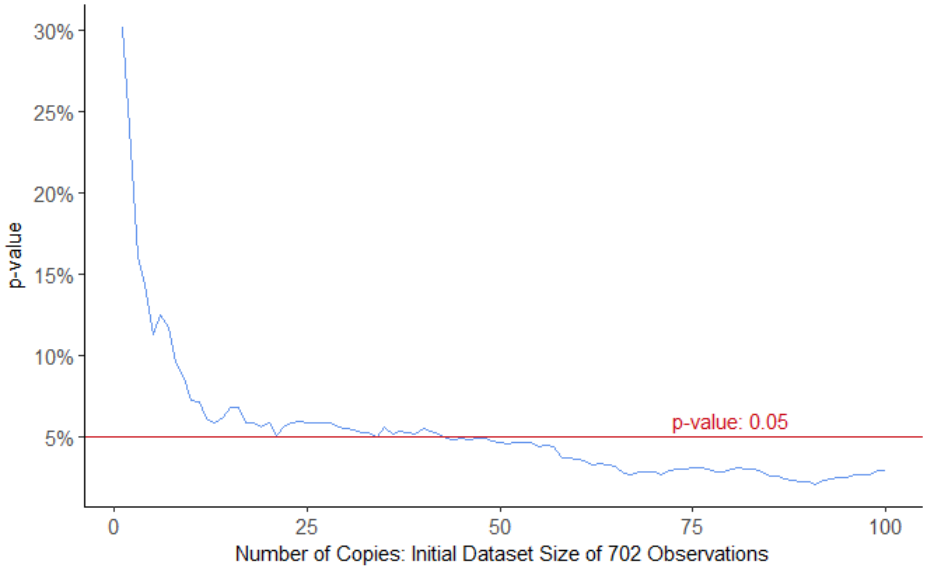
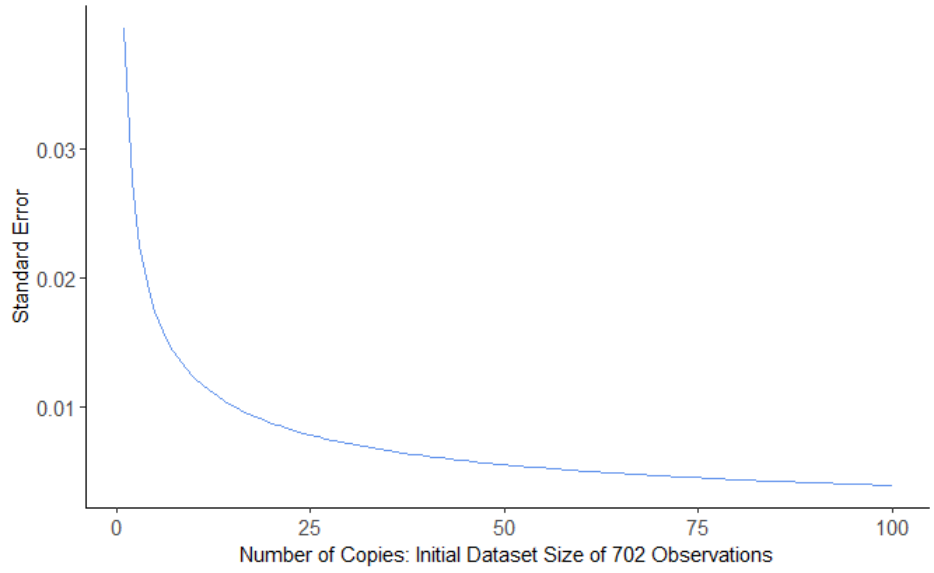


