Микроэконометрика

Модели для частотных данных

Потанин Богдан Станиславович

старший преподаватель, кандидат экономических наук

2021-2022

- Часто в исследованиях необходимо моделировать данные, отражающие частоту наступления некоторого события.

- Часто в исследованиях необходимо моделировать данные, отражающие частоту наступления некоторого события.
- Например, исследователь может моделировать частоту посещения индивидами врача, спортивного зала или приложения.

- Часто в исследованиях необходимо моделировать данные, отражающие частоту наступления некоторого события.
- Например, исследователь может моделировать частоту посещения индивидами врача, спортивного зала или приложения.
- Для анализа таких данных используются специальные модели, за основу которых часто берутся различные дискретные распределения, такие как распределение Пуассона.

- Предположим, что зависимая переменная y_i имеет распределение Пуассона с параметром λ_i, определяемым значениями регрессоров x_i:

$$P\left(y_i = k|x_i\right) = \frac{\lambda_i^k}{k!}e^{-\lambda_i}, \qquad E(y_i|x_i) = Var(y_i|x_i) = \lambda_i = e^{x_i\beta}$$

- Предположим, что зависимая переменная y_i имеет распределение Пуассона с параметром λ_i, определяемым значениями регрессоров x_i:

$$P\left(y_i = k|x_i\right) = \frac{\lambda_i^k}{k!}e^{-\lambda_i}, \qquad E(y_i|x_i) = Var(y_i|x_i) = \lambda_i = e^{x_i\beta}$$

- Например, y_i может отражать частоту посещения врача, а x_i – характеристики индивида, такие как доход, возраст, наличие заболеваний и т.д.

Пуассоновская регрессия

Формулировка, оценивание и интерпретация

- Предположим, что зависимая переменная y_i имеет распределение Пуассона с параметром λ_i, определяемым значениями регрессоров x_i:

$$P\left(y_i = k|x_i\right) = \frac{\lambda_i^k}{k!}e^{-\lambda_i}, \qquad E(y_i|x_i) = Var(y_i|x_i) = \lambda_i = e^{x_i\beta}$$

- Например, y_i может отражать частоту посещения врача, а x_i – характеристики индивида, такие как доход, возраст, наличие заболеваний и т.д.
- Коэффициенты β оцениваются при помощи метода максимального правдоподобия:

$$\ln L\left(\beta; y|X\right) = \sum_{i=1}^{n} y_i \ln\left(\lambda_i\right) - \lambda_i - y_i! = \sum_{i=1}^{n} y_i x_i\beta - e^{x_i\beta} - y_i!$$

- Предположим, что зависимая переменная y_i имеет распределение Пуассона с параметром λ_i, определяемым значениями регрессоров x_i:

$$P\left(y_i = k|x_i\right) = \frac{\lambda_i^k}{k!}e^{-\lambda_i}, \qquad E(y_i|x_i) = Var(y_i|x_i) = \lambda_i = e^{x_i\beta}$$

- Например, y_i может отражать частоту посещения врача, а x_i – характеристики индивида, такие как доход, возраст, наличие заболеваний и т.д.
- Коэффициенты β оцениваются при помощи метода максимального правдоподобия:

$$\ln L\left(\beta; y|X\right) = \sum_{i=1}^{n} y_i \ln\left(\lambda_i\right) - \lambda_i - y_i! = \sum_{i=1}^{n} y_i x_i\beta - e^{x_i\beta} - y_i!$$

- Знак предельного эффекта регрессора x_{ik} на ожидаемую частоту наступления события совпадает со знаком β_k, но величина эффекта зависит от λ_i:

$$\frac{\partial E(y_i|x_i)}{\partial x_{ik}} = \lambda_i\beta_k$$

Отрицательная биномиальная регрессия

Формулировка

- Нереалистичное допущение о равенстве математического ожидания и дисперсии (overdispersion) в распределении Пуассона мотивировало рассмотрение альтернативных моделей.

Отрицательная биномиальная регрессия

Формулировка

- Нереалистичное допущение о равенстве математического ожидания и дисперсии (overdispersion) в распределении Пуассона мотивировало рассмотрение альтернативных моделей.
- Наибольшую популярность приобрела отрицательная биномиальная регрессия, которая отличается от Пуассоновской регрессии включением случайной составляющей:

$$\lambda_i = e^{x_i\beta+\varepsilon_i}, \qquad e^{\varepsilon_i}|x_i \sim \Gamma\left(1/\theta, \theta\right) \text{ i.i.d.},$$
$$E(\varepsilon_i|x_i) = 1, \qquad Var(\varepsilon_i) = \theta$$

где ε_i отражает не наблюдаемые характеристики или гетерогенность в распределении ожидаемой частоты наступления события.

