2023

Trial Selection and Estimating Damages Equations

Keith N. Hylton
Boston University School of Law

Trial Selection and Estimating Damages Equations

Keith N. Hylton[*] and Sanghoon Kim[**]

October 2023

Forthcoming: *Review of Law and Economics*

Abstract: Many studies have employed regression analysis with data drawn from court opinions. For example, an analyst might use regression analysis to determine the factors that explain the size of damages awards or the factors that determine the probability that the plaintiff will prevail at trial or on appeal. However, the full potential of multiple regression analysis in legal research has not been realized, largely because of the sample selection problem. We propose a method for controlling for sample selection bias using data from court opinions.

Keywords: law and economics, empirical legal studies, damages equations estimation, sample selection bias

JEL Classifications: K10, K13, C51

[*] Boston University and Boston University School of Law.
[**] Department of Economics, University at Buffalo

1. Introduction

This paper addresses one of the most vexing problems of empirical legal research. Many studies have employed regression analysis with data drawn from court opinions (e.g., Kort, 1963); Eisenberg and Johnson, 1991; McChesney, 1999; Allison and Lemley, 1998). For example, an analyst might use regression analysis to determine the factors that explain the size of damages awards (Chang, Eisenberg, Ho and Wells, 2015), or the factors that determine the probability that the plaintiff will prevail at trial or on appeal (Studdert, Mello, Levy, Gruen, Dunn, Orav, and Brennan, 2007). This is an attractive approach to legal research because court opinions provide a great deal of information. Multiple regression analysis can be used to assess the factors that account for the key outcomes of litigation (e.g., verdict, dismissal, summary judgment). In addition, multiple regression analysis can be used to determine whether certain legal doctrines have actually been employed by courts to determine the outcomes of disputes within a specific area of litigation, such as medical malpractice or contract breach. Multiple regression holds the promise of providing a more rigorous method of assessing the relative importance of the factors that determine court outcomes compared to the traditional approach of lawyers, which is to read court opinions and to make judgmental assessments of the importance of the various factors examined by courts (cf., Fisher, 1958).

The full potential of multiple regression analysis in legal research has not been realized, largely because of one reason: the information contained in court opinions comes from a selected sample. The disputes that find their way into appellate court opinions are among the relatively small percentage that fail to settle at some point in the dispute process. Thus, if an analyst has a general regression model consisting of factors that he posits should explain the expected verdict for the plaintiff, or the expected damages award, the analyst's model generally should not be applied directly to a sample drawn from litigated cases unless some effort is made to correct for the bias due to sample selection. Of course, it is possible that the screening due to the selection process is entirely random and therefore imparts no bias to the regression analysis (Helland and Klerman, 2018), but that is unlikely to be true in general.

Heckman (1979) provided the most commonly used method of correcting for sample selection bias. To use Heckman's method with litigated cases, one must have data both on the litigated cases and on the settled cases – for example, a sample consisting of litigated automobile negligence disputes and settled automobile negligence disputes within a given jurisdiction. However, data on both litigated and settled cases are rarely available, except in a few special areas such as medical malpractice where insurance records provide the analyst with access to a substantial body of information on settled cases.[1] In most areas of legal research, the empirical legal analyst has access to court opinions based on litigated cases and no access to settled cases. Indeed, ordinarily the empirical legal analyst has access to information mostly from appellate court opinions, with only minimal information available from trial court decisions in the same set of disputes.

[1] Viscusi (1986) uses insurance records on products liability claims to estimate parameters influencing compensation levels, settlement, and the plaintiff win probability for trials.

We propose a method for controlling for sample selection bias in this paper that involves modifying the structural model to take selection due to settlement into account. Our approach seeks to enable the researcher to use regression analysis on a sample drawn exclusively from appellate court decisions.

Part 2 below provides a brief review of the literature using data from court opinions to estimate damages equations, or equations for the probability of a verdict for the plaintiff (or defendant). Part 3 discusses the limitations of court opinions as sources of data for regression analysis. Part 4 examines two simple models of the trial process involving appeals. Part 5 presents our model for estimating damages equations using data from appellate court opinions and controlling for sample selection. Part 6 shows the results of our estimation procedure. We use a data set that builds on the data used by Wriggins (2005) in her study of racial differences in wrongful death awards in Louisiana.

2. Literature Review

Although empirical legal studies is arguably still in its infancy, there are numerous papers that apply regression analysis to data drawn from court opinions.[2] Probably the first to do so is Kort's (1963) study of Supreme Court right-to-counsel decisions. Kort's regression analysis was an effort to improve upon an earlier contribution, Kort (1957), which developed an ad hoc estimation method that was criticized by Fisher (1958) for using more variables than observations and failing to have any theoretical basis for the empirical model. Fisher applauded the novelty of Kort's approach but worried that the new methodology had limited potential, and might retard empirical analysis in the legal field through the use of analytically unsound procedures. The second paper to use regression analysis is Segal (1984), who used a sample of U.S. Supreme Court decisions to examine the factors determining a finding that a search is reasonable. The third application is Eisenberg and Johnson (1991), who used a sample of appellate court sex discrimination cases to examine the factors that influence a court's finding of intentional discrimination.[3] Fourth in this series is McChesney (1993), examining the factors that influence a court's finding of limited liability in cases of defective incorporation.[4] Another early application is McChesney (1999), which examines the factors influencing a court's finding of tortious interference with contract. These early papers do not mention the sample selection bias problem.[5] However, gradually, the problem has received recognition in the papers that use

[2] For a survey of papers coding information from court opinions, many of which apply regression analysis, see Mark A. Hall & Ronald F. Wright (2008).

[3] Eisenberg and Johnson use a logistic regression model on data drawn from court opinions, which they described as a "largely untried technique," see Id. at 7.

[4] McChesney, at 519, mused that his article "may be the first to use multiple regression to discern the separate legal reasons for judicial decisions in a purely common-law domain." His article appears to have been the first to do so, but the distinction it draws with respect to earlier contributions is unimportant. The earlier papers use data from constitutional law or statutory law decisions. Whether constitutional law, statutory law, or common law, the judicial reasoning that determines the value of the dependent variable reflects and constitutes judge-made (i.e., "common") law.

