Московский государственный университет имени М. В. Ломоносова
Факультет космических исследований
Кафедра математической статистики и случайных процессов

Курсовая работа
студента 402 группы
Шигина Глеба Сергеевича

Трансформация данных в GLM моделях

Научный руководитель:
с.н.с., к.ф.-м.н.
Шкляев Александр Викторович

Москва, 2022

Содержание

1 Введение **2**

2 Трансформация данных **2**

2.1 Трансформация только зависимой переменной при помощи inverse response plot 2

2.1.1 Inverse response plot 3

2.1.2 Выбор степенного преобразования 4

2.2 Трансформация только зависимой переменной при помощи преобразования Бокса-Кокса 5

3 Заключение **5**

1 Введение

TBA

2 Трансформация данных

Существуют классы задач, для которых мы знаем, что матожидание $\mathbf{E}(Y|X)$ является линейной функцией от X. Это может быть теория, подкрепленная экспериментальными данными [Weisberg et al., 1978], либо статистическая информация о данных. Например, пусть y_i и x_i – выборки их нормальных распределенний со средними μ_X, μ_Y соответственно, дисперсиями σ_X, σ_Y соответственно и корреляцией ρ_{XY}. Тогда можно показать (см. [Berger and Casella, 2001, стр. 550]), что

$$y_i \mid x_i \sim N\left(\mu_Y - \rho_{XY}\frac{\sigma_Y}{\sigma_X}\mu_X + \rho_{XY}\frac{\sigma_Y}{\sigma_X}x_i, \sigma_Y(1-\rho_{XY}^2)\right). \tag{1}$$

Это можно переписать как $y_i \mid x_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$, где

$$\beta_0 = \mu_Y - \rho_{XY}\frac{\sigma_Y}{\sigma_X}\mu_X, \beta_1 = \rho_{XY}\frac{\sigma_Y}{\sigma_X}, \sigma^2 = \mathbf{D}(Y|X) = \sigma_Y^2(1-\rho_{XY}), \tag{2}$$

то есть мы получили линейную регрессию Y по X:

$$\mathbf{E}(Y|X = x_i) = \beta_0 + \beta_1 x_i. \tag{3}$$

Однако, у нас не всегда есть возможность узнать истинную зависимость целевой переменной от предикторов, и любое преобразование данных, которое мы используем, не более чем приближение, которое, как мы надеемся, является подходящим для рассматриваемой задачи. Это поднимает два важных вопроса: еак выбрать преобразование и подходит ли полученная модель к имеющимся данных?

Для удобства на данном этапе ограничимся одним **предиктором** X и **зависимой переменной** Y.

2.1 Трансформация только зависимой переменной при помощи inverse response plot

Предположим, что истинная регрессионная модель Y по X имеет вид:

$$Y = g(\beta_0 + \beta_1 X + \varepsilon), \tag{4}$$

где g – некоторая функция, вообще говоря, нам неизвестная. Модель (4) может быть приведена к линейному виду, преобразовав Y с помощью обратной функции g^{-1}:

$$g^{-1}(Y) = \beta_0 + \beta_1 X + \varepsilon. \tag{5}$$

Например, если $Y = \log(\beta_0 + \beta_1 X + \varepsilon)$, то $g(x) = \log(x)$, значит $g^{-1}(x) = \exp(x)$, и $\exp(Y) = \beta_0 + \beta_1 X + \varepsilon$.

В данной работе в качестве способа получения оценки g^{-1} предлагается рассмотреть метод обратного отклика (*inverse response plot*) с подбором функции из степенного семейства (*power family*), включающим в себя семейство преобразований Бокса-Кокса.

2.1.1 Inverse response plot

В работе [Cook and Weisberg, 1994] было показано, что если x имеет эллиптически симметричное распределение (что является менее жетским ограничением, чем нормальность), то g^{-1} можно оценить из точечного графика, где по горизонтальной оси откладываются истинные значения y, в по вертикальной – значения $\hat{y}$, полученные из регрессии на нетрансформированных данных: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Такой график называют **inverse response plot**.

