Extreme Incentives

Andrew Clausen, University of Edinburgh

22 July 2019

Introduction

- Classic moral hazard models predict harsh punishments, even for trivial offences such as not paying parking fees.
- These are bad predictions:
  - Positive: Actual punishments are proportional to the crime.
  - Normative: The predictions feel unjust. Perhaps there is a good reason why.
- Specifically:
  - Becker (1968) assumes traffic wardens' signals are perfect (but costly to acquire). The conclusion is unsurprising because there is no trade-off between insurance and incentives.
  - Mirrlees (1975) assumes traffic wardens' signals are noisy, but arbitrarily accurate signals occasionally surface. The conclusion is very surprising, because innocent people suffer harsh punishments.

Questions

- Does assuming that all evidence is flimsy lead to more moderate predictions?
- Is there an underlying methodical problem behind the bad predictions?

Contributions

1. I prove moral hazard models predict harsh punishments, even with flimsy evidence.
2. There is a methodical problem in all three versions of the model. I prove model predictions are a discontinuous function of the signal distribution.
3. Companion paper: I quantitatively evaluate other proposals using Edinburgh parking enforcement data.

Classic ingredients

- Grossman and Hart's (1983) formulation is most convenient.
- A risk-averse agent (e.g. drivers):
  - has a hidden action $a \in \{0, 1\}$, e.g. $a = 1$ for paying parking fees,
  - receives transfers t,
  - has ex-post utility $u(t) - ac$, where $u(t) \to -\infty$ as $t \to 0$.
- A risk-neutral principal (e.g. Edinburgh Council):
  - would like the agent to choose $a = 1$,
  - observes a noisy signal $\ell \sim f(\ell|a)$ with finite support,
  - promises to pay the agent $t(\ell)$.

Classic problem

- The principal's problem is:

$$W(f,c) = \max_{t(\cdot)} \lambda \sum_{\ell} f(\ell|1)u(t(\ell)) - \sum_{\ell} f(\ell|1)t(\ell)$$

$$\text{s.t. (IC)} \sum_{\ell} f(\ell|1)u(t(\ell)) - c \geq \sum_{\ell} f(\ell|0)u(t(\ell)).$$

- The model is isomorphic to having a voluntary participation constraint:
  - There is a Pareto frontier of all regimes that satisfy (IC).
  - Adjusting the welfare weight (λ) or the outside option traces out the same Pareto frontier.

Classic literature: Becker (1968)

- Idea: if traffic wardens are costly, then why not hire fewer wardens, and compensate with harsher punishments?
- Becker assumed that wardens never make false accusations:
  - The signal ℓ is either
    - 0 (acquit a guilty or innocent person), or
    - ∞ (convict a guilty person),
  - $f(\infty|1) = 0$, and
  - $f(\infty|0) > 0$.
- A punishment of $u(t(\infty)) = -\infty$ is never actually executed, so it deters crime without any social cost.
- There is a discontinuity between no wardens (i.e. $f(\infty|0) = 0$) and few wardens (i.e. $f(\infty|0) > 0$).

Classic literature: Becker (1968) continued

- But what if wardens sometimes make false accusations by mistake?
- Mirrlees (1975) considered this possibility.
- First, a detour about likelihood ratios.

Likelihood ratio reformulation, Kim (1995)

- The name ℓ of a signal realisation (good, black, etc.) does not matter:
  - Without loss of generality, assume ℓ is named after its likelihood ratio, i.e. $\ell = \frac{f(\ell|0)}{f(\ell|1)}$.
  - Let $f(\ell) = f(\ell|1)$. Since $f(\ell|0) = \ell f(\ell|1)$, we only need to know $f(\ell|1)$.
- Some useful properties of likelihood ratio distributions $f(\ell)$:
  - $mean(f) = 1$. Proof: $mean(f) = \sum_\ell \ell f(\ell) = \sum_\ell f(\ell|0) = 1$.
  - The the null signal $\emptyset$ has likelihood ratio distribution $\emptyset(\ell) = I(\ell = 1)$.
  - g can be obtained by discarding information from f if and only if f is a mean-preserving spread of g.
- The principal's problem in terms of likelihood ratio distributions is:

$$W(f,c) = \max_{t(\cdot)} \lambda \sum_\ell f(\ell)u(t(\ell)) - \sum_\ell f(\ell)t(\ell)$$

