STAT2014/6014 Lecture Week 6 GLM Introduction and exponential family

Lucy Hu
Email: yunxi.hu@anu.edu.au [1]

[1] Research School of Finance, Actuarial Studies and Statistics
Australian National University

Plan for Week 6:

Slides
- Week 6 GLM overview (1st hour)
- Week 7 $\boxed{T1}$ Binary regression (2nd hour) (modelling part) $Y = 0/1$

Exercise (self-study)
- A GLM Exercise - EFD / MLE, Binary regression
- R script for week7

- Regression: relation b/w Y & Xs
  - e.g: LR : SL relation μ_Y & Xs
  - e.g: GLM : non SL μ_Y & Xs
- Distribution: Y
  - e.g: Y ~ ND
  - e.g: Y ~ Poisson
    Y ~ Bernoulli
    ...

1. ANU STAT3015/7030 Lecture Notes
2. ANU STAT3008/7001 Lecture Notes
3. **Julian J.Faraway**
   Extending the Linear Model with R


Done!
Revision
w10
• Data (non-aggregate vs aggregate data)
• Diagnostics
• Deviance, dispersion and residuals
• Goodness of fit (Hypothesis 0, 5)
• Wald test – Hypothesis 1 (for 1 single beta)
• Drop in deviance (Hypothesis 2,3,4)
Hypothesis testing (Binomial, Poisson)
w11
• Motivation
• Link function and inverse link function
• Prediction
• Deviance and residual
• Hypothesis testing
Gamma regression with inverse link
w8
• Aggregate data vs non-aggregate data
• Binomial regression
• Binomial distribution
• Link function and inverse link function
• MLE and prediction
Binomial regression with logistic link
w9
• Binomial to Poisson distribution
• Poisson distribution
• Mean, variance
• Link function and inverse link function
• MLE and prediction
Poisson regression with log-linear link
w6
• Why to use GLM
• GLM components
• Linear predictor
• Link function and canonical link function
• Random component – Exponential family
• Different types of GLM
• Obtaining estimated parameters
GLM theory and exponential family
w7
• Mean, variance
• Link function and inverse link function
• Odds ratio and interpretation of coefficients
• MLE and prediction
• Deviance
• Hypothesis testing
Binary regression with logistic link
STAT6038
Linear Regression
(skip)

Week 6: GLM theory and exponential family

- GLM (Generalized linear model)
    - Why to use GLM
    - GLM components
        - The systematic component: Linear predictor
        - The link function and canonical link function
        - The random component: Exponential family
    - Different types of GLM
    - Obtaining estimated coefficients

(skip)

/

Why GLM ?

- From Week 1 to Week 6 in STAT6038/2008's classes, we learnt how to model data whose response variable $Y_i|X_{1_i} = x_{1_i}, ..., X_{k_i} = x_{k_i}$ **follows a** $N(\mu_{Y_i} = x_i^T\beta, \sigma^2)$ **distribution**.
- From Week 6 onwards in STAT6014/2014's classes, we will focus on modelling data whose response variable $Y_i|X_{1_i} = x_{1_i}, ..., X_{k_i} = x_{k_i}$ (or $Z_i|X_{1_i} = x_{1_i}, ..., X_{k_i} = x_{k_i}$) does **not necessarily follow a** $N(\mu_{Y_i} = x_i^T\beta, \sigma^2)$ **distribution** - we introduce Generalized linear model (GLM).

(skip)

We have covered SLR (Simple linear regression) so far.

For LR (Linear regression):

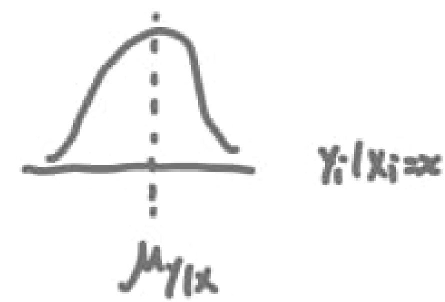
- Population regression line: $\mu\{Y_i|X_{1_i},...,X_{k_i}\} = \mu_{Y_i} = \beta_0 + \beta_1 X_{1_i} + ... + \beta_k X_{k_i} = x_i^T\beta$.
- Observation: $\{Y_i|X_{1_i},...,X_{k_i}\} = Y_i = \beta_0 + \beta_1 X_{1_i} + ... + \beta_k X_{k_i} + \varepsilon_i = x_i^T\beta + \varepsilon_i$, where $\varepsilon_i \sim N(0,\sigma^2)$, and $Y_i|X_{1_i} = x_{1_i},...,X_{k_i} = x_{k_i} \sim N(\mu_{Y_i} = x_i^T\beta, \sigma^2)$.

One of the underlying assumptions in LR model is that the response variable $Y_i|X_{1_i} = x_{1_i},...,X_{k_i} = x_{k_i}$ is **normally distributed**.

This means Y_i is continuous, and can be either greater or smaller than zero.

?

- L.R assumptions ?
  - $\varepsilon_i \sim$ idp, constant variance, ND
  - $\Downarrow$
  - $Y_i \mid X_i = x \sim$ ND
    - given X_i, regression
    - $Y_i \sim$ con. bell shaped, $\in \mathbb{R}$



- L.R: $Y_i \mid X_i = x \sim$ ND
  GLM: $Y_i \mid X_i = x \sim$ not necessarily ND $\to$ exponential family dsb (EFD)   subset

- e.g.:

LR: $Y \mid X_i = x$: exam grades $\rightarrow$ ND; x_i = gender   } lm(Y ~ X_1)

GLM: $Y \mid X_i = x$: whether gets an HD (0/1) $\rightarrow$ binary; x_i = gender   } glm(Y ~ X_1)

What if. . . $Y_i | X_i \sim$ Not ND ?

- The response variable Y_i is a categorical variable which only takes two possible values 0 and 1 (i.e. a binary variable). For example:
  1
  - In a study on the effectiveness of advertising, the response might be whether a given customer is willing to buy the new product.
  - In a study of home ownership, the response variable is whether a given individual owns a home.
- The response variable Y_i is a continuous variable which can only be non-negative. For example:
  2
  - Claim Severity. \$
  - Income per week.
- The response variable Y_i is a count variable which can only be $0, 1, 2, \ldots$ For example:
  3
  - The number of car accidents for Canberra over the past year.
  - The number of claims for an insurance company over the past one month (aggregate). #
  - The number of claims under one insurance policy over the past one month.