[5] The Fisher (1958) critique is exceptional in this regard. Fisher, at 330, provides an illuminating discussion of the sample selection problem in empirical work using Supreme Court opinions.

regression analysis on data drawn from court opinions. At this stage of development, papers acknowledge the sample selection problem, and recognize its limiting effect on the ability to draw inferences from the regression results, but continue to apply the regression methodology anyway without attempting to correct for sample selection.[6]

The empirical application of this paper's model is to wrongful death damages. There is now a substantial literature using data from court opinions to estimate damages equations (see Eisenberg, Eisenberg, Wells, and Zhang, 2015; Eisenberg and Heise, 2011; Flatscher-Thoni, Leiter, and Winner, 2013; Chang, Eisenberg, Ho, and Wells, 2015). Among the papers estimating damages equations, the closest to this paper's application is Chang, Eisenberg, Ho, and Wells, who study pain and suffering damages in wrongful death cases, drawing their data from trial court decisions in Taiwan. Closer in style to this paper is Eisenberg, Eisenberg, Wells, and Zhang, who develop a regression model for zero-value dependent variable observations, and apply their model to data drawn from court opinions.[7]

3. Court Opinions as Data Sources

Appellate court opinions in the U.S. offer a rich source of data for empirical legal scholarship. These opinions offer a detailed description of the facts of a dispute,[8] the parties to the dispute, and the legal issues and considerations involved in the court's resolution of the dispute. For example, if an analyst were attempting to estimate a regression model that explains the probability of a verdict for the plaintiff in a medical malpractice lawsuit, the analyst would find an invaluable quantity of information on the dispute in the appellate court opinion. If the analyst posits that certain demographic factors, such as the plaintiff's age or education level, enhance the likelihood of a verdict for a plaintiff, the analyst would likely find sufficient information to test the hypothesis in the appellate medical malpractice opinions. In addition, if the analyst posits that certain legal doctrines, such as rules on causation, affect the likelihood of a verdict for the

[6] See, e.g., Kent Barnett, Christina L. Boyd, and Christopher J. Walker (2018), at 605 ("Although these data allow us to speak confidently about how circuit courts apply administrative law deference doctrines, selection effects in how cases reach appellate courts limit our ability draw broader inferences about the legal system."); Hylton (2008), at 236 ("Given the possibility of sample selection bias, the probit regression results below must be interpreted with care. The regression estimates are reliable tests of the theory set out earlier in this paper if the marginal impacts of the independent variables … on the settlement decision are negligible. On the other hand, one could interpret the results as measuring the effects of the independent variables within the sample of litigated cases. In this case the estimated coefficients reflect a combination of direct effects on the preemption probability and effects on settlement, which is more difficult to interpret.").

[7] Although not attempting to estimate damages, there are some more recent papers that estimate "win probability" regressions without attempting to solve the underlying sample selection bias due to settlement and filing decisions. See, e.g., Muñoz Soro and Serrano-Cinca (2021).

[8] Some appellate judges have been criticized for distorting the facts for the purpose of making a more persuasive opinion. Indeed, dissenting judges have sometimes accused the majority authors of distorting or leaving out important parts of the factual record in their opinions (see e.g., State Farm v. Campbell, 538 U.S. 408, 431 (2003) (Ginsburg, J, dissenting, "In this regard, I count it significant that… there is a good deal more to the story than the Court's abbreviated account tells.")

plaintiff, he would find sufficiently detailed descriptions of the relevant causation law to enable coding and hypothesis testing in the appellate opinions.

Given the detailed information available in appellate court opinions, it is reasonable to ask why trial courts do not issue opinions with comparably rich information. There are several reasons. Trial courts decide a much larger number of disputes than do appellate courts, and consequently trial judges have less time available for writing accounts of their decisions and the reasoning behind them. Trial judges often operate with juries, and therefore tend to play a less prominent role in the decision making process. In addition, the incentives to write opinions are weak because trial decisions do not bind other trial courts. Finally, a customary practice of not writing opinions probably discourages trial judges from deviating from precedent. For all of these reasons, and probably others, trial courts decisions have not offered information on disputes comparable to that generated by the appellate courts.

One significant limitation of appellate court opinions as data sources is their relative paucity in comparison to other data sources, such as national surveys (e.g., Census).[9] The large-sample empirical analyses expected as the norm in economics today are generally infeasible with appellate court opinions as data sources.

Another important limitation of appellate court opinions as data sources – and the focus of this study – is that the disputes that appear in the appellate court opinions are not a random sample drawn from the underlying base of disputes. Many cases settle before reaching the appellate court. If, for example, all of the cases involving plaintiffs who are likely to prevail are screened out of the sample as a result of settlement, then the resulting sample would consist mostly of plaintiffs with weak cases, making it difficult to tease out the true effects of demographic factors on the probability of a verdict for the plaintiff.

We assume settlement selection occurs at two stages: "pre-trial," where cases settle before a trial verdict is issued, and "pre-appeal," which is after trial and before the appellate court verdict. When legal researchers use appellate court information for regression analysis, they are restricted to the set of disputes that have gone both to trial and to the appeal stage – disputes that have been described as reaching the apex of a "claims pyramid" (Miller and Sarat, 1981), shown in Figure 1.

[9] For example, in the year 2016 the Seventh Circuit issued roughly 600 opinions. http://media.ca7.uscourts.gov/opinion.html. If one drills further into any particular area of law (e.g., antitrust) then the number of opinions that can be used for data is much smaller.