$$\text{s.t. (IC)} \sum_\ell f(\ell)(1-\ell)u(t(\ell)) \geq c.$$

Classic solution

- The first-order condition with respect to $t(\ell)$ is:

$$\frac{1}{u'(t(\ell))} = \lambda + \mu(1 - \ell),$$

  where μ is the Lagrange multiplier on the (IC) constraint.
- When $\mu = 0$, this is the Borch (1962) equation of optimal insurance.
- The μ term looks more like statistics than Bayesian decision-making:
  - there is a likelihood ratio, ℓ,
  - this is not a posterior calculation,
  - if there were a posterior, it would be that the principal knows for sure that the agent plays his best-response, $a = 1$.

Classic literature: Mirrlees (1975) Unpleasant Theorem revisited

Claim: Consider any sequence of signals f_n.

- If $max(support(f_n)) = n$ then welfare $W(f_n, c)$ converges to the first best.
- If in addition $f_n(1) \to 1$, then $u(t_n(n)) \to -\infty$.

Proof:

1. $\mu_n \to 0$:
   - For $\ell = n$, the right side must be positive: $\frac{1}{u'(t_n(\ell))} = \lambda + \mu_n(1-\ell)$.
   - Rearrange: $\mu_n < \frac{\lambda}{n-1}$ for all n.
2. $W(\emptyset, 0) - W(f_n, c) \leq \mu_n c$ for all n:
   - $W(f_n, c) - W(f_n, 0) = \int_0^c W_c(f_n, \hat{c})\, d\hat{c}$.
   - By the envelope theorem, $W_c(f_n, c) = -\mu_n$.
   - W is concave in c, so $W_c(f_n, \hat{c}) \geq -\mu_n$ for all $\hat{c} \in [0, c]$.
3. Therefore, $W(f_n, c) \to W(\emptyset, 0)$, i.e. welfare converges to the first best.
4. Since $f_n(1) \to 1$, it follows that $u(t_n(n)) \to -\infty$.

Classic literature: Mirrlees (1975) Discussion

- Unlike Becker (1968), harsh punishments sometimes fall on innocent agents. So this prediction is even worse!
- My version of Mirrlees' theorem highlights the following discontinuity: even if f_n converges to an uninformative signal, welfare can converge to the first best.

Flimsy Evidence

- Becker and Mirrlees both assumed that overwhelming evidence is available.
- What if only moderate evidence is available, and the best evidence is only gathered rarely?
- Consider the signal $f(\ell) = (1-\varepsilon)g(\ell) + \varepsilon h(\ell)$, which is a mixture of
  - a signal g ("traffic wardens"), observed with probability $1-\varepsilon$, and
  - a stronger signal h ("traffic wardens plus a detective"), observed with probability ε.

Flimsy Evidence: Main result

- Assumptions:
  - Let $t(\ell, \varepsilon)$ and $\mu(\varepsilon)$ be the optimal transfers and Lagrange multiplier for ε.
  - Assume that for $\varepsilon = 0$ ("traffic wardens only"), the right side of the FOC

$$\frac{1}{u'(t(\ell, 0))} = \lambda + \mu(0)(1 - \ell)$$

    is negative for $\bar{\ell} = max(support(h))$, i.e. assume $\bar{\ell} > 1 + \frac{\lambda}{\mu(0)}$.
- *Theorem 1:* Compare $\varepsilon \to 0+$ versus $\varepsilon = 0$.
  - Welfare improves discontinuously: $lim_{\varepsilon \to 0} W((1-\varepsilon)g + \varepsilon h, c) > W(g, c)$,
  - Incentives become harsh: $lim_{\varepsilon \to 0} u(t(\bar{\ell}, \varepsilon)) = -\infty$.
- Proof sketch:
  - $\mu(\varepsilon)$ jumps downwards at $\varepsilon = 0+$ to satisfy the FOC at $\bar{\ell}$.
  - Therefore, welfare improves discontinuously at $\varepsilon = 0+$.
  - This is only possible with increasingly harsh punishments.

Flimsy Evidence: Optimal monitoring

- Suppose signal h costs P.
- Corollary: $sup_{\varepsilon\in[0,1]} W((1-\varepsilon)g+\varepsilon h, c) - P\varepsilon > W(g, c)$.
- Interpretation:
  - Wardens (g) check many cars $(1-\varepsilon)$ and issue small fines, and
  - Teams of wardens and detectives (h) to check few cars (ε) and issue harsh penalties.