- Examples that Y is <u>not</u> ND $\boxed{?}$

$Y_i \mid X_i = x \sim$ not ND

① Y_i = whether^{0,1} to buy a product ↗ 1 (Yes) / ↘ 0 (No) } binary (discrete)
X_i = gender

② Y_i = \$claim$^{\geq 0}$ made by P.H in 2025 (claim severity) } Gamma (ctn.)
X_i = age

③ Y_i = # claims$^{0,1,2,\ldots}$ made by P.H in 2025 (claim frequency) } Poisson (discrete)
X_i = age

$Y_i \mid X_i \sim$ EFD:

- In all of those cases, the response variable $Y_i|X_{1_i} = x_{1_i}, ..., X_{k_i} = x_{k_i}$ (or sometimes $Z_i|X_{1_i} = x_{1_i}, ..., X_{k_i} = x_{k_i}$) belongs to the **exponential family** (e.g. Normal, Binary/Bernoulli, Binomial, Poisson, Gamma, Exponential distributions.). Note that Y_i can be either discrete or continuous.

- As a result, the response variable $Y_i|X_{1_i} = x_{1_i}, ..., X_{k_i} = x_{k_i}$ is **no longer normally distributed**.

- Note: The distribution for $Y_i|X_{1_i} = x_{1_i}, ..., X_{k_i} = x_{k_i}$ is a **conditional distribution**. For convenience, we **omit** the given condition $(X_{1_i} = x_{1_i}, ..., X_{k_i} = x_{k_i})$ from now on.

exponential family distribution

- $Y_i | X_i \sim$ EFD (broad name)
  - e.g: ND, Binary, exponential dsb
  - Y_i = discrete / ctn.

LR: $Y_i | X_i = x \sim ND \Rightarrow Y_i \sim ND$ (omit given X_i)

GLM: $Y_i | X_i = x \sim EFD \Rightarrow Y_i \sim EFD$

How to build a GLM model $\boxed{?}$

GLM components

- We will use GLM model (instead of LR model) to model response variable Y_i (or Z_i, both given $X = x$), where Y_i (or Z_i, both given $X = x$) is **not normally distributed**.

- Assume the response variable Y_i (or Z_i, both given $X = x$) belongs to the **exponential family**
  - i.e. $Y_i \sim$ some distribution with population mean μ_{Y_i}.
  - Link function: $g(\mu_{Y_i}) = \beta_0 + \beta_1 X_{1_i} + ... + \beta_k X_{k_i} = x_i^T \beta$.

- There are three components in GLM:
  - **Component 1**: The systematic component
    - linear predictor: $g(\mu_{Y_i}) = \eta_i = \beta_0 + \beta_1 X_{1_i} + \ldots + \beta_k X_{k_i} = x_i^T \beta$.
  - **Component 2**: The link function $g(\cdot)$
    - Aims to connect μ_{Y_i} to the linear predictor $x_i^T \beta$.
    - Why?
  - **Component 3**: The random component: the specified distribution for Y_i (or Z_i)
    - Y_i (or Z_i) belongs to the **exponential family**.
    - Examples of expontial family distribution: Normal, Binary/Bernoulli, Binomial, Poisson, Gamma, Exponential distributions.

- $y_i \mid X_i = x \quad \sim EFD$
- How to build a GLM model $\boxed{?}$

linear predictor : $g(\mu_{y_i}) = \Sigma \beta X$

a fun of μ_y

linear predictor $= \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$

link fun $g(\cdot)$: link μ_y & $\Sigma \beta X$

y_i (response) $\sim$ EFD

In LR:

- $\mu_{Y_i} = \beta_0 + \beta_1 X_{1_i} + ... + \beta_k X_{k_i} = x_i^T \beta$, which may take values of $(-\infty, \infty)$.

In GLM:

- Define a link function $g(\cdot)$, which aims to connect μ_{Y_i} to the linear predictor $x_i^T \beta$.
- i.e. $g(\mu_{Y_i})$ = a function of $\mu_{Y_i} = \eta_i = \beta_0 + \beta_1 X_{1_i} + ... + \beta_k X_{k_i} = x_i^T \beta$.
- This means we connect μ_{Y_i} to the linear predictor $x_i^T \beta$ through a **link function** $g(\cdot)$.

(skip)

$$\text{LR}: \underset{\substack{\Uparrow \\ \mu_y = E(Y) \\ = \mu_{Y|X} = E(Y|X)}}{\mu_y} = \underbrace{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}_{\text{linear predictor}} = \Sigma \beta X = X_i^T \beta$$

$$\text{GLM}: \underbrace{g(\mu_y)}_{} = \Sigma \beta X$$

e.g. $g(\mu_y) = \mu_y$

link function

μ_y = inverse link function

e.g. $\boxed{?}$

$$\underbrace{g(\mu_y) = \ln(\mu_y) = \Sigma \beta X}_{\text{link fun}: g(\mu_y) = \Sigma \beta X} \longrightarrow \underbrace{\mu_y = \exp(\Sigma \beta X)}_{\text{inverse link}: \mu_y = \cdots}$$

Why link fun $\boxed{?}$

$$\begin{cases} LR: & \mu_y = \Sigma \beta X \\ GLM: & g(\mu_y) = \Sigma \beta X \end{cases}$$

To restrict the possible μ_y ranges

$\downarrow$ e.g.

Let Y_i denote an insurance company's claim frequency of over the past one month ($Y_i = 0, 1, ...$). * ≥ 0

Since Y_i is a discrete variable and cannot be negative, the mean value of Y_i (μ_{Y_i}) cannot be negative.

If we use LR, we have $\mu_{Y_i} = \beta_0 + \beta_1 X_{1_i} + ... + \beta_k X_{k_i} = x_i^T \beta$. $\leftarrow = \Sigma \beta x$

- The left hand side is μ_{Y_i}, which must be non-negative.
- However, the right hand side can be negative (since either the coefficients β_i or X_i might be negative).
- Inconsistency!

Solution?

- Suppose $Y_i \sim Poi(\mu_{Y_i})$, we may use Poisson regression with log-linear link (a type of GLM) to model Y_i.
- Set the link function: $\ln(\mu_{Y_i}) = \beta_0 + \beta_1 X_{1_i} + ... + \beta_k X_{k_i} = x_i^T \beta$. It takes every real number.
- The inverse link function is therefore: $\mu_{Y_i} = exp(\beta_0 + \beta_1 X_{1_i} + ... + \beta_k X_{k_i}) = exp(x_i^T \beta)$, which must be non-negative. This meets the previous requirement – the mean value of Y (μ_{Y_i}) cannot be negative. We will have a further discussion later.

