

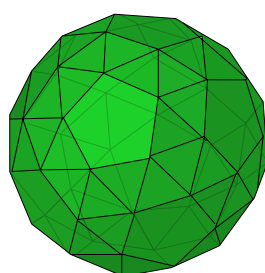
Projects in Mathematics and Applications

GENERALIZED LINEAR MODELS

Ngày 16 tháng 8 năm 2021

Bùi Nguyễn Đức Tân*
Nguyễn Thị Mai Anh[†]

*Châu Thanh Văn
[‡]Nguyễn Tư Thành Nhân



*Trường Phổ Thông Năng Khiếu, ĐHQG TP HCM

[†]Trường THPT chuyên Phan Bội Châu, Nghệ An

[‡]Trường Đại học Công nghệ Thông tin, ĐHQG TP HCM

Lời cảm ơn

Lời nói đầu tiên, chúng em xin gửi lời cảm ơn sâu sắc đến Ban tổ chức cũng như các đơn vị tài trợ trại hè Toán học và Ứng dụng PiMA 2021. Dù tình hình dịch bệnh diễn biến phức tạp, trại hè diễn ra trực tuyến nhưng chúng em đã có cơ hội tiếp cận với các bài giảng kiến thức mới cũng như cải thiện các kỹ năng cần thiết như làm việc nhóm, nghiên cứu,... một cách hiệu quả nhất. Đồng thời, PiMA cũng tổ chức các buổi trò chuyện cực kỳ bổ ích với các anh, chị và thầy có kinh nghiệm trong ngành. Thông qua những cơ hội đó, chúng em đã có thêm được nhiều góc nhìn mới đối với việc học môn Toán trong môi trường giáo dục cao hơn, cũng như việc tìm kiếm định hướng cho bản thân trong tương lai. Chúng em muốn gửi lời cảm ơn chân thành đến các anh chị mentor, đặc biệt là anh Nam Trung và Thế Anh, đã luôn sát sao, tận tình hướng dẫn, giúp đỡ nhóm trong quá trình nghiên cứu và hoàn thành dự án này.

Chúng mình cũng xin cảm ơn các bạn trại sinh đã tích cực tham gia, góp phần vào thành công của PiMA 2021. Trại hè PiMA 2021 tuy đã kết thúc nhưng những trải nghiệm trong 2 tuần qua sẽ theo chân chúng em trên con đường tương lai sắp tới. Chúc PiMA tiếp tục thành công trong những năm sắp tới.

Mặc dù đã có nhiều cố gắng nhưng không thể tránh khỏi sai sót, chúng em mong nhận được sự góp ý của quý bạn đọc để hoàn thiện bản báo cáo này.

Việt Nam, Ngày 16 tháng 8 năm 2021

Nhóm tác giả.

Tóm tắt nội dung

Supervised Learning (SL) hay **Học có giám sát** một trong các hướng nghiên cứu chính của Machine Learning. Trong số các thuật toán thuộc nhóm SL, một trong những thuật toán nổi bật và được sử dụng rộng rãi nhất là **Linear Regression (hay Hồi quy tuyến tính)**. Tuy nhiên, cùng với sự phát triển của xã hội, nhu cầu nghiên cứu của con người dần tiến xa hơn những gì Linear Regression có thể thực hiện hiệu quả. Để khắc phục điều đó, một phiên bản tổng quát hoá và đa dụng hơn của mô hình này đã được nghiên cứu mang tên **Generalized Linear Model**.

Trong paper này, chúng ta sẽ tìm hiểu Generalized Linear Model thông qua các mục sau:

1. Linear Regression dưới góc nhìn Ước lượng tham số và động lực của Generalized Linear Model (GLM).
2. Các thành phần và cách đánh giá một mô hình GLM.
3. Ước lượng tham số trong GLM.
4. Thực hành áp dụng GLM lên dữ liệu dạng Count.

Mục lục

1	Linear Regression (Hồi quy tuyến tính)	1
1.1	Ví dụ mở đầu	1
1.2	Định nghĩa	1
1.3	Linear Regression dưới góc nhìn ước lượng tham số:	1
2	Generalized Linear Model	4
2.1	Motivation	4
2.2	Các thành phần trong Generalized Linear Model:	5
2.3	Generalized Linear Model dưới góc nhìn ước lượng tham số	7
2.4	Đánh giá mô hình	14
3	Count data	15
3.1	Generalized Linear Model cho Count Data:	15
3.2	Tính toán MLE đối với Count data	16
3.3	Đánh giá mô hình	16
3.4	Overdispersion	17
3.5	Cài đặt	19
3.6	Ví dụ minh họa	20
4	Áp dụng mô hình	23
5	Kết luận đánh giá	23

1 Linear Regression (Hồi quy tuyến tính)

1.1 Ví dụ mở đầu

Giả sử ta cần trả lời câu hỏi: Các yếu tố gì ảnh hưởng đến điểm trung bình của một lớp học? Nếu như ta quan sát được được số học sinh, số tiền phụ huynh đóng và số tiết học một tuần của lớp đó thì có công cụ gì để đánh giá mức độ ảnh hưởng của ba yếu tố đó lên điểm trung bình không? Trong toán học, ta có thể nghiên cứu mối quan hệ giữa các điều nói trên bằng cách đặt chúng thành các biến như sau:

- y là điểm trung bình của lớp học.
- x_1 là số học sinh.
- x_2 là số tiền phụ huynh đóng góp cho lớp học.
- x_3 là số tiết học một tuần.

Từ các thông tin trên, câu hỏi trở thành bài toán đánh giá mối quan hệ tương quan giữa biến y và các biến x_1, x_2, x_3 . Ta gọi dạng bài toán này là “Hồi quy.” Có nhiều phương pháp phân tích hồi quy khác nhau, nhưng bằng cách giả sử một mối quan hệ tuyến tính giữa các biến, ta có một trong những phương pháp hiệu quả và đơn giản nhất **Linear Regression**, hay mô hình **Hồi quy tuyến tính**. Cụ thể, trong ví dụ này,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

với $\beta_0, \beta_1, \beta_2, \beta_3$ là các tham số hồi quy, ϵ là biến lỗi.

1.2 Định nghĩa

Trong bài toán hồi quy, ta có một số khái niệm, ký hiệu sau

- Y là biến phản hồi

Cho một data set $\{y_i, x_{i1}, \dots, x_{ip}\}_{i=1}^n$ với n là số lượng quan sát và p là số các biến độc lập x . Linear Regression hay Hồi quy tuyến tính là một mô hình thể hiện mối quan hệ giữa giá trị scalar y_i và tập hợp các biến $\{x_1, x_2, \dots, x_p\}$ thông qua một hàm tuyến tính với tập hợp tham số $\beta = \{\beta_0, \beta_1, \dots, \beta_p\}$ và giá trị lỗi ϵ . Mối quan hệ đó được thể hiện như sau:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \epsilon_i, i = 1 \dots n$$

Khi xếp chồng các giá trị của y_i thành một vector cột, ta có công thức tổng quát sau:

$$y = \mathbf{X}\beta + \epsilon$$

Trong đó: $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{12} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$

1.3 Linear Regression dưới góc nhìn ước lượng tham số:

Có hai cách được sử dụng phổ biến để ước lượng tham số β sao cho phù hợp:

- Phương pháp **Least Square Estimate**: Tìm β sao cho giá trị các biến ϵ_i nhỏ nhất có thể.
- Phương pháp **Maximun Likelihood Estimate**. Ở đây ta sẽ quan tâm đến phương pháp này.

1.3.1 Likelihood Function và Maximum Likelihood Estimation:

Với một data set được giả sử tuân theo một phân phối tham số hoá, **hàm Likelihood** được dùng để tính xác suất mô hình với tham số được chọn có thể sinh ra được mẫu data set đã quan sát được. Giá trị trả về của hàm càng cao thì dữ liệu sinh ra từ mô hình càng gần với dữ liệu quan sát được. Nói cách khác, hàm Likelihood có thể đánh giá mức độ **phù hợp** của tham số trong mô hình. Với một biến X ngẫu nhiên có hàm phân phối xác suất là f phụ thuộc vào tham số θ , gọi hàm Likelihood của biến là L , ta có:

$$L(\theta|X = \mathbf{x}) = f(X = \mathbf{x}|\theta)$$

Khi có mẫu dữ liệu $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, giả sử các biến độc lập và phân phối như nhau, ta có công thức:

$$L(\theta|X = \mathbf{x}_1, \dots, \mathbf{x}_n) = f(\mathbf{x}_1|\theta) \dots f(\mathbf{x}_n|\theta)$$

$$\Leftrightarrow L(\theta|X = \mathbf{x}_1, \dots, \mathbf{x}_n) = \prod_{i=1}^n f(\mathbf{x}_i|\theta)$$

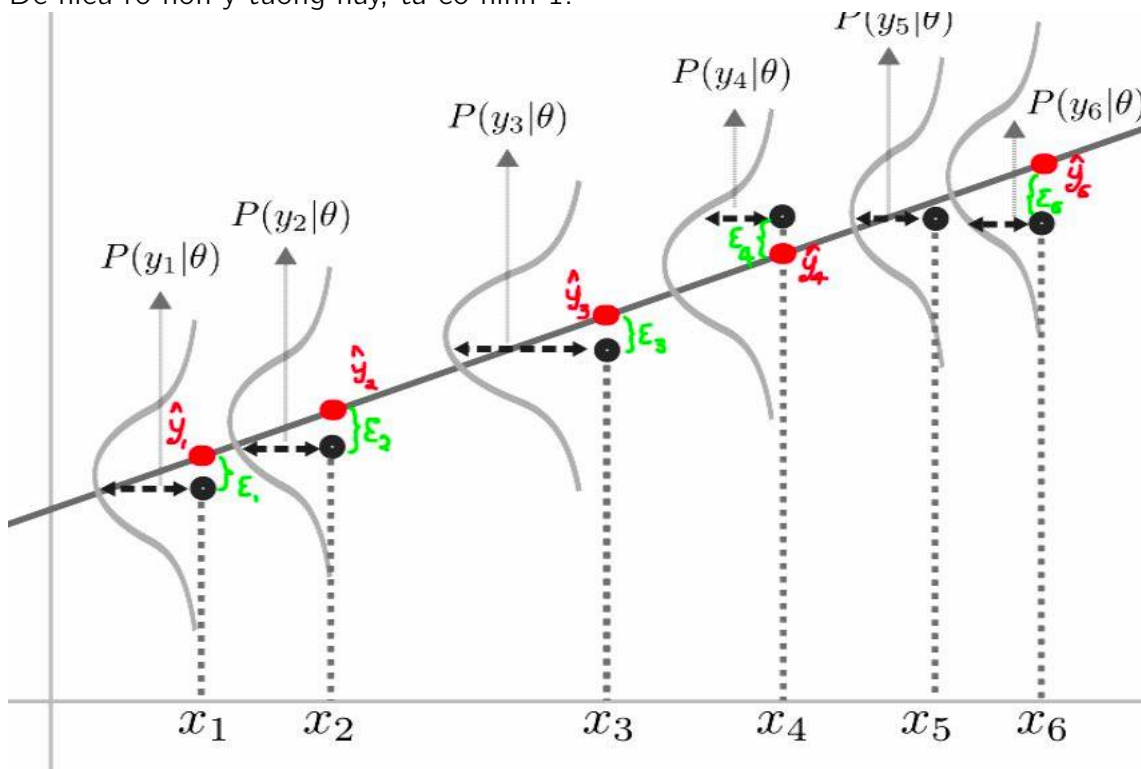
Bằng cách tối đa hoá hàm Likelihood, ta sẽ tìm được tham số phù hợp nhất với mô hình. Đây chính là nền tảng cho phương pháp **Maximum Likelihood Estimation**.

