Physician Models of Prescribing and Topcoding

Lucas

July 2022

1 A Threshold Model with Random Effects

1.1 Basic Model

Doctor i treats patient b in setting s and year t if the expected net utility is positive:

$$T_{ibts} = 1 \iff (\underbrace{h_s(X_{bts})\beta_{is} + \alpha_{ts} + \eta_{ibts}}_{\text{patient's utility as perceived by the doctor}}) \frac{1}{\theta_{is}} + \pi_{ibts} \geq 0$$

where

- X_{bts} are patient characteristics and $h_s(\cdot)$ maps the characteristics to health gains from treatment. The mapping is different from setting to setting
- π_{ibts} is the doctor's monetary reward for treating the patient.
- α_{ts} are setting-specific time trends in doctors' perception of treatment effect on patient health.
- η_{ibts} is the component of patient's utility observed by the doctor but not the econometrician.
- β_{is} is the doctor's perception of how a patient's health affects her utility.
- θ_{is} is an "inverse altruism" parameter governing how much the doctor cares about the patient's utility.

The probability of treatment is:

$$Pr\{T_{ibts} = 1\} = Pr\{h_s(X_{bts})\beta_{is} + \alpha_{ts} + \pi_{ibts}\theta_{is} + \eta_{ibts} \geq 0\} =: Pr\{I_{ibts} \geq 0\}$$

We can think of I_{ibts} as the 'propensity' to treat: the higher it is, the more likely that the doctor crosses the treatment threshold.

1.2 Aggregate Model

The model-implied expected number of treatments Y_{ibts} by doctor i at time t and setting s given patient set $\mathcal{B}_{its}$ is:

$$E[Y_{its}|\mathcal{B}_{its}] = \sum_{b \in \mathcal{B}_{its}} Pr\{I_{ibts} \geq 0\}$$

However, we do not observe the individual patients treated by the doctor, only the aggregate number of treatments and patients. To make some headway, assume:

1. The unobserved component of patient utility is distributed independently of b: $\eta_{ibts} = \eta_{its}$.
2. Assuming we know the distribution $G_x(X_{ibts}), G_\pi(\pi_{ibts})$, which is something that can probably be inferred from the doctor's case mix, we can integrate out the b:

$$E[Y_{its}|\mathcal{B}_{its}] = |\mathcal{B}_{its}| \int_\pi \int_X Pr\{I_{ibts} \geq 0\} dG(X_{ibts}) dG_\pi(\pi_{ibts}) \quad (1)$$

The first assumption is probably innocuous, the second assumption is problematic if we only observe marginal distributions of the characteristics and not the joint.

1.3 Adding Random Effects

Ultimately, we want to examine what may be responsible for the correlations between doctor behavior across settings. To this end, we assume $(\{\beta_{is}, \theta_{is}\}_s)$ are random variables and estimate their covariance structure.

1.3.1 Distributional Assumption

We assume each doctor i draws $(\beta_{is}, \theta_{is})$ independently for each setting s from a joint normal distribution:

$$(\beta_{is}, \theta_{is})_s \sim \mathcal{N}(\mathbf{0}, \mathbf{\Lambda})$$

If there are S settings, then we have $2 \times S$ jointly normal random variables.

The covariance matrix $\mathbf{\Lambda}$ cannot be non-parametrically identified given our data. The intuition is that without something linking the settings together, the fact that these settings are mutually exclusive means we cannot recover most of the off-diagonal covariance terms.

For example, there is no instance of decision-making where $\beta_{i,opioid}$ and $\beta_{i,benzo}$ both play a role, unless we observe the doctor prescribing both drugs at the same time to the same patient. Yet this is required to recover their covariance. However, we can identify the covariance between $(\beta_{is}, \theta_{is})$ using a

setting where both the patient's health and the profit incentive matters to the decision-making process. MIPS is such a setting.

Finally, $_{its}$ is assumed iid Type I Extreme Value and independent of all other variables.

1.3.2 Parametrization

To proceed with our question, we must parametrize the random effects as follows:

1. $\beta_{is} = \bar{\beta} + \beta_i + w_{is}^{\beta}$, one for each setting s. Assume $(\beta_i, w_{is}^{\beta})$ are independently distributed normal for all s.
2. $\theta_{is} = \bar{\theta} + \theta_i + w_{is}^{\theta}$, one for each setting s. Assume $(\theta_i, w_{is}^{\theta})$ are independently distributed normal for all s.