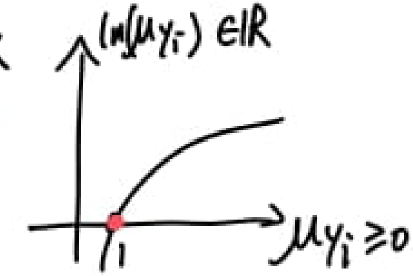
Data $\begin{cases} y_i = \text{claim frequency } (\#) \to \begin{cases} y_i = 0, 1, 2, \dots \\ y_i \geq 0 \to \mu_{y_i} \geq 0 \end{cases} \\ x_i = \text{gender, age, } \dots \end{cases}$

LR $\begin{cases} y_i \mid x_i = x \sim ND \\ \underbrace{\mu_{y_i}}_{\geq 0} = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k = \underbrace{\sum \beta x}_{\in \mathbb{R}} \to \text{not match} \end{cases}$

Solution?

$y_i \mid x_i = x \sim Poi(\mu_{y_i}) \to$ Poi regression with log-linear link

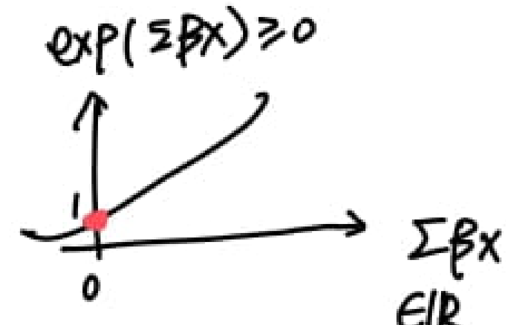
link fun: $g(\mu_{y_i}) = \underbrace{\ln(\mu_{y_i})}_{\in \mathbb{R}} = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k = \underbrace{\sum \beta x}_{\in \mathbb{R}}$



$\downarrow$ exp.

inverse link fun: $\underbrace{\mu_{y_i}}_{\geq 0} = \underbrace{\exp(\sum \beta x)}_{\geq 0}$

Match



/ So far:

- L.R: $Y_i \mid X_i = x \sim ND$

  $\mu_{Y_i} = \Sigma \beta X$

  (SL relation)

  special case: $g(\mu_Y) = \mu_Y = \Sigma \beta X$

- GLM: $Y_i \mid X_i = x \sim EFD$

  $g(\mu_{Y_i}) = \Sigma \beta X$ (link fun)

  (non-SL relation)

$\rightarrow \mu_{Y_i} = \cdots$ (inverse link fun)

/

GLM Terminologies

Y_i (or Z_i) $\sim$ EFD

- n: the number of observations (rows) in the given data (regardless of aggregate/non-aggregate data).
- k: the number of slopes $(\beta_1, ..., \beta_k)$.
- $p = k + 1$: the number of parameters $(\beta_0, ..., \beta_k)$.

- M_i: the number of trials/experiments under 1 observation (row) (for aggregate data only).
- Y_i: the response variable for non-aggregate data (continuous/discrete).
- Z_i: the response variable for aggregate data (if Y_i is discrete). $Z_i = \sum_{i=1}^{M_i} Y_i$. The number of success over M_i number of trials under 1 observation (row).

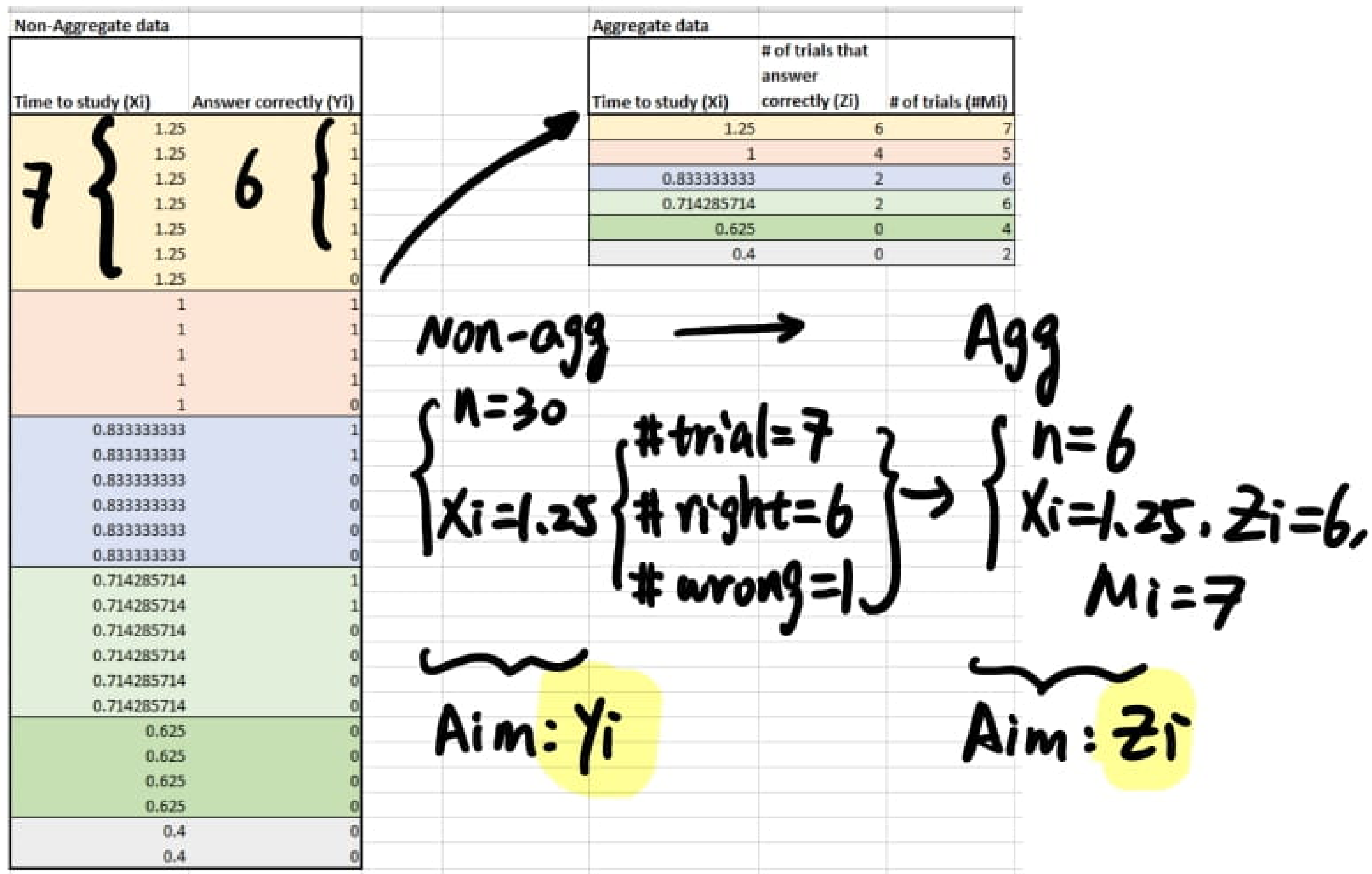
- LR $\begin{cases} n = \#\text{ rows} \\ k = \#\text{ slopes } (\beta_1, \beta_2, \cdots, \beta_k) \\ p = \#\text{ parameters } = k+1 = (\beta_0, \beta_1, \cdots, \beta_k) \end{cases}$

$$\Rightarrow \mu_{Y_i|X_i} = \mu_{Y_i} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