1.3.2 Áp dụng Maximum Likelihood Estimation cho Linear Regression:

Để ước lượng tham số β dưới góc nhìn Ước lượng tham số sử dụng Maximum Likelihood Estimator, trước tiên ta cần giả sử một số điều sau:

- X, Y là các biến ngẫu nhiên.
- Biến ϵ cũng là một biến ngẫu nhiên và tuân theo **phân phối chuẩn** với $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Từ đây ta có thể hiểu rằng: các giá trị của biến ngẫu nhiên Y khi biết $X = \mathbf{x}$ cũng tuân theo phân phối chuẩn, hay $Y|X = \mathbf{x} \sim \mathcal{N}(\mathbf{x}^\top \beta, \sigma^2)$.

Để hiểu rõ hơn ý tưởng này, ta có hình 1:



Có thể thấy, với mỗi giá trị x xác định, giá trị của biến ngẫu nhiên Y tuân theo phân phối chuẩn có dạng như trên hình. Ở đây, giá trị trung bình chính là giá trị y dự đoán (điểm màu đỏ nằm trên đường thẳng best-fit): $\mu = \hat{y} = \beta_i x_{ij}$. Giá trị y thật sự là điểm màu đen, bằng giá trị dự đoán cộng với giá trị lỗi. Khoảng cách từ điểm y thật đến đồ thị của phân phối chuẩn thể hiện xác suất xuất hiện của y , hay $P(y_i|\theta)$. Xác suất y nằm trên đường thẳng best-fit là cao nhất, càng xa đường thẳng (hay lỗi càng lớn) thì xác suất xuất hiện y càng nhỏ.

1.3.3 Tính toán MLE trong mô hình Linear Regression

Như chúng ta đã được tìm hiểu, ta có:

$$p(y_n|\mathbf{x}_n, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (y_n - \boldsymbol{\beta}^\top \mathbf{x}_n)^2\right)$$

Từ đó, ta xây dựng được hàm likelihood như sau:

$$L(\boldsymbol{\beta}) = \prod_n p(y_n|\mathbf{x}_n, \boldsymbol{\beta}, \sigma^2) = \prod_n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (y_n - \boldsymbol{\beta}^\top \mathbf{x}_n)^2\right)$$

Mục đích của bài toán MLE là tính toán $\boldsymbol{\beta}$ để hàm Likelihood nhận giá trị lớn nhất, tuy nhiên ở đây việc tính toán giá trị lớn nhất của hàm Likelihood một cách trực tiếp sẽ rất khó giải quyết, bởi vì việc tính đạo hàm theo tích các biến tương đối phức tạp. Nên thay vì vậy, ta sẽ tính $\log(\text{logarit cơ sở tự nhiên})$ của hàm Likelihood và tính $\boldsymbol{\beta}$ để làm Log-likelihood đạt giá trị lớn nhất. Ta có thể thấy rằng hàm \log là một hàm đồng biến, nên $\arg \max_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) = \arg \max_{\boldsymbol{\beta}} LL(\boldsymbol{\beta})$. Từ đó, ta tính hàm Log-likelihood như sau:

$$\begin{aligned} LL(\boldsymbol{\beta}) &= \sum_n \left[\left(\log \frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{1}{2\sigma^2} (y_n - \boldsymbol{\beta}^\top \mathbf{x}_n)^2 \right] \\ &= -\frac{1}{2\sigma^2} \sum_n (y_n - \boldsymbol{\beta}^\top \mathbf{x}_n)^2 - \frac{n}{2} \log(2\pi) - n \log \sigma \end{aligned}$$

Ta có, tổng phần dư bình phương (RSS - Residual Sum of Squares) được tính như sau:

$$RSS(\boldsymbol{\beta}) = \sum_n (y_n - \boldsymbol{\beta}^\top \mathbf{x}_n)^2$$

Suy ra

$$LL(\boldsymbol{\beta}) = \text{const} - \frac{1}{2\sigma^2} RSS(\boldsymbol{\beta})$$

Do đó, ta có thể thấy rằng việc tối đa hóa giá trị Likelihood cũng tương đương với việc tối thiểu giá trị RSS, nên ta cũng có thể xem 2 bài toán MLE và OLS Estimation (Ordinary Least Square Estimation) ở đây là tương đương nhau

Gọi

$$\hat{\boldsymbol{\beta}} = \arg \max_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) = \arg \max_{\boldsymbol{\beta}} LL(\boldsymbol{\beta})$$

Suy ra

$$\nabla_{\boldsymbol{\beta}} LL(\hat{\boldsymbol{\beta}}) = \mathbf{0}$$

Ta tính toán như sau:

$$\begin{aligned} \nabla_{\boldsymbol{\beta}} LL(\boldsymbol{\beta}) &= -\frac{1}{2\sigma^2} \nabla_{\boldsymbol{\beta}} \sum_n (y_n - \boldsymbol{\beta}^\top \mathbf{x}_n)^2 = -\frac{1}{2\sigma^2} \nabla_{\boldsymbol{\beta}} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 \\ &= -\frac{1}{2\sigma^2} \nabla_{\boldsymbol{\beta}} (\mathbf{X}\boldsymbol{\beta} - \mathbf{y})^\top \nabla_{(\mathbf{X}\boldsymbol{\beta} - \mathbf{y})} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 \end{aligned}$$

Vì

$$\nabla_{\mathbf{x}} \|\mathbf{x}\|_2^2 = \nabla_{\mathbf{x}} \sum_i x_i^2 = \left(\frac{\partial \sum_i x_i^2}{\partial x_1}, \frac{\partial \sum_i x_i^2}{\partial x_2}, \dots, \frac{\partial \sum_i x_i^2}{\partial x_n} \right) = (2x_1, 2x_2, \dots, 2x_n) = 2\mathbf{x}$$

Nên

$$\nabla_{\beta} LL(\beta) = -\frac{1}{2\sigma^2} (\mathbf{X}^T \cdot 2(\mathbf{X}\beta - \mathbf{y})) = -\frac{1}{\sigma^2} (\mathbf{X}^T \mathbf{X}\beta - \mathbf{X}^T \mathbf{y})$$

Vì $\nabla_{\beta} LL(\hat{\beta}) = 0$, suy ra $\mathbf{X}^T \mathbf{X}\hat{\beta} = \mathbf{X}^T \mathbf{y}$

Do đó, $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$

2 Generalized Linear Model

2.1 Motivation

Mô hình hồi quy tuyến tính (Linear Regression) rất đơn giản những đã được chứng minh vô cùng hữu ích trong một lượng lớn các tình huống. Tuy nhiên, có một số trường hợp thì đã thấy được mặt hạn chế của mô hình này. Ta xét các ví dụ sau :

Ví dụ 2.1. Dự đoán mong muốn học tiếp sau đại học (có hay không) của sinh viên Việt Nam. Ta có thể quy ước biến có giá trị 1 nếu câu trả lời là có, và nhận giá trị 0 nếu câu trả lời là không mong muốn học tiếp sau đại học. Có nhiều yếu tố ảnh hưởng đến quyết định có học tiếp hay không, ví dụ như điểm trung bình học tập, số tiết cúp học trong kì gần nhất, số tiền hiện có cho việc học, ... Ví dụ ở đây, ta sẽ xét điểm trung bình học tập và số tiền hiện đầu tư cho học tập.

- Input: Tập dữ liệu gồm bộ 3 số: điểm trung bình học tập, số tiền, 1 số 1 hoặc 0.
- Output: Một mô hình mà khi biết điểm, số tiền sẽ dự đoán được có mong muốn học tiếp hay không.

Nhận xét 2.2. • *Biến phản hồi có thể là biến ngẫu nhiên Bernoulli vì nó chỉ nhận 1 trong 2 giá trị là 0 hoặc 1. Do đó, dùng phân phối chuẩn sẽ không phù hợp vì biến phản hồi trong phân phối chuẩn nhận mọi giá trị trên \mathbb{R}*

- *Nếu dùng mô hình hồi quy tuyến tính, giá trị kì vọng khi biết số điểm và số tiền sẽ là một tổ hợp tuyến tính của 2 yếu tố trên và có giá trị trên toàn bộ số thực. Do đó, ta cần xây dựng một hàm f sao cho giá trị đó qua hàm f sẽ nằm trên $(0, 1)$.*

Ta xét ví dụ tiếp theo, không phải dữ liệu thuộc kiểu nhị phân nữa mà là kiểu đếm.

Ví dụ 2.3. Dự đoán số ca dương tính với COVID-19 tại TPHCM.

Cũng tương tự như ví dụ trên, ta xét yếu tố ảnh hưởng đến số ca nhiễm bệnh (y) là số lượng người đã tiêm vaccine phòng chống COVID-19 (x_1), số người ra ngoài bị phạt (x_2), số người rời khỏi thành phố (x_3).

- Input: Tập dữ liệu gồm (x_1, x_2, x_3, y)
- Output: Một mô hình mà khi biết x_1, x_2, x_3 sẽ dự đoán được y

Nhận xét 2.4. • *Bởi vì biến phụ thuộc thuộc kiểu dữ liệu đếm, tức là chỉ nhận giá trị là các số nguyên không âm, phân phối Poisson là một phân phối có vẻ phù hợp hơn phân phối chuẩn (phân phối chuẩn chỉ dùng cho biến ngẫu nhiên liên tục).*

- *Cũng tương tự như ví dụ 1, ta cũng cần tìm một hàm f để cho giá trị kì vọng của biến phản hồi nhận giá trị không âm, chứ không phải nhận mọi giá trị thực.*

Ví dụ cuối cùng, ta xét một kiểu dữ liệu khác: dữ liệu tỉ lệ (proportion).

Ví dụ 2.5. Dự đoán tỉ lệ học sinh lớp 12 thích Toán của một trường.

Các yếu tố ảnh hưởng lên tỉ lệ học sinh lớp 12 thích Toán (y) mà chúng ta xét ở đây là tỉ lệ học sinh thi HSG môn toán cấp trường (x_1), điểm trung bình bài kiểm tra 1 tiết gần nhất (x_2).

- Input: Tập dữ liệu gồm (x_1, x_2, y)
- Output: Một mô hình khi nhập vào các giá trị x_1, x_2 sẽ cho dự đoán giá trị của y .

Nhận xét 2.6. • Nếu dùng mô hình hồi quy tuyến tính thì biến phản hồi sẽ tuân theo phân phối chuẩn, cũng không phù hợp vì biến phản hồi là một tỉ lệ nên nằm trong đoạn $(0, 1)$.