In words, the random effects are made up of (1) an individual-persistent component and (2) an orthogonal deviation from it.

1.3.3 Restrictions on the Covariance Structure

We assume the following restriction on the covariances, in addition to the ones above:

1. $cov(\beta_i, w_{is}^{\theta}) = cov(\theta_i, w_{is}^{\beta}) = 0$ for all s. There is no correlation between the persistent component of a taste parameter and the setting-specific deviation component of the other taste parameter.
2. Group the settings into three categories {Part D, Part B, MIPS}. Then for all s, s' in any two different categories:

$$cov(w_{is}^{\theta}, w_{is'}^{\theta}) = cov(w_{is}^{\beta}, w_{is'}^{\beta}) = cov(w_{is}^{\theta}, w_{is'}^{\beta}) = 0$$

   So that any cross-category correlation comes from the persistent components (β_i, θ_i). Note that there is no restriction if s, s' belong to the same category, e.g. if $s = opioid$ and $s' = benzo$.

3. After simulating: We may need to restrict $var(w_{is}^{\theta}) = var(w_i^{\theta})$ and $var(w_{is}^{\beta}) = var(w_i^{\beta})$ for all s since the RE parameters are hard to identify....

Any other covariance is unrestricted.

1.4 Identification of REs

See the simulation section below for a visual proof of identification. To fix intuition, note that under our model, any correlation between doctor's behavior

across settings s, s' must be attributed to the correlation between the propensity to treat I_{its} in each setting. Conditioning on $X_{ibts} = x$ and $\pi_{ibts} = \pi_s$:

$$
\begin{aligned}
&cov(I_{its}, I_{its'}) = \\
&cov(\eta_{its} + h_s(x)(\beta_i + w^{\beta}_{is}) + \pi_s(\theta_i + w^{\theta}_{is}), \eta_{its'} + h_{s'}(x)(\beta_i + w^{\beta}_{is'}) + \pi_{s'}(\theta_i + w_{is'}\theta)) \\
&= \quad \underbrace{cov(\eta_{its}, I_{its'}) + cov(I_{its}, \eta_{its'})}_{= 0 \text{ due to the distributional assumptions of } \eta} \\
&+ h_s(x)h_{s'}(x)cov(\beta_i + w^{\beta}_{is}, \beta_i + w^{\beta}_{is'}) + \pi_s\pi_{s'}cov(\theta_i + w^{\theta}_{is}, \theta_i + w^{\theta}_{is'}) \\
&+ h_{s'}(x)\pi_s cov(\theta_i + w^{\theta}_{is}, \beta_i + w^{\beta}_{is'}) + h_s(x)\pi_{s'}cov(\beta_i + w^{\beta}_{is}, \theta_i + w^{\theta}_{is'}) \\
&= h_s(x)h_{s'}(x)[var(\beta_i) + cov(w^{\beta}_{is}, w^{\beta}_{is'})] + \pi_s\pi_{s'}[var(\theta_i) + cov(w^{\theta}_{is}, w^{\theta}_{is'})] \\
&+ [h_s(x)\pi_{s'} + h_{s'}(x)\pi_s]cov(\theta_i, \beta_i)
\end{aligned}
$$

where we use our restrictions to cancel out some terms and omit writing the fixed effects because their addition contributes nothing to the covariances.

We first show that $\{var(\theta_i), var(\beta_i), cov(\theta_i, \beta_i)\}$ are identified as long as we observe at least one setting per category. The identification argument goes as follows:

1. If $s \in$ Part D and $s' \in$ Part B, then the propensity covariance is just the last term of the final equality, since there is no variation in $h_{s'}(x)$ in Part B to pin down w^{β}_{is} and no variation in π_s in Part D to pin down $w^{\theta}_{is'}$. In other words, under our model's assumptions, $cov(\beta_i, \theta_i)$ is responsible for the correlation between Part D and Part B behavior.

   So $cov(\theta_i, \beta_i)$ is identified from variation in $h_s(x), \pi_{s'}$ as long as we use Part B and Part D data.

