Увеличение числа наблюдений в регрессии. Асимптотические свойства.

16 апреля 2020 г.

Есть несколько стандартных методов увеличения числа наблюдений n в регрессии. Самый простой из них — когда измерения проводятся много раз в одних и тех же точках. Второй — когда измерения проводятся на одном и том же отрезке (по x), но с шагом, стремящемся к нулю (например, в точках вид i/n). Третий, когда измерения проводятся в точках $1, 2, 3, \ldots, n$ (обычно это время). Это три случая, когда регрессоры не случайны. Случай, когда данные порождены случайным вектором (повторные наблюдения), самый часто встречающийся, но его мы рассмотрим позже.

Основной источник — Демиденко. Линейная и нелинейная регрессии. 1981.

1 Детерминированные регрессоры. Условия состоятельности и асимптотической нормальности

Рассмотрим $Y = \mathbf{X}B + \mathcal{E},\ \mathrm{E}\varepsilon_i = 0,\ \mathrm{D}\varepsilon_i = \sigma^2, \{\varepsilon_i\}$ независимы, $i = 1, \ldots, n$, матрица $\mathbf{X}$ детерминирована, то есть x_{ij} не случайны, $\mathrm{rank}\mathbf{X} = m$.

Определение. *Слабая состоятельность:* $P(|\hat{\Theta}_n - \Theta| > \varepsilon) \to 0,\ n \to \infty$.

Определение. *Сильная состоятельность:* $P\{\lim_{n\to\infty} \hat{\Theta}_n = \Theta\} = 1$.

Определение. *Состоятельность в среднеквадратическом:* $\mathrm{E}(\hat{\Theta}_n - \Theta)^2 \to 0,\ n \to \infty$.

Состоятельность в среднеквадратическом и сильная состоятельность влекут состоятельность в слабом смысле.

1.1 Условия для состоятельности

Определение. *Условия сильной регулярности:*

$$\lim_{n\to\infty} \frac{1}{n}\mathbf{X}_n^{\mathrm{T}}\mathbf{X}_n = \mathbf{A}$$

где $\mathbf{A}$*—невырожденная матрица* $m \times m$.

Теорема (Теорема 1.4)**.** *Если матрицы* $\mathbf{X}_n$ *сильно регулярны, то оценка МНК состоятельна в среднеквадратическом.*

Это условие очевидно выполняется, если проводить эксперименты в одних и тех же точках много раз ($\mathbf{X}_n$ состоит из матрицы данных $\mathbf{X}$, повторенной несколько (много) раз).

Оно же будет выполнено для все более частых измерений на одном отрезке. Однако в случае временных рядов оно не выполняется. Сформулируем другое условие.

Теорема (Теорема 1.5, условие Эйкера)**.** *Условие* $\lambda_{min}(\mathbf{X}_n^{\mathrm{T}}\mathbf{X}_n) \to \infty,\ n \to \infty$ *эквивалентно состоятельности оценки МНК в среднеквадратическом.*

Заметим, что $\lambda_{\min}(\mathbf{X}_n^{\mathrm{T}}\mathbf{X}_n)$ есть квадрат минимальной длины вектора, являющегося линейной комбинацией вектор-столбцов $x_1, x_2, \ldots, x_m$ матрицы $\mathbf{X}_n$. Таким образом, $\lambda_{\min}(\mathbf{X}_n^{\mathrm{T}}\mathbf{X}_n)$ можно трактовать как показатель вырожденности $\mathbf{X}_n^{\mathrm{T}}\mathbf{X}_n$ или как меру линейной зависимости переменных $x_1, x_2, \ldots, x_m$.

Условие $\lambda_{\min}(\mathbf{X}_n^{\mathrm{T}}\mathbf{X}_n) \to \infty$ проверять на практике весьма сложно. Можно предложить более простой критерий состоятельности оценки МНК.

Построим для матрицы $\mathbf{X}_n$ матрицу сопряженности $\mathbf{R}_n$:

$$(\mathbf{R}_n)_{ij} = \frac{\sum\limits_t x_{ti}x_{tj}}{\sqrt{\sum\limits_t x_{ti}^2 \sum\limits_t x_{tj}^2}}, \quad i,j = 1,\dots,m.$$

$(\mathbf{R}_n)_{ij}$ — это косинус угла между векторами (признаками) x_i и x_j в евклидовом пространстве $\mathbb{R}^n$. Матрица сопряженности похожа на матрицу корреляций $x_1, x_2, \dots, x_m$. Отличие лишь в том, что в матрице корреляций рассматриваются отклонения x_{ti} от соответствующего среднего $\bar{x}_i$.