- Tương tự, ta cũng cần tìm một hàm f

Không chỉ ba ví dụ trên, mà còn nhiều ví dụ khác cho thấy nhiều tình huống mô hình hồi quy tuyến tính không còn phù hợp nữa. Để hoàn thiện mô hình này, người ta đã tổng quát hóa lên thành **Generalized Linear Model (GLM)** bằng cách:

- Ngoài phân phối chuẩn, biến phản hồi có thể tuân theo một kiểu phân phối khác, ví dụ như là phân phối Poisson hay phân phối Bernoulli.
- Xây dựng một hàm f giữa biến quan sát và tổ hợp tuyến tính của biến độc lập.

2.2 Các thành phần trong Generalized Linear Model:

Như ta đã biết, Generalized Linear Model (GLM) là một trường hợp tổng quát hoá của Linear Model. Trong GLM, quan hệ giữa Y và X không còn chắc chắn là tuyến tính nữa, hơn nữa, phân phối của giá trị lỗi không còn nhất thiết phải tuân theo phân phối chuẩn mà có thể sẽ phức tạp hơn. Khi Y và X không còn chắc chắn tuyến tính, ta cần thêm một thành phần khác. Như vậy, trong GLM, ta sẽ có tổng cộng 3 thành phần: **Random Component (Thành phần ngẫu nhiên)**, **Linear Predictor (Dự đoán tuyến tính)**, và **Link Function (Hàm liên kết)**.

2.2.1 Random Component:

Cho $(X_n, Y_n) \in \mathbb{R}^p \times \mathbb{R}$, $n = 1, \dots, i$ là các cặp biến ngẫu nhiên độc lập sao cho hàm phân phối của Y_n với điều kiện biết $X_n = \mathbf{x}_n$ có dạng là một họ hàm exponential (hay exponential family):

$$f(y_n; \theta, \phi) = \exp \left[\frac{y_n \theta_n - b(\theta)}{a_n(\phi)} + c(y_n, \phi) \right]$$

Trong đó:

- θ được gọi là **canonical parameter** hay **natural parameter**. Tham số này chính là μ của phân phối chuẩn, $\log\left(\frac{p}{1-p}\right)$ trong phân phối nhị thức (Binominal Distribution), hay $\log(\mu)$ trong phân phối Poisson, vân vân.
- ϕ được gọi là **dispersion parameter**. Thông thường hàm $a(\phi) = 1$, khi đó, họ hàm sẽ chuyển thành dạng đơn giản hơn: $f(y_n; \theta) = h(y_n) \exp[y_n \theta_n - b(\theta)]$.
- $b(\theta)$ còn được gọi là **cummulant function**.

Bằng việc giới hạn y theo họ hàm exponential, ta sẽ có thể:

- Thu được general expression cho phương trình likelihood của mô hình, phân phối tiệm cận (asymptotic distribution) của các estimator cho tham số của mô hình, thuật toán để fit mô hình.
- Các giá trị trong hàm sẽ được sử dụng để tìm được $E(y_n)$ và $\text{Var}(y_n)$.

Một số ví dụ về các phân phối thuộc exponential family:

- Phân phối biến ngẫu nhiên liên tục gồm: Phân phối chuẩn, Phân phối Gamma,...
- Phân phối biến ngẫu nhiên rời rạc gồm: Phân phối Poisson, Phân phối Binomial,...

2.2.2 Linear Predictor:

Thành phần này dùng để dự đoán mối quan hệ tuyến tính giữa các biến độc lập x so với mô hình. Ta ký hiệu thành phần này là η với:

$$\eta_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij}, i = 1 \dots n$$

2.2.3 Link Function:

Cho tới bây giờ ta đã thấy sự xuất hiện của Y và X , tuy nhiên ta vẫn chưa liên kết được chúng với nhau. Để thực hiện được việc đó, ta dùng một **hàm liên kết (link function)**.

Ta sử dụng Link Function để liên kết η_i và giá trị kỳ vọng, hay giá trị trung bình, của y , ký hiệu $\mu = E(y_n)$. Link Function của mô hình GLM được ký hiệu:

$$\eta = g(\mu)$$

Một Link Function của GLM phải thoả mãn **hai điều kiện**:

- Hàm này là hàm đơn điệu.
- Hàm này khả vi.

Đặc biệt: Nếu như Link Function có thể liên kết μ và canonical parameter θ với dạng $\eta = g(\mu) = \theta$, ta gọi hàm đó là **Canonical Link Function**. Sử dụng được Canonical Link Function sẽ có một số tác dụng sau:

- **Đảm bảo thống kê đủ (Sufficient Statistic):** Một thống kê được gọi là đủ khi “nếu việc thêm bất kỳ thống kê nào được tính toán từ cùng mẫu (sample) không cung cấp thêm thông tin mới về tham số của mô hình đó.” **Đối với Canonical Link Function**, phương trình có dạng $g(\mu) = \theta = \mathbf{X}\boldsymbol{\beta}$ sẽ cho phép $\mathbf{X}^T \mathbf{Y}$ là một thống kê đủ cho $\boldsymbol{\beta}$.
- **Đơn giản hoá việc tính toán, đơn giản hoá quá trình ước lượng tham số.**

Bảng 1: Bảng: GLM cho các phân phối khác họ exponential.

Tên phân phối	Normal	Binominal	Gamma
Ký hiệu	$\mathcal{N}(\mu, \sigma^2)$	$\mathcal{B}(m, \pi)/m$	$\mathcal{G}(v, \mu)$
Link Function	$\mathbf{X}\beta = \mu$	$\mathbf{X}\beta = \log\left(\frac{\mu}{n - \mu}\right)$	$\mathbf{X}\beta = -\mu^{-1}$
Tham số ϕ	$\phi = \sigma^2$	$\phi = \frac{1}{m}$	$\phi = v^{-1}$
Hàm $b(\theta)$	$\frac{\theta^2}{2}$	$\log(1 + e^\theta)$	$-\log(-\theta)$
Hàm $c(y, \phi)$	$-\frac{1}{2}\left(\frac{y^2}{\phi} + \log(2\pi\phi)\right)$	$\log C_{my}^m$	$v \log(vy) - \log(y)$

2.3 Generalized Linear Model dưới góc nhìn ước lượng tham số

2.3.1 Generalized Least Square (GLS) và mối liên hệ với Ordinary Least Squares (OLS) và Weighted Least Squares (WLS)

Least Square Estimation là phương pháp để xác định tham số của một mô hình thống kê sao cho phù hợp với nhất với tập dữ liệu, bằng cách tối thiểu Residual Sum of Squares (RSS), hay còn được biết tới là Sum of Squared Residuals (SSR), Squared Estimate of Errors (SSE), tổng phần dư bình phương.

1. OLS estimation

Ta có: linear model (mô hình tuyến tính) như sau,

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

Trong đó,

- y_i là biến ngẫu nhiên
- β_i là biến không ngẫu nhiên, nhưng chúng ta chưa xác định được
- x_{ij} là biến ngẫu nhiên, là dữ liệu chúng ta quan sát được
- ϵ_i là biến ngẫu nhiên và được gọi là biến lỗi ngẫu nhiên

Ta có giả định Gauss-Markov như sau,

- $E[\epsilon_n] = 0, \forall n$
- $\text{Var}[\epsilon_n] = \sigma^2 < \infty, \forall n$
- $\text{Cov}[\epsilon_n, \epsilon_m] = 0, \forall n \neq m$

Lần lượt đặt, $\mathbf{y} = \mathbf{X}\hat{\beta} + \mathbf{e}, \hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ Gọi OLS estimator là

$$\hat{\beta}_{OLS} = \arg \min_{\beta} \text{RSS}(\beta) \Rightarrow \nabla_{\beta} \text{RSS}(\hat{\beta}_{OLS}) = 0$$

Lại có,

$$\text{RSS}(\hat{\beta}) = \|\mathbf{e}\|_2^2 = \|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2$$

Suy ra,

$$\nabla_{\beta} \|\mathbf{X}\hat{\beta} - \mathbf{y}\| = \nabla_{\beta} (\mathbf{X}\hat{\beta} - \mathbf{y}) \nabla_{(\mathbf{X}\hat{\beta} - \mathbf{y})} \|\mathbf{X}\hat{\beta} - \mathbf{y}\|_2^2 = 2(\mathbf{X}^T \mathbf{X}\hat{\beta} - \mathbf{X}^T \mathbf{y})$$

Do đó, ta có OLS estimator như sau

$$\hat{\beta}_{OLS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

2. GLS estimation

Tương tự, ta cũng có:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\epsilon}$$

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

Giả định là $E[\boldsymbol{\epsilon}] = 0$, $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{V}$, trong đó \mathbf{V} là ma trận được xác định. Ta có, $\sigma^2 \mathbf{V}$ là ma trận hiệp phương sai, trong đó chúng ta xác định được \mathbf{V} là ma trận đối xứng và khả nghịch, khi đó, chúng ta có thể định nghĩa

$$\mathbf{V} = \mathbf{K}^\top \mathbf{K} = \mathbf{K} \mathbf{K}$$

với \mathbf{K} được gọi là ma trận căn bậc hai của \mathbf{V}

Khi đó, định nghĩa,

- $\mathbf{y}' = \mathbf{K}^{-1} \mathbf{y}$
- $\mathbf{X}' = \mathbf{K}^{-1} \mathbf{X}$
- $\boldsymbol{\epsilon}' = \mathbf{K}^{-1} \boldsymbol{\epsilon}$
- $\mathbf{e}' = \mathbf{K}^{-1} \mathbf{e}$

Lúc đó, chúng ta cũng xác định được,

- $E[\boldsymbol{\epsilon}'] = E[\mathbf{K}^{-1} \boldsymbol{\epsilon}] = \mathbf{K}^{-1} E[\boldsymbol{\epsilon}] = \mathbf{K}^{-1} \times 0 = 0$
- $\text{Var}[\boldsymbol{\epsilon}'] = \text{Var}[\mathbf{K}^{-1} \boldsymbol{\epsilon}] = \mathbf{K}^{-1} \text{Var}[\boldsymbol{\epsilon}] \mathbf{K}^{-1} = \mathbf{K}^{-1} \sigma^2 \mathbf{V} \mathbf{K}^{-1} = \sigma^2 \mathbf{K}^{-1} \mathbf{K} \mathbf{K}^{-1} = \sigma^2 \mathbf{I}$

Để ước lượng được $\hat{\boldsymbol{\beta}}$ tốt nhất chúng ta cần tối thiểu RSS điều chỉnh như sau:

$$\begin{aligned} \text{RSS}(\hat{\boldsymbol{\beta}}) &= \mathbf{e}'^\top \mathbf{e}' \\ &= (\mathbf{y}' - \mathbf{X}'\hat{\boldsymbol{\beta}})^\top (\mathbf{y}' - \mathbf{X}'\hat{\boldsymbol{\beta}}) \\ &= (\mathbf{K}^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}))^\top (\mathbf{K}^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})) \\ &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top \mathbf{K}^{-1} \mathbf{K}^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= (\mathbf{y}^\top - \hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top) \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{y}^\top \mathbf{V}^{-1} \mathbf{y} - 2\hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{y} + \hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X} \hat{\boldsymbol{\beta}} \end{aligned}$$