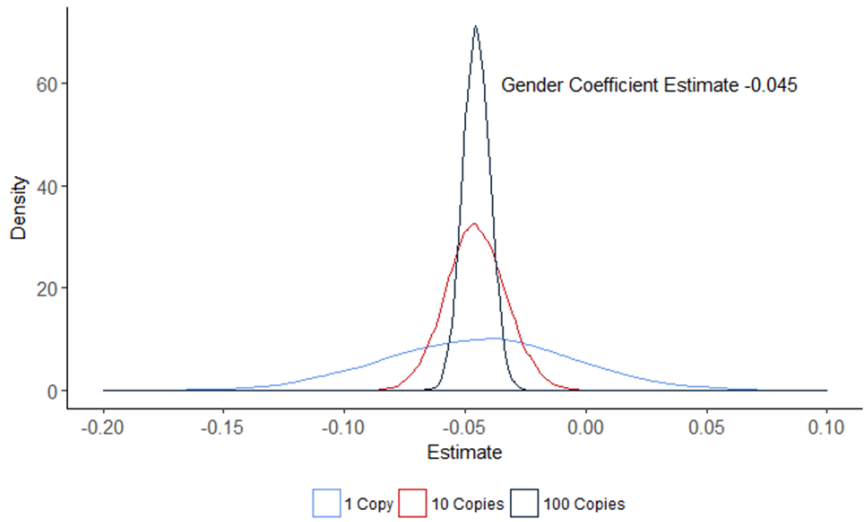
Figure 2: Simlulated p-value and Standard Error over Expanded Observation Size



[75] See Preliminaries. Recall that $t^* = \frac{b_k}{s\{b_k\}}$. Where t^* is the critical t value which determines the p-value for the hypothesis test. Also, $s\{b_k\}$ is calculated as $s\{b_k\} = MSE * (X'X)^{-1}$ and MSE is calculated as $\frac{SSE}{n-p}$. Therefore, by adjusting the number of observations the relationship is re-enforced and driving the $s\{b_k\}$ to be smaller as $n - p$ is increased. Although this simulation is not likely with real data, the effect of adding more observations to a regression model can drive the b_k estimate to statistically significant based on the number of observations analyzed even though the effect size of b_k is small.

Figures 1 and 2 demonstrate that simply by increasing the sample size even though the underlying data is exactly the same, the p-value decreases as sample size is increased. This shows that running linear regression on a large dataset could lead to significant results just based on the sheer size without consideration of any underlying facts. This occurs because as more observations are added to the regression the additional observations reinforce the relationship and the additional observations drive the standard error down.[76] This in effect tightens the confidence intervals and makes significance more likely to appear. However, this example shows that the p-value is simply decreasing as a pure function of sample size. This leads into the next hypothetical about increasing the power of the test. Simply increasing the number of observations will increase the power of the test (because the confidence interval is getting tighter).

Figure 3: Distribution Around Gender Coefficient Estimate by Simulated Copies of the Data



As Figure 3 shows,[77] if the relationship is real and exists throughout the data then obtaining more data will ultimately lead to higher power results. However, as more observations are added, it is

[76] Effectively, if the relationship exists in a smaller subset of the data, the additional observations added to the model re-enforce the relationship.

[77] See Preliminaries. This is additional evidence for how the number of observations can influence the distribution around the estimate of b_k. In the initial regression with 4,086 observations the b_k for Gender is -0.045, with a standard error of 0.015.

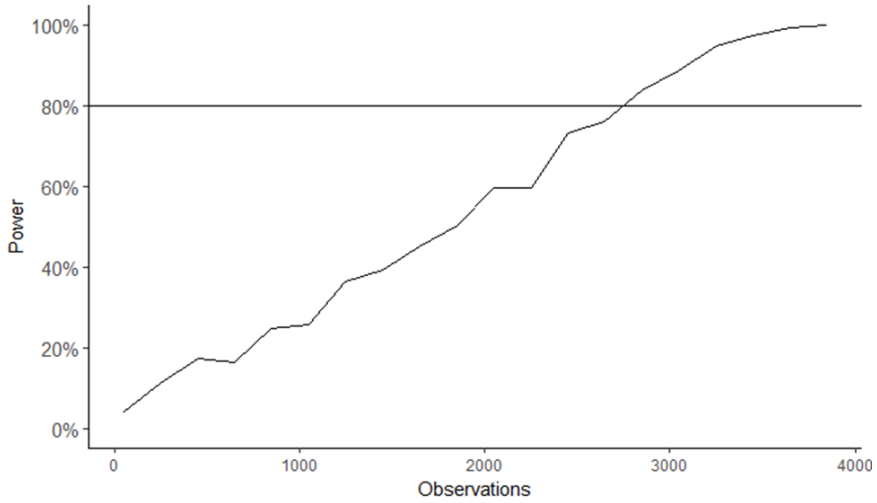
easier to detect a statistical significant difference based on a "bright line" threshold. The Power of the test should be an important consideration in hypothesis testing, but it does not seem to be a focus of the courts when determining the efficacy of a hypothesis test. Let us take our prior hypothetical and conduct a Power test. Typically, post-hoc Power tests are not recommended because researchers should make sample size decisions before conducting hypothesis tests. However, in the context of litigation, a post-hoc Power test can illuminate how the hypothesis test is performing and if the number of observations used for the analysis is sufficient. The Power of the test is a useful statistic that determines the probability that the hypothesis test will reject the a truly false hypothesis. A low Power suggests that the hypothesis has little probative value, and it could be likely that the Null Hypothesis is being falsely rejected.