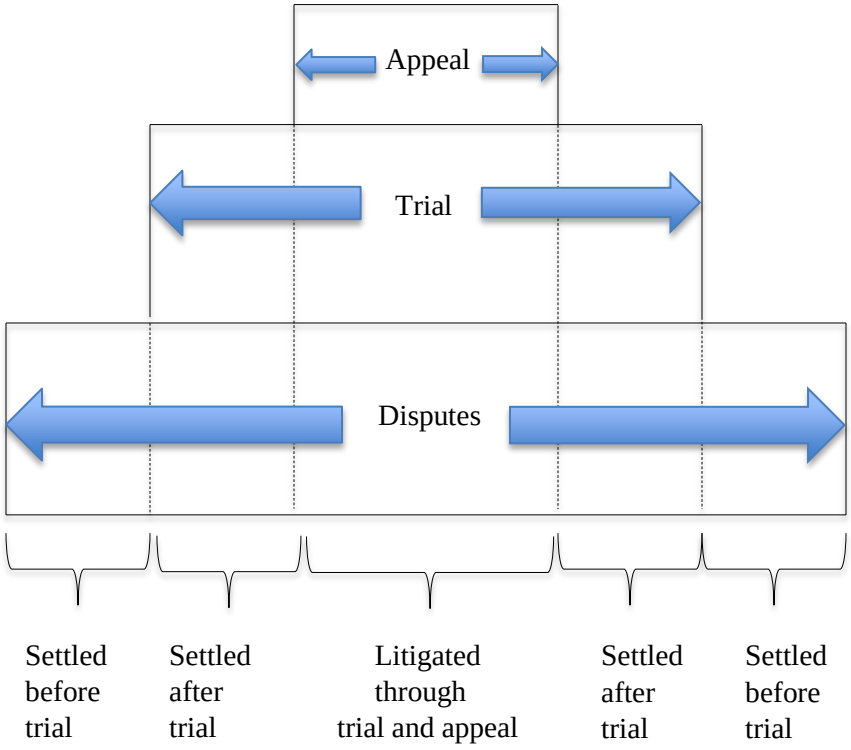

Figure 1: Dispute pyramid with trial and appeal

4. Models of the Trial-Appeal Settlement Process

Although settlement can occur at any time before the trial verdict, we simplify matters by considering only two periods for settlement. The first time period, Stage 1 or "pre-trial," refers to settlements that happen before the trial verdict. The second time period, Stage 2 or "pre-appellate decision," refers to settlements that occur before the appellate decision is issued.

Multi-stage litigation is examined in Bebchuk (1996), though he considers a single trial with several embedded phases of litigation, each phase representing a decision on some dispositive motion such as dismissal or summary judgment. In this model, by contrast, we examine two separate proceedings, trial and appeal,[10] and the win probabilities vary across the two stages.

[10] One might at first think that the Bebchuk model would be inappropriate for the examination of trial and appeal, because in the Bebchuk model the plaintiff remains the party seeking relief in each stage of the trial. In this model,

Litigation costs in stage *i* are given by $\{c_p^i, c_d^i\}$ and the expected win probabilities are $\{p_p^i, p_d^i\}$. Let *v* represent the award (injury loss) to the plaintiff. In each period, the plaintiff has an expected payoff from litigation $\pi_p^i = p_p^i v - c_p^i, i = 1, 2$, and the defendant has expected cost $\pi_d^i = p_d^i v + c_d^i$.

This structure assumes that the defendant does not have to pay the judgment at the end of the first stage (trial) if he loses. This assumption is consistent with practice. In all jurisdictions, the defendant can stay the trial judgment and file an appeal, and need only file an appeal bond.

We consider below two approaches to modeling litigation incentives. The first (Rationality Model) assumes that the parties take the anticipated outcomes in both stages of litigation into account in determining whether to litigate in the first stage. The other approach we consider is a "Myopia Model," which assumes that the parties consider only the current stage of litigation (trial or appeal) in choosing whether to settle or litigate. Both models assume that neither party possesses an informational advantage in predicting the trial outcome. Trial outcome predictions are determined by inconsistent beliefs or expectations, leading to litigation due to mutual optimism (Shavell, 1982; Hylton, 2023): $p_p^i - p_d^i > 0, i = 1, 2$.

4.1. Rationality Model

The Rationality Model employs backward induction to analyze settlement decisions in multi-stage litigation. At the second stage (appeal), the parties proceed in litigation if and only if $\pi_p^2 > \pi_d^2$, or (equivalently) $v > \frac{c_p^2 + c_d^2}{p_p^2 - p_d^2}$. Assuming litigation in the second stage, the plaintiff's expected reward in the first stage is $p_p^1 p_p^2 v + \left(1 - p_p^1\right) p_p^2 v - c_p^1 - c_p^2 = p_p^2 v - \left(c_p^1 + c_p^2\right)$. Note that this is independent of the plaintiff's first stage trial outcome prediction, p_p^1. The defendant's expected cost in the first stage is $p_d^2 v + (c_d^1 + c_d^2)$, also independent of his first-period trial outcome prediction. Therefore, the litigation condition in the first stage is $v > \frac{(c_p^1 + c_p^2) + (c_d^1 + c_d^2)}{p_p^2 - p_d^2}$.

Now consider the sequence where settlement would be rational in the second stage, that is, $v < \frac{c_p^2 + c_d^2}{p_p^2 - p_d^2}$. Following Bebchuk, we assume the settlement amount in the second stage takes the average of plaintiff and defendant's settlement bids: $S_2^* = \frac{1}{2}\left(p_p^2 + p_d^2\right)v + \frac{1}{2}(c_d^2 - c_p^2)$. Given this expectation, should the plaintiff choose to litigate in the first stage, he expects to receive $S_2^* - c_p^1$. The defendant, on the other hand, will pay no more than $S_2^* + c_d^1$ to settle. Therefore, the parties will settle in the first stage for $S_1^* = \frac{1}{2}(S^H + S^L) = S_2^* + \frac{1}{2}\left(c_d^1 - c_p^1\right) > S_2^* - c_p^1$. Thus, given the expectation of settlement in the second stage, the parties always settle in the first stage.

---

however, the defendant may be the party seeking relief in the second stage (appeal). But this difference is more apparent than real. With appeal available, the plaintiff must petition to receive his award in both periods.

Summarizing: (1) *Litigation occurs when* $v > \frac{(c_p^1+c_p^2)+(c_d^1+c_d^2)}{p_p^2-p_d^2}$. (2) *Disputes either litigate in both stages or settle in the first stage – there are no cases of litigation to trial verdict followed by settlement before appellate verdict*. (3) *Only the win probabilities at the final stage matter in determining the decision to litigation*.