Từ đó, ta suy ra được

$$\begin{aligned} \nabla_{\boldsymbol{\beta}} \text{RSS}(\hat{\boldsymbol{\beta}}_{OLS}) &= 0 \\ \Leftrightarrow \nabla_{\boldsymbol{\beta}} \left[\mathbf{y}^\top \mathbf{V}^{-1} \mathbf{y} - 2\hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{y} + \hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X} \hat{\boldsymbol{\beta}} \right] &= 0 \\ \Leftrightarrow -2\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{y} + 2\mathbf{X}^\top \mathbf{V}^{-1} \hat{\boldsymbol{\beta}} &= 0 \\ \Leftrightarrow \hat{\boldsymbol{\beta}} &= (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{y} \end{aligned}$$

Do đó, ta có GLS estimator

$$\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{y}$$

3. WLS estimation

Một trong những hạn chế của GLS là chúng ta phải chỉ định ma trận \mathbf{V} . Ma trận này thường chúng ta sẽ không biết được và xác định thường rất khó khăn

WLS là trường hợp đặc biệt của GLS, và nó có những giả định đơn giản hơn là đối với ma trận \mathbf{V} . Ta giả định là ϵ không tương quan với nhau và các phương sai ở đây cũng không bằng nhau, $\text{Var}[\epsilon] = \sigma^2 \mathbf{W}$ với ma trận \mathbf{W} là ma trận đường chéo với các phương sai. Các quan sát phương sai nhỏ hơn thì sẽ có trọng số lớn hơn, các quan sát có phương sai lớn hơn thì sẽ có trọng số nhỏ hơn. Do đó,

$$\hat{\beta}_{WLS} = (\mathbf{X}^\top \mathbf{W}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}^{-1} \mathbf{y}$$

Trong đó, \mathbf{W} còn được gọi là ma trận trọng số

2.3.2 Tính toán MLE trong mô hình Generalized Linear Model

Xác định bài toán

Như chúng ta đã được biết,

$$f(y_n; \theta_n, \phi) = \exp \left(\frac{y_n \theta_n - b(\theta_n)}{a_n(\phi)} + c(y_n, \phi) \right)$$

Ta sẽ xây dựng hàm Likelihood như sau,

$$L(\beta) = \prod_n f(y_n; \theta_n, \phi) = \prod_n \exp \left(\frac{y_n \theta_n - b(\theta_n)}{a_n(\phi)} + c(y_n, \phi) \right)$$

Cũng như bài toán tính MLE ở trong Linear Regression, ta cũng sẽ đi tính hàm Log-likelihood,

$$LL(\beta) = \sum_n \left(\frac{y_n \theta_n - b(\theta_n)}{a_n(\phi)} + c(y_n, \phi) \right)$$

Gọi

$$LL_n = \frac{y_n \theta_n - b(\theta_n)}{a_n(\phi)} + c(y_n, \phi)$$

Khi đó,

$$\frac{\partial LL_n}{\partial \theta_n} = \frac{y_n - b'(\theta_n)}{a_n(\phi)} \quad \frac{\partial^2 LL_n}{\partial \theta_n^2} = \frac{-b''(\theta_n)}{a_n(\phi)}$$

Hai kết quả của General Likelihood:

$$E \left[\frac{\partial LL_n}{\partial \theta_n} \right] = 0 \quad E \left[\frac{\partial^2 LL_n}{\partial \theta_n^2} \right] + E \left[\left(\frac{\partial LL_n}{\partial \theta_n} \right)^2 \right] = 0$$

Ta sẽ chứng minh lần lượt hai công thức này,

Gọi

$$L_n = f(y_n; \theta_n; \phi) = \exp LL_n$$

$$\begin{aligned} E \left[\frac{\partial LL_n}{\partial \theta_n} \right] &= E \left[\frac{\partial \log L_n}{\partial \theta_n} \right] = E \left[\frac{1}{L_n} \frac{\partial L_n}{\partial \theta_n} \right] \\ &= \int \frac{1}{L_n} \frac{\partial L_n}{\partial \theta_n} L_n dy \\ &= \frac{\partial}{\partial \theta_n} \int L_n dy \\ &= \frac{\partial}{\partial \theta_n} \int f(y_n; \theta_n; \phi) dy \\ &= \frac{\partial}{\partial \theta_n} 1 = 0 \end{aligned}$$

$$\begin{aligned}
E \left[\frac{\partial^2 LL_n}{\partial \theta_n^2} \right] + E \left[\left(\frac{\partial LL_n}{\partial \theta_n} \right)^2 \right] &= E \left[\frac{\partial}{\partial \theta_n} \left(\frac{1}{L_n} \frac{\partial L_n}{\partial \theta_n} \right) \right] + E \left[\left(\frac{1}{L_n} \frac{\partial L_n}{\partial \theta_n} \right)^2 \right] \\
&= E \left[\frac{L_n \frac{\partial^2 L_n}{\partial \theta_n^2} - \frac{\partial L_n}{\partial \theta_n} \frac{\partial L_n}{\partial \theta_n}}{L_n^2} \right] + E \left[\frac{\frac{\partial L_n}{\partial \theta_n} \frac{\partial L_n}{\partial \theta_n}}{L_n^2} \right] \\
&= E \left[\frac{\frac{\partial^2 L_n}{\partial \theta_n^2}}{L_n} \right] \\
&= \int \frac{\frac{\partial^2 L_n}{\partial \theta_n^2}}{L_n} L_n dy \\
&= \frac{\partial^2}{\partial \theta_n^2} \int L_n dy \\
&= \frac{\partial^2}{\partial \theta_n^2} \int f(y_n; \theta_n; \phi) dy \\
&= \frac{\partial^2}{\partial \theta_n^2} 1 = 0
\end{aligned}$$

Do đó, ta có:

$$\begin{cases} \frac{E[y_n - b'(\theta_n)]}{a_n(\phi)} = 0 \\ \frac{b''(\theta_n)}{a_n(\phi)} = \frac{E[(y_n - b'(\theta_n))^2]}{a_n^2(\phi)} \end{cases} \Leftrightarrow \begin{cases} E[y_n] = b'(\theta_n) \\ E[(y_n - b'(\theta_n))^2] = b''(\theta_n) a_n(\phi) = \text{Var}[y_n] \end{cases}$$

Ta rút ra được hai kết quả chính như sau

$$\begin{cases} E[y_n] = b'(\theta_n) = \mu_n \\ \text{Var}[y_n] = b''(\theta_n) a_n(\phi) = \sigma^2 \end{cases}$$

Tiếp tục với bài toán MLE trong mô hình Generalized Linear Model

Áp dụng công thức Chain rule:

$$\frac{\partial LL_n}{\partial \beta_m} = \frac{\partial LL_n}{\partial \theta_n} \frac{\partial \theta_n}{\partial \mu_n} \frac{\partial \mu_n}{\partial \eta_n} \frac{\partial \eta_n}{\partial \beta_m}$$

Trong đó,

$$\begin{aligned}
\frac{\partial LL_n}{\partial \theta_n} &= \frac{y_n - b'(\theta_n)}{a_n(\phi)} = \frac{y_n - \mu_n}{a_n(\phi)} \\
\frac{\partial \mu_n}{\partial \theta_n} &= b''(\theta_n) = \frac{\text{Var}[y_n]}{a_n(\phi)} \Rightarrow \frac{\partial \theta_n}{\partial \mu_n} = \frac{a_n(\phi)}{\text{Var}[y_n]} \\
\frac{\partial \mu_n}{\partial \eta_n} &\text{ không tính toán được, do phụ thuộc vào link function } g(\mu_n) = \eta_n \\
\frac{\partial \eta_n}{\partial \beta_m} &= \frac{\partial \sum_m \beta_m x_{nm}}{\partial \beta_m} = x_{nm}
\end{aligned}$$

Do đó,

$$\frac{\partial LL_n}{\partial \beta_m} = \frac{y_n - \mu_n}{a_n(\phi)} \frac{a_n(\phi)}{\text{Var}[y_n]} \frac{\partial \mu_n}{\partial \eta_n} x_{nm} = \frac{(y_n - \mu_n) x_{nm}}{\text{Var}[y_n]} \frac{\partial \mu_n}{\partial \eta_n}$$

Do đó, chúng ta cần tìm: $\hat{\beta} = \arg \max_{\beta} L(\beta) = \arg \max_{\beta} LL(\beta)$ và nó xảy ra khi $\nabla_{\beta} LL(\hat{\beta}) = \mathbf{0}$

Gọi $\nabla_{\beta} LL(\beta) = \mathbf{0}$ là đẳng thức Likelihood

Đẳng thức Likelihood chúng ta cần tính là

$$\frac{\partial LL(\beta)}{\partial \beta_m} = \sum_n \frac{(y_n - \mu_n) x_{nm}}{\text{Var}[y_n]} \frac{\partial \mu_n}{\partial \eta_n} = 0, \forall m$$

Gọi \mathbf{V} là ma trận đường chéo của các phương sai quan sát được,

\mathbf{D} là ma trận đường chéo của các phần tử $\frac{\partial \mu_n}{\partial \eta_n}$

Khi đó, đẳng thức Likelihood trở thành

$$\mathbf{X}^\top \mathbf{D} \mathbf{V}^{-1} (\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0}$$

Trong đó, $\boldsymbol{\beta}$ được nằm trong công thức $\mu_n = g^{-1}(\boldsymbol{\beta}^\top \mathbf{x}_n)$ Do các phương trình xảy ra ở trên thường là phi tuyến tính, nên ở đây chúng ta sẽ cùng nhau tìm hiểu các phương pháp sử dụng vòng lặp để tính toán MLE

1. Newton-Raphson method

Nhắc lại bài toán: Chúng ta đang cần đi tìm $\boldsymbol{\beta}$ để tối đa Likelihood nhưng bởi vì đẳng thức Likelihood chúng ta có:

$$\mathbf{X}^\top \mathbf{D} \mathbf{V}^{-1} (\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0}$$

không phải lúc nào cũng có thể giải quyết một cách dễ dàng như trong bài toán Linear Regression nên ta có phương pháp chung như sau:

Gọi

$$\mathbf{g}(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} \text{LL}(\boldsymbol{\beta})$$

là gradient của $\text{LL}(\boldsymbol{\beta})$,

$$\mathbf{H}(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}}^2 \text{LL}(\boldsymbol{\beta})$$

là ma trận Hessian của $\text{LL}(\boldsymbol{\beta})$

Xấp xỉ Taylor đến đạo hàm cấp 2 của $\text{LL}(\boldsymbol{\beta})$ tại $(\boldsymbol{\beta}^{(t)})$:

$$\text{LL}(\boldsymbol{\beta}) \approx \text{LL}(\boldsymbol{\beta}^{(t)}) + \mathbf{g}(\boldsymbol{\beta}^{(t)})^\top (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) + \frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)})^\top \mathbf{H}(\boldsymbol{\beta}^{(t)}) (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)})$$

Giải phương trình:

$$\nabla_{\boldsymbol{\beta}} \text{LL}(\boldsymbol{\beta}) \approx \mathbf{g}(\boldsymbol{\beta}^{(t)}) - \mathbf{H}(\boldsymbol{\beta}^{(t)}) (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) = \mathbf{0}$$

Ta xây dựng được vòng lặp như sau:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \left(\mathbf{H}(\boldsymbol{\beta}^{(t)}) \right)^{-1} \mathbf{g}(\boldsymbol{\beta}^{(t)})$$

Với giá trị khởi tạo $\boldsymbol{\beta}^{(0)}$ là một vector bất kì. Điều kiện dừng của chúng ta là khoảng giá trị giữa các chu kỳ liên tiếp tương đối nhỏ ($\boldsymbol{\beta}$ gần như hội tụ) hoặc là \mathbf{g} gần tiến đến 0. Khi $\boldsymbol{\beta}$ hội tụ. Giá trị đó cũng chính là $\boldsymbol{\beta}$ làm cho Likelihood chúng ta cực đại mà chúng ta cần tìm

2. Fisher-scoring method

Ở đây cũng thực hiện tương tự như trên, thay vì sử dụng ma trận Hessian (*thông tin quan sát được*) như Newton-Raphson method, thì nó sẽ sử dụng giá trị kì vọng của nó, được gọi là *thông tin kì vọng*.

