Introduction to Generalized Linear Models

Victor Huang

Data Scientist

victor@ritual.co

August 8, 2019

Overview

1. Theory Behind GLM
   - Simple Linear Regression Model
   - Exponential Family
   - Link Function
   - Iterative Weighted Lease Square
2. A Numerical Example About Poisson Regression
   - Problem Discription
   - Solution
   - Interpretation

Simple Linear Regression Model

In stats 101, we learned that simple linear regression can be written in this way (Assume we have only three variables):
$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$
And based on Gauss Markov Theorem, we have the following properties about this model, in both one more multi variables regression model.

- $E(\epsilon_i) = 0$
- $Var(\epsilon_i) = \sigma^2$
- $Cov(\epsilon_i, \epsilon_j) = 0$

In general, the error follows the Gaussian distribution and are identical and independent distributed (i.i.d).

Solution of Simple Linear Regression

Let $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{m \times k}$, the general problem is

$$\min_{x \in \mathbb{R}^{n \times k}} \|AX - B\|_2$$

By leveraging the power of computer, we can solve simple linear regression easily today. The analytical solution looks like this:

$$X = A^{\dagger}B + (I - A^{\dagger}A)Y$$

$A^{\dagger}$ is penrose pseudo inverse. And when A is full rank, it becomes the solution we are familar with.

Potential Limitations of Simple Linear Regression

Distribution of Dependent Variable

Even though simple linear regression's independent variables can be continuous, categorical or mixed, its dependent variable must be continuous and follow Gaussian distribution.

Relationship Between Mean and Variance of Dependent Variable

Gaussian distribution is very special because the mean and variance are dependent. If $y \sim N(\mu, \sigma^2)$, then $E(y) = \mu$ And $Var(y) = \sigma^2$. But for binomial distribution, $Var(y) = \mu(1 - \mu)$.

Generalized Linear Model

To solve the problem listed above, this paper and the following people developed a model fitting process.

1. Model Specification
   - probability distribution of the response variable
   - an equation linking the response and explanatory variables
2. Parameter Estimation
3. Adequecy Checking – how well it fits or summarize the data
4. Inference – For classical or freqeuntist inference this involves calculating confidence intervals, testing hypotheses about paramteers in the model and interpreting the results.

Exponential Family Distribution

- The PDF for normal distribution looks like this:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

- And the PDF for Poisson Distribution:
  $P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$

And here is the general form of exponential familty distribution, include normal distribution, gamma distribution, poisson distribution and binomial distribution.

$$f(y;\theta) = e^{a(y)b(\theta)+c(\theta)+d(y)}$$

And $a(y) = y$, the distribuition called canonical form. For normal distribition, all paramters are the following:

- $b = \frac{\mu}{\sigma^2}$
- $c = -\frac{\mu^2}{2\sigma^2} - \frac{1}{2}log(2\pi\sigma^2)$
- $d = -\frac{y^2}{2\sigma^2}$

Exponential Family Distribution

Proof.

The probability density function of normal distribution is

$$P(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/2\sigma^2}$$

And it is equal to

$$P(y) = (\frac{1}{2\pi\sigma^2})^{\frac{1}{2}} exp[-\frac{1}{2\sigma^2}(y-\mu)^2] = (\frac{1}{2\pi\sigma^2})^{\frac{1}{2}} exp[-\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}]$$

This can be rewritten as

$$f(y;\mu) = exp[-\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}log(2\pi\sigma^2)]$$

□

Similarly, Binomial and Poisson distribution can be rewritten into this family.

Link Function

Link function connects response and independent variables. And the table below summarized various models and its corresponding link functions. And the random part with systematic part.