Пример

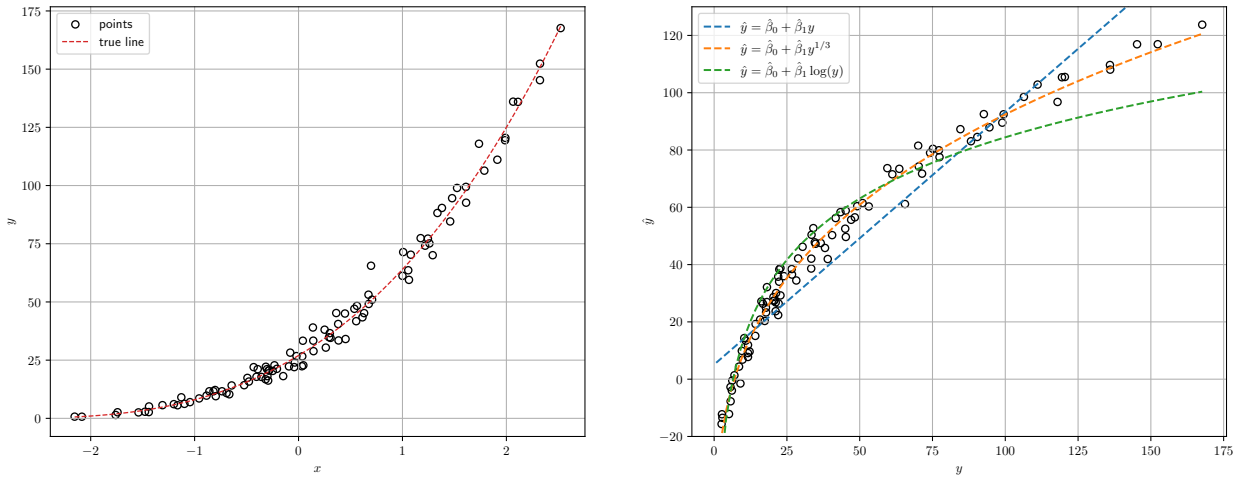

Рис. 1: Scatter plot Y от X (слева) и inverse response plot (справа)

Пусть $X \sim N(0,1)$, $\varepsilon \sim N(0,0.1)$, $Y = (3 + X + \varepsilon)^3$. Была построена выборка размера $N = 100$. Получившуюся зависимость можно увидеть на Рис.1. Построим линейную регрессию y по x без преобразования данных. Получим некоторые оценки $\hat{y}$. Inverse response plot можно также увидеть на Рис.1. Помимо точек, на графике присутствуют три пунктирные кривые – они показывают результаты линейных регрессий $\hat{y}$ по y, $y^{1/3}$ и $\log(y)$ соответсвенно. Видно, что кривая $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 y^{1/3}$ наиболее близка к исходным точкам. Это ожидаемо, так как из построения Y следует, что искомое нами преобразование для Y имеет вид $g^{-1}(Y) = Y^{1/3}$.

2.1.2 Выбор степенного преобразования

Семейство преобразований (*transformation family*) – это параметризованное множество преобразований, где, варьируя параметры, мы можем получить любого представителя семейства.

Одним из таких семейств является **степенное семейство (*power family*)**. Оно определено для положительных X и имеет вид:

$$\psi(Y, \lambda) = \begin{cases} Y^\lambda, & \lambda \neq 0 \\ \log(Y), & \lambda = 0 \end{cases}. \tag{6}$$

Семейство параметризовано числом λ, и значение $\lambda = 0$ принимается не за тождественную единицу (ведь $Y^0 \equiv 1$), а за логарифмическое преобразование $\log(Y)$. Здесь возникает проблема, затрудняющая работу с этим семейством – его представители $\psi(Y, \lambda)$ не являются непрерывным по λ. Поэтому удобнее работать с так называемым **нормированным (или отмастабированным) степенным семейством (*scaled power family*)**:

$$\psi_S(Y, \lambda) = \begin{cases} \dfrac{Y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(Y), & \lambda = 0 \end{cases}. \tag{7}$$

Несложно видеть, что $\psi_S(Y, \lambda)$ и $\psi(Y, \lambda)$ отличаются только преобразованием сдвига и масштаба, и линейная регресиия будет давать аналогичные результаты при двух этих преобразованиях, отличатся будут только веса предикторов. Также, функция $\psi_S(Y, \lambda)$ непрерывна по λ, и $\log(Y)$ является естественным представителем семейства, так как $\lim_{\lambda \to 0} \psi_S(Y, \lambda) = \log(Y)$.