Gọi \mathcal{J} là ma trận biểu thị thông tin kì vọng, thì \mathcal{J} sẽ có các phần tử là $-E \left[\frac{\partial^2 \text{LL}_n}{\partial \beta_m \partial \beta_p} \right]$. Hay $\mathcal{J} = -E[\mathbf{H}]$

Do đó, ta xác định vòng lặp như sau:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left(\mathcal{J}(\boldsymbol{\beta}^{(t)}) \right)^{-1} \mathbf{g}(\boldsymbol{\beta}^{(t)})$$

Với giá trị khởi tạo $\beta^{(0)}$ là một vector bất kì. Điều kiện dừng là khoảng giá trị giữa các chu kì liên tiếp tương đối nhỏ (β gần như hội tụ) hoặc là \mathbf{g} gần tiến đến 0

Từ kết quả của General Likelihood ở trên:

$$E \left[-\frac{\partial^2 LL_n}{\partial \beta_m \partial \beta_p} \right] = E \left[\left(\frac{\partial LL_n}{\partial \beta_m} \right) \left(\frac{\partial LL_n}{\partial \beta_p} \right) \right]$$

Mặt khác, ta có

$$\frac{\partial LL_n}{\partial \beta_m} = \frac{(y_n - \mu_n) x_{nm}}{\text{Var}[y_n]} \frac{\partial \mu_n}{\partial \eta_n}$$

Thế vào phương trình trên

$$-E \left[\frac{\partial^2 LL_n}{\partial \beta_m \partial \beta_p} \right] = E \left[\frac{(y_n - \mu_n) x_{nm}}{\text{Var}[y_n]} \frac{\partial \mu_n}{\partial \eta_n} \frac{(y_n - \mu_n) x_{np}}{\text{Var}[y_n]} \frac{\partial \mu_n}{\partial \eta_n} \right] = \frac{x_{nm} x_{np}}{\text{Var}[y_n]} \left(\frac{\partial \mu_n}{\partial \eta_n} \right)^2$$

Gọi \mathbf{W}^{-1} trận đường chéo của các phần tử $w_n = \left(\frac{\partial \mu_n}{\partial \eta_n} \right)^2 \frac{1}{\text{Var}[y_n]}$ Với ma trận mô hình \mathbf{X} , ta xây dựng được ma trận *thông tin kì vọng* như sau

$$\mathcal{J} = \mathbf{X}^\top \mathbf{W}^{-1} \mathbf{X}$$

Thông thường người ta sẽ thường sử dụng Fisher-scoring method thay vì việc sử dụng Newton-Raphson method bởi ma trận Hessian chúng ta chưa xác định được, còn ở đây chúng ta đã chéo hóa được ma trận *thông tin kì vọng*, nên việc tính toán và tính khả nghịch của nó sẽ đơn giản hơn rất nhiều

Sử dụng Canonical Link

Ta có kết quả của Canonical Link như sau:

$$\eta_n = \sum_m \beta_m x_{nm} = \theta_n$$

Khi đó,

$$\frac{\partial \mu_n}{\partial \eta_n} = \frac{\partial \mu_n}{\partial \theta_n} = \frac{\partial b'(\theta_n)}{\partial \theta_n} = b''(\theta_n)$$

Ta lại có

$$\frac{\partial LL_n}{\partial \beta_m} = \frac{(y_n - \mu_n) x_{nm}}{\text{Var}[y_n]} \frac{\partial \mu_n}{\partial \eta_n} = \frac{(y_n - \mu_n) x_{nm}}{b''(\theta_n) a(\phi)} b''(\theta_n) = \frac{(y_n - \mu_n) x_{nm}}{a_n(\phi)}$$

$$\frac{\partial^2 LL_n}{\partial \beta_p \partial \beta_m} = -\frac{x_{nm}}{a_n(\phi)} \frac{\partial \mu_n}{\partial \beta_p}$$

Do giá trị trên không phụ thuộc vào y , nên ta được

$$E \left[\frac{\partial^2 LL_n}{\partial \beta_p \partial \beta_m} \right] = \frac{\partial^2 LL_n}{\partial \beta_p \partial \beta_m}$$

Do đó, $\mathbf{H} = -\mathcal{J}$, nên 2 phương pháp Newton-Raphson và Fisher scoring hoàn toàn giống nhau

3. Iteratively Reweighted Least Squares method

Nhắc lại WLS estimator:

$$\hat{\beta}_{WLS} = (\mathbf{X}^\top \mathbf{W}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}^{-1} \mathbf{y}$$

Ta có kết quả của Fisher Scoring như sau,

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left(\mathcal{J}(\boldsymbol{\beta}^{(t)}) \right)^{-1} \mathbf{g}(\boldsymbol{\beta}^{(t)})$$

Mục tiêu của chúng ta là tham số hóa sao cho Fisher Scoring trông giống như WLS estimator

Ta đã có:

$$\mathcal{J} = \mathbf{X}^\top \mathbf{W}^{-1} \mathbf{X}$$

với \mathbf{W}^{-1} là ma trận đường chéo của các phần tử $w_n = \left(\frac{\partial \mu_n}{\partial \eta_n} \right)^2 \frac{1}{\text{Var}[y_n]}$

Ngoài ra,

$$\frac{\partial LL_n}{\partial \beta_m} = \frac{(y_n - \mu_n) x_{nm}}{\text{Var}[y_n]} \frac{\partial \mu_n}{\partial \eta_n} = x_{nm} \left(\frac{\partial \mu_n}{\partial \eta_n} \right)^2 \frac{1}{\text{Var}[y_n]} (y_n - \mu_n) \frac{\partial \eta_n}{\partial \mu_n} = x_{nm} w_n (y_n - \mu_n) \frac{\partial \eta_n}{\partial \mu_n}$$

Gọi $\tilde{\mathbf{y}}$ là vector có các phần tử

$$(y_n - \mu_n) \frac{\partial \eta_n}{\partial \mu_n}$$

Khi đó Iteratively Reweight Least Squares method được xác định bởi vòng lặp như sau,

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - (\mathbf{X}^\top \mathbf{W}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}^{-1} \tilde{\mathbf{y}}$$

Trong đó, \mathbf{W}^{-1} là ma trận đường chéo của các phần tử $w_n = \left(\frac{\partial \mu_n}{\partial \eta_n} \right)^2 \frac{1}{\text{Var}[y_n]}$

$\tilde{\mathbf{y}}$ là vector có các phần tử $(y_n - \mu_n) \frac{\partial \eta_n}{\partial \mu_n}$

Với giá trị khởi tạo $\boldsymbol{\beta}^{(0)}$ là một vector bất kì. Điều kiện dừng là khoảng giá trị giữa các chu kỳ liên tiếp tương đối nhỏ

2.3.3 Ước lượng dispersion parameter ϕ

Mặc dù ϕ không cần thiết để ước lượng $\boldsymbol{\beta}$, tuy nhiên nó sẽ cần thiết để kiểm tra giả thiết khoảng tin cậy. Vì vậy, trừ khi ϕ xác định trước, nó phải được ước lượng. Các phương pháp ước lượng thường được dùng nhất sẽ mô tả ở phần này.

1. Maximum Likelihood Estimator của ϕ

Về nguyên tắc, chúng ta có thể sử dụng MLE để ước tính ϕ như cách chúng ta ước tính $\boldsymbol{\beta}$. Tuy nhiên MLE của ϕ sẽ rất chệch lệch, trừ khi n phải rất lớn so với r (trong đó r là số tham số chưa biết)

Modified profile log-likelihood (MPL) được định nghĩa như sau:

$$LL^0(\phi) = \frac{r}{2} \log \phi + LL(\boldsymbol{\beta})$$

Giá trị ϕ để $LL^0(\phi)$ đạt cực đại được gọi là MPL estimator

Gọi $\hat{\mu}$

Ví dụ. Trong Linear Regression MPL là

$$LL^0(\sigma^2) = \frac{r}{2} \log \phi - \frac{1}{2} \sum_n \frac{\log 2\pi\sigma^2}{w_n} - \frac{1}{2\sigma^2} \sum_n w_n (y_n - \hat{\mu}_n)^2$$

Ta lần lượt tính đạo hàm $LL^0(\sigma^2)$ theo σ^2 , rồi cho nó bằng 0, và giải phương trình đó ra thì ta có:

$$(\hat{\sigma}^2)^2 = \frac{1}{n-r} \sum_{i=1}^n w_i (y_i - \hat{\mu}_i)^2$$

2. Mean Deviance Estimator của ϕ

$$\bar{\phi} = \frac{D(y, \hat{\mu})}{n - r}$$

Trong đó, $D(y, \hat{\mu})$ được gọi là Deviance, được tính theo công thức như sau

$$D(y, \hat{\mu}) = 2 [\text{LL}(\text{tính theo } y) - \text{LL}(\text{tính theo } \hat{\mu})]$$

3. Person Estimator của ϕ

$$\tilde{\phi} = \frac{X^2}{n - r}$$

Trong đó, X^2 được gọi là Person statistic, được tính theo công thức như sau

$$X^2 = \sum_n \frac{w_i (y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

Nhận xét. Mean Deviance estimator và Person estimator rất tiện lợi, vì chúng có sẵn từ độ lệch và phần dư tương ứng. Mean Deviance estimator phải hoạt động tốt khi điểm yên ngựa gần như giữ nguyên. Trong đó, Person estimator dường như được áp dụng phổ biến hơn bởi vì $(y - \mu)^2/V(\mu)$ thường sẽ không bị lệch đối với ϕ . Tuy nhiên thông thường $\hat{\phi}^0, \tilde{\phi}, \bar{\phi}$ sẽ giống hệt nhau.

