RA Sample Exercise

February 26, 2020

The following is an research note excerpt from a project on star ratings for General Practitioners (GPs) in the UK. The note regards how a Bayesian consumer should interpret the expected quality of a GP given that the UK quality ratings website only shows the average star rating (rounded to the nearest half-star) and the number of ratings that the GP has received. One thing that should be obvious to anyone who uses ratings (e.g. Yelp) is that consumers should place more weight on average ratings for GPs that have received a large sample of ratings in the past.

1 The Model

1.1 How Reviews Are Determined

GP j has some true underlying quality q_j^*. When reviewer r arrives, she draws some experienced quality $q_{rj}^* \sim \mathcal{N}(q_j^*, 1)$.[1] She then gives the GP a rating, q_{rj}, according to an ordered-probit

[1]Note that I have not specified the arrival process for reviewers. Independence of q_{rj}^* makes the arrival process irrelevant to interpreting the information from ratings.

with cut points $\{\underline{q}_c\}_{1\leq c\leq 4}$.[2] Formally:

$$q_{rj} = \begin{cases} 5 & \text{if } q^*_{rj} \in (\underline{q}_4, \infty) \\ 4 & \text{if } q^*_{rj} \in (\underline{q}_3, \underline{q}_4] \\ 3 & \text{if } q^*_{rj} \in (\underline{q}_2, \underline{q}_3] \\ 2 & \text{if } q^*_{rj} \in (\underline{q}_1, \underline{q}_2] \\ 1 & \text{if } q^*_{rj} \in (-\infty, \underline{q}_1] \end{cases} \tag{1}$$

1.2 How Ratings are Interpreted By Patients Choosing their GP

Consider that prospective patient i observes only $\tilde{q}_{jt} := round(\bar{q}_{jt}, .5)$ and N_{jt}, where $\bar{q}_{jt}$ is the average q_{rj} for the N_{jt} reviewers who arrived within the two years prior to t, and $round(x, .5)$ signifies rounding x to the nearest half-integer. Assuming that N_{jt} is independently of everything else, prospective patient i computes a posterior:[3]

$$p(q^*_j|\tilde{q}_{jt}, N_{jt}) = \frac{p(\tilde{q}_{jt}|N_{jt}, q^*_j)p(q^*_j)}{p(\tilde{q}_{jt}|N_{jt})} = \frac{p(\tilde{q}_{jt}|N_{jt}, q^*_j)p(q^*_j)}{\int p(\tilde{q}_{jt}|N_{jt}, q^*_j)dP(q^*_j)} \tag{2}$$

The value for $p(\tilde{q}_{jt}|N_{jt}, q^*_j)$ can be computed analytically from the ordered probit, and measure $P(q^*_j)$ is a prior. The only challenging part of computing (2) is determining which combinations of N_{jt} ratings could yield rounded average rating $\tilde{q}_{jt}$:

$$R(\tilde{q}_{jt}, N_{jt}) := \left\{\{q_{rt}\}_{r\leq N_{jt}} \ : \ \tilde{q}_{jt} = round(\bar{q}_{jt}, .5)\right\} \subseteq \{1,2,3,4,5\}^{N_{jt}} \tag{3}$$

We need this in order to compute:

$$p(\tilde{q}_{jt}|N_{jt}, q^*_j) = \sum_{\{q_{rt}\}_{r\leq N_{jt}} \in R(\tilde{q}_{jt}, N_{jt})} P(\{q_{rt}\}_{r\leq N_{jt}}|N_{jt}, q^*_j). \tag{4}$$

[2]Note that the cut points are the same for all j.

[3]One could try to incorporate a relationship between N_{jt} and q^*_j by not requiring $P(q^*_j|N_{jt}) = P(q^*_j)$.

Note that while, $P(q_{rt}|q_j^*)$ is straightforward to compute from the Normal CDF and the cutpoints from the ordered probit, you have to be a little thoughtful about the combinatorics in constructing $P(\{q_{rt}\}_{r\leq N_{jt}}|N_{jt},,q_j^*)$ from $P(q_{rt}|q_j^*)$.

1.3 Task

This is a real RA task I gave to an undergrad RA. I have since added additional details to the model above and some hints below in order to make this more easily/quickly feasible. Still, this is a challenging task, so let me know if you need hints. I will be extremely impressed if you're able to get all of this without any help.

Your goal is to produce a function that computes $\mathbb{E}[q_j^*|\tilde{q}_{jt}, N_{jt}]$, the expected quality of a GP given observed rounded rating $\tilde{q}_{jt}$ and number of reviews N_{jt}. The function obviously needs to take $\tilde{q}_{jt}, N_{jt}$ as variable inputs. Some fixed inputs (i.e. inputs that won't ever change) that you also need are:

- A vector of random draws from the prior $P(q_j^*)$. This is so that computing $\int f(q_j^*)dP(q_j^*)$ is as simple as taking an average of $f(\cdot)$ over the vector of random draws.
- A vector of 4 cutpoints for the ordered probit.

Some hints:

- Start by building the simplest functions. For example, $P(q_{rt}|q_j^*)$ is really easy.
- Think carefully through the combinatorics when computing $R(\tilde{q}_{jt}, N_{jt})$.
  - This is one of the hardest (if not the hardest) parts, especially if you're trying to do so efficiently.
  - I suggest using recursion. This shows how to do so for an equivalent problem: `https://www.techiedelight.com/total-ways-sum-with-n-throws-dice-having-k-faces/`. Memoization can also really help.

- An excellent "checkpoint" is computing $p(\tilde{q}_{jt}|N_{jt}, q_j^*)$. If you can compute $p(\tilde{q}_{jt}|N_{jt}, q_j^*)$, you're basically done with the hard part. Let me know when you can compute $p(\tilde{q}_{jt}|N_{jt}, q_j^*)$. I may tell you that the code gives me enough information and you don't need to finish the exercise.

- Test your functions along the way. For example: do the probabilities sum to 1 when you integrate your probability over the whole state space?

Note that I only need this to work up to $N_{jt} = 20$. High performance is desirable but not strictly necessary here. Clean and clear code is important. I would strongly prefer that you use Julia. (Though, feel free to work out the logic in your preferred language first and then translate it over.)