Flimsy Evidence: Limited liability

- Similar logic applies if there is a limited liability constraint, $t(\ell) \geq b$.
- Now, the FOC
$$\frac{1}{u'(t(\ell))} = \lambda + \mu(1 - \ell)$$
fails if the right side falls below $u'(b)$.
- When $\varepsilon \to 0$, the FOC fails, giving the boundary solution $t(\ell) = b$.
    - Interpretation: with limited liability, moderately reliable evidence leads to the worst possible punishment, b.

Flimsy Evidence: Conclusion

Even if the signal g leads to moderate incentives, the theory still fails on all three counts:

- Positive, normative: If there is a cheap way to expand the support of g, then it is optimal to do so and use extreme punishments.
- Methodical: The predictions are discontinuous. The signal g leads to very different predictions than the signal $(1-\varepsilon)g+\varepsilon h$, even when $\varepsilon \to 1$.

Preliminary: What is missing?

Note: this is preliminary work that I plan to test empirically before developing the theory.

There are two separate problems:

1. The model predicts harsh punishments if the evidence is overwhelming, regardless of the probability of getting caught.
2. Small checking probabilities can be compensated by large punishments.

I propose two new ingredients:

1. All evidence is flimsy, e.g. because people make innocent mistakes.
2. The principal must be deterred from extorting the agent (“hand over your money, or I will check your car very carefully”).

“Random” monitoring: model amendments

- Investigations:
  - Each possible investigation $i \in I$ has likelihood ratio distribution f_i.
  - The monitoring regime $p \in \Delta(I)$ costs $M(p)$.
  - Transfers now depend on (i, ℓ).
- The utility function is bounded above by 0 (e.g. CRRA with $\rho > 1$).
- I will work with a participation constraint with outside option u_0.

“Random” monitoring: principal’s problem

$$\min_{p, t_i(\ell)} M(p) + \sum_{i,\ell} p_i f_i(\ell) t_i(\ell)$$

s.t. (VP) $\sum_{i,\ell} p_i f_i(\ell) u(t_i(\ell)) - c \geq u_0,$

(IC-a) $\sum_{i,\ell} p_i f_i(\ell)(1-\ell) u(t_i(\ell)) \geq c,$

(IC-p) $\sum_{\ell} f_i(\ell) u(t_i(\ell)) = \sum_{\ell} f_j(\ell) u(t_j(\ell))$ for all $i, j \in$ support(p).

"Random" monitoring: analysis

Claim. The (VP) and (IC-p) constraints imply

$$u(t_i(\ell)) \geq \frac{u_0 + c}{f_i(\ell)}.$$

- Without loss of generality, assume $p \in interior(\Delta)$.
- Let U_i be the agent's expected utility under investigation i.
- (IC-p) says all $U_i = U_j$ are equal.
- (VP) then implies $U_i \geq u_0 + c$ for all i.
- So $f_i(\ell)u(t_i(\ell)) + (1 - f_i(\ell))0 \geq u_0 + c$ for all (i, ℓ).

Literature

- Related work:
  - Bolton (1987) has a special case of my Corollary 1, where the background information is the null signal.
  - Kim (1995) and Jewitt (2007) developed the likelihood ratio approach.
  - Moroni and Swinkels (2014) study a different moral hazard setting in which extreme punishments arise.

Literature, continued

- Efficient crime:
  - Polinsky and Shavell (1979): Fines should be low, so people speed to the hospital in emergencies.
  - Kaplow and Shavell (1994): You can tell the police about emergencies. Large fines are for dishonesty.
  - But what if you forget to tell the police?

Literature, continued

- Dishonest policing (“subjective performance evaluation”):
  - In Bull (1987), evidence is soft and the principal cannot commit to honesty.
  - In MacLeod (2003), courts can enforce contracts based on soft messages, but not hard evidence.
  - But some evidence is hard, such as recordings.

Literature, continued

- Continuous time:
  - Holmstrom and Milgrom (1987) and Sannikov (2008) study moral hazard in continuous time.
  - All actions and all information is small:
    - Does not accommodate big actions, such as whether to comply with design regulations.
    - Does not accommodate big information, such as investigations, whistleblowers, etc.

Conclusion

- Even with flimsy evidence, moral hazard model predictions are bad on positive, normative, and methodical grounds.
- Perhaps the principal's incentives to randomise are the key?