Микроэконометрика

Модели с неслучайным отбором

Потанин Богдан Станиславович

старший преподаватель, кандидат экономических наук

2021-2022

Неслучайный отбор

Мотивация

- Иногда значение зависимой переменной наблюдается лишь при соблюдении некоторого условия (правила).
- Например, зарплата наблюдается лишь для работающих индивидов, а затраты на покупку в приложении лишь для тех, кто им пользуется.
- В отличие от модели Тобина модели с неслучайным отбором предполагают, что правило, определяющее попадение наблюдений в выборку, моделируется отдельно.
- Например, в моделях с неслучайным отбором одновременно моделируется как зарплата, так и занятость индивида. Причем различные факторы могут по разному (в том числе с разным знаком) влиять на ожидаемую зарплату и вероятность занятости.

Усеченное двумерное нормальное распределение

Частный случай с односторонним усечением одной компоненты

- Случайные величины X_1 и X_2 имеют совместное нормальное распределение:

$$(X_1, X_2) \sim \mathcal{N}\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}\right)$$

- Напомним, что:

$$E(X_1|X_2 = t) = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(t - \mu_2), \quad Var(X_1|X_2 = t) = (1 - \rho^2)\sigma_1^2$$

- При усечении X_2 получаем:

$$E\left(X_1|X_2 > t\right) = \mu_1 + \rho\sigma_1\sigma_2\lambda(t^*), \quad E\left(X_1|X_2 < t\right) = \mu_1 - \rho\sigma_1\sigma_2\lambda(-t^*)$$

$$Var\left(X_1|X_2 > t\right) = \sigma_1^2\left(1 - \rho^2\delta(t^*)\right), \quad Var\left(X_1|X_2 < t\right) = \sigma_1^2\left(1 - \rho^2\delta(-t^*)\right)$$

$$\lambda(t^*) = \frac{\phi\left(t^*\right)}{1 - \Phi\left(t^*\right)}, \quad \delta(t^*) = \lambda(t^*)(\lambda(t^*) - t^*), \quad t^* = \frac{t - \mu_2}{\sigma_2}$$

- Имеется два уравнения:

$$\text{Целевое уравнение: } y_i^* = x_i\beta + \varepsilon_i$$
$$\text{Уравнение отбора: } z_i^* = w_i\gamma + u_i$$

- Значение y_i^* наблюдается лишь при соблюдении определенного условия (правила):

$$z_i = \begin{cases} 1, \text{ если } z_i^* \geq 0 \\ 0, \text{ в противном случае} \end{cases} \qquad y_i = \begin{cases} y_i^*, \text{ если } z_i = 1 \\ \text{не наблюдаем }, \text{ в противном случае} \end{cases}$$

- Например, y_i^* может отражать потенциальную заработную плату индивида, которая наблюдается лишь в случае, когда индивид работает, то есть $z_i = 1$.
- Для простоты предположим, что случайные ошибки имеют совместное нормальное распределение:

$$(u_i, \varepsilon_i) \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{bmatrix}\right), \text{i.i.d.}$$

- Поскольку y_i наблюдается лишь при $z_i = 1$, то:

$$E\left(y_i|w_i, x_i\right) = E\left(y_i^*|z_i = 1, w_i, x_i\right) = x_i\beta + E\left(\varepsilon_i|u_i \geq -w_i\gamma, w_i, x_i\right),$$

где по свойствам усеченного двумерного нормального распределения:

$$E\left(\varepsilon_i|u_i \geq -w_i\gamma, w_i, x_i\right) = \frac{\phi\left(w_i\gamma\right)}{\Phi\left(w_i\gamma\right)} = \rho\sigma\lambda_i(w_i\gamma) = \rho\sigma\lambda_i,$$

- Отметим, что λ_i часто именуют лямбдой Хекмана.
- В результате регрессионное уравнение может быть записано как:

$$y_i = x_i\beta + \rho\sigma\lambda_i + v_i, \qquad v_i = \varepsilon_i - \rho\sigma\lambda_i \implies E(v_i|x_i, w_i) = 0$$

- Без учета λ_i при $\rho \neq 0$ и наличии корреляции между λ_i и x_i МНК оценки коэффициентов β окажутся несостоятельными вследствие проблемы пропущенной переменной.