2. If $s \in$ Part D and $s' \in$ MIPS, then the propensity between the covariance is the first and last term. Given that we have pinned down $cov(\theta_i, \beta_i)$ in step 1, we can back out the sum of the covariances in the first term. Furthermore, since s, s' are in different categories, $cov(w^{\beta}_{is}, w^{\beta}_{is'}) = 0$ under our restriction.

   So $var(\beta_i)$ is identified from variation in $h_s(x), h_{s'}(x)$ as long as we also have MIPS data to use with Part D.

3. By symmetrical logic, $var(\theta_i)$ is identified from variation in $\pi_s, \pi_{s'}$ as long as we use MIPS and Part B data together.

Next, we should that if we do have two settings within a category, the within-category covariances between the parameters are identified.

1. If $s, s' \in$ Part D (e.g. benzo and opioid prescription), then the covariance is just the first term. Given that $var(\beta_i)$ is identified, $cov(w^{\beta}_{is}, w^{\beta}_{is'})$ is identified from variation in $h_s(x), h_{s'}(x)$.

2. If $s, s' \in$ Part B, then the covariance is just the first term. Given that $var(\theta_i)$ is identified, $cov(w^{\theta}_{is}, w^{\theta}_{is'})$ is identified from variation in $\pi_s, \pi_{s'}$.

3. If $s, s' \in$ MIPS, then the covariance is all three terms. Given that we have identified $\{var(\theta_i), var(\beta_i), cov(\theta_i, \beta_i)$, variation in $h_s(x), h_{s'}(x)$ identifies $cov(w^{\beta}_{is}, w^{\beta}_{is'})$ and variation in $\pi_s, \pi_{s'}$ identifies $cov(w^{\theta}_{is}, w^{\theta}_{is'})$.

1.4.1 What do we do with this information?

Once we have identified all the covariances, we can test our theories as explanations for the empirical correlations. For example:

- A weak correlation between Part D and Part B behavior may be attributable to a weak $cov(\theta_i, \beta_i)$.
- A strong within-Part D correlation may be due to a strong $cov(w^{\beta}_{is}, w^{\beta}_{is'})$ or a high $var(\beta_i)$. We can test for their relative magnitudes to see what drives the correlation.

  Note that a high $var(\beta_i)$ is consistent with a strong correlation between Part D and MIPS, so if we do not see this in the data, it's likely the latter.
- My hypothesis is that the within-category covariances far outweighs the variances because we have strong within-categories correlations but extremely weak ones across categories.

1.5 Simulations

1.5.1 Data Simulation

First I simulate data (x_{its}, π_{its}) for $N = 10,000$ NPIs, over $T = 4$ years and across $S = 4$ settings corresponding to opioids, benzos, upcoding and MIPS, each with $B = 100$ patients. For simplicity I directly draw $h(x_{ibts})$ instead of x_{ibts}. I draw them all independently from normal distributions.

1.5.2 Specify the Parameters

For the fixed effects, I parametrize $\alpha_{ts} = t\alpha_s$. There are $S = 4$ fixed effects parameters, one for each setting.

For the covariances, recall that the $2 \times S = 8$ variables we have assumed to be jointly normal are:

$$(\beta_{is}, \theta_{is})_s = \boldsymbol{\beta}_{full} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}_{full})$$

Let:

$$\boldsymbol{\beta}_{full} = \begin{pmatrix} \beta_o \\ \beta_b \\ \beta_u \\ \beta_m \\ \theta_o \\ \theta_b \\ \theta_u \\ \theta_m \end{pmatrix}$$

However, note that $(\theta_o, \theta_b, \beta_u)$ are not identified, so we look at the associated joint distributions of the following 5 variables:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_o \\ \beta_b \\ \beta_m \\ \theta_u \\ \theta_m \end{pmatrix}$$

Under the restrictions, the 5×5 covariance matrix has the following triangular form. This is incidentally a visual proof of identification.