- In addition:

$\begin{cases} Z_i = \text{response for aggregate data} \\ Y_i = \text{. . . . . . . non-aggregate data} \end{cases}$ → Y_i (or Z_i) $\sim$ EFD

M_i = # trails under 1 obs. (aggregate only)



Non-Aggregate data

Time to study (Xi)	Answer correctly (Yi)
1.25	1
1.25	1
1.25	1
1.25	1
1.25	1
1.25	1
1.25	0
1	1
1	1
1	1
1	1
1	0
0.833333333	1
0.833333333	1
0.833333333	0
0.833333333	0
0.833333333	0
0.833333333	0
0.714285714	1
0.714285714	1
0.714285714	0
0.714285714	0
0.714285714	0
0.714285714	0
0.625	0
0.625	0
0.625	0
0.625	0
0.4	0
0.4	0

Aggregate data

Time to study (Xi)	# of trials that answer correctly (Zi)	# of trials (#Mi)
1.25	6	7
1	4	5
0.833333333	2	6
0.714285714	2	6
0.625	0	4
0.4	0	2

Figure 2: Data conversion

non-aggregate (idv-level record)

$n = \#$ student $= 30$

X_i = time to study math (in hours)

Y_i = answer 1 quiz question correctly or not : $\begin{cases} 1 \sim \text{Yes} \\ 0 \sim \text{No} \end{cases}$

aggregate level (group-level record)

Z_i = # students who answer the quiz correctly given the same X_i (hours to study)

M_i = # students attend the quiz given the same X_i (hours to study)

- $\underbrace{\text{Non-aggregate}}_{Y_i}$ data $\qquad$ $\underbrace{\text{Aggregate data}}_{Z_i}$

Y_i (or Z_i) $\sim$ EFD

GLM Types

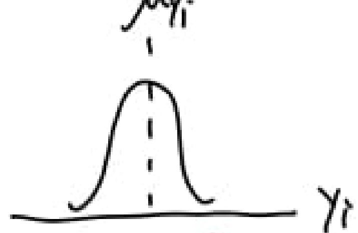
$$\underbrace{\text{Non-aggregate data}}_{Y_i} \longrightarrow \underbrace{\text{Aggregate data}}_{Z_i}$$

Y_i (or Z_i) $\sim$ EFD

$(y_i | x_i \sim ND)$

For x & y (or z)

- T_0 : $y_i \sim ND \rightarrow$



(show slides)

$\rightarrow$ linear regression

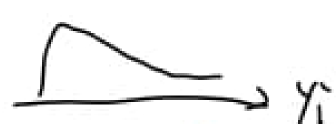
- T_1: $y_i \sim$ Binary $\rightarrow y_i = 0/1$ (show slides) $\rightarrow$ binary regression
  (Bernoulli)

- T_2: $z_i \sim$ Binomial $\rightarrow z_i = \sum_{i=1}^{M_i} y_i = 0, 1, 2, \cdots$ $\rightarrow$ binomial regression

- T_3: $z_i \sim$ Poisson $\rightarrow$ (show slides) $\rightarrow$ Poisson regression

- T_4: $y_i \sim$ Categorical $\rightarrow$ e.g: x_i = gender
  y_i = exam grade $\rightarrow$ multicategorical regression
  (HD, D, CR, P, F ⋯)
  no need modelling detail for exam

- T_5: $y_i \sim$ Gamma



(show slides)

$\rightarrow$ gamma regression

Component 2 & 3: Link function, Exponential family - Different types of GLM

Australian National University

We will introduce 6 types of exponential family distribution for Y_i (or Z_i) in this course:

- $\boxed{\text{T0}}$ (Type 0): Linear regression (Normal distribution with identity link)
  - $Y_i \sim N(\mu_{Y_i}, \sigma^2)$ (Normal distribution). ← bell shaped
  - Y_i is continuous, and $Y_i \in (-\infty, \infty)$.
  - Non-aggregate data: (X_i, Y_i).
  - i.e. This is LR (LR is a special case of GLM.). ← $\begin{cases} \text{LR}: \mu_Y = \Sigma\beta X \\ \text{GLM}: g(\mu_Y) = \Sigma\beta X \end{cases}$
- $\boxed{\text{T1}}$ (Type 1): Binary regression with logistic link
  - $Y_i \sim Bin(M_i = 1, \pi_i) = Bern(\pi_i)$ (Bernoulli/Binary distribution).
  - Y_i is discrete (categorical), which has two levels.
  - $Y_i = 0$ (a failure under one trial) or 1 (a success under one trial).
  - Non-aggregate data: (X_i, Y_i).
  - i.e. The Bernoulli distribution a special case of a Binomial distribution (when there is only one trial/experiment).

- $\boxed{\text{T2}}$ (Type 2): Binomial regression with logistic link
  - $Z_i \sim Bin(M_i, \pi_i)$ (Binomial distribution).
  - $Z_i = \sum_{i=1}^{M_i} Y_i$ = the total number of success (counts) over M_i number of trials.
  - Y_i is discrete (categorical), which has two levels: 0 or 1.
  - Aggregate data: (X_i, Z_i, M_i).

$M_i \to \infty$

$M_i \to NG$

- $\boxed{\text{T3}}$ (Type 3): Poisson regression with log-linear link
  - $Z_i \sim Poi(\mu_{Z_i})$ (Poisson distribution).
  - Z_i is the total number of success (counts) (the same as T2), but M_i (the total number of trials) is unknown (but goes to infinity).
  - Aggregate data: (X_i, Z_i).

e.g: claim frequency #

- T4 (Type 4): Multicategory regression
  - $Y_i \sim$ **Categorical distribution** $(\pi_i, .., \pi_C)$.
  - Y_i is discrete (categorical), which has C levels (level 1,2,. . . ,C).
  - Non-aggregate data: (X_i, Y_i).
  - *(Not examinable)* ← the modelling part is not needed
- T5 (Type 5): Gamma regression with inverse link (default: inverse link function)
  - $Y_i \sim Gam(\alpha_i, \beta_i)$ (Gamma distribution).

    e.g: claim severity $
  - Y_i is continuous, positively skewed and $Y_i \in [0, \infty)$.
  - Non-aggregate data: (X_i, Y_i).

Note that those are the response types and/or GLM (canonical) link functions we will focus on in this course. In reality there may be other response types and/or GLM link functions (e.g. $Z_i \sim Bin(M_i, \pi_i)$ but uses Binomial probit regression).