Теорема (Теорема 1.6)**.** *Если:*
1. $\sum\limits_{i=1}^{n} x_{ij}^2 \to \infty,\ n \to \infty,\ \ \forall j = 1 \dots m,$
2. $\mathbf{R}_n \to \mathbf{R},\ n \to \infty,\ |\mathbf{R}| \neq 0,$

то оценка МНК состоятельна в среднеквадратическом.

Рассмотрим регрессию на время $y_t = at + b + \varepsilon_t,\ t = 1, 2, \dots, n.$

Утверждение. *Оценка МНК в регрессии на время $y_t = at + b + \varepsilon_t,\ t = 1, 2, \dots, n$ состоятельна в среднеквадратическом.*

Доказательство. Условие 1 теоремы 1.6, очевидно, выполнено. Пользуясь формулами

$$\sum_{t=1}^{n} t = \frac{n(n+1)}{2}, \quad \sum_{t=1}^{n} t^2 = \frac{n(n+1)(2n+1)}{6},$$

проверим выполнимость условия 2.

$$\mathbf{X}_n^{\mathrm{T}}\mathbf{X}_n = \begin{pmatrix} \frac{n(n+1)(2n+1)}{6} & \frac{n(n+1)}{2} \\ \frac{n(n+1)}{2} & n \end{pmatrix}.$$

$$\mathbf{R}_n = \begin{pmatrix} 1 & \frac{\sqrt{6}}{2}\sqrt{\frac{(n+1)}{2n+1}} \\ \frac{\sqrt{6}}{2}\sqrt{\frac{(n+1)}{2n+1}} & 1 \end{pmatrix} \to \begin{pmatrix} 1 & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & 1 \end{pmatrix} = \mathbf{R}, \quad |\mathbf{R}| \neq 0.$$

Таким образом, оценка МНК состоятельна в среднеквадратическом. □

1.2 Условия для асимптотической нормальности

Теорема (Теорема 1.9)**.** *Если $\{\varepsilon_i\}$ независимы и одинаково распределены, матрицы $\mathbf{X}_n$ сильно регулярны и $\forall j = 1, \dots, m\ \frac{1}{n}\max\limits_{i=1\dots n} x_{ij}^2 \to 0,\ n \to \infty$, тогда оценка МНК асимптотически нормальна, более того,*

$$\sqrt{n}(\hat{B}_{MHK}^{(n)} - B) \rightrightarrows N(0, \sigma^2\mathbf{A}^{-1}).$$

Как и ранее, эти условия не подходят для временных рядов. Для них есть след.теорема.

Теорема (Теорема 1.10)**.** *Если $\{\varepsilon_i\}$ независимы и одинаково распределены и выполнено:*
1. $\mathbf{R}_n \to \mathbf{R},\ n \to \infty,\ |\mathbf{R}| \neq 0,$
2. $\max\limits_i x_{ij}^2 / \sum\limits_i x_{ij}^2 \to 0,\ n \to \infty,\ \ \forall j = 1 \dots m,$

то оценка МНК асимптотически нормальна.

Утверждение. *Оценка МНК в регрессии на время $y_t = at + b + \varepsilon_t,\ t = 1, 2, \dots, n$ асимптотически нормальна.*

Доказательство. Условие 1 теоремы 1.10 уже было проверено. Рассмотрим условие 2:

$$j = 1, \quad \frac{\max\limits_i x_{ij}^2}{\sum\limits_i x_{ij}^2} = \frac{6n^2}{n(n+1)(2n+1)} \to 0,\ n \to \infty,$$

$$j = 2, \quad \frac{\max\limits_i x_{ij}^2}{\sum\limits_i x_{ij}^2} = \frac{1}{n} \to 0,\ n \to \infty.$$

Условия теоремы 1.10 выполнены, значит, оценка МНК асимптотически нормальна. □

Свойство асимптотической нормальности делает возможным при больших n и при определенных условиях на независимые переменные считать распределение оценки МНК приблизительно нормальным. Это позволяет строить доверительные интервалы и проверять гипотезы относительно параметров регрессии.