2.4 Đánh giá mô hình

Quá trình fitting một mô hình dữ liệu có thể xem là quá trình thay thế các giá trị y bởi một tập hợp fitted values $\hat{\mu}$ thu được từ mô hình mà thường là có tương đối ít các tham số. Nhìn chung thì các giá trị μ không bằng y , nên chúng ta cần quan tâm hai giá trị này lệch nhau bao nhiêu. Nếu mà chúng lệch nhau ít, thì có thể chấp nhận được, nhưng nếu lệch nhau lớn thì đây hẳn không phải là một mô hình tốt.

Có nhiều cách để đo độ lệch này, nhưng chúng ta ở đây chủ yếu quan tâm đến cách tính được hình thành từ hàm logarit của tỉ số giá trị hàm likelihood, và được gọi là *sự sai lệch* (deviance), là độ lệch giữa mô hình cần đánh giá và mô hình mà fit hoàn toàn so với biến quan sát (được gọi là saturated model). Ta có công thức tính (tổng) độ lệch của mô hình M_0 với kì vọng $\hat{\mu} = E[Y|\hat{\beta}_0]$ như sau:

$$D(y, \hat{\mu}) = 2(\log(p(y|\hat{\beta}_s)) - \log(p(y|\hat{\beta}_0)))$$

trong đó, $\hat{\beta}_0$ kí hiệu là fitted values của tham số trong mô hình M_0 , còn $\hat{\beta}_s$ là fitted tham số trong mô hình saturated.

Ta có ví dụ cách xây dựng công thức tính sự sai lệch của phân phối Poisson:

Hàm likelihood của n biến quan sát độc lập Y , với tham số μ tuân theo phân phối Poisson là tích của các xác suất:

$$p(Y = y) = \frac{e^{-\mu} \mu^y}{y!}$$

Lấy logarit cơ số e của tích trên ta thu được:

$$\log L(\beta) = \sum [y_i \log(\mu_i) - \mu_i]$$

trong đó $\log(\mu_i) = x_i^T \cdot \beta$

Thay vào công thức, ta có công thức tính độ lệch là:

$$D = 2 \sum \left[y_i \log \left(\frac{y_i}{\mu_i} \right) - (y_i - \mu_i) \right].$$

Phân phối	Công thức tính sự lệch
Chuẩn	$\sum (y_i - \hat{\mu}_i)^2$
Poisson	$2 \sum \left[y \log \left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right]$
Nhị thức	$2 \sum \left[y \log \left(\frac{y_i}{\hat{\mu}_i} \right) + (m - y_i) \log \left(\frac{m - y_i}{m - \hat{\mu}_i} \right) \right]$
Gamma	$2 \sum \left[-\log \left(\frac{y_i}{\hat{\mu}_i} \right) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right]$

Chỉ số này càng nhỏ thì có thể xem là mô hình càng tốt. Tùy vào loại phân bố mà chúng ta có hàm likelihood khác nhau. Công thức tính sự sai lệch của một số loại phân bố có thể xem ở bảng sau, trong đó $i = 1, 2, \dots, n$.

Bên cạnh đó, có một cách quan trọng để đo độ lệch nữa là *the generalized Pearson X^2 statistic*, có dạng là:

$$X^2 = \sum \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)},$$

trong đó, $V(\hat{\mu}_i)$ là hàm ước lượng phương sai cho phân phối được dùng đến. Đối với phân phối chuẩn, X^2 chính là tổng bình phương phần dư (residual sum of squares).

3 Count data

3.1 Generalized Linear Model cho Count Data:

3.1.1 Random Component:

Đối với count data, Y tuân theo phân phối Poisson

$$p(y_n | \mathbf{x}_n, \boldsymbol{\beta}) = \frac{e^{-\mu_n} \mu_n^{y_n}}{y_n!}$$

$$\Rightarrow p(y_n | \mathbf{x}_n, \boldsymbol{\beta}) = \exp[y_n \log \mu - \mu_n - \log(y_n!)]$$

Từ công thức tổng quát của họ hàm exponential, ta thấy: $\theta_n = \log(\mu_n)$, $b(\theta_n) = \mu_n = \exp(\theta_n)$, $a(\phi) = 1$, $c(y_n, \phi) = \log(y_n!)$.

$$\Rightarrow f(y_n, \mu_n) = \exp[y_n \theta_n - \exp(\theta_n) - \log(y_n!)]$$

Ngoài ra, với phân phối Poisson ta còn có tính chất: $E(y_i) = \text{Var}(y_i) = \mu$.

3.1.2 Linear Predictor và Link Function:

Dựa theo công thức trên, ta có **Canonical Link Function** sau:

$$\eta_i = \theta_i = \log(\mu_i) = \beta_0 + \sum_{j=1}^p \beta_j x_{ij}$$

Hàm liên kết được sử dụng ở đây còn gọi là **Log Function** vì $\log(\mu) = \eta$. Hàm nghịch đảo của nó là: $\mu = \exp(\eta)$.

3.2 Tính toán MLE đối với Count data

Ta xét:

$$p(y_n | \mathbf{x}_n, \boldsymbol{\beta}) = \frac{\mu_n^{y_n} e^{-\mu_n}}{y_n!}$$

Trong đó, $\mu_n = \exp(\boldsymbol{\beta}^\top \mathbf{x}_n)$

Khi đó, hàm Likelihood là

$$L(\boldsymbol{\beta}) = \prod_n \frac{\mu_n^{y_n} e^{-\mu_n}}{y_n!}$$

Tương tự, ta cũng sẽ xét hàm Log-likelihood như sau:

$$\begin{aligned} LL(\boldsymbol{\beta}) &= \sum_n \log \left(\frac{\mu_n^{y_n} e^{-\mu_n}}{y_n!} \right) = \sum_n [y_n \log \mu_n - \mu_n - \log y_n!] \\ &= \sum_n y_n [(\boldsymbol{\beta}^\top \mathbf{x}_n) - \exp(\boldsymbol{\beta}^\top \mathbf{x}_n) - \log y_n!] \end{aligned}$$

Theo đó, ta có:

$$\frac{\partial LL(\boldsymbol{\beta})}{\partial \beta_m} = \sum_n x_{nm} (y_n - \exp(\boldsymbol{\beta}^\top \mathbf{x}_n))$$

Để giải phương trình được $\frac{\partial LL(\boldsymbol{\beta})}{\partial \beta_m} = 0$ có dạng như trên thì tương đối khó khăn nên ta có thể sử dụng Newton-Raphson method hoặc Fisher-scoring method để tính $\hat{\boldsymbol{\beta}}$. Vì ở đây Poisson Regression sử dụng Canonical Link, nên hai phương pháp Newton-Raphson và Fisher-scoring là tương đương nhau.

Algorithm 1 Newton-Raphson method trong Count Data

Input: Tập dữ liệu $\mathbf{X} \in \mathbb{R}^{N \times d}$ và tập nhãn tương ứng $\mathbf{y} \in \mathbb{R}^{N \times 1}$

Output: $\boldsymbol{\beta}$

```

1: Khởi tạo  $\boldsymbol{\beta}$  bất kì
2: while  $\boldsymbol{\beta}$  chưa hội tụ do
3:   for  $i = 1$  to  $N$  do
4:      $g_i \leftarrow \sum_m x_{mi} (y_i - \exp(\boldsymbol{\beta}^\top \mathbf{x}_i))$ 
5:    $\mathbf{g} \leftarrow (g_1, g_2, \dots, g_N)^\top$ 
6:   for  $i = 1$  to  $N$  do
7:     for  $j = 1$  to  $N$  do
8:        $h_{ij} \leftarrow - \sum_n x_{ni} x_{nj} \exp(\boldsymbol{\beta}^\top \mathbf{x}_n)$ 
9:    $\mathbf{H} \leftarrow \begin{pmatrix} h_{11} & \dots & h_{1N} \\ \vdots & \ddots & \vdots \\ h_{N1} & \dots & h_{NN} \end{pmatrix}$ 
10:   $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} + \mathbf{H}^{-1} \mathbf{g}$ 
11: return  $\boldsymbol{\beta}$ 

```

3.3 Đánh giá mô hình

Công thức tính độ lệch là:

$$D = 2 \sum_n \left[y_n \log \left(\frac{y_n}{\mu_n} \right) - (y_n - \mu_n) \right].$$

Một cách khác để kiểm tra sự tốt của một mô hình là sử dụng Pearson's chi-squared, được định nghĩa như sau:

$$\chi_p = \sum \frac{(y_n - \mu_n)^2}{\mu_n}$$

3.4 Overdispersion

Đối với phân phối Poisson, phương sai $\text{Var}(Y) = \mu$. Tuy nhiên, trên thực tế thì sai phương của dữ liệu thường vượt quá μ , tức là $\text{Var}(Y) > E(Y)$, và hiện tượng này được gọi là phân tán quá mức (overdispersion). Underdispersion cũng xảy ra nhưng không phổ biến.

3.4.1 Nguyên nhân

Hiện tượng này có thể xảy ra do nhiều nguyên nhân như là:

- Giá trị kì vọng μ vẫn còn thay đổi, không phải cố định dù được sinh bởi x_i cố định.
- Trong assumption của GLM, các biến Y là độc lập với nhau, nhưng có thể ở đây Y vẫn có mối quan hệ nào đó. Ví dụ: sự không làm việc, không nộp report đúng hạn, sự không theo hướng dẫn của mentor đều là các yếu tố tạo nên tai nạn sự cố, 3 hành vi tội tệ này đều không độc lập với nhau, như mô hình Poisson yêu cầu mà có mối quan hệ tương quan đồng biến. Điều này dẫn đến sự phân bố của các dữ liệu đếm mà quá thay đổi/thay đổi quá rộng đối với mô hình Poisson.

3.4.2 Hậu quả

Overdispersion có thể có hoặc không ảnh hưởng lên việc ước lượng tham số β , phụ thuộc vào bản chất của việc phân tán quá mức, nhưng sẽ đánh giá thấp sai số chuẩn (underestimated) và làm sai lệch kết luận cho sự hồi quy tham số (regression parameter).

Overdispersion có thể được phát hiện nhờ việc đánh giá mô hình, nếu deviance và Pearson statistic lớn hơn nhiều so với residual degrees of freedom (số chiều của biến độc lập X), thì hoặc là fitted model bị thiếu hoặc dữ liệu bị phân tán quá. Khi mà số đếm nhỏ, xấp xỉ tiệm cận của deviance và Pearson statistic không đáng tin, lúc này khó mà đánh giá overdispersion có xảy ra hay không

3.4.3 Giải pháp

Đầu tiên, chúng ta sẽ tìm hiểu quasi-likelihood cũng như extended quasi-likelihood.