4.2 Myopia Model

An alternative to the Rationality Model assumes that the parties are myopic, in the sense that their incentives to settle are based entirely on the payoffs relevant to the stage in which they find themselves.

In the Myopia Model, the plaintiff in Stage 1 examines only the payoff from Stage 1 litigation, ignoring the likelihood of appeal to the second stage. The following conclusions apply: (1) if $\frac{c_p^1+c_d^1}{p_p^1-p_d^1} > \frac{c_p^2+c_d^2}{p_p^2-p_d^2}$, then all settlements occur in the first stage and the probability of settlement is equal to the probability that $v < \frac{c_p^1+c_d^1}{p_p^1-p_d^1}$. (2) On the other hand, if $\frac{c_p^1+c_d^1}{p_p^1-p_d^1} < \frac{c_p^2+c_d^2}{p_p^2-p_d^2}$, then it is possible (contrary to the Rationality Model) to have disputes that litigate in the first stage and settle in the second stage. The probability of settlement in the first stage is equal to the probability that $v < \frac{c_p^1+c_d^1}{p_p^1-p_d^1}$, and the probability of settlement in the second stage is equal to the probability that $\frac{c_p^1+c_d^1}{p_p^1-p_d^1} \leq v < \frac{c_p^2+c_d^2}{p_p^2-p_d^2}$.

Since the Rationality Model, (1) and (2) of Myopia Model result in different regression specifications, the three models could be put in competition with one another through specification tests on the associated regression equations.

5. Econometrics of Legal Analysis

5.1 Description of Problem

In this part, we discuss a structural econometric model for estimation using appellate court data. We assume that the analyst has a basic theoretical regression model to explain some particular dependent variable. For example, the legal analyst might develop a linear regression model that explains the amount of damages awarded in a tort lawsuit. The independent variables in the model are drawn from information provided in the case, such as the age or education level of the plaintiff. However, a simple linear regression of the dependent variable on the independent variables drawn from court reports is likely to be biased because of selection.

Consider an equation for damages estimation. The theoretical model that the analyst has designed aims to explain the expected damages award v_i for each case i. Specifically, the theoretical structural model is $E(v_i|x_i) = \beta' x_i$, where x_i is the vector of independent variables of individual case i that the analyst posits should explain the damages awarded in a lawsuit, and β is

the vector of coefficients for the related independent variables. The corresponding regression model is $v_i = \beta' x_i + \varepsilon_i$, where $E(\varepsilon_i|x_i) = 0$ and $Var(\varepsilon_i|x_i) = \sigma^2$.

The results from the estimation procedure just described are likely to be biased by the selection of disputes for litigation. The theoretical structural model is based on the analyst's belief that all realizations v_i are determined in expectation by the structural form $\beta' x_i$. But the analyst never sees all realizations v_i that occur in the population. He sees only the realizations that have not been screened out as a result of the settlement process. Because of the screening due to settlement, the direct estimation of the theoretical structural model is likely to result in biased estimates. Given the likely bias resulting from sample selection when using the theoretical structural model, we derive an alternative structural model that incorporates the selection process below.

5.2 Information Structure

The information structure assumed in the litigation model influences the selection of disputes into the settlement (or litigation) process. In the previous part, we examined settlement incentives under the assumption that inconsistent beliefs could determine trial outcome predictions. An alternative to this approach, not explored here, would assume informational asymmetry generates litigation (Png, 1987, Bebchuk 1984, Nalebuff, 1987). Sieg (2000) estimates the parameters of an asymmetric information model of litigation using medical malpractice data.

In the parts below, we will assume inconsistent beliefs, in particular, *mutual optimism* as the basis for litigation. We assume $p_p^k,\ p_d^k$ are the subjective beliefs of the parties in stage *k* (trial=1, appeal=2), where $p_p^k > p_d^k$. This is the optimism model proposed in Shavell (1982) and formalized in Hylton (2023).

We adhere to (2) of the Myopia Model of multi-stage litigation described previously because the model allows settlements at any stage, which is more representative of real-world cases. Under the model, litigation occurs when

$$v \geq \bar{v} = \frac{c_p^2 + c_d^2}{p_p^2 - p_d^2}.$$

That is, the appellate court data is left-truncated, and the truncation threshold depends on the plaintiff's and defendant's litigation costs and their predictions of the litigation outcome at the appellate court.

5.3. Truncated Regression Model with Stochastic and Unobserved $\bar{v}$

As we discuss above, when the analyst estimates damages using appellate court data, damages awards observations are left-truncated at $\bar{v}_i$ for each case i. However, each threshold $\bar{v}_i$ is

unobserved in the data. To deal with this issue, we apply the truncated regression model with stochastic and unobserved thresholds as described in Maddala (1983, 257-91):

$$v_i = \beta_1{}' x_{1i} + u_{1i}$$

$$\bar{v}_i = \beta_2{}' x_{2i} + u_{2i}$$

where v_i and $\bar{v}_i$ each denotes the damages amount and truncation threshold for individual case i. x_{1i} and x_{2i} are the vectors of explanatory variables for v_i and $\bar{v}_i$, respectively. v_i, x_{1i}, and x_{2i} are observable in appellate court data if and only if $v_i > \bar{v}_i$. $\bar{v}_i$ is unobservable and stochastic. We assume that (u_{1i}, u_{2i}) are distributed bivariate normal with mean vector zero and covariance matrix $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$. The likelihood function for this model is

$$\mathcal{L}(\beta_1, \beta_2, \Sigma) = \prod_{i=1}^{N} \frac{1}{\Phi\left(\frac{\beta_1' x_{1i} - \beta_2' x_{2i}}{\sqrt{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}}\right)} \int_{-\infty}^{v_i - \beta_2' x_{2i}} f(v_i - \beta_1' x_{1i}, u_2) du_2$$

where N is the number of observations, $\Phi(\cdot)$ is the cumulative distribution function of the standard normal, and $f(\cdot,\cdot)$ is the joint density of u_{1i} and u_{2i}. After simplifying the above equation as in Maddala (1983), [11] we can write down the log-likelihood function as

$$\log \mathcal{L} = -N \log \sigma_1 - \frac{1}{2\sigma_1^2} \sum_{i=1}^{N} (v_i - \beta_1' x_{1i})^2 + \sum_{i=1}^{N} \log \Phi \left(\frac{v_i - \beta_2' x_{2i} - \frac{\sigma_{12}}{\sigma_1^2}(v_i - \beta_1' x_{1i})}{\sqrt{\sigma_2^2 - \frac{\sigma_{12}^2}{\sigma_1^2}}} \right)$$

$$- \sum_{i=1}^{N} \log \Phi \left(\frac{\beta_1' x_{1i} - \beta_2' x_{2i}}{\sqrt{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}} \right).$$