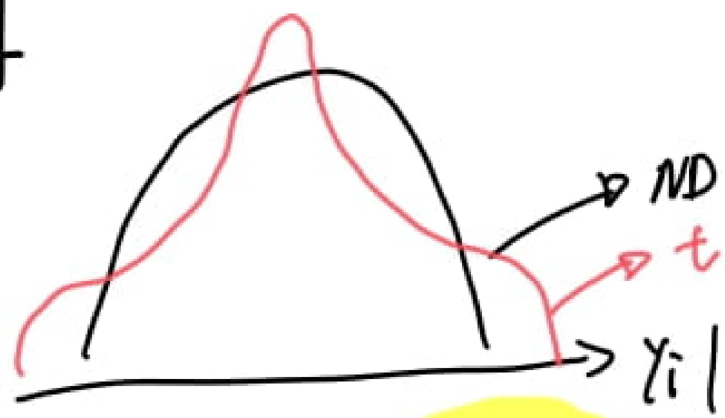
Y_i (or Z_i) $\sim$ EFD $\longleftrightarrow$ GLM

- Y_i is continuous
  - $Y_i \in (-\infty, \infty)$ and Y_i's histogram is normally distributed ($Y_i \sim ND$) $\Rightarrow$ T0 (Linear regression).
  - $Y_i \in [0, \infty)$ and Y_i's histogram is positively skewed ($Y_i \sim Gam$) $\Rightarrow$ T5 (Gamma regression with inverse link).
- Y_i (or Z_i) is discrete
  - $Y_i = 0/1$ (2 levels) $\Rightarrow Y_i \sim Bin(M_i = 1, \pi_i) = Bern(\pi_i) \Rightarrow$ T1 (Binary regression with logistic link).
  - $Z_i = \sum_{i=1}^{M_i} Y_i =$ the number of counts, where M_i (the number of trials) is known $\Rightarrow$ $Z_i \sim Bin(M_i, \pi_i) \Rightarrow$ T2 (Binomial regression with logistic link).
  - $Z_i = \sum_{i=1}^{M_i} Y_i =$ the number of counts, where M_i (the number of trials) is unknown & M_i goes to infinity $\Rightarrow Z_i \sim Poi(\mu_{Z_i}) \Rightarrow$ T3 (Poisson regression with log-linear link).
- Not all types of distribution for Y_i fit for the GLM framework.
- e.g. $Y_i \in (-\infty, \infty)$, Y_i's symmetric and has heavy tails on both sides ($Y_i \sim t$)

(skip)

$\begin{cases} Y_i \sim ND \rightarrow L.R \ (\text{for } x/y) \\ Y_i \sim \text{not necessarily } ND \xrightarrow{\text{subset.}} EFD \Leftrightarrow GLM \ (\text{for } x/y) \end{cases}$

$\boxed{\text{e.g}}$



$\begin{cases} Y_i | X_i = x \sim t \text{ dsb} \rightarrow \notin EFD \rightarrow \text{can't use } GLM \\ Y_i | X_i = x \rightarrow \in EFD \rightarrow \text{can use } GLM \end{cases}$

$\boxed{Q}$

Y_i (or Z_i) $\sim$ EFD $\longleftrightarrow$ GLM for Y_i & X_i

How to check $\boxed{?}$

EFD

=

Y_i (or Z_i) $\sim$ an exponential family distribution (EF), which has a general density (i.e. PDF or PMF):

$$f(y_i|\theta, \phi) = \exp\{\frac{y_i\theta - b(\theta)}{\alpha(\phi)} + c(y_i, \phi)\}$$

for some specified functions $\alpha(\cdot)$,$b(\cdot)$ and $c(\cdot)$. $b(\theta_i)$ is the cumulant generating function.

- θ: natural parameter (or canonical parameter), represents location.
  → ND: μ (e.g)
- ϕ: dispersion parameter, represents scale.
  → ND: σ^2 (e.g)

By changing the forms of the functions $b(\cdot)$ and $c(\cdot)$, we can define various members of the exponential family (e.g. Y_i (or Z_i) $\sim$ Normal, Binary/Bernoulli, Binomial, Poisson, Gamma, Exponential distributions.).

- $E(Y)$ (or $E(Z)$) $= \mu = \mu_Y$ (or μ_Z) $= b'(\theta)$, where $b'(\theta) = \frac{\partial b(\cdot)}{\partial \theta}$.
- $V(Y)$ (or $V(Z)$) $= \alpha(\phi)b''(\theta) = \alpha(\phi)V(\mu_Y)$ (or $\alpha(\phi)V(\mu_Z)$),
    - The **variance function** $= b''(\theta) = V(\mu_Y)$ (or $V(\mu_Z)$), which describes how the variance $V(Y)$ (or $V(Z)$) relates to the mean $E(Y)$ (or $E(Z)$).
    - $b''(\theta) = \frac{\partial^2 b(\cdot)}{\partial \theta^2}$.

The mean is a function of the location parameter θ only.

The variance is a function of both the location parameter θ and the dispersion parameter ϕ.

$\rightarrow$ general steps

link fun: $g(\mu_Y) = \Sigma \beta X$

$\rightarrow$ canonical link fun: $g(\mu_Y) = \Sigma\beta X = \theta$

Set

- $\theta =$ the canonical link function $= g(\mu_Y)$ (or $g(\mu_Z)$) $= g(b'(\theta)) = \beta_0 + \beta_1 X_{1_i} + ... + \beta_k X_{k_i} = x_i^T \beta$.
- $\phi = \phi_{assumed}$
  - $\phi = 1$ for T1 (Binary)/2(Binomial)/3(Poisson)
  - $\phi = MSE = \sigma^2$ for T0 (LR)
  - $\phi = cv = 1/\text{shape parameter} = \frac{1}{\alpha}$ for T5 (Gamma). Note that cv means the coefficients of variation.
- $\alpha(\phi) = \phi$ for T0/1/2/3, and $\alpha(\phi) = -\phi$ for T5.
- Variance function: $b''(\theta) = V(\mu_Y)$ (or $V(\mu_Z)$) $= \frac{V(Y)}{\alpha(\phi)}$ (or $\frac{V(Z)}{\alpha(\phi)}$).

given $a(\phi)$

Aim: Show Y_i (or Z_i) $\sim$ EFD

Steps: Y_i (or Z_i) $\sim$ --- dsb

$\downarrow$

$f(y_i) =$ --- (given pdf / pmf)

$\downarrow$

express $f(y_i)$ in the form of $f(y_i) = \exp\left\{\frac{y_i \cdot \theta - b(\theta)}{a(\phi)} + c(y_i, \phi)\right\}$