- Нереалистичное допущение о равенстве математического ожидания и дисперсии (overdispersion) в распределении Пуассона мотивировало рассмотрение альтернативных моделей.
- Наибольшую популярность приобрела отрицательная биномиальная регрессия, которая отличается от Пуассоновской регрессии включением случайной составляющей:

$$\lambda_i = e^{x_i\beta+\varepsilon_i}, \qquad e^{\varepsilon_i}|x_i \sim \Gamma\left(1/\theta, \theta\right) \text{ i.i.d.},$$
$$E(\varepsilon_i|x_i) = 1, \qquad Var(\varepsilon_i) = \theta$$

  где ε_i отражает не наблюдаемые характеристики или гетерогенность в распределении ожидаемой частоты наступления события.
- При нулевой дисперсии ε_i данная модель сводится к Пуассоновской регрессии. Поэтому, если $H_0 : \theta = 0$ не отвергается, то можно применять Пуассоновскую регрессию, вместо отрицательной биномиальной (тест на overdispersion).

Отрицательная биномиальная регрессия

Формулировка

- Нереалистичное допущение о равенстве математического ожидания и дисперсии (overdispersion) в распределении Пуассона мотивировало рассмотрение альтернативных моделей.
- Наибольшую популярность приобрела отрицательная биномиальная регрессия, которая отличается от Пуассоновской регрессии включением случайной составляющей:

$$\lambda_i = e^{x_i\beta+\varepsilon_i}, \qquad e^{\varepsilon_i}|x_i \sim \Gamma\left(1/\theta, \theta\right) \text{ i.i.d.},$$
$$E(\varepsilon_i|x_i) = 1, \qquad Var(\varepsilon_i) = \theta$$

  где ε_i отражает не наблюдаемые характеристики или гетерогенность в распределении ожидаемой частоты наступления события.
- При нулевой дисперсии ε_i данная модель сводится к Пуассоновской регрессии. Поэтому, если $H_0 : \theta = 0$ не отвергается, то можно применять Пуассоновскую регрессию, вместо отрицательной биномиальной (тест на overdispersion).
- Можно показать, что $E(y_i|x_i) = e^{x_i\beta}$, поэтому предельные эффекты на ожидаемую частоту наступления события считаются по аналогии с Пуассоновской регрессией.

Модель с инфляцией нулей (Zero-inflated model)

Формулировка

- Иногда зависимая переменная может с большой вероятностью принимать нулевое значение.

Модель с инфляцией нулей (Zero-inflated model)

Формулировка

- Иногда зависимая переменная может с большой вероятностью принимать нулевое значение.
- Например, при изучении факторов, влияющих на частоту полетов на самолете, могут часто встречаться люди, которые ни разу за рассматриваемый период не пользовались этим видом транспорта.

Модель с инфляцией нулей (Zero-inflated model)

Формулировка

- Иногда зависимая переменная может с большой вероятностью принимать нулевое значение.
- Например, при изучении факторов, влияющих на частоту полетов на самолете, могут часто встречаться люди, которые ни разу за рассматриваемый период не пользовались этим видом транспорта.
- Предполагается, что существует отдельный процесс, определяющий, будет ли значение нулевым или может принимать иные значения. Этот процесс задается моделью бинарного выбора.

$$z_i^* = w_i\gamma + u_i, \qquad z_i = \begin{cases} 1, \text{ если } z_i^* \geq 0 \\ 0, \text{ в противном случае} \end{cases} \qquad u_i \sim \mathcal{N}(0,1) \text{ i.i.d. (пробит модель)}$$

Модель с инфляцией нулей (Zero-inflated model)

Формулировка

- Иногда зависимая переменная может с большой вероятностью принимать нулевое значение.
- Например, при изучении факторов, влияющих на частоту полетов на самолете, могут часто встречаться люди, которые ни разу за рассматриваемый период не пользовались этим видом транспорта.
- Предполагается, что существует отдельный процесс, определяющий, будет ли значение нулевым или может принимать иные значения. Этот процесс задается моделью бинарного выбора.

