Никольский И.М., Фурманов К.К.
Москва, МГУ им М.В. Ломоносова, ЦЭМИ РАН

ПРИБЛИЖЕННЫЙ РАСЧЁТ КОЭФФИЦИЕНТОВ СОГЛАСОВАННОСТИ ИСТИННЫХ И ОЦЕНЁННЫХ РАНЖИРОВОК ПРЕДПРИЯТИЙ ПО НЕЭФФЕКТИВНОСТИ В ПРОСТОЙ МОДЕЛИ СТОХАСТИЧЕСКОЙ ГРАНИЦЫ

Мы рассматриваем простую модель стохастической границы, предложенную в работе [1]:

$$\ln y_i = x_i'\beta + \upsilon_i - u_i, \qquad (1)$$

где y_i — выпуск предприятия i, x_i' — вектор-строка объясняющих переменных (затрат факторов производства, обычно в логарифмированном виде), β — вектор коэффициентов при объясняющих переменных, υ_i — случайный шок, u_i — показатель неэффективности предприятия i.

Предполагается, что наблюдения независимы, объясняющие переменные детерминированы, случайные шоки υ_i и показатели неэффективности u_i независимы, причём $\upsilon_i \sim N(0, \sigma_\upsilon^2)$, $u_i \sim N^+(0, \sigma_u^2)$. На практике кроме полунормального распределения используются ещё показательное (экспоненциальное) и усечённое нормальное распределение, но в настоящем докладе основное внимание уделяется полунормальному распределению.

Оценивание параметров модели проводится обычно методом максимального правдоподобия, который позволяет получить состоятельные и асимптотически эффективные оценки коэффициентов β и параметров распределения случайных составляющих σ_υ^2, σ_u^2, после чего можно рассчитать оценки $\hat{u}_1, \ldots, \hat{u}_n$ — компонент неэффективности отдельных предприятий. Эти оценки свойством состоятельности не обладают и даже неограниченное увеличение объёма выборки не гарантирует их приемлемую точность.

В докладе предлагается способ приближённого расчёта коэффициентов ранговой корреляции Харрелла и Кендалла между истинными и модельными значениями компонент неэффективности для больших выборок.

Как правило, для измерения согласованности ранжировок исследователи используют коэффициент ранговой корреляции Спирмена (см., например, [2] и [3]), но мы предпочитаем коэффициенты Харрелла и Кендалла из-за интерпретируемости их значений. Применительно к задаче измерения неэффективности его можно определить следующим образом. Пусть из генеральной совокупности случайно отбираются пара субъектов. Обозначим истинные значения их компонент неэффективности u_1, u_2, а оценки этих компонент соответственно $\hat{u}_1, \hat{u}_2$. Коэффициент согласованности Харрелла C задаётся выражением ([4]):

$$C = P(\hat{u}_1 < \hat{u}_2 | u_1 < u_2) \quad (2)$$

Значение C лежит в пределах $[0; 1]$ и отражает долю случаев, в которых модель верно выделяет относительно эффективные предприятия, среди всех пар предприятий. Например, значение $C = 0.5$ соответствует полной неспособности отличить более эффективный субъект от менее эффективного, а при $C = 1$ оценённая модель ранжирует предприятия безошибочно.

Коэффициент Кендалла τ связан с коэффициентом Харрелла равенством[1] $\tau = 2C - 1$, принимает значения из отрезка $[-1; 1]$ и показывает, насколько вероятность правильно выбрать более эффективное предприятие больше вероятности ошибиться:

$$\tau = P(\hat{u}_1 < \hat{u}_2 | u_1 < u_2) - P(\hat{u}_1 > \hat{u}_2 | u_1 < u_2).$$

Непосредственный расчёт даже выборочных значений коэффициентов корреляции невозможен, потому что показатели неэффективности предприятий u_i ненаблюдаемы.

Покажем, что формула (2) на больших выборках сводится к функции от дисперсий неэффективности и случайных шоков. Вывод будет опираться на следующие утверждения.

1) Метод максимального правдоподобия даёт состоятельные оценки $\hat{\beta}$ коэффициентов β уравнения (1). Далее будем пренебрегать различием между оценками и истинными коэффициентами.

2) Оценка показателя неэффективности $\hat{u}_i$ есть строго убывающая функция от $\left(\ln y_i - x_i'\hat{\beta}\right)$, то есть остатка модели (1).