Then, we estimate the parameters of the model with the standard maximum likelihood estimation. As in Muthén and Jöreskog (1983), we assume that $\sigma_1^2 = \exp(\widetilde{\sigma_1})$ and $\sigma_2^2 - \frac{\sigma_{12}^2}{\sigma_1^2} = 1$, and we estimate $\widetilde{\sigma_1}$ and σ_{12} to ensure positive values for σ_1^2 and σ_2^2. [12]

As noted in Maddala (1983), we do not know whether the above equation is well-behaved so that it has a unique global maximum. Thus, starting from different initial values and using different optimization algorithms in the maximum likelihood estimation may help in empirical applications.

Another issue in the truncated regression model with stochastic and unobserved thresholds is the reliability of the estimated parameters β_2 for x_2. Monte-Carlo simulation results

[11] For the detailed derivation, see Appendix A.

[12] Note that σ_2^2 is restricted as $1 + \frac{\sigma_{12}^2}{\sigma_1^2}$.

in Muthén and Jöreskog (1983) show that the estimated parameters of β_2 could be unreliable while it can still correct for selectivity bias in β_1.[13] Thus, this model cannot be used to directly estimate the selection criterion parameters β_2. The model, however, still corrects for selectivity bias in β_1 even with the unreliable estimates of β_2. Thus, we can use the model in estimating theoretical structural models in empirical legal research to correct for sample selection bias from appellate court opinions.

5.4. Explanatory Variables for $\bar{v}$

Here we consider the explanatory variables for $\bar{v}$. As we discuss in Section 5.2, $\bar{v}$ depends on the total litigation costs for both parties and the difference in the parties' predictions of the litigation outcome.

The data for total litigation costs are generally unavailable, and certainly not for any historical series of appellate cases. However, the number of attorneys involved are generally available in most court opinions, even in reports from more than one hundred years ago. Additionally, total litigation costs vary depending on the difficulty of each case. We assume that more difficult cases will result in longer appellate opinions. Thus, we use the number of pages in each appellate decision as a proxy for the case's difficulty. We use those two types of data as the explanatory variables for the total cost of litigation in our empirical application in the following section.

Another problem in choosing the explanatory variables for $\bar{v}$ is the part for the probability prediction (or expectations) differential. It requires the identification of variables that would tend to generate plaintiff optimism (plaintiff believing he has a strong case) and simultaneous defendant optimism (defendant believing plaintiff has a weak case). We use three dummy variables: one for whether the defendant was subject to vicarious or direct liability, another for whether the decedent was claimed to be at fault or not, and the third for whether the plaintiff won or lost at trial. Vicarious liability means that the defendant is liable for the fault of another party (e.g., an employer being liable for the negligence of an employee), and direct liability means, as the label suggests, the actor is liable for his own fault. An employer, for example, can be held directly liable if it negligently hired an incompetent employee, whose foreseeable incompetence led to the death of the decedent. Allegations of direct fault, whether against the defendant or against the decedent, might tend to generate or exacerbate mutual optimism, through hindsight bias or correspondence bias (fundamental attribution error) – and their associated influences on tastes for risk bearing. According to the correspondence bias theory, individuals tend to evaluate responsibility for harmful actions of others by referring to or invoking personal traits rather than situational factors, while doing the opposite in evaluating their own harmful actions (Flick and Schweitzer, 2021). Such biases can generate divergent expectations on the probability of a finding of a fault.[14] After the trial court decision, the plaintiff

[13] Maddala (1983)

[14] For experimental evidence on fundamental attribution error in the perception of fault, see Flick and Schweitzer (2021) on perceptions of negligence, and Kassin and Sukel (1997) on perceptions of specific intent.

and defendant adjust their probability predictions; the probability prediction differential is larger or smaller if the plaintiff won or lost at trial, respectively.

6. Application

In this part we present an application of the method developed in the previous parts of this paper. We consider the empirical application as largely a "proof of concept."

6.1 Data

Our data consists of wrongful death and survival action appellate decisions in the state of Louisiana from the year 1901 to the year 1950. We attempted to get every informative (i.e., discussing damages and litigant characteristics) appellate decision on record.[15] Wrongful death actions are lawsuits brought by the survivors of a decedent for the loss in financial support resulting from the death of the decedent due to the defendant's tortious conduct. Many states, such as Louisiana during the period of our data set, permit recovery for the emotional suffering of survivors as well. Survival claims, by contrast, are for the losses the decedent could have brought for injuries personally suffered from the moment of injury until his death, which include lost wages and emotional suffering. Damages awards in Louisiana courts did not, as a general matter, separate these separate grounds for damages in the final award.[16]

[15] We began with a data set compiled by Jennifer Wriggins for her study of racial differences in wrongful death awards in Louisiana (Wriggins, 2005). We added cases from an additional year (1950), and combed through the cases for data to create additional variables.

[16] For an excellent discussion of wrongful death and survival lawsuits, see Wriggins (2005), at 113-114.