Model	**Random**	**Link**	**Systematic**
Linear Regression	Normal	Identity	Continuous
ANOVA	Normal	Identity	Categorical
ANCOVA	Normal	Identity	Mixed
Logistic Regression	Binomial	Logit	Mixed
Loglinear	Poisson	Log	Categorical
Poisson Regression	Poisson	Log	Mixed
Multinomial response	Multinomial	Generalized Logit	Mixed

Simple Example

The best known special case of generalized linear might be simple linear model.
$E(Y_i) = \mu_i = \mathbf{x}_i^T \beta;\ Y_i \sim N(\mu_i, \sigma^2)$
Where $Y_1, ..., Y_N$ are independent. In this case, link function is the identity function. $g(\mu_i) = \mu_i$. And this model can be rewritten to the form mentioned in the beginning, which is

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

- systematic part: $\mu = X\beta$
- random part: ϵ

Estimation

In GLM, people use maximum of likelihood to estimate parameters. The complete derivation is time consuming so I just list the initial step and the result here. And I will elaborate this by a numerical example.
To estimate parameters, we need likelihood function. For each Y_i, we have the log-likelihood function

$$l_i = y_i b(\theta_i) + c(\theta_i) + d(\theta_i)$$

And with one full page derivation, we have the iterative equation:

$$X^T W X b^{(m)} = X^T W z$$

And the method is called iterative weighted least square.

Poisson Regression Example

Some fake data for testing purposes.

y_i	2	3	6	7	8	9	10	12	15
x_i	-1	-1	0	0	0	0	1	1	1

Based on some simple calulcation, we use Poisson distribution to model it.
For Poisson distribution, the mean equals to the variance.
For simplicity, we use linear relationship:
$E(Y_i) = \mu_i = \beta_1 + \beta_2 x_i = \mathbf{x_i^T}\beta$
Where $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$ and $\mathbf{x_i} = \begin{bmatrix} 1 \\ x_i \end{bmatrix}$ for $i = 1, 2, ..N$. In this case, we take the identity link function $g(\mu_i)$ to be the identity function.

$$g(\mu_i) = \mu_i = \mathbf{x_i^T}\beta = \eta_i$$

.
For this reason, $\frac{\partial \mu_i}{\partial \eta_i} = 1$.

Poisson Regression Example Continued

Some useful results:

- Information Matrix

$$\mathfrak{J} = \mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{X}$$

- $\mathbf{W}$ is $N \times N$ matrix with elements

$$w_{ii} = \frac{1}{var(Y_i)}(\frac{\partial \mu_i}{\partial \eta_i})^2$$

- $z_i = \sum_{k=1}^{p} x_{ik} b_k^{m-1} + (y_i - \mu_i)(\frac{\partial \eta_i}{\partial \mu_i})$
- For Poisson distribution, $E(Y_i) = Var(Y_i)$

Poisson Regression Example Continued

Based on previous results, in this specific problem, w_{ii} can be simplified:

$$w_{ii} = \frac{1}{var(Y_i)}(\frac{\partial \mu_i}{\partial \eta_i})^2 = \frac{1}{var(Y_i)} = \frac{1}{\beta_1 + \beta_2 x_i}$$

By applying $\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$ for β, we got

$$z_i = b_1 + b_2 x_i + (y_i - b_1 - b_2 x_i) = y_i$$

Also, we have

$$\mathfrak{J} = \mathbf{X^T W X} = \begin{bmatrix} \sum_{n=1}^{N} \frac{1}{b_1 + b_2 x_i} & \sum_{n=1}^{N} \frac{x_i}{b_1 + b_2 x_i} \\ \sum_{n=1}^{N} \frac{x_i}{b_1 + b_2 x_i} & \sum_{n=1}^{N} \frac{x_i^2}{b_1 + b_2 x_i} \end{bmatrix}$$

Poisson Regression Example Continued

also

$$\mathbf{X^T W z} = \begin{bmatrix} \sum_{n=1}^{N} \dfrac{y_i}{b_1 + b_2 x_i} \\ \sum_{n=1}^{N} \dfrac{x_i y_i}{b_1 + b_2 x_i} \end{bmatrix}$$

