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

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 зависимой переменной методом обратной функции

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 $\exp(Y) = \beta_0 + \beta_1 X + \varepsilon$.

Существует несколько способов получения оценки g^{-1}, например, с помощью графика обратного отклика (*inverse response plot*) [Cook and Weisberg, 1994], или с помощью подбора функции из степенного семейства (*power family*), включающим в себя семейство преобразований Бокса-Кокса [Box and Cox, 1964]. В данной работе мы сосредоточимся на последнем методе.

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.