Variables	(1) Mean	(2) Standard Deviation	(3) Min	(4) Max
Appellate damages award	6,872	4,502	0	25,900
Race, white=1	0.846	0.363	0	1
Gender, male=1	0.765	0.426	0	1
Widow with children/child	0.272	0.447	0	1
Monthly wage	65.37	128.5	0	875
Age at death	31.35	21.09	0.167	80
Occupation, railroad	0.0662	0.250	0	1
Occupation, driver	0.0441	0.206	0	1
Occupation, farmer	0.0441	0.206	0	1
Occupation, business	0.140	0.348	0	1
Occupation, labor	0.213	0.411	0	1
Occupation, other[17]	0.493	0.502	0	1
Vicarious/direct liability	0.588	0.494	0	1
Decedent was claimed to be at fault	0.632	0.484	0	1
Plaintiff won at trial	0.743	0.439	0	1
Number of plaintiffs	2.132	1.403	1	8
Region, Acadiana	0.221	0.416	0	1
Region, central LA	0.0809	0.274	0	1
Region, north LA	0.235	0.426	0	1
Region, Florida Parishes	0.103	0.305	0	1
Region, Great New Orleans	0.360	0.482	0	1
Number of attorneys involved	3.449	1.321	2	8
Number of pages in the appellate decision	6.059	2.544	2	14
Number of observations		136		

[17] It indicates cases that the decedent was a child, housewife, or retiree.

6.2 Estimation

We start with the simplest regression model in Table 1. The first column shows the most basic OLS result. The dependent variable is the amount awarded in damages by the appellate court. The second column shows the result of the model in Section 5.3. The estimated coefficients for the truncation threshold equation are not reported.

The most striking difference between the two columns is the enhanced race effect in the second column. In both columns, the race effect is highly statistically significant, with the result in the first column showing that the survivors of white decedents, after controlling for the income of the decedent, status of the plaintiff-survivor (a widow with children or not), and occupation of the decedent, received roughly $4,006 more in compensation than did the plaintiff-survivors of black decedents. The second column finds that this race premium was about 53.6 percent higher, at $6,152. In the second column, a one-dollar increase in the monthly wage leads to a $10.34 increase in the award to plaintiff-survivors. Wriggins' (2005) finding, based on an examination of average awards, that Louisiana courts used race as basis for discounting awards to survivors of black decedents receives considerably stronger support in the new regression.[18]

The variables coding for cases where the decedent worked for a railroad and as a driver are significant and positive. The baseline of the variables is the cases where the decedent was a child, housewife, or retiree. The awards were on the order of $4,126 and $3,484 higher for survivors of decedents who worked for a railroad and as a driver, respectively. This is interesting given that we have already controlled for the wage at the time of death. The occupation control must therefore convey something other than the effect of the decedent's compensation on the court's award. The most plausible explanation is that railroad jobs were more stable and secure than other occupations. In light of this, a court was more likely to perceive the railroad-employed decedent as a greater source of support for family members than decedents of other occupations. Many of the decedents listed as laborers, for example, worked seasonally and moved from job to job, leading to substantial fluctuations in income over time.

Table 2 repeats the same comparison between the simple OLS and the model in Section 4.3 but includes regional controls. The regional controls separate the regions of the state of Louisiana where the appellate court that rendered the judgment sits. These regions also generally hold the trial courts from which the case was appealed. We have 29 separate parishes in the state of Louisiana over this period. We grouped these parishes into 5 separate regions: Acadiana (Cajun Country approximately), Florida Parishes, North Louisiana (Sportsmen's Paradise), Greater New Orleans, and Central Louisiana (Crossroads).[19] These regions are understood to be culturally different,[20] and we posited that these cultural differences might influence the way

[18] Wriggins (2005), at 117-118, reports that the average award black family members, $3,559, was less than half of the average award to white family members, $8,245. The difference, $4,686, is close to the differential we find in the OLS regression.

[19] For a map of these regions, see https://en.wikipedia.org/wiki/Central_Louisiana.

[20] http://www.louisianafolklife.org/LT/Articles_Essays/la_3_folk_reg.html; http://microsite.smithsonianmag.com/ads/louisiana/plan-your-trip/regions.html.

courts view these cases. The results indicate that North Louisiana courts typically give greater awards. The reasons for these North Louisiana premium are not obvious; perhaps the lower prevalence of slavery in its history may have generated different perceptions of the value of a work-life. The results of the other variables in Table 2 largely replicate those in Table 1.

Table 3 provides a fuller regression equation, including age, age-squared, and the number of plaintiff-survivors. Regional controls were included in this regression, though their coefficient estimates are excluded from the table to reduce clutter. The results of the regional controls were consistent with those of Table 2. The award goes up about $576.5 for every additional plaintiff-survivor. Age and age-squared show that the award to survivors generally goes up with work experience (proxied by age) but at a declining rate. For every year of age on the decedent, the award rises by roughly $176 dollars, but the estimated coefficient of the age-squared variable is about -2.07. This implies that the maximum contribution on average was observed at age 42.5. This is consistent with the rapidity of declining health for workers during this time period.

Table 1

Dependent variable = appellate damages award VARIABLES	Column 1 OLS	Column 2 MLE
Race, white=1	4,006***	6,152***
	(739.1)	(1,185)
Gender, male=1	139.8	384.4
	(744.9)	(971.3)
Widow with children/child	4,093***	4,545***
	(731.5)	(843.8)
Monthly Wage	10.73***	10.34***
	(2.583)	(2.821)
Occupation, railroad	3,580***	4,126***
	(1,203)	(1,379)
Occupation, driver	2,609*	3,484**
	(1,419)	(1,652)
Occupation, farmer	555.6	1,125
	(1,362)	(1,684)
Occupation, business	84.08	594.9
	(1,115)	(1,303)
Occupation, labor	581.8	931.6
	(897.9)	(1,090)
Constant	1,051	-1,925
	(804.9)	(1,386)
Observations	136	136

Standard errors in parentheses. The MLE results show only the parameters of the explanatory variables for appellate damages award.
*** p<0.01, ** p<0.05, * p<0.1