To determine the Power, I performed a simulation by applying Model 1 to the data created for Hypothetical 1. First, I randomly sampled the data at various numbers of observations between 50 and 4,000 and performed the regression analysis for each sample. I conducted the simulation 500 times for each amount of sampled observations. Next, I calculated the estimate and p-value of the Gender coefficient for each iteration of the simulation. Last, I determined which of the simulations produced a significant result using a 0.05 p-value threshold. The Power is the number of times the simulation produced the expected result.[78]

However, when the data is sampled to 702 observations, the b_k for Gender in the sample is still -0.045, but the standard error is 0.040. The standard error (s.e.) is defined as $\frac{\sigma}{\sqrt{n}}$, therefore with a standard error of 0.040 and 702 observations the variance is 1.12. This is derived as $0.040 = \frac{\sigma}{\sqrt{702}} \rightarrow \sigma = \sqrt{702} * 0.040 \rightarrow \sigma = 1.06 \rightarrow \sigma^2 = 1.12$. So simply increasing the number of observations from 702 to 7,020 lowers the standard error from 0.040 to 0.012. $s.e. = \frac{1.06}{\sqrt{7020}} = 0.012$. The increase in observations tightens the interval around the b_k estimate for Gender.

[78] I denoted 80% power with a horizontal line. A power of 80% was initially devised by Cohen for when an "investigator has no other basis for setting the desired power value." See Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. Hove: Lawrence Erlbaum Associates, 1988, pg. 56.

Figure 4: Power of the Hypothesis Test on the Gender Coefficient
Over the Number of Observations in the Model



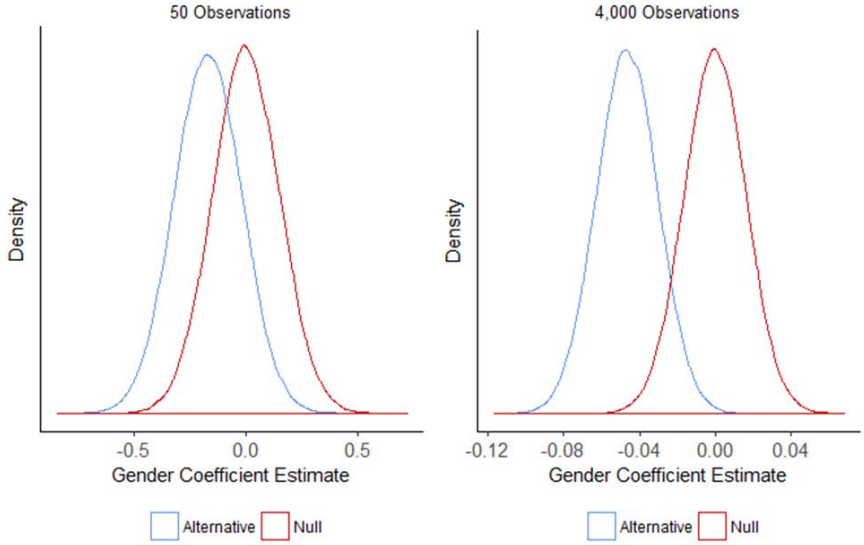
As Figure 4 shows when the regression analysis has a small number of observations an observed statistically significant result has little Power,[79] but as the number of observations increases the Power of the test increases. This suggests that if a regression analysis has a small number of observations, the p-value on the Gender coefficient in the regression may have little probative value.[80] Power can also be visualized by showing the distributions of the Null and Alternative hypotheses. In Figure 5, the distributions of Gender coefficient estimate are plotted under the Null Hypothesis and the Alternative Hypothesis.[81]

[79] The power plot may look odd based on the number of observations and the power. The low power at what is a relatively large number of observations is expected based on the way I designed the data for the simulations. The Gender effect in these models was designed to be very noisy, therefore it takes a significant amount of observations to show a real effect.