Суммируя все вышесказанное, для нахождения оценки g^{-1} мы рассматриваем модели вида

$$\mathbf{E}(\hat{y}|Y = y) = \hat{\beta}_0 + \hat{\beta}_1 \psi_S(y, \lambda). \tag{8}$$

При фиксированном λ модель (8) представляет собой простую линейную регрессию с предиктором $\psi_S(y, \lambda)$ и зависимой переменной $\hat{y}$. Оптимальным параметром $\hat{\lambda}$ предлагается считать тот, который минимизирует **остаточную сумму квадратов (*residual sum of squares*)**:

$$RSS(\lambda) = \sum_{i=1}^{n} \left(\hat{y}_i - \hat{\beta}_0 - \hat{\beta}_1 \psi_S(y_i, \lambda) \right). \tag{9}$$

В примере пункта 2.1.1 такой подход дает значение $\hat{\lambda} = 0.31$.

2.2 Трансформация только зависимой переменной при помощи преобразования Бокса-Кокса

В своей работе [Box and Cox, 1964] Бокс и Кокс рассматривали модифицированное семейство степенных преобразований:

$$\psi_M(Y,\lambda) = GM(Y)^{1-\lambda} \cdot \psi_S(Y,\lambda) = \begin{cases} GM(Y)^{1-\lambda} \cdot \dfrac{Y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ GM(Y) \cdot \log(Y), & \lambda = 0 \end{cases}, \tag{10}$$

где $GM(Y) = \sqrt[n]{y_1 y_2 \dots y_n}$ – среднее геометрическое выборки.

Метод Бокса-Кокса основывается на предположении, что для некоторого неизвестного λ после преобразования зависимая переменная $\psi_M(Y,\lambda)$ такова, что $\psi_M(y_i,\lambda)$ – независимые нормально распределенные сл.в. с постоянной дисперсией σ^2 и матожиданием

$$\mathbf{E}(\psi_M(Y,\lambda)|X = x) = \beta_0 + \beta_1 x. \tag{11}$$

Из этих предположений предлагается брать такое λ, которое максимизировало бы правдоподобие. В нормальной модели логарифм правдоподобия имеет вид:

$$\log(L) = \ell = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = \\ = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}RSS \tag{12}$$

Оценка максимума правдоподобия дает нам $\hat{\sigma}^2 = RSS/n$. Отсюда получаем:

$$\ell = -\frac{n}{2}\log(2\pi) - \frac{n}{2} - -\frac{n}{2}\log(RSS/n). \tag{13}$$

Можно показать, что Якобиан замены преобразования Бокса-Кокса $\psi_M(Y,\lambda)$ равен 1 при любом λ. А значит при фиксированном λ после трансформации зависимой переменной вид функции правдоподобия останется прежним:

$$\ell(\lambda) = -\frac{n}{2}\log(2\pi) - \frac{n}{2} - -\frac{n}{2}\log(RSS(\lambda)/n). \tag{14}$$

Так как только последний член выражения (14) зависит от λ, то решение задачи максимизации правдоподобия (или логарифма правдоподобия) по λ эквивалентна задаче минимизации RSS по λ.

3 Заключение

TBA

Список литературы

[Berger and Casella, 2001] Berger, R. and Casella, G. (2001). Statistical Inference. Duxbury Press, Florence, AL, 2 edition.

[Box and Cox, 1964] Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. Journal of the Royal Statistical Society. Series B (Methodological), 26(2):211–252.

[Cook and Weisberg, 1994] Cook, R. D. and Weisberg, S. (1994). Transforming a response variable for linearity. Biometrika, 81(4):731–737.

[Weisberg et al., 1978] Weisberg, H., Beier, E., Brody, H., Patton, R., Raychaudhuri, K., Takeda, H., Thern, R., and Van Berg, R. (1978). s-dependence of proton fragmentation by hadrons. ii. incident laboratory momenta 30-250 gev/c. Phys. Rev. D, 17:2875–2887.