$$z_i^* = w_i\gamma + u_i, \qquad z_i = \begin{cases} 1, \text{ если } z_i^* \geq 0 \\ 0, \text{ в противном случае} \end{cases} \qquad u_i \sim \mathcal{N}(0,1) \text{ i.i.d. (пробит модель)}$$

- Предположим, что в Пуассоновской регрессии (по аналогии для отрицательной биномиальной) наблюдается частота наступления события лишь при $z_i = 1$:

$$y_i^* \sim Pois(\lambda_i), \qquad y_i = \begin{cases} y_i^*, \text{ если } z_i = 1 \\ 0, \text{ в противном случае} \end{cases}$$

Модель с инфляцией нулей (Zero-inflated model)

Формулировка

- Иногда зависимая переменная может с большой вероятностью принимать нулевое значение.
- Например, при изучении факторов, влияющих на частоту полетов на самолете, могут часто встречаться люди, которые ни разу за рассматриваемый период не пользовались этим видом транспорта.
- Предполагается, что существует отдельный процесс, определяющий, будет ли значение нулевым или может принимать иные значения. Этот процесс задается моделью бинарного выбора.

$$z_i^* = w_i\gamma + u_i, \qquad z_i = \begin{cases} 1, \text{ если } z_i^* \geq 0 \\ 0, \text{ в противном случае} \end{cases} \qquad u_i \sim \mathcal{N}(0,1) \text{ i.i.d. (пробит модель)}$$

- Предположим, что в Пуассоновской регрессии (по аналогии для отрицательной биномиальной) наблюдается частота наступления события лишь при $z_i = 1$:

$$y_i^* \sim Pois(\lambda_i), \qquad y_i = \begin{cases} y_i^*, \text{ если } z_i = 1 \\ 0, \text{ в противном случае} \end{cases}$$

$$P(y_i = 0|x_i, w_i) = P(z_i = 0|x_i, w_i) + P(y_i = 0|z_i = 1, x_i, w_i)P(z_i = 1|x_i, w_i) = 1 - \Phi(w_i\gamma) + e^{-\lambda_i}\Phi(w_i\gamma)$$

- Иногда зависимая переменная может с большой вероятностью принимать нулевое значение.
- Например, при изучении факторов, влияющих на частоту полетов на самолете, могут часто встречаться люди, которые ни разу за рассматриваемый период не пользовались этим видом транспорта.
- Предполагается, что существует отдельный процесс, определяющий, будет ли значение нулевым или может принимать иные значения. Этот процесс задается моделью бинарного выбора.

$$z_i^* = w_i\gamma + u_i, \qquad z_i = \begin{cases} 1, \text{ если } z_i^* \geq 0 \\ 0, \text{ в противном случае} \end{cases} \qquad u_i \sim \mathcal{N}(0,1) \text{ i.i.d. (пробит модель)}$$

- Предположим, что в Пуассоновской регрессии (по аналогии для отрицательной биномиальной) наблюдается частота наступления события лишь при $z_i = 1$:

$$y_i^* \sim Pois(\lambda_i), \qquad y_i = \begin{cases} y_i^*, \text{ если } z_i = 1 \\ 0, \text{ в противном случае} \end{cases}$$

$$P(y_i = 0|x_i, w_i) = P(z_i = 0|x_i, w_i) + P(y_i = 0|z_i = 1, x_i, w_i)P(z_i = 1|x_i, w_i) = 1 - \Phi(w_i\gamma) + e^{-\lambda_i}\Phi(w_i\gamma)$$

$$P(y_i = k|x_i, w_i) = P(y_i = k|z_i = 1, x_i, w_i)P(z_i = 1|x_i, w_i) = \frac{\lambda_i^k}{k!}e^{-\lambda_i}\Phi(w_i\gamma), \text{ где } k \geq 1$$

- Ожидаемая частота наступления события будет иметь вид:

$$E(y_i|x_i,w_i)=E(y_i|z_i=1,x_i,w_i)P(z_i=1|x_i,w_i)+\underbrace{E(y_i|z_i=0,x_i,w_i)}_{\text{равно нулю}}P(z_i=0|x_i,w_i)=$$

$$=E(y_i|z_i=1,x_i,w_i)P(z_i=1|x_i,w_i)=\lambda_i\Phi\left(w_i\gamma\right)$$

- Ожидаемая частота наступления события будет иметь вид:

$$E(y_i|x_i,w_i) = E(y_i|z_i=1,x_i,w_i)P(z_i=1|x_i,w_i) + \underbrace{E(y_i|z_i=0,x_i,w_i)}_{\text{равно нулю}}P(z_i=0|x_i,w_i) =$$
$$= E(y_i|z_i=1,x_i,w_i)P(z_i=1|x_i,w_i) = \lambda_i\Phi\left(w_i\gamma\right)$$