$$\begin{aligned} \boldsymbol{\Lambda} &= \begin{pmatrix} \sigma_o^\beta & & & & \\ \sigma_{ob}^\beta & \sigma_b^\beta & & & \\ \sigma_{om}^\beta & \sigma_{bm}^\beta & \sigma_m^\beta & & \\ \sigma_{ou}^{\beta\theta} & \sigma_{bu}^{\beta\theta} & \sigma_{mu}^{\beta\theta} & \sigma_u^\theta & \\ \sigma_{om}^{\beta\theta} & \sigma_{bm}^{\beta\theta} & \sigma_{mm}^{\beta\theta} & \sigma_{mu}^\theta & \sigma_m^\theta \end{pmatrix} \\ &= \begin{pmatrix} var(\beta) + var(w_o^\beta) & & & & \\ var(\beta) + cov(w_o^\beta, w_b^\beta) & var(\beta) + var(w_b^\beta) & & & \\ var(\beta) & var(\beta) & var(\beta) + var(w_m^\beta) & & \\ cov(\beta,\theta) & cov(\beta,\theta) & cov(\beta,\theta) & var(\theta) + var(w_u^\theta) & \\ cov(\beta,\theta) & cov(\beta,\theta) & cov(\beta,\theta) + cov(w_m^\beta, w_m^\theta) & var(\theta) & var(\theta) + var(w_m^\theta) \end{pmatrix} \end{aligned}$$

In total, there are $K_r = 10$ random effect parameters and $K_f = 4$ fixed effect parameters to estimate.

I specify the parameters and use them to compute the variance-covariance matrix of the random effects. Make sure that the resulting matrix is positive-definite!

1.5.3 Drawing the Random Effects

I draw $N = 10,000$ draws of (β_i, θ_i) from a multivariate normal with mean $\mathbf{0}$ and variance $\begin{pmatrix} var(\beta) & cov(\beta,\theta) \\ cov(\beta,\theta) & var(\theta) \end{pmatrix}$. I then draw $N = 10,000$ draws of $\begin{pmatrix} w_o^\beta \\ w_b^\beta \\ w_m^\beta \\ w_u^\theta \\ w_m^\theta \end{pmatrix}$ with mean $\mathbf{0}$ and variance $\begin{pmatrix} var(w_o^\beta) & & & & \\ cov(w_o^\beta, w_b^\beta) & var(w_b^\beta) & & & \\ 0 & 0 & var(w_m^\beta) & & \\ 0 & 0 & 0 & var(w_u^\theta) & \\ 0 & 0 & cov(w_m^\beta, w_m^\theta) & 0 & var(w_m^\theta) \end{pmatrix}$.

I use the random effects and the data to compute the propensity I_{ibts}, for each doctor-patient-year-setting. Then I generate a binary variable indicating treatment, i.e. whether $I_{ibts} + \Delta\eta \geq 0$. Finally, I collapse the $(Y_{ibts}, h_{ibts}, \pi_{ibts})$ to the doctor-year-setting level by taking their means to imitate the aggregate data set we have.

1.5.4 Estimation: Simulated MLE

The assumption that η_{ibts} is iid Type I Extreme Value is useful. Given a setting s and time t, assuming that treatment is a binary decision[1] and the outside option has a normalized utility of zero, the probability of treatment conditional on a realization of $(\beta_{is}, \theta_{is})_s$ is:

$$p_{its} = \int_\pi \int_X \frac{\exp(h_s(X_{bts})\beta_{is} + \pi_{ibts}\theta_{is} + t\alpha_s)}{1 + \exp(h_s(X_{bts})\beta_{is} + \pi_{ibts}\theta_{is} + t\alpha_s)} dG_x(X_{ibts}) dG_\pi(\pi_{ibts})$$

For each simulated $(\beta_{is}^m)_s$, the associated probability is:

$$p_{its}^m := \int_\pi \int_X \frac{\exp(h_s(X_{bts})\beta_{is}^m + \pi_{ibts}\theta_{is} + t\alpha_s)}{1 + \exp(h_s(X_{bts})\beta_{is}^m + \pi_{ibts}\theta_{is}^m + t\alpha_s)} dG_x(X_{ibts}) dG_\pi(\pi_{ibts})$$

But this is computationally very slow. Alternatively, we can approximate the above equation using the observed means $(\bar{X}_{bts}, \bar{\pi}_{bts})$ if not confident about our estimated distributions:

$$p_{its}^m \approx \frac{\exp(h_s(\bar{X}_{its})\beta_{is}^m + \bar{\pi}_{its}\theta_{is} + t\alpha_s)}{1 + \exp(h_s(\bar{X}_{its})\beta_{is}^m + \bar{\pi}_{its}\theta_{is}^m + t\alpha_s)}$$