[80] See Preliminaries. The Power is defined as $1-\beta$ where β is defined as $= P(\bar{X} \leq \bar{x}; H_a) = P\left(\frac{\bar{X}-\mu_a}{\sqrt{\sigma^2/n)}} < \frac{\bar{x}-\mu_a}{\sqrt{\sigma^2/n)}}\ ; H_a\right)$. This equation determines the likelihood that the test is correctly rejecting the Null hypothesis. If a test has low Power, this suggests that there is a higher likelihood that the Null could be rejected even though the Null is correct. From the equation for β, it is clear that the number of observations is present in the denominator. Therefore, increasing the number of observations will lower β and increase the Power. As seen in Preliminary Figure 3, the Power is expected to increase as the number of observations increases.

[81] Recall Preliminary Figure 1. β is the area under the distribution that falls under the Alternative distribution to the right of the point estimate that falls under the Null distribution. For the distribution based on 50 observations, the Null and Alternative distributions overlap significantly more than the Null and Alternative distributions based on 4,000 observations. This shows that that model with 50 observations has less Power than the model with 4,000 observations. Additionally, as the number of observations increases, the distributions around the estimate get tighter. This further increases the Power as less of the Alternative distribution can fall under the Null distribution.

Figure 5: Distribution of the Gender Estimates Under the Null and Alternative Hypothesis based on Different Number of Observations



XVI. Hypothetical 2 – Intentional Pay Differences

Let us suppose the company had a policy to discriminate against female software developers.[82] However, all occupations were included in the litigation. The analyst naturally includes all occupations in the regression analysis based on Model 1's equation, and the analyst produces a summary regression model as follows:

Table 6: Regression Based on Modified American Community Survey Data with Pay Differences Assigned to Software Developers

Coefficient	Estimate	Error	T-Stat	p-Value	Lower	Upper
(Intercept)	10.29	0.08	121.72	0.000	10.125	10.456
Gender Female	-0.02	0.02	-1.60	0.109	-0.055	0.006
Highschool Or Equivalent	0.01	0.09	0.14	0.885	-0.156	0.180
Some College	0.23	0.08	2.83	0.005	0.071	0.388
Bachelors	0.50	0.08	6.24	0.000	0.344	0.658
Masters	0.70	0.08	8.66	0.000	0.543	0.861
Professional or Ph.D	0.85	0.09	9.84	0.000	0.681	1.019
Computer Programmers (1010)	0.18	0.04	4.38	0.000	0.099	0.259
Computer System Analysts (1006)	0.23	0.04	5.83	0.000	0.154	0.310
Financial Managers (0120)	0.30	0.03	9.85	0.000	0.242	0.362
General and Operations Managers (0020)	0.08	0.04	2.16	0.031	0.007	0.151
Information Systems Managers (0110)	0.41	0.04	10.96	0.000	0.334	0.479
Marketing and Sales Managers (0050)	0.32	0.03	10.25	0.000	0.259	0.382
Misc. Engineers (1530)	0.26	0.03	7.38	0.000	0.188	0.324

[82] For this analysis, I adjusted the Wage variable for Female Software Developers downward. I calculated the average Wage for Male Software Developers and adjusted all Female Software Developers to a certain percentage below the average. The adjustment was made only on the Software Developer occupation.

Misc. Managers (0430)	0.24	0.02	10.12	0.000	0.191	0.283
Software Developers (1020)	0.38	0.03	14.23	0.000	0.324	0.427
San Diego	0.10	0.02	5.42	0.000	0.064	0.136
San Francisco	0.31	0.02	17.54	0.000	0.273	0.342
Tenure	0.01	0.00	20.78	0.000	0.011	0.014

As Table 6 shows there is a negative coefficient on Female that suggests an approximate 2% pay difference between females and males. If this is the only evidence presented to the court, the court may reject the prima facie evidence because the p-value on the Gender coefficient is not less than 0.05. However, the intentional pay disparity may or may not have been known through the evidence gathering process. More analysis and inspection may be called for considering what the statistics on the Gender coefficient were. Rejecting the case because of the initial prima facie evidence not achieving a 0.05 p-value may not have been proper in this case. Now let us suppose the analyst was more careful and analyzed each occupation separately. The following regression equation was used while filtering the dataset to each occupation and running the regression separately for each occupation.[83]

$$\boldsymbol{Log(Pay) = \beta_o + \beta_1 Gender + \beta_2 Education + \beta_3 Tenure + \beta_4 Location + \varepsilon}$$

Table 7 shows the statistics for only the Gender coefficient if each occupation was isolated in the analysis.