- **Идея метода**: истинное значение λ_i исследователю неизвестно, поскольку зависит от неизвестных параметров γ. Однако, оценив параметры γ можно получить состоятельную оценку $\hat{\lambda}_i$ и использовать ее вместо λ_i для того, чтобы избежать смещения в оценках вследствие пропущенной переменной.
- Двухшаговая процедура оценивания:
  - **Первый шаг**: при помощи пробит модели оцениваются параметры γ. В силу инвариантности ММП оценок состоятельная оценка λ_i рассчитывается как $\hat{\lambda}_i = \lambda_i\left(w_i\hat{\gamma}\right)$.
  - **Второй шаг**: В регрессионное уравнение для y_i подставляется $\hat{\lambda}_i$ в качестве дополнительного регрессора с коэффициентом $\rho\sigma$. Затем β и $\rho\sigma$ оцениваются при помощи МНК.
- Состоятельные оценки $\hat{\sigma}^2$ и $\hat{\rho}$ можно получить как:

$$\hat{\sigma}^2 = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 + \left(\hat{\rho\sigma}\right)^2\hat{\lambda}_i\left(\lambda_i - w_i\gamma\right), \qquad \hat{\rho} = \frac{\widehat{\rho\sigma}}{\sqrt{\hat{\sigma}^2}}$$

Метод Хекмана: двухшаговая процедура

Тестирование гипотез

- Случайные ошибки гетероскедастичны, поскольку по свойствам усеченного двумерного нормального распределения:

$$E\left(v_i^2|z_i=1,w_i,x_i\right)=\sigma^2\left(1-\rho^2\delta_i\right),\qquad \delta_i=\lambda_i\left(\lambda_i+w_i\gamma\right)$$

- Для коррекции ковариационной матрицы оценок регрессионных коэффициентов необходимо учесть как гетероскедастичность, так и то, что вместо истинного значения λ_i используется его оценка, зависящая от $\hat{\gamma}$, откуда:

$$\widehat{\text{As.Cov}}\left(\hat{\beta}^*\right)=\hat{\sigma}^2\left(X_*^TX_*\right)^{-1}\left(\hat{A}_1+\hat{A}_2\right)\left(X_*^TX_*\right)^{-1}$$

$$\hat{A}_1=X_*^T\left(I-\hat{\rho}^2\widehat{\triangle}\right)X_*,\quad \hat{A}_2=\hat{\rho}^2\left(X_*\widehat{\triangle}W\right)\widehat{\text{As.Cov}}\left(\hat{\gamma}\right)\left(X_*\widehat{\triangle}W\right)^T,$$

где $\beta_*=(\beta,\rho\sigma)^T$ и X_* является матрицей регрессоров, полученной за счет присоединения столбца $\hat{\lambda}$ к матрице X справа. Также, $\widehat{\triangle}$ – диагональная матрица, такая, что $\widehat{\triangle}_i=1-\rho^2\delta_i$. Элементы $\hat{A}_1$ и $\hat{A}_2$ позволяют учесть гетероскедастичность и использование оценок λ_i вместо истинных значений соответственно.

Ограничения исключения (exclusion restrictions)

- Лямбда Хекмана $\lambda(w_i\gamma)$ крайне близка к линейной функции при $w_i\gamma \in (-\infty, 2)$, то есть в данном диапазоне $\lambda(w_i\gamma) \approx cw_i\gamma$, где $c \in R_{++}$.
- Из-за этого при сильном сходстве между x_i и w_i может возникнуть сильная коллинеарность между $\lambda(w_i\gamma)$ и x_i.
- Эта коллинеарность часто приводит к существенному снижению в эффективности оценок.
- Для смягчения проблемы коллинеарности, как правило, исследователи пытаются обеспечить наличие **ограничений исключения**: регрессоров, входящих в w_i и не входящих в x_i.
- Например, исследователь может предположить, что количество детей влияет на вероятность занятости (входит в w_i), но не влияет на зарплату индивида (не входит в x_i).
- В качестве более устойчивой к отсутствию ограничений исключения альтернативы вместо двухшаговой процедуры можно воспользоваться методом максимального правдоподобия.

Метод Хекмана: метод максимального правдоподобия

Оценивание

- Оценки параметров β, ρ и σ можно также получить за счет максимизации функции правдоподобия:

$$L(\beta, \rho, \sigma; y, z|X, W) = \prod_{i:z_i=1} f_{y_i|x_i,w_i}(y_i)P(z_i = 1|y_i, x_i, w_i) \prod_{i:z_i=0} P(z_i = 0|x_i, w_i) =$$

$$= \prod_{i:z_i=1} \frac{1}{\sigma}\phi\left(\frac{y_i - x_i\beta}{\sigma}\right)\Phi\left(\frac{\rho\left(y_i - x_i\beta\right)/\sigma + w_i\gamma}{\sqrt{1-\rho^2}}\right)\prod_{i:z_i=0}\left(1 - \Phi\left(w_i\gamma\right)\right)$$

- Оценки ММП метода более эффективны, чем оценки двухшаговой процедуры.
- Недостаток ММП заключается в сложности технической реализации, связанной с возможностью наличия несколько локальных максимумов функции правдоподобия.
- Для тестирования гипотезы о наличии неслучайного отбора достаточно проверить $H_0 : \rho = 0$ для ММП или $H_0 : \rho\sigma = 0$ для двухшаговой процедуры. Если нулевая гипотеза отвергается, то МНК оценки несостоятельны, что мотивирует применение метода Хекмана.