2 Случайные регрессоры, случай повторной выборки

2.1 Общий случай

В общем случае, когда регрессоры случайные (т.е. Y — случайный вектор, $\mathbf{X}$ — случайная матрица, условия записываются через условные мат.ожидания:

$$\mathbb{E}(Y|\mathbf{X}) = \mathbf{X}\mathring{B},$$

$$\operatorname{cov}(Y|\mathbf{X}) = \sigma^2\mathbf{I}.$$

Если определить $\bar{\varepsilon} = Y - \mathbf{X}\mathring{B}$, то получим обычные условия

$$\mathbb{E}(\bar{\varepsilon}|\mathbf{X}) = \mathbb{O},$$

$$\operatorname{cov}(\bar{\varepsilon}|\mathbf{X}) = \sigma^2\mathbf{I}.$$

Теорема. *В условиях, сформулированных выше, оценка МНК $\hat{B}$ является несмещенной и $\operatorname{cov}(\hat{B}) = \sigma^2\mathbb{E}(\mathbf{X}^\top\mathbf{X})^{-1}$.*

Другой подход, когда формулируется модель $Y^{(n)} = \mathbf{X}^{(n)}B + \varepsilon^{(n)}$ со случайной матрицей $\mathbf{X}^{(n)}$. Тогда получают такой общий (и слабый) результат:

Теорема. *Пусть существуют пределы (по вероятности):*

1. $$\operatorname*{plim}_{n\to\infty}\left[\frac{1}{n}\boldsymbol{X}^{(n)\top}\boldsymbol{X}^{(n)}\right] = \mathbf{A},$$

   причем $det(\mathbf{A}) \neq 0$ с вероятностью 1,

2. $\operatorname{plim}_{n\to\infty}\left[\frac{1}{n}\boldsymbol{X}^{(n)\top}\bar{\varepsilon}^{(n)}\right] = \mathbb{O}_m$.

Тогда

$$\operatorname*{plim}_{n\to\infty}\hat{B}^{(n)}_{MHK} = \mathring{B}$$

2.2 Случай повторной выборки

Пусть имеются $k+1$ случайных величин $\eta, \xi_1, \xi_2, ..., \xi_k$. Эти величины имеют свою функцию распределения, математические ожидания, дисперсии и т. д. (считаем, что они конечны). Обозначим

$$\mu_0 = \mathbb{E}(\eta), \mu_i = \mathbb{E}(\xi_i), i = 1, \ldots, k,$$

$$\sigma^2(\eta) = \sigma_0^2 = \mathbb{E}(\eta - \mu_0)^2 > 0,$$

$$\mathbf{C} = \operatorname{cov}(\boldsymbol{\xi}) = \mathbb{E}((\boldsymbol{\xi} - \mu)(\boldsymbol{\xi} - \mu)^\top)$$

— положительно определена и $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_k)^\top, \mu = (\mu_1, \mu_2, ..., \mu_k)^\top$. Относительно случайных величин η, $\boldsymbol{\xi}$ предположим, что

$$\mathbb{E}(\eta|\boldsymbol{\xi}) = \mathbb{E}(\eta|\xi_1, \xi_2, \ldots, \xi_k) = b_0 + b_1\xi_1 + b_2\xi_2 + ... + b_m\xi_m = B^\top\boldsymbol{\xi}, \tag{1}$$

где $B = (b_1, b_2, \ldots, b_k)^\top$ — вектор неизвестных коэффициентов. Таким образом, регрессия η на $\boldsymbol{\xi}$ линейна и неизвестна с точностью до своих коэффициентов. Предположим, что разброс регрессии постоянен, т. е.

$$\sigma^2(\eta|\boldsymbol{\xi}) = \mathbb{E}_{\boldsymbol{\xi}}(\mathbb{E}(\eta - \mathbb{E}(\eta|\boldsymbol{\xi}))^2|\boldsymbol{\xi})) = const = \sigma^2$$

и σ также неизвестно. Из случайной величины $(\eta, \boldsymbol{\xi}) \in \mathbb{R}^{k+1}$ производится случайная выборка $(y_t, \mathbf{x}_t)$, $t = 1, \ldots, n$, т. е. $(y_t, \mathbf{x}_t)$ независимы и одинаково распределены. Наблюдения образуют вектор $Y \in \mathbb{R}^n$ и матрицу $\mathbf{X}^{n\times m}$ (если добавить столбец из 1). В силу независимости $(y_t, \mathbf{x}_t)$ равенство (1) эквивалентно уравнению $\mathbb{E}(y|\mathbf{X}) = \mathbf{X}B$.

Теорема. *Оценка МНК в схеме случайно выборки асимптотически нормальна.*

$$\sqrt{n}(\hat{B}_{MHK} - B) \xrightarrow[n\to\infty]{} N(0, \sigma^2(\boldsymbol{C} + \mu\mu^\top)^{-1}),$$

где $\hat{b}$ *— оценка, полученная Методом Наименьших Квадратов.*

Теорема. *В схеме случайной выборки из нормального распределения оценки МНК и ММП(MLE) совпадают.*