Table 7: Regression Based on Modified American Community Survey Data with Pay Differences Assigned to Software Developers. Occupation is Isolated in each Regression

Occupation	Gender Coefficient Estimate	Error	T-Stat	p-Value	Lower	Upper
Misc. Managers (0430)	0.00	0.03	-0.01	0.993	-0.062	0.061
Financial Managers (0120)	0.03	0.05	0.54	0.592	-0.069	0.121
Information Systems Managers (0110)	0.01	0.07	0.17	0.868	-0.125	0.149
Computer System Analysts (1006)	-0.03	0.05	-0.58	0.562	-0.122	0.066
Computer Programmers (1010)	-0.03	0.13	-0.20	0.843	-0.289	0.237
Marketing and Sales Managers (0050)	-0.01	0.04	-0.15	0.883	-0.092	0.079
Misc. Engineers (1530)	-0.01	0.07	-0.08	0.937	-0.136	0.126
Software Developers (1020)	-0.19	0.05	-4.09	0.000	-0.281	-0.099
General and Operations Managers (0020)	0.00	0.06	0.05	0.958	-0.122	0.128
Accountants (0800)	0.00	0.03	0.07	0.944	-0.066	0.071

[83] This model is referred to as Model 3.

In effect, the other observations with no pay differences between male and females crowd out the real effect of the pay differences for software developers. By isolating occupation, the coefficient for Software Developers show a statistically significant result and an approximate 19% pay difference. Since courts have adopted "bright line" p-value thresholds, the initial prima facie regression analysis may have ended the litigation, even though a real issue likely existed for a specific occupation.

XVII. Hypothetical 3 – Omitted Variable Bias

A significant issue with this type of regression analysis is the selection of variables to include or exclude from the model. Based on variable selection, it is very easy to manipulate the p-value result. Let us look at several examples based on the data simulated for this analysis. Table 8 shows the results from using Model 1 and adjusting the wages so that there is no pay disparity based on the full model.[84]

Table 8: Regression Based on Modified American Community Survey Data with No Statistical Significant Pay Differences

Coefficient	Coefficient Estimate	Error	T-Stat	p-Value	Lower	Upper
(Intercept)	10.26	0.08	122.34	0.000	10.097	10.426
Gender Female	-0.03	0.02	-1.92	0.054	-0.060	0.001
Highschool Or Equivalent	0.01	0.09	0.11	0.912	-0.157	0.176
Some College	0.23	0.08	2.86	0.004	0.072	0.387
Bachelors	0.51	0.08	6.37	0.000	0.351	0.663
Masters	0.71	0.08	8.84	0.000	0.553	0.869
Professional or Ph.D	0.86	0.09	10.00	0.000	0.689	1.025
Computer Programmers (1010)	0.20	0.04	4.89	0.000	0.119	0.278
Computer System Analysts (1006)	0.24	0.04	6.13	0.000	0.165	0.319
Financial Managers (0120)	0.31	0.03	10.15	0.000	0.249	0.368
General and Operations Managers (0020)	0.09	0.04	2.36	0.018	0.014	0.157
Information Systems Managers (0110)	0.42	0.04	11.39	0.000	0.347	0.491
Marketing and Sales Managers (0050)	0.33	0.03	10.62	0.000	0.269	0.390
Misc. Engineers (1530)	0.27	0.03	7.94	0.000	0.206	0.341
Misc. Managers (0430)	0.25	0.02	10.57	0.000	0.200	0.291
Software Developers (1020)	0.43	0.03	16.52	0.000	0.381	0.484
San Diego	0.10	0.02	5.57	0.000	0.066	0.137
San Francisco	0.27	0.02	15.53	0.000	0.236	0.304
Tenure	0.01	0.00	22.23	0.000	0.012	0.015

This regression, based on all the available variables, shows no prima facie evidence of a pay difference using a 0.05 p-value threshold. However, selectively choosing the variables that are

[84] I adjusted wages by applying negative uniformly distributed random variables to the Wage field for Females in San Francisco.