Table 2

Dependent variable = appellate damages award VARIABLES	Column 1 OLS	Column 2 MLE
Race, white=1	4,073***	6,151***
	(748.5)	(1,150)
Gender, male=1	379.4	586.5
	(749.8)	(941.1)
Widow with children/child	3,850***	4,293***
	(754.9)	(845.0)
Monthly Wage	11.23***	11.08***
	(2.641)	(2.833)
Occupation, railroad	3,500***	4,170***
	(1,224)	(1,386)
Occupation, driver	2,625*	3,551**
	(1,427)	(1,629)
Occupation, farmer	506.0	1,077
	(1,405)	(1,700)
Occupation, business	-77.96	345.9
	(1,130)	(1,278)
Occupation, labor	607.5	960.0
	(898.8)	(1,055)
Region, Acadiana	1,030	1,692
	(1,139)	(1,361)
Region, North LA	2,294**	3,011**
	(1,094)	(1,311)
Region, Florida Parishes	1,178	1,585
	(1,245)	(1,532)
Region, Great New Orleans	1,358	2,161
	(1,083)	(1,315)
Constant	-510.2	-4,028**
	(1,271)	(1,852)
Observations	136	136

Standard errors in parentheses. The MLE results show only the parameters of the explanatory variables for appellate damages awards.
*** p<0.01, ** p<0.05, * p<0.1

Table 3

Dependent variable = appellate damages award VARIABLES	Column 1 OLS	Column 2 MLE
Race, white=1	4,004***	5,776***
	(737.2)	(1,066)
Gender, male=1	1,270	1,899*
	(807.8)	(1,023)
Widow with children/child	2,674***	2,821***
	(823.9)	(896.4)
Monthly Wage	9.963***	9.532***
	(2.602)	(2.747)
Age	139.2**	176.0**
	(64.73)	(79.89)
Age-squared	-1.690**	-2.068**
	(0.807)	(0.986)
Number of plaintiffs	510.5**	576.5**
	(215.6)	(233.5)
Occupation, railroad	2,696**	2,994**
	(1,327)	(1,486)
Occupation, driver	1,553	2,059
	(1,483)	(1,654)
Occupation, farmer	-158.0	153.2
	(1,455)	(1,694)
Occupation, business	-728.1	-655.8
	(1,197)	(1,344)
Occupation, labor	-399.7	-383.9
	(1,012)	(1,170)
Constant	-3,151**	-7,070***
	(1,500)	(2,090)
Observations	136	136

Standard errors in parentheses. The MLE results show only the parameters of the explanatory variables for appellate damages awards.
*** p<0.01, ** p<0.05, * p<0.1
Note: Regional controls included in the regression, but not shown in the table.

7. Conclusion

In this paper, we present a method for controlling for sample selection bias due to settlement when empirical legal researchers conduct regression analysis using data from appellate court decisions. Based on a model of the trial-appeal settlement process, we correct sample selection bias by utilizing a truncated regression model with unobserved and stochastic settlement thresholds. In an empirical application using data from wrongful death appellate decisions in the state of Louisiana, we demonstrate the differences in estimate results between the standard OLS, which fails to correct for sample selection bias, and our method, which corrects the bias. Our approach aims to broaden empirical legal research based on data from appellate court decisions. Future research employing the approach in this paper might address limited dependent variable regressions using appellate court data.

References

John R. Allison; Mark A. Lemley, Empirical Evidence on the Validity of Litigated Patents, 26 AIPLA Quarterly Journal (1998), 185-276.

Kent Barnett, Christina L. Boyd, and Christopher J. Walker, The Politics of Selecting Chevron Deference, 15 Journal of Empirical Legal Studies (2018), 597–619.

Lucian A. Bebchuk, Litigation and Settlement Under Imperfect Information, 15 RAND Journal of Economics (1984), 404-15.

Lucian A. Bebchuk, A New Theory Concerning the Credibility and Success of Threats to Sue, 25 J. Legal Stud. (1996), 1-25.

Yun-chien Chang, Theodore Eisenberg, Han-Wei Ho, and Martin T. Wells, Pain and Suffering Damages in Wrongful Death Cases: An Empirical Study, 12 Journal of Empirical Legal Studies (2015), 128–160.

Theodore Eisenberg, Thomas Eisenberg, Martin T. Wells, and Min Zhang, Addressing the Zeros Problem: Regression Models for Outcomes with a Large Proportion of Zeros, with an Application to Trial Outcomes, 12 Journal of Empirical Legal Studies (2015), 161–186.

Theodore Eisenberg and Michael Heise, Judge-Jury Difference in Punitive Damages Awards: Who Listens to the Supreme Court? 8 Journal of Empirical Legal Studies (2011), 325–357.

Theodore Eisenberg and Sheri L. Johnson, The Effects of Intent: Do We Know How Legal Standards Work?, 76 Cornell Law Review (1992), 1151-1197.

Franklin M. Fisher, The Mathematical Analysis of Supreme Court Decisions and Abuse of Quantitative Methods, 53 American Political Science Review (1958), 321-338.

Cassandra Flick and Kimberly Schweitzer, Influence of the Fundamental Attribution Error on Perceptions of Blame and Negligence, 68 Experimental Psychology (2021), 175-188.

Magdalena Flatscher-Thöni, Andrea M. Leiter, and Hannes Winner, Pricing Damages for Pain and Suffering in Court: The Impact of the Valuation Method, 10 Journal of Empirical Legal Studies (2013), 104–119.

Mark A. Hall and Ronald F. Wright, Systematic Content Analysis of Judicial. Opinions, 96 California Law Review (2008), 63-122.

James J. Heckman, Sample Bias as a Specification Error, 47 Econometrica (1979), 153-162.

Eric Helland, Daniel Klerman, and Yoon-Ho Alex Lee, Maybe there Is No Bias in the Selection of Disputes for Litigation, 174 Journal of Institutional and Theoretical Economics (2018), 143-170.

Keith N. Hylton, Preemption and Products Liability: A Positive Theory, 16 Supreme Court Economic Review (2008), 205-249.

Keith N. Hylton, Mutual Optimism and Risk Preferences in Litigation, 76 International Review of Law and Economics 106157 (2023).

Saul M. Kassin and Holly Sukel, Coerced Confessions and the Jury: An Experimental Test of the "Harmless Error" Rule, 21 Law and Human Behavior (1997), 27–46.