3) Пусть u_1 и u_2 независимы, $u_i \sim N^+(0, \sigma_u^2)$. Тогда их разность распределена приближённо нормально: $u_1 - u_2 \sim N\left(0, 2\sigma_u^2\left(1 - \frac{2}{\pi}\right)\right)$. Выражение для дисперсии следует из формулы дисперсии полунормального распределения.

Утверждения 1 и 2 точные, утверждение 3 приблизительное, поэтому и результат исследования оказывается приблизительным. Рис. 1 иллюстрирует утверждение 3 — на графике представлена гистограмма разности полунормальных величин в искусственно сгенерированной выборке из миллиона наблюдений вместе с графиком плотности нормального распределения.

[1] Это равенство выполняется только для непрерывных случайных величин. Общий случай описан в статье [5].

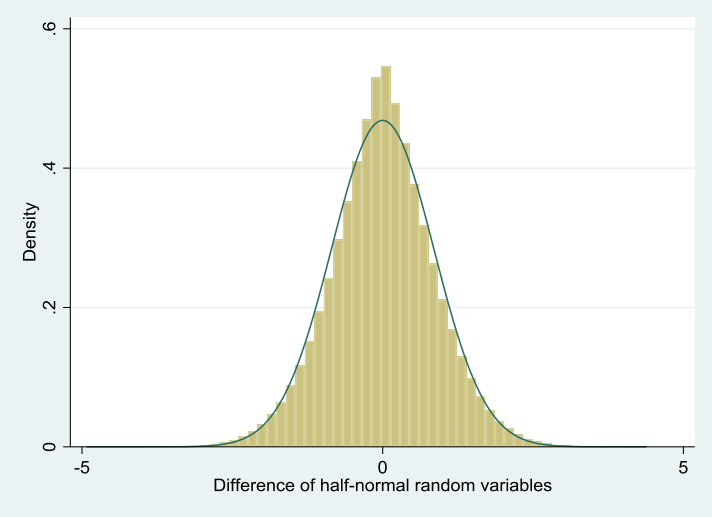

Рис.1. Гистограмма разности полунормальных случайных величин и нормальная плотность.

Из рис.1 видно, что распределения в целом схожи, хотя плотность $u_1 - u_2$ заметно больше нормальной в окрестности нуля.

Из утверждения 2 следует, что событие $\{\hat{u}_1 < \hat{u}_2\}$ эквивалентно событию $\{\ln y_1 - x_1'\hat{\beta} > \ln y_2 - x_2'\hat{\beta}\}$. Из утверждения 1 следует, что на больших выборках $\ln y_i - x_i'\hat{\beta} \approx \ln y_i - x_i'\beta = u_i - \upsilon_i$, так что выражение (2) можно переписать следующим образом:

$$C = P(\hat{u}_1 < \hat{u}_2 | u_1 < u_2) \approx P\big((u_1 - \upsilon_1) - (u_2 - \upsilon_2) < 0 | u_1 - u_2 < 0\big).$$

Обозначим $\xi = u_1 - u_2$ и $\eta = \upsilon_1 - \upsilon_2$. Тогда

$$C \approx P(\xi - \eta < 0 | \xi < 0) \qquad (3)$$

Если считать (согласно утверждению 3), что $\xi \sim N\left(0, 2\sigma_u^2\left(1 - \frac{2}{\pi}\right)\right)$, то величины $\xi - \eta$ и ξ имеют совместное нормальное распределение, хотя это верно лишь приблизительно. Математические ожидания этих величин равны нулю. Можно выразить коэффициент корреляции этих величин через параметры σ_u^2 и σ_υ^2 случайных компонент модели (1):

$$Corr(\xi, \xi - \eta) = \sqrt{\frac{\sigma_u^2\left(1 - \frac{2}{\pi}\right)}{\sigma_u^2\left(1 - \frac{2}{\pi}\right) + \sigma_\upsilon^2}}.$$

Обратим внимание на то, что под корнем находится доля дисперсии неэффективности u_i в дисперсии суммарной случайной составляющей $u_i + \upsilon_i$. Обозначим её θ, так что $Corr(\xi, \xi - \eta) = \sqrt{\theta}$.