Khi chỉ xác định được hàm phương sai $V(\cdot)$, tồn tại một hàm xác suất quasi thỏa mãn:

$$\frac{\partial \log \bar{P}(y; \mu, \phi)}{\partial \mu} = \frac{y - \mu}{\phi V(\mu)}.$$

Giả sử ta có các giá trị quan sát y_i , $E[y_i] = \mu_i$ và $\text{Var}[y_i] = \phi V(\mu_i)/\omega_i$. Và giả sử ta cũng có một link-linear predictor cho μ_i liên kết với tổ hợp tuyến tính của β_j , thỏa mãn GLM. Khi đó, hàm quasi-likelihood được xác định như sau:

$$\mathcal{Q}(y, \mu) = \sum_{i=1}^n \log \bar{P} \left(y_i; \mu_i, \frac{\phi}{\omega_i} \right)$$

Khi định nghĩa quasi-likelihood, chúng ta xét đạo hàm của $\log \bar{P}$ đối với μ chứ không phải ϕ . Do đó, hàm xác suất quasi chỉ xác định đối với những hạng tử không chứa μ . Để kết luận một hàm xác suất quasi hoàn thiện, the saddle point approximation có thể được sử dụng:

$$\log \tilde{P}(y; \mu, \phi) = -\frac{1}{2} \log[2\pi\phi V(y)] - \frac{d(y, \mu)}{2\phi}$$

Phương trình trên được gọi là hàm extended quasi-log-probability. Và extended quasi-likelihood được xác định như sau:

$$Q^+(y; \mu, \phi/\omega) = \sum_{i=1}^n \log \tilde{\mathcal{P}}(y_i; \mu_i, \phi/\omega_i)$$

Việc giải $\frac{dQ^+(y; \mu, \phi/\omega)}{d\mu} = 0$ nghiệm theo μ cũng giống với quasi-likelihood nhưng extended quasi-likelihood có 1 lợi thế là giải phương trình trên sẽ cho ước lượng độ lệch trung bình của ϕ .

Quay trở lại bài toán overdispersion. Có rất nhiều mô hình cụ thể khác nhau được xây dựng để giải quyết vấn đề này, và chúng ta có thể chia các cách tiếp cận thành hai nhóm:

1. Giả sử có thêm một dạng chung cho hàm phương sai, có thể bằng cách thêm các tham số mới.
2. Giả sử có một mô hình 2-bước cho biến phản hồi. Tức là, giả sử bản thân tham số cũng tuân theo một phân phối nào đó.

Mô hình thuộc loại 1 sẽ không tương ứng với một phân phối xác suất cụ thể nào của biến phản hồi, nhưng được xem như là một mở rộng hữu ích của mô hình cơ bản. Việc ước lượng tham số hồi quy có thể sử dụng phương pháp quasi-likelihood (sẽ được trình bày ở phần sau). Mô hình thuộc loại 2 dẫn đến một mô hình xác suất phức hợp cho biến phản hồi và nói chung, tất cả các tham số đều có thể ước lượng nhờ sử dụng MLE. Tuy nhiên, nhìn chung, phân phối phức hợp thu được thường có dạng khá phức tạp, và thường sẽ sử dụng phương pháp ước lượng xấp xỉ.

Mô hình và ước lượng:

Giả sử ta có các biến ngẫu nhiên là $Y_i, i = \overline{1, n}$ khi biết X_i , với giá trị kì vọng μ_i . Mô hình tiêu chuẩn dùng phân phối Poisson sẽ giả sử $Y_i \sim \mathcal{P}(\mu_i)$ với hàm phương sai $\text{Var}[Y_i] = \mu_i$. Khi xảy ra phân tán quá mức, ta cần một hàm phương sai mà dự đoán một sự biến đổi lớn hơn (greater variability). Bằng cách thêm một tham số ϕ , ta có một mô hình đơn giản là: $\text{Var}[Y_i] = \phi\mu_i$ và chúng ta sẽ dùng quasi-likelihood để ước lượng.

1. Negative binomial type variance:

Một mô hình 2-bước giả sử rằng $Y_i \sim \mathcal{P}(\theta_i)$, trong đó, θ_i là các biến ngẫu nhiên với $E(\theta_i) = \mu_i, \text{Var}[\theta_i] = \sigma_i^2$. Và ta luôn có: $E(Y_i) = \mu_i, \text{Var}[Y_i] = \mu_i + \sigma_i^2$ cho một mô hình phân tán quá mức. Hơn nữa, nếu θ_i có hệ số biến thiên không đổi, σ^2 , sẽ dẫn đến $\text{Var}[Y_i] = \mu_i + \sigma^2\mu_i^2$, một hàm bậc hai của phương sai. Để mô hình xác định cụ thể, một giả sử phổ biến thường dùng là θ_i tuân theo phân phối $\Gamma(k, \lambda_i)$, dẫn đến Y_i tuân theo phân phối negative binomial với $E[Y_i] = \frac{k}{\lambda_i} = \mu_i$ và $\text{Var}[Y_i] = \mu_i + \frac{\mu_i^2}{k}$.

Với mỗi giá trị k cố định, MLE cho phân phối negative binomial cũng tương tự như các phân phối khác thuộc họ exponential và ước lượng tham số hồi quy β có thể dùng thuật toán chuẩn iteratively re-weighted least-squares (IRLS) cho GLM. Để ước lượng k , chúng ta có thể dùng Newton-Raphson cho score equation, và lặp đi lặp lại quá trình ước lượng β và k , ta có MLE chung cho cả mô hình.

Sự tiệm cận độc lập (asymptotic independence) của $\hat{\beta}$ và \hat{k} đồng nghĩa là độ lệch chuẩn của $\hat{\beta}$ sau khi fit IRLS là chính xác. Sử dụng hàm extended quasi-likelihood (EQL) cho hàm phương sai của phân phối negative binomial để biểu diễn sự nhập nhằng do các cách phân tích khác nhau của $\text{Var}[Y_i] = \phi_i V(\mu_i)$, có 3 khả năng xảy ra sau:

$$(a) \phi_i = 1, V(\mu_i) = \mu_i + \frac{\mu_i^2}{k}$$

$$(b) \phi_i = 1 + \frac{\mu_i}{k}, V(\mu_i) = \mu_i$$

$$(c) \phi_i = \mu_i + \frac{\mu_i^2}{k}, V(\mu_i) = 1$$

Cả ba trường hợp trên đều dẫn đến phương trình ước lượng khác nhau của β , xác định các phương án lập khác nhau nhưng chúng đều cho ước lượng như nhau, và một cách tinh tế để tiếp cận là sử dụng quasi-likelihood với hàm phương sai của negative binomial. Tuy nhiên, việc ước lượng k lại không đơn giản và mỗi công thức khác nhau cho một ước lượng khác nhau. Nếu sử dụng khả năng a thì phương trình ước lượng của k sẽ có dạng của negative binomial score equation. Trong trường hợp b, c k xuất hiện trong tham số tỉ lệ (scale parameter) và ta thu được phương trình ước lượng gamma. Ở b, công thức tính độ lệch Poisson được sử dụng với y là biến, trong khi ở c, chúng ta sử dụng Poisson Pearson residual, với phương trình ước lượng có dạng như là:

$$\sum_{i=1}^n \left[\frac{(y_i - \mu_i)^2}{\mu_i(1 + \frac{\mu_i}{k})} - 1 \right] \frac{\partial \log(1 + \frac{\mu_i}{k})}{\partial k} = 0$$

Phương pháp moment (moment method) cho phương trình ước lượng không bị chệch:

$$\sum_{i=1}^n \left[\frac{(y_i - \mu_i)^2}{\mu_i(1 + \frac{\mu_i}{k})} - 1 \right] = 0$$

Từ phương trình này có thể dễ dàng giải ra k nhờ Newton-Raphson.

2. Poisson-normal và những mô hình liên quan:

Chúng ta xem xét thêm vào yếu tố ngẫu nhiên trong linear predictor. Sử dụng mô hình Poisson log-linear và yếu tố ngẫu nhiên tuân theo phân phối chuẩn ta thu được mô hình Poisson-normal. Hàm phương sai của mô hình này có dạng $\text{Var}[Y_i] = \mu_i + k'\mu_i^2$, tương tự của phân phối negative binomial.

3. Hàm phương sai tổng quát:

$$\text{Var}[Y_i] = \mu_i \{1 + \phi \mu_i^\delta\}$$

Công thức tổng quát này có thể dùng để cung cấp thông tin likelihood cho các tham số thêm vào ϕ và δ , cái mà cho phép ta so sánh giữa các hàm phương sai.

3.5 Cài đặt

Trong Python, có hai thư viện chính đã cài sẵn mô hình Poisson để hỗ trợ người dùng cài đặt trực tiếp lên dataset là thư viện statsmodels. với hàm GLM. và thư viện scikit-learn. với hàm PoissonRegressor.

3.5.1 Thư viện statsmodels

```
1 # X là dữ liệu và y là nhãn tương ứng
2 import statsmodels as sm
3 X = sm.add_constant(X)
4 model = sm.GLM(y, X, family=sm.families.Poisson()).fit()
```

3.5.2 Thư viện `scikit-learn`

Hàm `PoissonRegressor` chỉ xuất hiện từ phiên bản 0.23, do đó, cần đảm bảo thư viện đã được cập nhật phiên bản mới nhất trước khi sử dụng hàm này.

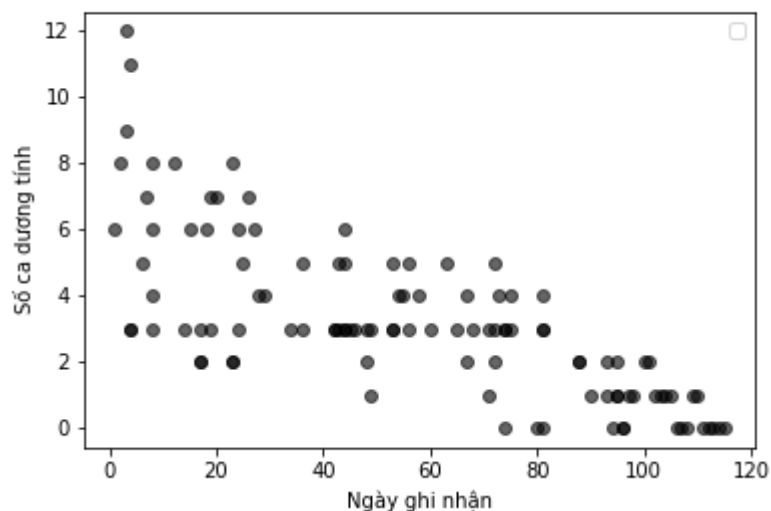
```
1 from sklearn.linear_model import PoissonRegressor
2 NUM_ITERS = 1000
3 model = PoissonRegressor(max_iter=NUM_ITERS).fit(X, y)
```

3.6 Ví dụ minh họa

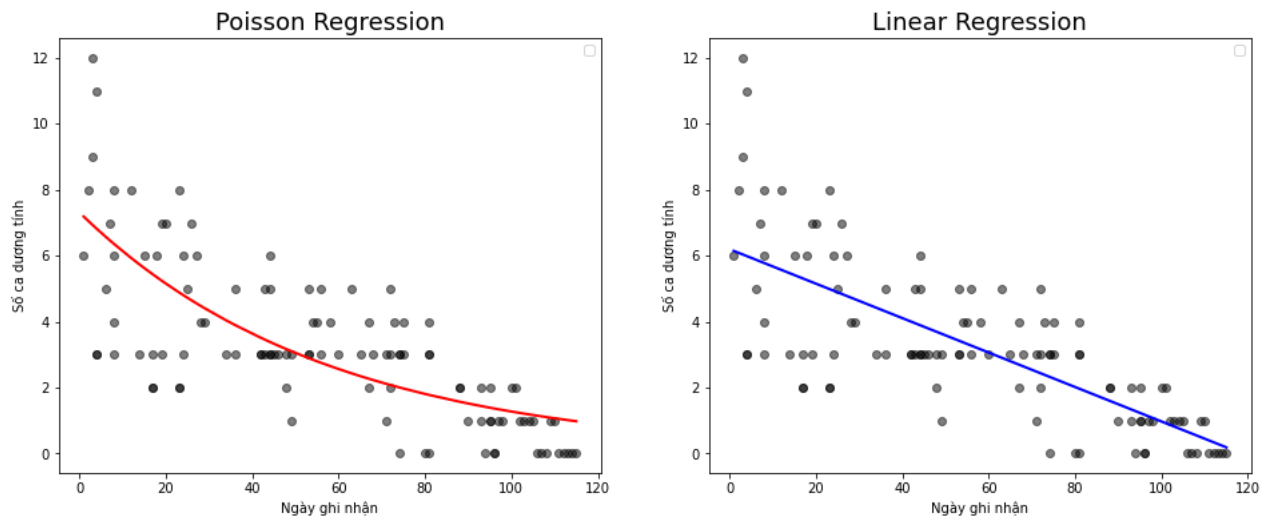
Trong các ví dụ dưới đây, bên cạnh Poisson Regression, ta cũng sẽ cài đặt Linear Regression để so sánh hai mô hình này và đưa ra những nhận xét về các ưu, khuyết điểm của từng mô hình với loại dữ liệu count data.