Then, the MLE are obtained iteratively from the equations

$$(X^T W X)^{m-1} b^{(m)} = X^T W z^{m-1}$$

Poisson Regression Example Continued

For these data, N = 9

$$\mathbf{y} = \mathbf{z} = \begin{bmatrix} 2 \\ 3 \\ \dots \\ 15 \end{bmatrix}$$

and

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \dots \\ \mathbf{x}_9^T \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & -1 \\ \dots & \dots \\ 1 & 1 \end{bmatrix}$$

Poisson Regression Example Continued

In this example, we choose initial estimates $b_1^{(1)} = 7$ and $b_2^{(1)} = 5$ And therefore,

$$(\mathbf{X}^T\mathbf{W}\mathbf{X})^{(1)} = \begin{bmatrix} 1.8214 & -0.75 \\ -0.75 & 1.25 \end{bmatrix} \qquad (\mathbf{X}^T\mathbf{W}\mathbf{z})^{(1)} = \begin{bmatrix} 9.869048 \\ 0.583333 \end{bmatrix}$$

So

$$\mathbf{b}^{(2)} = [(\mathbf{X}^T\mathbf{W}\mathbf{X})^{(1)}]^{-1}(\mathbf{X}^T\mathbf{W}\mathbf{z})^{(1)} = \begin{bmatrix} 0.729167 & 0.4375 \\ 0.4375 & 1.0625 \end{bmatrix} \begin{bmatrix} 9.8690 \\ 0.5833 \end{bmatrix}$$

$$= \begin{bmatrix} 7.4514 \\ 4.9375 \end{bmatrix}$$

The process is continued untile it converges.

Poisson Regression Example Continued

And the successive appromate for the regression coefficients looks like this

m	1	2	3	4
$b_1^{(m)}$	7	7.45139	7.45163	7.45163
$b_2^{(m)}$	5	4.93750	4.93531	4.93530

And the information matrix is

$$\tilde{\jmath}^{-1} = \begin{bmatrix} 0.7817 & 0.4166 \\ 0.4166 & 1.1863 \end{bmatrix}$$

for that reason, the standard errors $b_1 = \sqrt{0.7817} = 0.8841$ and $b_2 = \sqrt{1.1863} = 1.0892$

By using the language R, we get the following results:

	Dependent variable:
	y
x	4.935^{***}
	(1.089)
Constant	7.452^{***}
	(0.884)
Observations	9
Log Likelihood	-18.004
Akaike Inf. Crit.	40.008
Note:	*p<0.1; **p<0.05; ***p<0.01