Теперь получим выражение для коэффициента Харрелла (3) согласно нормальному приближению:

$$C \approx P(\xi - \eta < 0 | \xi < 0) = \frac{P(\{\xi - \eta < 0\} \cap \{\xi < 0\})}{P(\xi < 0)} = 2\Phi(0, 0, \sqrt{\theta}). \qquad (4)$$

Здесь $\Phi(0{,}0, \sqrt{\theta})$ — значение функции совместного распределения двух центрированных и нормированных совместно нормальных величин с коэффициентом корреляции $\sqrt{\theta}$ в точке $(0{,}0)$. Величины $\xi - \eta$ и η не нормированы, но в нашем случае это не имеет значения, так как значение функции распределения в точке $(0{,}0)$ не зависит от дисперсий.

Формула (4) может быть использована для приближённого расчёта коэффициента Харрелла между истинными и оценёнными показателями неэффективности. В таблице 1 сопоставлены значения, рассчитанные по формуле (4), и оценки коэффициента Харрелла, полученные в результате серии экспериментов Монте-Карло с генерированием случайных выборок разного объёма (для каждой клетки таблицы генерировалась 1000 выборок, объёмы выборок n указаны в таблице) и подгонкой к этим данным стохастической границы, соответствующей производственной функции Кобба-Дугласа.

Как показывают эксперименты, погрешность формулы (4) невелика — наибольшее расхождение между симулированными и рассчитанными по формуле значениями в таблице 1 составляет 0.018 для выборок в 100 наблюдений и 0.012 для выборок в 1000 и 10000 наблюдений. Таким образом, формула (4) даёт исследователю возможность измерить точность оценок неэффективности, опираясь на соотношение между дисперсиями неэффективности и случайных шоков. По желанию исследователя, она может быть использована для расчёта более популярной меры связи — коэффициента корреляции τ Кендалла — при этом погрешность вырастет в два раза, так как $\tau = 2C - 1$.

Конечно, коэффициенты Харрелла и Кендалла не измеряют близость истинных значений и их оценок, а только согласованность их ранжировок. В статье [6] предлагается следующий пример: если все оценки превосходят истинные значения в десять раз, коэффициенты ранговой корреляции покажут идеальную согласованность, хотя такие оценки совсем не точные. Однако в нашем случае такая ситуация практически невозможна, так как параметры распределений случайных составляющих оцениваются состоятельно и асимптотически эффективно, поэтому порядок этих величин можно надёжно установить — это и позволяет исследователю применить формулу (4), рассчитывая вклад неэффективности в суммарную дисперсию θ, опираясь не на ненаблюдаемые истинные значения σ_u^2 и σ_υ^2, а на их оценки.

Таблица 1

Значения коэффициента согласованности Харрелла между истинными и модельными ранжировками, рассчитанные по приближённой формуле (4) и в результате статистических экспериментов.

θ	0.1	0.2	0.3	0.4	0.5
Приближённые значения	0.602	0.648	0.685	0.718	0.750
Симуляции, $n = 100$	0.598	0.641	0.675	0.708	0.737
Симуляции, $n = 1000$	0.599	0.642	0.678	0.710	0.740
Симуляции, $n = 10000$	0.599	0.643	0.678	0.710	0.741
θ	0.6	0.7	0.8	0.9	
Приближённые значения	0.782	0.815	0.852	0.898	
Симуляции, $n = 100$	0.769	0.800	0.834	0.880	
Симуляции, $n = 1000$	0.771	0.803	0.840	0.886	
Симуляции, $n = 10000$	0.771	0.804	0.840	0.886	

Список использованной литературы:

1. Aigner D., Lovell C.A.K., and Schmidt P. “Formulation and estimation of stochastic frontier function models”. Journal of Econometrics, 6, 1977, pp. 21–37.
2. Feng D., Wang C., and Zhang X. “Estimation of inefficiency in stochastic frontier models: a Bayesian kernel approach”. Journal of Productivity Analysis, 51(1), 2019, pp. 1–19.
3. Малахов Д.И., Пильник Н.П. “Методы оценки показателя эффективности в моделях стохастической производственной границы”. Экономический журнал Высшей Школы Экономики, 17(4), 2013, с. 660–686.
4. Harrell F.E. Jr., Califf R.M., Pryor D.B., Lee K.L., and Rosati R.A. “Evaluating the yield of medical tests”. Journal of the American Medical Association, 247(18), 1982, pp. 2543–2546.
5. Newson R.B. “Parameters behind “nonparametric” statistics: Kendall’s tau, Somers’ D and median differences”. The Stata Journal, 2(1), 2002, pp. 45–64.
6. Румянцева Е.В., Фурманов К.К. “Использование вневыборочных остатков Кокса–Снелл при прогнозировании наступления событий”. Бизнес-Информатика, 15(1), 2021, с. 7–18.