3.6.1 Bộ dữ liệu `students`

Bộ dữ liệu `students` mô tả số học sinh dương tính với một loại truyền nhiễm tính theo từng ngày kể từ khi bùng phát dịch. Dataset này gồm 2 thông số là `day` và `cases` lần lượt mô tả ngày được ghi nhận và số ca dương tính tại thời điểm đó. Biểu đồ dưới đây mô tả phân bố của các điểm dữ liệu theo thời gian:



Cài đặt mô hình Poisson Regression và Linear Regression trên dữ liệu trên, ta thu được best fit line như sau:



3.6.2 Bộ dữ liệu DoctorVisits

Bộ dữ liệu DoctorVisits mô tả các thông số về tuổi tác (age), bệnh lý (illness), năng suất hoạt động (reduced),... Ta cần dự đoán số lần đi khám bác sĩ trong 2 tuần vừa qua của mỗi đối tượng.

	visits	gender	age	income	illness	reduced	health	private	freepoor	freerepat	nchronic	lchronic
1	1	1	0.19	0.55	1	4	1	1	0	0	0	0
2	1	1	0.19	0.45	1	2	1	1	0	0	0	0
3	1	0	0.19	0.90	3	0	0	0	0	0	0	0
4	1	0	0.19	0.15	1	0	0	0	0	0	0	0
5	1	0	0.19	0.45	2	5	1	0	0	0	1	0
...
5186	0	1	0.22	0.55	0	0	0	0	0	0	0	0
5187	0	0	0.27	1.30	0	0	1	0	0	0	0	0
5188	0	1	0.37	0.25	1	0	1	0	0	1	0	0
5189	0	1	0.52	0.65	0	0	0	0	0	0	0	0
5190	0	0	0.72	0.25	0	0	0	0	0	1	0	0

5190 rows x 12 columns

Để kiểm tra độ chính xác của từng mô hình, ta chia bộ dữ liệu thành 2 phần: một phần chứa $\frac{3}{4}$ kích thước dữ liệu gốc để fit với mô hình, phần còn lại sử dụng để đánh giá mô hình. Cài đặt mô hình Linear Regression, Poisson Regression (scikit-learn) và GLM (statsmodels) trên dữ liệu trên và tính độ lệch so với nhãn chuẩn, ta thu được kết quả sau:

Mô hình	Deviance
Linear Regression	0.824381
Poisson Regression (scikit-learn)	0.824381
GLM (statsmodels)	0.879162

Nhìn chung, độ lệch của cả 3 mô hình so với nhãn thực tế đều không quá lớn để bị đánh giá là thiếu hiệu quả. Đây là vì biến phản hồi visits có các giá trị: $E[y] = 0.3, \text{var}[y] = 0.64$ - chênh lệch của hai thông số này nhỏ nên hiện tượng overdispersion không gây ảnh hưởng lớn đến sự sai lệch của mô hình Poisson

3.6.3 Bộ dữ liệu crab

Bộ dữ liệu `crab` được xây dựng để tìm những đặc tính ảnh hưởng đến số lượng sam đực bám vào cặp sam trong quá trình giao phối. Một số đặc tính được đề cập trong bộ dữ liệu gồm: màu của sam (C), tình trạng đuôi gai (S), chiều dài mai (W) và khối lượng (Wt).

	C	S	W	Wt	Sa
Obs					
1	2	3	28.3	3.05	8
2	3	3	26.0	2.60	4
3	3	3	25.6	2.15	0
4	4	2	21.0	1.85	0
5	2	3	29.0	3.00	1
...
169	2	3	28.3	3.20	0
170	2	3	26.5	2.35	4
171	2	3	26.5	2.75	7
172	3	3	26.1	2.75	3
173	2	2	24.5	2.00	0

173 rows × 5 columns

Ta thực hiện các bước tương tự ở bộ dữ liệu `DoctorVisits` để tính độ sai lệch của từng mô hình với bộ dữ liệu đã cho. Kết quả thu được như sau:

Mô hình	Deviance
Linear Regression	3.100807
Poisson Regression (scikit-learn)	3.100807
GLM (statsmodels)	3.114424

Độ lệch của cả ba mô hình đều tăng so với dữ liệu ban đầu. Đây là vì biến phản hồi `Sa` có các giá trị: $E[y] = 2.92$, $\text{Var}[y] = 9.85$ - chênh lệch của hai thông số này tăng lên đáng kể và hiện tượng overdispersion cũng đóng vai trò lớn hơn ảnh hưởng đến sai lệch của mô hình.

4 Áp dụng mô hình

Generalized Linear Model được ứng dụng trong nhiều lĩnh vực đa dạng như nông nghiệp, kinh tế, giáo dục, địa lý, sinh học, y văn. Ngoài ví dụ thực hành trên, một số ứng dụng khác của GLM bao gồm:

1. **Ứng dụng GLM trong phân tích xử lý nước uống:** Xem xét các biến như độ đục, màu của nước, nồng độ kiềm, độ pH và nhiệt độ nước. Biến phản hồi theo phân phối Bernoulli với giá trị 1 có nghĩa “Áp dụng vôi để lọc nước” và 0 là ngược lại. Ở đây hàm liên kết được sử dụng là hàm logit
2. **Áp dụng GLM cho phân phối hình học (Geometric Distribution):** Nghiên cứu mô hình dự đoán lần tử vong đầu tiên của trẻ sơ sinh dựa theo thứ tự sinh với biến phản hồi tuân theo Geometric Distribution. Các yếu tố được nghiên cứu bao gồm: trình độ giáo dục của người mẹ, chỉ số giàu nghèo, giới tính sinh học của trẻ, tuổi khi sinh của người mẹ, y văn. Ở đây, hai hàm liên kết được đưa ra để so sánh là **canonical function rút ra từ công thức của Geometric Distribution** và **hàm log-link** với kết luận hàm log-link (non-canonical trong trường hợp này) thể hiện tốt hơn.
3. **Áp dụng trong lĩnh vực sinh tin (Bioinfo):** Nghiên cứu về single cell RNA sequencing. Cụ thể trong đó có nghiên cứu về việc biểu diễn 1 cell (tế bào) bằng mức biểu hiện gen (gene expression) (tức là 1 cell bằng 1 vector N chiều ứng với gene expression của N cell). Mức biểu hiện gen là số lượng gen đó đếm được trong tế bào đó và mỗi cell thì lại được phân lớp vào kiểu tế bào T (T-cell), tế bào hồng cầu,... GLM được dùng để fit vào một tập dữ liệu nhiều cell, để xem có thể rút được thông tin gì về liên kết giữa loại tế bào và các gen mà nó biểu hiện không.

5 Kết luận đánh giá

Một trong những ưu điểm lớn nhất của GLM là biến lỗi ϵ không phải tuân theo phân phối chuẩn. Phương pháp ước lượng Least Square cho tham số xuất phát từ việc tối ưu hoá tổng của lỗi bình phương, tuy nhiên việc này đòi hỏi chúng ta phải giả sử trước phân phối chuẩn và phương sai nhất định cho các biến lỗi và biến phản hồi. Đối với một số loại data, biến phản hồi có thể theo một kiểu phân phối khác. Để ước lượng được tham số, các mô hình trước GLM cần phải biến đổi (transform) biến y sao cho y theo phân phối chuẩn. Trong thực tế, việc tìm cách biến đổi y để đảm bảo phân phối và phương sai nhất định của một mô hình không hề dễ dàng. Cách biến đổi để đạt phân phối chuẩn tốt nhất thường không nhất quán với cách biến đổi để đạt được phương sai xác định. Hơn nữa, các phương pháp biến đổi cũng có thể không phù hợp với boundary của không gian mẫu, chẳng hạn như áp dụng biến đổi log lên biến có giá trị 0.

Đối với GLM, sự lựa chọn link function và random component là độc lập với nhau. Phương pháp Maximum Likelihood Estimate không giới hạn với một phân phối cụ thể nào mà sử dụng được với toàn bộ phân phối họ exponential, trong khi việc lựa chọn link function chỉ nhằm mục đích đưa ra được mối quan hệ tuyến tính giữa hai thành phần với nhau. Ngoài ra, sử dụng hàm thuận và ngược của link function còn giúp chúng ta tìm ra sự ảnh hưởng từ biến độc lập lên giá trị trung bình của biến phản hồi và **ngược lại**. GLM đã đưa ra một lý thuyết tổng quát hoá cho đa số các mô hình quan trọng, ứng dụng được với cả biến phản hồi liên tục và rời rạc.

Tài liệu

- [1] Alan Agresti. *Foundations of Linear and Generalized Linear Models*. 2015.
- [2] Peter K. Dunn and Gordon K. Smyth. *Generalized Linear Models With Examples in R*. 2018.
- [3] Borhan Siddika và M. Ataharul Islam Farzana Jahan. *An Application of Generalized Linear Model for the Geometric Distribution*.
- [4] William Fleshman. *Linear Regression article*.
- [5] Christophe Hurlin. *Lecture Maximum Likelihood Estimation*. University of Orleans.
- [6] P. McCullagh and John A. Nelder. *Generalized Linear Models*. 1989.
- [7] Andrew Rothman. *Generalized Least Squares (GLS): Relations to OLS & WLS parts*. Towards Data Science Website.
- [8] Andrew Rothman. *Generalized Linear Models parts*. Towards Data Science Website.
- [9] John Hinde và Clarice G.B. Demétrio. *Overdispersion: Models and estimation*.
- [10] Zuma Cepeda và Edilberto Cepeda. *Application of Generalized Linear Model to Data Analysis in Water Treatment*.