- Предельный эффект переменной x_{ik} на обычное математическое ожидание имеет такой же вид, как в случае с обычной линейной регрессией:

$$\frac{\partial E\left(y_i^*|x_i\right)}{\partial x_{ik}} = \beta_k$$

- Предельный эффект на условное математическое ожидание рассчитывается как:

$$\frac{\partial E\left(y_i^*|z_i = 1\right)}{\partial x_{ik}} = \beta_k - \gamma_* \rho\sigma\delta_i,$$

  где γ_* является коэффициентом при x_{ki} в уравнении отбора, если x_{ki} входит в w_i. В противном случае $\gamma_* = 0$.

- Предельный эффект на условное математическое ожидание целевой переменной складывается из предельного эффекта на безусловное математическое ожидание β_k и части, обусловленной наличием неслучайного отбора $\gamma_* \rho\sigma\delta_i$.

Метод Ньюи

Оценивание

- При нарушении допущения о совместном нормальном распределении случайных ошибок оценки метода Хекмана могут оказаться несостоятельными.
- В качестве альтернативы допущению о конкретной форме совместного распределения случайных ошибок условное математическое ожидание случайной ошибки основного уравнения можно аппроксимировать при помощи полинома k-й степени:

$$E\left(\varepsilon_i|z_i=1,w_i,x_i\right)\approx\sum_{t=0}^{k}\tau_t g(w_i\gamma)^t,\quad \tau=(\tau_1,...,\tau_k),$$

  где $g(w_i\gamma)$ является произвольно выбираемой сглаживающей функцией, в качестве которой, как правило, рассматривают $g(w_i\gamma)=w_i\gamma$ или $g(w_i\gamma)=\lambda(w_i\gamma)$.
- На **первом шаге** параметры γ оцениваются при помощи полупараметрической модели бинарного выбора (например, метода Галланта и Ннчки), а на **втором шаге** все k переменных $g(w_i\hat{\gamma})^t$ подставляются в целевое уравнение в качестве регрессоров (в дополнении к x_i), в котором параметры β и τ оцениваются с помощью МНК.
- Оценки данного метода состоятельные и асимптотически нормальные. Для тестирования гипотез и оценивания асимптотической ковариационной матрицы оценок регрессионных коэффициентов, как правило, применяют бутстрап.

Метод Ньюи

Кросс-валидация

- По мере увеличения степени полинома k растет точность аппроксимации, что позволяет снизить смещение оценок. Однако, вместе с ростом k увеличивается и число оцениваемых параметров, а также усугубляется проблема коллинеарности, что приводит к росту дисперсии оценок.
- Оптимальная степень полинома k, как правило, подбирается с помощью leave-one-out кросс-валидации.
- Создается n (объем исходной выборки) выборок объема $n-1$, каждая из которых формируется за счет исключения из исходной выборки одного (каждый раз разного) наблюдения.
- На каждой из этих выборок при заданном k методом Ньюи оцениваются параметры модели, а затем с их помощью предсказывается $\hat{y}_i$ – значение исключенного из выборки наблюдения y_i.
- Рассчитывается $\text{RMSE}_k = \sqrt{\frac{1}{n}\sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2}$ мера качества модели при данной степени полинома.
- К счастью существует аналитическая формула, позволяющая рассчитать RMSE_k без необходимости n раз оценивать параметры модели.
- Выбирается степень k с наименьшим RMSE_k.
- При использовании бутстрапа кросс-валидацию необходимо проводить каждую итерацию.

- Помимо метода Ньюи существуют и иные подходы к ослаблению допущения о совместном нормальном распределении случайных ошибок в моделях с неслучайным отбором. Например, можно воспользоваться методом Галланта и Нички для аппроксимации соответствующего совместного распределения и получить оценки за счет максимизации функции квази-правдоподобия.
- Во многих исследованиях рассматриваются альтернативные механизмы неслучайного отбора наблюдений. Например, в качестве уравнения отбора можно использовать мультиномиальную логит модель или порядковую пробит модель. Также, рассматриваются модели с несколькими правилами отбора, когда, например, наблюдения по зарплате доступны лишь для работающих индивидов (первое правило), согласившихся ответить на вопрос о зарплате (второе правило).