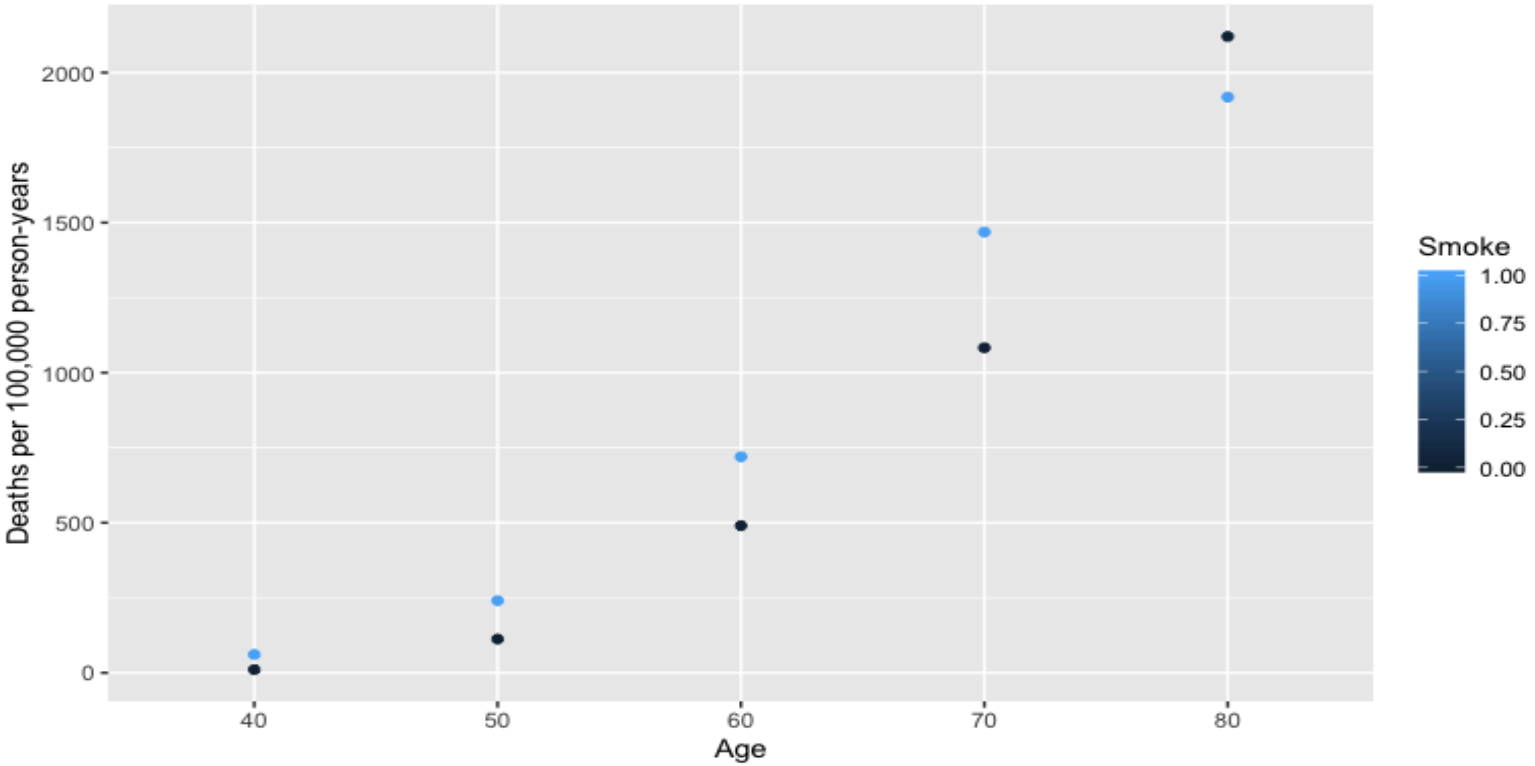
Problem

In 1951, all British doctors were sent a brief questionnaire about whether they smoked tobacco. Since then information about their deaths has been collected. The table below shows the numbers of deaths from coronary heart disease among make doctors 10 years after the survey. It also shows the total number of person years of observations at the time of the analysis.

	Age	Smoke	Person-Year	Deaths
1	40	0.00	18790.00	2.00
2	50	0.00	10673.00	12.00
3	60	0.00	5710.00	28.00
4	70	0.00	2585.00	28.00
5	80	0.00	1462.00	31.00
6	40	1.00	52407.00	32.00
7	50	1.00	43248.00	104.00
8	60	1.00	28612.00	206.00
9	70	1.00	12663.00	186.00
10	80	1.00	5317.00	102.00

Interesting Questions

- Is the death rate higher for smokers than non-smokers?
- If so, by how much?
- Is the differential effect related to age?



Deaths rates from coronary heart disease per 100,00 person-years for smokers(blue) and non-smokers(dark)

Model Approach

Various models can be used to solve this problem. Here is one simple approach. The form is

$log(deaths_i) =$

$log(personyears_i) + \beta_1 + \beta_2 smoke_i + \beta_3 age_i + \beta_4 agesq_i + \beta_5 smkage_i$

For different variables, their meanings are:

- personyear: how many people are in this age group
- smoke: boolean variable, Whether this is a smoker or non-smoker
- age: the age category
- agesq: sqaure of age
- smkage: equal to age for smokers and 0 for non-smokers

Model Interpretation

This chart stores all the data generated by this model (R/Python code is very straight forawrd based on the fomular i described before)

Thanks For Your Time