- Отсюда нетрудно получить выражения для предельных эффектов на ожидаемую частоту. Рассмотрим переменную v_i, которая входит в x_i и w_i с коэффициентами β_v и γ_v соответственно:

$$\frac{\partial E(y_i|x_i,w_i)}{v_i} = \lambda_i\Phi\left(w_i\gamma\right)\beta_v + \lambda_i\phi\left(w_i\gamma\right)\gamma_v =$$
$$= \Phi\left(w_i\gamma\right)\frac{\partial E(y_i|z_i=1,x_i,w_i)}{\partial v_i} + \lambda_i\frac{\partial P(z_i=1|x_i,w_i)}{\partial v_i}$$

Модель с инфляцией нулей (Zero-inflated model)

Предельные эффекты

- Ожидаемая частота наступления события будет иметь вид:

$$E(y_i|x_i,w_i)=E(y_i|z_i=1,x_i,w_i)P(z_i=1|x_i,w_i)+\underbrace{E(y_i|z_i=0,x_i,w_i)}_{\text{равно нулю}}P(z_i=0|x_i,w_i)=$$

$$=E(y_i|z_i=1,x_i,w_i)P(z_i=1|x_i,w_i)=\lambda_i\Phi\left(w_i\gamma\right)$$

- Отсюда нетрудно получить выражения для предельных эффектов на ожидаемую частоту. Рассмотрим переменную v_i, которая входит в x_i и w_i с коэффициентами β_v и γ_v соответственно:

$$\frac{\partial E(y_i|x_i,w_i)}{v_i}=\lambda_i\Phi\left(w_i\gamma\right)\beta_v+\lambda_i\phi\left(w_i\gamma\right)\gamma_v=$$

$$=\Phi\left(w_i\gamma\right)\frac{\partial E(y_i|z_i=1,x_i,w_i)}{\partial v_i}+\lambda_i\frac{\partial P(z_i=1|x_i,w_i)}{\partial v_i}$$

- Таким образом данная переменная влияет на ожидаемую частоту как через условное математическое ожидание, так и через вероятность попадания в ненулевой режим.

Модель с инфляцией нулей (Zero-inflated model)

Дополнительные примечания

- Иногда наличие отдельного механизма, определяющего возникновением нулевых значений, интерпретируют как существование латентного класса.

Модель с инфляцией нулей (Zero-inflated model)

Дополнительные примечания

- Иногда наличие отдельного механизма, определяющего возникновением нулевых значений, интерпретируют как существование латентного класса.
- Например, могут существовать люди, которые вообще не пользуются авиатранспортом и лишь при переходе этих людей в категорию тех, кто все же пользуется авиатранспортом, мы потенциально сможем наблюдать их частоту.

Модель с инфляцией нулей (Zero-inflated model)

Дополнительные примечания

- Иногда наличие отдельного механизма, определяющего возникновением нулевых значений, интерпретируют как существование латентного класса.
- Например, могут существовать люди, которые вообще не пользуются авиатранспортом и лишь при переходе этих людей в категорию тех, кто все же пользуется авиатранспортом, мы потенциально сможем наблюдать их частоту.
- Часто бинарное уравнение в данной модели формулируют обратным образом, то есть меняя единицы и нули местами. При этом вместо пробит модели можно применять любую иную модель бинарного выбора.

Модель с инфляцией нулей (Zero-inflated model)

Дополнительные примечания

- Иногда наличие отдельного механизма, определяющего возникновением нулевых значений, интерпретируют как существование латентного класса.
- Например, могут существовать люди, которые вообще не пользуются авиатранспортом и лишь при переходе этих людей в категорию тех, кто все же пользуется авиатранспортом, мы потенциально сможем наблюдать их частоту.
- Часто бинарное уравнение в данной модели формулируют обратным образом, то есть меняя единицы и нули местами. При этом вместо пробит модели можно применять любую иную модель бинарного выбора.
- Для проверки гипотезы о наличии инфляции нулей можно применить тест Вьонга (Vuong test).