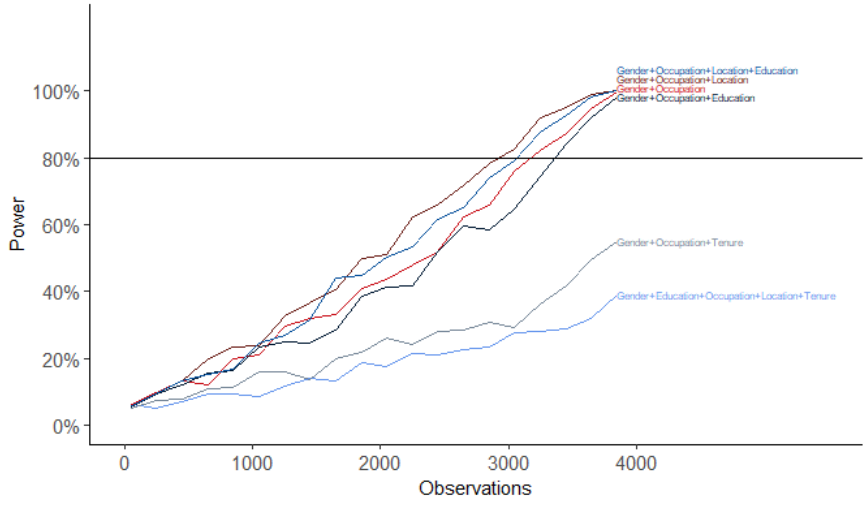
included in the model can change the p-value on the Gender coefficient and show a statistically significant result.

Table 9: Regression Based on Modified American Community Survey Data Gender Coefficient Statistics based on Various Models

Model	Gender Coefficient Estimate	Error	T-Stat	p-Value	Lower	Upper
Occupation+Location	-0.05	0.02	-2.85	0.004	-0.084	-0.015
Occupation+Location+Education	-0.04	0.02	-2.73	0.006	-0.076	-0.012
Occupation	-0.05	0.02	-2.62	0.009	-0.083	-0.012
Occupation+Education	-0.04	0.02	-2.51	0.012	-0.075	-0.009
Occupation+Tenure	-0.04	0.02	-2.05	0.040	-0.070	-0.002
Full Model	-0.03	0.02	-1.92	0.054	-0.060	0.001

Table 9 shows the effect of using different variables in the model. The other models, which are not the full model, shows a statistically significant result assuming a 0.05 p-value threshold. However, if you consider the 95% confidence interval, the Occupation + Tenure and the Occupation + Education models are very close to 0. A post hoc Power simulation highlights how Power reacts to the number of observations while considering each of the model parameters. Very clearly, the Full Model has weak Power (as expected). Even though the Occupation + Tenure model shows statistical significance at the 0.05 level, the model shows weak Power through all amounts of observations. Even the Occupation + Education model has the lowest Power of the other models that eventually reach 80% Power threshold line.

Figure 6: Power of the Hypothesis Test on the Gender Coefficient
Over the Number of Observations by Model



In this hypothetical, the analyst can play with the data to produce either a statistical significant result or a non-statistically significant result based on the variables selected. Furthermore, in certain circumstances, the statistical significant result has very weak Power even on the full data. The confidence intervals also demonstrate that even though the overall result may be statistically significant, the interval contains values that could be considered practically non-significant.

XVIII. Conclusion

As demonstrated, a p-value can be manipulated to produce a statistical result or a non-statistical result by adjusting the variables included or excluded from the model, or by adjusting the number of observations to include. Analysts should take care when performing these types of regressions and consider other statistics such as the confidence interval of the Gender coefficient and the overall Power of the hypothesis being conducted. There are two extensions to the preceding analyses presented in this thesis. The first is to study and determine what, if any, impact a Bayesian approach may have on the conclusions drawn from regression analysis and the use of “bright line” tests. Second, is to perform a deeper inspection of the analysis that was touch on in Figure 1 and Figure 2. As Figure 1 and Figure 2 suggested, putting more observations into a model could show a statistical significant result. A deeper inspection of use of classical regression techniques, “Big” data, and its use in litigation is warranted.

Bibliography

1. Castaneda v. Partida, 430 U.S. 482 (1977). (n.d.). Retrieved from https://supreme.justia.com/cases/federal/us/430/482/ (Last Accessed December 28, 2017)

2. Hazelwood Sch. Dist. v. United States, 433 U.S. 299 (1977). (n.d.). Retrieved from https://supreme.justia.com/cases/federal/us/433/299/ (Last Accessed December 28, 2017)

3. David H. Kaye, Is Proof of Statistical Significance Relevant? Penn State Law, 1986.

4. https://definitions.uslegal.com/b/bright-line-rule/ (Last Accessed November 24, 2018).

