The benefits of big data for selecting venues

Dr. Jerome Krief

Senior Data Scientist , Premonition LLC

September 25, 2016

Introduction

- Frequently, the plaintiff has several possible venues to choose from (residence, business location, or the location where matter arose)

- According to standard economic theory, plaintiff accepts a settlement offer if the amount exceeds the expected monetary damage net of litigation costs.

- Maximizing expected damage permits to optimize settlement outcome for plaintiffs.

- Percentage increase in expected damage substantial even if slight differences in odds of winning because likelihood of sucess in courts typically not high.

- A lawyer can now take advantage of big data to detect statistically the court maximizing the expected damage for a particular case.

Separating legalistic factors from judges preferences

- In law, it is believed that judges reach decisions based on legal materials.
- Empirically, given a civil case type, substantial differences in win rates for plaintiffs across circuit courts (A. Broscheid, Law and Society Review, 2011)
- According to political scientists, differences in outcomes across circuit courts are simply explained by differences in judges' preferences.
- Even though the political sciences view is appealing, it is based on the wrong premise that cases are randomly allocated, in the statistical sense, across courts.

Limitations of naive methods with legal data

- Data: select the court offering the greatest historical win rate for a particular field of the law

- Misleading: omitted factors driving decisions of judges and sample selection are not controlled for

Limitations of naive method with legal data

- Omitted bias: socio-economic attributes of plaintiffs do vary across counties and can impact on a plaintiffs ability to win a trial.
- Hence, poor win rates can be observed without any variation in judges' preferences.
- Example in auto negligence: Court A has win rate 20 percentage points higher than court B. Are judges in court A more favorable to plaintiffs if substantial fraction of cases in court B involves plaintiffs lacking health insurance?

Limitations of naive methods with legal data

- Sample selection: overwhelming majority of cases either settled or dismissed after discovery. However, win-loss outcomes are only available for cases which go to trial. Selection is a problem if only 'strong' cases make it to courts.

- Example: Court C displays a higher win rate and plaintiffs tend to be less motivated in court C. Furthermore, suppose 'motivated' plaintiffs win more often and are more likely to refuse settlements...

- Higher win rate in court C is an artifact of the selection process which leaves out unrepresentative motivated plaintiffs.

The benefits of Big Data

• Large number of attributes extracted from court documents: solves omitted variable problem and selection.

• Statistical model: predicts from historical data the probability of wining for each court keeping possible confounding factors constant

• As a result, a lawyer can determine whether a court is more favorable to plaintiffs per se. That is, given the case at stake, a lawyer can select the court maximizing the chance of winning

• This principle of data analysis can also be employed to select courts offering the best 'outcome' (i.e. monetary damage, duration of litigation).

Illustration: estimating the court maximizing the probability of winning an auto negligence case

• Possible venues: two circuits courts located in south FL, Collier's circuit court and another court which we shall name court D.

• Data: 1013 cases of auto negligence. Let Win=1 if plaintiff wins and Win=0 otherwise. Also, Let Collier=1 if trial takes place in Collier's circuit court and Collier=0 otherwise.

• Model: $Proba[Win = 1|Collier, else] = a + b * collier + else$ where else contains control systematic factors (such as plaintiff's-defendants' race, win rates of attorneys, plaintiff's-defendants's income bracket, plaintiff's-defendant,'s age, duration of litigation etc...) and random legal factors unique to the case.

• b is the causal effect of Collier's circuit court on the probability of winning.

Results

- Output with STATA: Win=0.06+0.12*Collier+ else
- Keeping other factors constant, the probability of wining an auto negligence case in Collier's circuit court is estimated to be 0.12 higher than in court D.
- t-stat=3.23: Collier variable is statistically significant. Hence, estimate is not an artifact of data but reflects the fact that judges in Collier are more favorable to plaintiffs for auto negligence cases.
- The 95 percent confidence interval has an upper bound of 0.18: choosing Collier's circuit court may increase the probability of wining by as much as 0.18.

Economic implication

- Assuming for simplicity damage is fixed at the same level in both courts if plaintiff wins: choosing Colllier's circuit court increases expected damage by about $(0.12/P_d) * 100$ percent where P_d is the probability of winning that case in circuit court D.
- For instance, if circuit court D offers a 10 percent chance given the attributes of a case then selecting Collier's circuit court increases the expected damage by 120 percent.
- More bargaining power during settlement if Collier's circuit court selected.

Conclusion

- Big Data as a tool to reveal the most attrative venue for optimizing expected damages and thus settlements.
- Statistical method can be extended to forecast key outcomes of interest in a given court.
- Equipped with informative forecasts about the courts, lawyers can better enhance their strategic decisions