Fred Kort, Predicting Supreme Court Cases Mathematically: Analysis of the Right to Counsel Cases, 57 American Political Science Review (1957), 1-12.

Fred Kort, Content Analysis of Judicial Opinions and Rules of Law, in JUDICIAL DECISION-MAKING, G. Schubert (Ed.), Glencoe, Ill.: Free Press (1963), 133-196.

G. S. Maddala, Limited-Dependent and Qualitative Variables in Econometrics, New York: Cambridge University Press (1983).

Fred S. McChesney, Doctrinal Analysis and Statistical Modeling in Law: The Case of Defect. Incorporation, 71 Wash. U.L.Q. (1993), 493-534.

Fred S. McChesney, Tortious Interference with Contract Versus "Efficient Breach": Theory and Empirical Evidence, 28 Journal of Legal Studies (1999), 131-86.

Miller, Richard, and Austin Sarat, Grievances, Claims, and Disputes: Assessing the Adversary Culture, 15 Law & Society Review (1981), 525–65.

Jose Felix Muñoz Soro and Carlos Serrano-Cinca, A model for predicting court decisions on child custody, 16 PLoS ONE (2021): e0258993. https://doi.org/10.1371/journal.pone.0258993.

Barry Nalebuff, Credible Pretrial Negotiation. 18 RAND Journal of Economics (1987), 198–210.

Png, I. P. L., Litigation, Liability, and Incentives for Care, 34 Journal of Public Economics (1987), 61-85.

Jeffrey A. Segal, Predicting Supreme Court Cases Probabilistically: The Search and Seizure Cases, 1962-1981, 78 The American Political Science Review (1984), 891-900.

Steven Shavell, Suit, Settlement, and Trial: A Theoretical Analysis Under Alternative Methods for the Allocation of Legal Costs, 11 Journal of Legal Studies (1982), 55-81.

Holger Sieg, Estimating a Bargaining Model with Asymmetric Information: Evidence from Medical Malpractice Disputes, 108 Journal of Political Economy (2000), 1006-1021.

David M. Studdert, Michelle M. Mello, Marin K. Levy, Russell L. Gruen, Edward J. Dunn, E. John Orav, and Troyen A. Brennan, Geographic Variation in Informed Consent Law: Two Standards for Disclosure of Treatment Risks, 4 Journal of Empirical Legal Studies (2007), 103–124.

W. Kip Viscusi, The Determinants of the Disposition of Product Liability Claims and Compensation for Bodily Injury, 15 Journal of Legal Studies (1986), 321-46.

Jennifer B. Wriggins, Torts, Race, and the Value of Injury, 1900–1949, 49 Howard Law Journal (2005), 99–138.

Appendix A

In this appendix, we repeat the detailed derivations of equations in Section 5.3, which are well-documented in Maddala (1983). For each observation, we know that $v_i = \beta_1{}' x_{1i} + u_{1i}$ and that $\bar{v}_i < v_i$. That is, $u_{1i} = v_i - \beta_1{}' x_{1i}$ and $u_{2i} < v_i - \beta_2{}' x_{2i}$. Note that $(u_2 - u_1) \sim N(0, \sigma_1^2 + \sigma_2^2 - 2\sigma_{12})$, and thus

$$P(v_{2i} < v_{1i}) = P(u_{2i} - u_{1i} < \beta_1{}' x_{1i} - \beta_2{}' x_{2i}) = \Phi\left(\frac{\beta_1' x_{1i} - \beta_2' x_{2i}}{\sqrt{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}}\right)$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal. Also note that we only have observations such that $v_{2i} < v_{1i}$. Therefore, the likelihood function of the model is

$$\mathcal{L}(\beta_1, \beta_2, \sigma_1, \sigma_2, \sigma_{12}) = \prod_{i=1}^{N} \frac{1}{\Phi\left(\frac{\beta_1' x_{1i} - \beta_2' x_{2i}}{\sqrt{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}}\right)} \int_{-\infty}^{v_i - \beta_2' x_{2i}} f(v_i - \beta_1' x_{1i}, u_2) du_2$$

where $f(\cdot,\cdot)$ is the joint density of u_{1i} and u_{2i}. Note that $f(u_1, u_2) = f(u_1) \cdot f(u_2|u_1)$. Also note that $u_1 \sim N(0, \sigma_1^2)$ and that $u_2|u_1 \sim N\left(\frac{\sigma_{12}}{\sigma_1^2} u_1, \sigma_2^2 - \frac{\sigma_{12}^2}{\sigma_1^2}\right)$. Thus,

$$\int_{-\infty}^{v_i - \beta_2' x_{2i}} f(v_i - \beta_1' x_{1i}, u_2) du_2 = f(v_i - \beta_1' x_{1i}) \int_{-\infty}^{v_i - \beta_2' x_{2i}} f(u_2|u_1 = v_i - \beta_1' x_{1i}) du_2$$

$$= \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_1^2}(v_i - \beta_1' x_{1i})^2\right) \Phi\left(\frac{v_i - \beta_2' x_{2i} - \frac{\sigma_{12}}{\sigma_1^2}(v_i - \beta_1' x_{1i})}{\sqrt{\sigma_2^2 - \frac{\sigma_{12}^2}{\sigma_1^2}}}\right).$$

Therefore, we can write the log-likelihood function as

$$\log \mathcal{L} = -N \log \sigma_1 - \frac{1}{2\sigma_1^2} \sum_{i=1}^{N} (v_i - \beta_1' x_{1i})^2 + \sum_{i=1}^{N} \log \Phi\left(\frac{v_i - \beta_2' x_{2i} - \frac{\sigma_{12}}{\sigma_1^2}(v_i - \beta_1' x_{1i})}{\sqrt{\sigma_2^2 - \frac{\sigma_{12}^2}{\sigma_1^2}}}\right)$$

$$- \sum_{i=1}^{N} \log \Phi\left(\frac{\beta_1' x_{1i} - \beta_2' x_{2i}}{\sqrt{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}}\right).$$