$\downarrow$

find out
$$\begin{cases} \theta = \cdots \\ a(\phi) = \text{(given)} \\ c(y_i, \phi) = \cdots \\ b(\theta) = \cdots \end{cases}$$

$b(\theta) = \cdots \rightarrow$
$$\begin{cases} b'(\theta) = E(Y) \\ b''(\theta) = \text{variance function} \\ a(\phi) \cdot b''(\theta) = V(Y) \end{cases}$$

Recall
$$\begin{cases} \text{link fun: } g(\mu_Y) = \Sigma \beta X \\ \downarrow \text{e.g} \\ \text{canonical link fun: } g(\mu_Y) = \Sigma \beta X = \theta \\ \text{inverse canonical link fun: solve for } \mu_Y \end{cases}$$

$a(\phi)$: given

- show slides examples of $\boxed{\text{To}}$

- In a **general** case, we define the link function $g(\cdot)$:
  - $g(\mu_{Y_i})$ = a function of $\mu_{Y_i} = \eta_i = \beta_0 + \beta_1 X_{1_i} + ... + \beta_k X_{k_i} = x_i^T \beta$ (or $g(\mu_{Z_i})$).
  - We have defined $E(Y)$ (or $E(Z)$) $= \mu = \mu_Y$ (or μ_Z) $= b'(\theta)$.
- The canonical link function is a **special** case of the link function, where we set the link function to **equal to the location parameter** θ, i.e.:
  - θ = the canonical link function = $g(\mu_Y)$ (or $g(\mu_Z)$)
    $= g(b'(\theta)) = \eta_i = \beta_0 + \beta_1 X_{1_i} + ... + \beta_k X_{k_i} = x_i^T \beta$.
  - For the purpose of this course, we will ~~only~~ mainly focus on **canonical** link functions.
- Note that the canonical link is not necessarily the most appropriate choice for a given dataset.

Show that $Y \sim N(\mu, \sigma^2)$

belongs to EFD $\boxed{?}$

Australian National University

Note that we use Y_i since this is non-aggregate data.

$Y_i \sim N(\mu_Y, \sigma^2) \rightarrow E(Y) = \mu_Y = \mu, V(Y) = \sigma^2.$

$$
\begin{aligned}
f(y_i|\theta,\phi) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right) \\
&= \exp\left(\frac{y_i\mu - \mu^2/2}{\sigma^2} - \underbrace{\frac{1}{2}\{\frac{y_i^2}{\sigma^2} + \ln(2\pi\sigma^2)\}}\right) \\
&= \exp\left(\frac{y_i\theta - b(\theta)}{\alpha(\phi)} + \underbrace{c(y_i,\phi)}\right)
\end{aligned}
$$

given

So

$\alpha(\phi) = \phi = \sigma^2, \theta = \mu, b(\theta) = \mu^2/2, c(y_i,\phi) = -\frac{1}{2}\{\frac{y_i^2}{\sigma^2} + \ln(2\pi\sigma^2)\} = -\frac{1}{2}\{\frac{y_i^2}{\phi} + \ln(2\pi\phi)\}$

Given $\theta = \mu$

- $\rightarrow b(\theta) = \mu^2/2 = \theta^2/2.$
- $\rightarrow E(Y) = b'(\theta) = \theta = \mu.$
- $\rightarrow V(Y) = \alpha(\phi)b''(\theta) = \phi \times 1 = \phi = \sigma^2.$

canonical link fun:

$g(\mu_y) = \Sigma\beta X = \theta = \mu_y$

(above)

$\Rightarrow g(\mu_y) = \mu_y$

$\Rightarrow$ identity link

$\therefore g(\mu_y) = \mu_y = \Sigma\beta X$

- link fun: $g(\mu_y) = \Sigma\beta X \rightarrow$

- Recall LR: $\mu_y = \Sigma\beta X$

$\Downarrow \quad \Downarrow$

LR is a special case of GLM.
where $Y \sim N(\mu, \sigma^2)$ & link fun at identity link

(canonical link fun)

(to continue)

- $\rightarrow$ The canonical link function $= \theta = \mu = \mu_Y = x_i^T \beta$.

Component 3: Exponential family - T2: Binomial regression with logistic link

Note that we use Z_i since this is aggregate data.

$Z_i \sim Bin(M, \pi) \rightarrow E(Z) = \mu_Z = M \times \pi, V(Z) = M \times \pi \times (1 - \pi)$.

$$f(z_i|\theta, \phi) = \binom{M}{z_i} \pi^{z_i}(1-\pi)^{M-z_i} = \exp\left(z_i \ln(\pi) + (M - z_i)\ln(1-\pi) + \ln\binom{M}{z_i}\right)$$

$$= \exp\left(z_i \ln(\frac{\pi}{1-\pi}) + M\ln(1-\pi) + \ln\binom{M}{z_i}\right)$$

$$= \exp\left(z_i \ln(\frac{\mu_Z}{M-\mu_Z}) + M\ln(\frac{M-\mu_Z}{M}) + \ln\binom{M}{z_i}\right)$$

$$= \exp\left(\frac{z_i\theta - b(\theta)}{\alpha(\phi)} + c(z_i, \phi)\right)$$

Note that for (T1) Binary:$\mu_Y = E(Y) = \pi$, and for (T2) Binomial $E(Z) = M\pi$.

Therefore $\frac{\pi}{1-\pi} = \frac{\mu_Z}{M-\mu_Z}$ (multiple by M) and $1 - \pi = 1 - \mu_Z/M = \frac{M-\mu_Z}{M}$).

So $\alpha(\phi) = \phi = 1, \theta = \ln(\frac{\pi}{1-\pi}) = \ln(\frac{\mu_Z}{M-\mu_Z}), b(\theta) = -M \times \ln(1-\pi) = -M \times \ln(\frac{M-\mu_Z}{M}) =$ $... = M \times \ln(1+\exp(\theta)),\ c(y_i, \phi) = \ln \begin{pmatrix} M \\ z_i \end{pmatrix}$.

Given $\theta = \ln(\frac{\pi}{1-\pi})$

- $\rightarrow b(\theta) = M \times \ln(1+\exp(\theta))$.
- $\rightarrow E(Z) = b'(\theta) = M \times \pi$.
- $\rightarrow V(Z) = \alpha(\phi)b''(\theta) = 1 \times M \times \pi \times (1-\pi) = M \times \pi \times (1-\pi) = M\mu_Y(1-\mu_Y) = \frac{\mu_Z(M-\mu_Z)}{M}$.
- $\rightarrow$ Variance function $=V(\mu_Z) = b''(\theta) = M \times \pi \times (1-\pi) = M \times \mu_Y \times (1-\mu_Y) = \frac{\mu_Z(M-\mu_Z)}{M}$.
- $\rightarrow$ The canonical link function $=\theta = \ln(\frac{\pi}{1-\pi}) = \ln(\frac{\mu_Y}{1-\mu_Y}) = x_i^T\beta$.