So for each simulation we model Y_{its} using a binomial distribution with parameter p_{its}^m. Let B_{its} be the total number of patients/office visits. Then the likelihood of observing Y_{its} is given by:

$$\mathcal{L}_{its}^m(\boldsymbol{\gamma}) = \binom{B_{its}}{Y_{its}} {p_{its}^m}^{Y_{its}} (1 - p_{its}^m)^{B_{its} - Y_{its}}$$

We must average the simulations at the level at which the parameter p_{its} changes! Here it is its. The simulated likelihood for each i in year t and setting s is given by:

$$\mathcal{L}_{its}^{sim}(\boldsymbol{\gamma}) = \frac{1}{M} \sum_{m=1}^{M} \binom{B_{its}}{Y_{its}} {p_{its}^m}^{Y_{its}} (1 - p_{its}^m)^{B_{its} - Y_{its}}$$

The simulated joint likelihood is:

$$\mathcal{L}^{sim}(\boldsymbol{\gamma}) = \prod_{i=1}^{N} \prod_{t \in T(i)} \prod_{s \in S(T(i),i)} \mathcal{L}_{its}^{sim}(\boldsymbol{\gamma})$$

So the simulated log-likelihood over all i is:

$$\ell(\boldsymbol{\gamma}) = \sum_{i=1}^{N} \sum_{t \in T(i)} \sum_{s \in S(T(i),i)} \log(\mathcal{L}_{its}^{sim})$$

[1] This could easily be extended to accommodate a multinomial logit problem if we observe other treatment options.

Notes of logsumexp: $\log(\mathcal{L}_{its}^{sim})$ could be written in terms of logSumExp it it helps:

$$\begin{aligned}\log(\mathcal{L}_{its}^{sim}) &= -\log(M) + \\ &\log(\sum_{m=1}^{M} \exp(\log\binom{B_{its}}{Y_{its}} + Y_{its}\log(p_{its}) + (B_{its} - Y_{its})\log(1 - p_{its}^m))) \\ &= -\log(M) + \log(\sum_{m=1}^{M} \exp(\ell_{its}^m)) \\ &= -\log(M) + \text{logSumExp}(\ell_{its}^m)\end{aligned}$$

The steps go as follows:

1. (Optional) Draw $(h(x_{ibts}), \pi_{itbs})$ M times for each its, based on the distributions that are either known or estimated from the observed patient mix $h(\bar{X}_{its})$ and mean profit $\bar{\pi}_{its}$. We can also just use the observed means directly.
2. Draw N draws of 5 joint standard normals $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Repeat $M = 1000$ times for Monte Carlo integration.
3. Optimization! Fix a candidate set of fixed and random effect parameters $\boldsymbol{\gamma}_c$. Maybe just do the Choelsky matrix and restrict diagonal to be positive by estimaing the log of the diagonal.
   (a) Compute the implied covariance matrix $\boldsymbol{\Lambda}_c$
   (b) Perform Cholesky decomposition on the joint normals drawn in (2) to rescale their variances and covariances:
   $$\boldsymbol{\beta}_c^m = chol(\boldsymbol{\Lambda}_c)\boldsymbol{Z}^m$$
   where $chol(\boldsymbol{\Lambda}_c)$ refers to the lower-triangular Cholesky matrix.
   (c) Compute $p_{its}^m(\boldsymbol{\gamma})$ by plugging in the $\boldsymbol{\beta}_c^m$.
   (d) Calculate the associated likelihood $\ell^m(\boldsymbol{\gamma})$.
   (e) Repeat M times to average to the simulated likelihood.

The biggest drawback of MLE is that we are approximating the sum of differently-distributed Bernoullis with a Binomial. Thus we are guaranteed to have misspecification error (and we do, even in simulations).