5. https://www.law.cornell.edu/wex/prima_facie) (Last Accessed November 24, 2018).

6. Correspondence to Dr. Sander Greenland, Department of Epidemiology, School of Public Health, University of California, 2017, American Journal of Epidemiology Vol 186, No. 6 DOI: 10.1093/aje/kwx259.

7. Daniel J. Benjamin Et. Al in Nature Human Behaviour, 2017 (DOI: 10.1038/s41562-017-0189-z).

8. Valentin Amrhein Et. Al in Nature Human Behaviour, 2017 (DOI: 10.1038/s41562-017-0224-0).

9. Phillip Johnson, Edward Leamer, and Jeffrey Leitzinger, Statistical Significance and Statistical Error in Antitrust Analysis, Antitrust Law Journal, American Bar Association, 2017 Volume 81 Issue 2.

10. Everitt, Brian. The Cambridge Dictionary of Statistics. Cambridge, UK New York: Cambridge University Press. 4th Edition, 2010.

11. Hogg, Robert V, et al. Probability and Statistical Inference. Pearson, 2015.

12. Lehmann, Erich Leo., and Joseph P. Romano. Testing Statistical Hypotheses. Springer, 2008.

13. Johannes Lenhard, Models and Statistical Inference: The Controversy between Fisher and Neyman–Pearson, Brit. J. Phil. Sci. 57 (2006).

14. S.P. Bloomberg, Power Analysis Using R, January 7, 2014

15. Kutner, Michael H. Applied Linear Statistical Models, McGraw-Hill Irwin, 2005.

16. https://www.law.cornell.edu/wex/trier_of_fact (Last Accessed November 24, 2018).

17. https://www.law.cornell.edu/wex/discovery (Last Accessed November 24, 2018).

18. https://www.archives.gov/education/lessons/civil-rights-act (Last Accessed November 24, 2018).

19. Hansen, Bruce. Econometrics, University of Wisconsin, 2018.

20. Henry W. Segar, et al. v. William French Smith, Attorney General, et al., Appellants.henry W. Segar et al., Cross-appellants v. William French Smith, Attorney General, et al, 738 F.2d 1249 (D.C. Cir. 1984). (n.d.). Retrieved from https://law.justia.com/cases/federal/appellate-courts/F2/738/1249/135384/.

21. David H. Kaye, The Dynamics of Daubert: Methodology, Conclusions, and Fit in Statistical and Econometric Studies, 87 Va. L. Rev. 1933 (2001).

22. Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993). (n.d.). Retrieved from https://supreme.justia.com/cases/federal/us/509/579/.

23. https://definitions.uslegal.com/d/daubert-challenge/ (Last Accessed November 24, 2018).

24. Kumho Tire Co. v. Carmichael, 526 U.S. 137 (1999). (n.d.). Retrieved from https://supreme.justia.com/cases/federal/us/526/137/.

25. https://www.law.cornell.edu/rules/fre/rule_702 (Last Accessed November 24, 2018)

26. General Electric Co. v. Joiner, 522 U.S. 136 (1997). (n.d.). Retrieved from https://supreme.justia.com/cases/federal/us/522/136/.

27. https://simplystatistics.org/2013/07/31/the-researcher-degrees-of-freedom-recipe-tradeoff-in-data-analysis/ (Last Accessed November 24, 2018).

28. https://legaldictionary.net/preponderance-of-evidence/ (Last Accessed November 24, 2018).

29. Significant Statistics: The Unwitting Policy Making of Mathematically Ignorant Judges, Michael I. Meyerson & William Meyerson, Pepperdine Law Review Vol. 37:771, 2010.

30. https://www.dol.gov/whd/flsa/ (Last Accessed November 24, 2018).

31. https://www.eeoc.gov/laws/statutes/titlevii.cfm (Last Accessed November 24, 2018).

32. https://www.census.gov/programs-surveys/acs/ (Last Accessed November 24, 2018).

33. http://mcdc.missouri.edu/websas/geocorr12.html (Last Accessed November 24, 2018).

34. https://www.bls.gov/soc/ (Last Accessed November 24, 2018).

35. Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hove: Lawrence Erlbaum Associates.

36. Center, F. J. Reference Manual on Scientific Evidence. National Academies Press, 2011.