Note that (T1) Binary is a special case of (T2) Binomial, where we have non-aggregate data. We may replace Z_i by Y_i to the above equations and set $M_i = 1$.

Note that we use Z_i since this is aggregate data.

$Z_i \sim Poi(\mu_Z = \lambda) \rightarrow E(Z) = V(Z) = \mu_Z = \mu$

$$f(z_i|\theta,\phi) = \frac{e^{-\mu}\mu^{z_i}}{z_i!} = \exp\left(z_i \ln(\mu) - \mu - ln(z_i!)\right)$$
$$= \exp\left(\frac{z_i\theta - b(\theta)}{\alpha(\phi)} + c(z_i,\phi)\right)$$

So $\alpha(\phi) = \phi = 1, \theta = \ln(\mu), b(\theta) = \mu = e^{\theta}, c(z_i, \phi) = -ln(z_i!)$.

Note that since $\theta = \ln(\mu_Z)$, we have $\mu_Z = \mu = e^{\theta}$.

Given $\theta = \ln(\mu)$

- $\rightarrow b(\theta) = e^{\theta}$.
- $\rightarrow E(Z) = b'(\theta) = e^{\theta} = e^{\ln(\mu)} = \mu = \lambda$.
- $\rightarrow V(Z) = \alpha(\phi)b''(\theta) = 1 \times e^{\theta} = e^{\theta} = e^{\ln(\mu)} = \mu = \lambda$.
- $\rightarrow$ Variance function $= V(\mu_Z) = b''(\theta) = e^{\theta} = e^{\ln(\mu)} = \mu = \lambda$.

(to continue)

- $\rightarrow$ The canonical link function $= \theta = \ln(\mu) = x_i^T \beta$.

Note that we use Y_i since this is non-aggregate data.

$Y_i \sim Gam(\alpha, \beta) \rightarrow E(Y) = \frac{\alpha}{\beta}, V(Y) = \frac{\alpha}{\beta^2}$.

$$f(y_i|\theta, \phi) = \frac{1}{\Gamma(\alpha)} \beta^\alpha y_i^{\alpha-1} e^{-\beta y_i}$$

Hence

$$\ln(f(y_i|\theta, \phi)) = \ln(\beta^\alpha y_i^{\alpha-1} e^{-\beta y_i}) - \ln(\Gamma(\alpha)) = (\alpha - 1)\ln(y_i) + \alpha \ln(\beta) - \beta y_i - \ln(\Gamma(\alpha))$$

$$\exp(ln(f(y_i|\theta, \phi))) = f(y_i|\theta, \phi) = \exp[(\alpha - 1)\ln(y_i) + \alpha \ln(\beta) - \beta y_i - \ln(\Gamma(\alpha))]$$

$$f(y_i|\theta, \phi) = \exp[\{\frac{\frac{\beta}{\alpha} y_i - \ln(\beta)}{-1/\alpha}\} + (\alpha - 1)\ln(y_i) - \ln(\Gamma(\alpha))]$$

Let $\theta = 1/E(Y) = \frac{\beta}{\alpha} \rightarrow \beta = \theta\alpha$

Let $\phi = 1/\alpha \rightarrow \beta = \theta/\phi$

Hence $\ln(\beta) = \ln(\theta) - \ln(\phi)$

$$f(y_i|\theta,\phi) = \exp[\{\frac{\frac{\beta}{\alpha}y_i - \ln(\beta)}{-1/\alpha}\} + (\alpha-1)\ln(y_i) - \ln(\Gamma(\alpha))]$$

$$= \exp[\{\frac{\theta y_i - \ln(\theta)}{-\phi}\} - \ln(\phi)/\phi + (1/\phi - 1)\ln(y_i) - \ln(\Gamma(1/\phi))]$$

$$= \exp\left(\frac{y_i\theta - b(\theta)}{\alpha(\phi)} + c(y_i,\phi)\right)$$

So $\alpha(\phi) = -\phi = -1/\alpha$, $\alpha = 1/\phi$.
$\theta = \frac{\beta}{\alpha}, b(\theta) = \ln(\theta), c(y_i,\phi) = -\ln(\phi)/\phi + (1/\phi - 1)\ln(y_i) - \ln(\Gamma(1/\phi))$

Given $\theta = \frac{\beta}{\alpha} = 1/E(Y)$

- $\rightarrow b(\theta) = \ln(\theta)$.
- $\rightarrow E(Y) = b'(\theta) = 1/\theta = \frac{\alpha}{\beta}$.
- $\rightarrow V(Y) = \alpha(\phi)b''(\theta) = -1/\alpha \times (-(\frac{\alpha}{\beta})^2) = \frac{\alpha}{\beta^2}$.
- $\rightarrow$ Variance function $= V(\mu_Y) = b''(\theta) = -\theta^{-2} = -(\frac{\beta}{\alpha})^{-2} = -(\frac{\alpha}{\beta})^2$.
- $\rightarrow$ The canonical link function $=\theta = 1/E(Y) = 1/\mu_Y = x_i^T\beta$.

Note that this is not the unique parameterization method for (T5) Gamma regression.

- Using this logic, we can

show Y_i (or Z_i) $\sim$ ND (T0) / Binary (T1) / Binomial (T2) / Poisson (T3) / Gamma (T5) $\sim$ EFD(GLM)

obtain the canonical link function equations

link function: $g(\mu_y) = \Sigma \beta X$

canonical link function: .... $= \theta$ = a fun of μ_y from proof

(self-study for EFD of T1/T2/T3/T5)

- show canonical link / inverse canonical link fun. summary

(below)

$g(\mu_{y,\text{or } z}) = \Sigma \beta x = \theta = \cdots$

- T0: Linear regression (Normal distribution with identity link)
  - $Y_i \sim N(\mu_{Y_i}, \sigma^2) \to E(Y) = \mu_{Y_i} = \mu_i, V(Y) = \sigma^2$.
  - [1] The canonical link function $= \theta_i = \mu_{Y_i} = x_i^T \beta$.
  - [2] The inverse canonical link function is the same as the canonical link function: $\mu_{Y_i} = \theta_i = x_i^T \beta$.
  - Non-aggregate Data: (X_i, Y_i).
- T1: Binary regression with logistic link
  - $Y_i \sim Bern(\pi_i) \to E(Y) = \mu_{Y_i} = \pi_i, V(Y) = \pi_i \times (1 - \pi_i)$.
  - [1] The canonical link function $= \theta_i = \ln(\frac{\pi_i}{1-\pi_i}) = \ln(\frac{\mu_{Y_i}}{1-\mu_{Y_i}}) = x_i^T \beta$.
  - [2] The inverse canonical link function: $\mu_{Y_i} = \frac{e^{\theta_i}}{1+e^{\theta_i}}$.
  - Non-aggregate Data: (X_i, Y_i).