1.5.5 MLE with Poisson

If we decide to approximate the Bernoulli sum with a Poisson distribution instead[2], our log-likelihood will be:

$$\ell^m(\boldsymbol{\gamma}^{Poisson}) = \sum_{its}[Y_{its}\log(p_{its}^m) - p_{its}^m]$$

[2] https://projecteuclid.org/journals/pacific-journal-of-mathematics/volume-10/issue-4/An-approximation-theorem-for-the-Poisson-binomial-distribution/pjm/1103038058.full

1.5.6 MLE with Homogeneous Patients

Assume that patients b within each its are identical, so $\pi_{ibts} = \bar{\pi}_{its}$ and $h_{ibts} = \bar{h}_{its}$. Conditioned on a fixed set of parameters and an its, the only randomness from one patient to the next comes from the Type I Extreme Value error, which is iid. This means that the expression for probability of treatment is exact:

$$p_{its}^m = \frac{\exp(h_s(\bar{X}_{its})\beta_{is}^m + \bar{\pi}_{its}\theta_{is} + t\alpha_s)}{1 + \exp(h_s(\bar{X}_{its})\beta_{is}^m + \bar{\pi}_{its}\theta_{is}^m + t\alpha_s)}$$

And because our aggregation involves a sum iid Bernoullis, when conditioned on a draw of random effects, our Binomial log likelihood is also exact:

$$\ell^m(\boldsymbol{\gamma}^{homog}) = \sum_{its}[Y_{its}^{homog}\ln(p_{its}^m) + (B_{its} - Y_{its}^{homog})\ln(1-p_{its}^m) + \ln\binom{B_{its}}{Y_{its}^{homog}}]a$$

The difference between $\boldsymbol{\gamma}^{homog}$ and $\boldsymbol{\gamma}$ measures the aggregation bias of our MLE method.

1.5.7 Something wrong with the model

In simulation, the treatment propensity I is not correlated with the random effects but is correlated with the whole sum...

We also need to estimate the means of the persistent random effects otherwise if we center them at zero, we have the implication that on average, neither profit nor patient welfare affects propensity to treat!

What exactly identifies the random effect?

In simulation: Seems like $cov(I_{[}PartD], I_{PartB})$ does match $cov(\theta, \beta)$ in sign and magnitude. $cov(Y_{[}PartD], Y_{PartB})$ is extremely unrelated. $cov(Y, I)$ is positive however.

I guess this all depends on the variance.

1.5.8 Estimation: Simulated GMM

Let p_{its}^{sim} be the model-implied proportion of times the doctor chooses to treat at year t and setting s:

$$p_{its}^{sim}(\boldsymbol{\gamma}) = \frac{1}{M}\sum_{m=1}^{M}\frac{\exp(h_s(\bar{X}_{its})\beta_{is}^m + \bar{\pi}_{its}\theta_{is} + t\alpha_s)}{1 + \exp(h_s(\bar{X}_{its})\beta_{is}^m + \bar{\pi}_{its}\theta_{is}^m + t\alpha_s)}$$

Let y_{its} be the observed percentage of times treatment occurs. Then the following moment satisfies the key condition:

$$E[g(\boldsymbol{\gamma}; Y, X, \Pi)] = E[p_{its}^{sim}(\boldsymbol{\gamma}) - y_{its}] = 0$$

That is, we expect the average prediction error to be zero. Note that in the model, we have two types of errors: aggregation and simulation. That is, the error comes from using the means of X_{its} and π_{its} as 'representative' even though

we are dealing with a non-linear aggregation. The error also comes from having to simulate β_{is}.

We have $K = 14$ parameters and we $L \geq K$ moments. So we need instruments Z that's uncorrelated with the above errors:

1. Using year dummies, of which we have 4. The simulation error is not correlated with years.
2. The sum of the average patient characteristics of the other doctors $\bar{X}_{i'ts}$.
3. The sum of the average profits of the other doctors. $\bar{\pi}_{i'ts}$.
4. Doctors' own characteristics.
5. Sum of other doctors' characteristics.
6. 2-5 can be interacted with year dummies probably, which would give us 12 moments.
7. Can also interact with setting dummies conditional on Part D. Aggregation bias coming from X is likely different from aggregation bias coming from π.

Another possible moment is that the average correlation between any pair of doctors' behavior for a given year is zero. This is because the doctors draw their taste parameters independently, so once we condition on the time trend, there should be no correlation between their behavior. So basically just correlation-across settings, sum over all pairs of doctors, interacted with year dummies.