· $\pi_i = P(Y_i = 1 | X_i = x) = P(S) = P_i$
· $\pi_i = \mu_{Y_i} = E(Y_i | X_i = x) = E(Y_i)$

- T2: Binomial regression with logistic link
  - $Z_i \sim Bin(M_i, \pi_i) \rightarrow E(Z) = \mu_{Z_i} = M_i \times \pi_i, V(Z) = M_i \times \pi_i \times (1 - \pi_i)$.
  - [1] The canonical link function $=\theta_i = \ln(\frac{\pi_i}{1-\pi_i}) = \ln(\frac{\mu_{Y_i}}{1-\mu_{Y_i}}) = x_i^T \beta$.
  - [2] The inverse canonical link function: $\mu_{Y_i} = \frac{e^{\theta_i}}{1+e^{\theta_i}}$.
  - Therefore $\mu_{Z_i} = \mu_{Y_i} \times M_i$.
  - Aggregate Data: (X_i, Z_i, M_i).
- T3: Poisson regression with log-linear link
  - $Z_i \sim Poi(\mu_{Z_i} = \lambda_i) \rightarrow E(Z) = V(Z) = \mu_{Z_i} = \mu_i$.
  - [1] The canonical link function $=\theta_i = \ln(\mu_{Z_i}) = x_i^T \beta$.
  - [2] The inverse canonical link function: $\mu_{Z_i} = e^{\theta_i}$.
  - Aggregate Data: (X_i, Z_i)

- T5: Gamma regression with inverse link
  - $Y_i \sim Gam(\alpha_i, \beta_i) \rightarrow E(Y) = \frac{\alpha_i}{\beta_i}, V(Y) = \frac{\alpha_i}{\beta_i^2}$ (The beta here is the rate parameter).
  - [1] The canonical link function $=\theta_i = 1/\mu_{Y_i} = x_i^T \beta$ (The betas here are the coefficients).
  - [2] The inverse canonical link function: $\mu_{Y_i} = \frac{1}{\theta_i}$.
  - Non-aggregate Data: (X_i, Y_i).

MLE

- In LR, how to calculate b_i manually $\boxed{?}$

(Recall:

$$\begin{aligned} b_0 &= \hat{\beta}_0 \longrightarrow \beta_0 \\ b_1 &= \hat{\beta}_1 \longrightarrow \beta_1 \\ &\vdots \qquad\qquad \vdots \\ b_k &= \hat{\beta}_k \longrightarrow \beta_k \end{aligned}$$

)

Assume there are n number of observations in the sample, each observation is independent.

In **LR model**, we use **OLSE (Ordinary Least Square Estimator)** method to obtain estimated coefficients $\widehat{\beta}_i$s ($\widehat{\beta}_0, ..., \widehat{\beta}_k$).

- *SSE* $= \sum_{i=1}^{i=n}(Y_i - \widehat{Y}_i)^2$. $\leftarrow$ sub $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k$
- To minimize SSE: $\frac{\partial SSE}{\partial \widehat{\beta}_0} = 0, ..., \frac{\partial SSE}{\partial \widehat{\beta}_k} = 0$.
- Solve for $\widehat{\beta}_0, ..., \widehat{\beta}_k$.

MLE

① (x_i, y_i) (n#)

↓

GLM

↓

MLE to obtain b_is

↓

Here ☆

② Y_i only (n#)

e.g: $Y_i \sim N(\mu, \sigma^2)$

↓

Not "regression"

(Y only)

↓

MLE to obtain parameters of Y_i dsb

(e.g: $\hat{\mu} = \cdots$, $\hat{\sigma}^2 = \cdots$)

Detail:

MLE (for GLM) ①

- n# (x_i, y_i)
- aim: solve for bi (i.e., $b_0 \sim b_k$)
- How?

$\boxed{S1}$ $\max L = \prod_{i=1}^{n} f(y_i)$ ← y_i (or z_i) $|x$ ~EFD (omit)

$f(y_i) = \exp\left\{\frac{y_i \cdot \theta - b(\theta)}{a(\phi)} + c(y_i, \phi)\right\}$

θ = canonical link fun $= g(\mu_y) = \sum \beta x$

$\therefore f(y_i)$ = fun of βs

↓

$\max \ln(L) = l$

↓

$\boxed{S2}$ $\frac{\partial l}{\partial b_i} = 0$ $(i = 0 \sim k)$

↓

$\boxed{S3}$ $b_i = \cdots$ ($b_0 \sim b_k$ each)

⇒ slides

MLE (for y_i only) ②

- n # obs y_i, each y_i ~ Some distributions (θ, γ) ← some parameters; can extend ...

  ~ density: $f(y_i)$ ← given

- aim: Solve for θ and γ = ?

- $\boxed{S1}$: $L = \prod_{i=1}^{n} f(y_i) = \cdots\cdots \rightarrow \max L$
- $\boxed{S2}$ $\ln L = l = \ln\left(\prod_{i=1}^{n} f(y_i)\right) = \sum_{i=1}^{n} \ln(f(y_i)) \rightarrow \max l$
- $\boxed{S3}$ $\left\{\begin{array}{l} \frac{\partial l}{\partial \theta} = 0 \\ \frac{\partial l}{\partial \gamma} = 0 \end{array}\right\} \rightarrow \left\{\begin{array}{l} \hat{\theta} = \cdots \\ \hat{\gamma} = \cdots \end{array}\right.$

⇒ GLM exercise (MLE)

(1)

In **GLM**, we use **MLE (Maximum Likelihood Estimator)** method instead to obtain estimated coefficients $\widehat{\beta}_i$s ($\widehat{\beta}_0, ..., \widehat{\beta}_k$).

- Maximum likelihood estimator: estimate the parameters by those values which make the likelihood function as large as possible for our particular set of observed data. These estimators are the ones which have the maximum likelihood of having produced the observed data.

- Since Y_i (or Z_i) follows the exponential family, we may obtain the density of Y_i (or Z_i), i.e. $f(Y_i = y_i) = f(y_i)$.
  - $$f(y_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\alpha(\phi)} + c(y_i, \phi)\right)$$
    .
  - As discussed, θ_i is the canonical link function, where $\theta_i = g(\mu_{Y_i}) = x_i^T\beta$.
  - The density is either a PDF (probability density function for Y_i is continuous) or a PMF (probability mass function for Y_i is discrete/categorical).

Show steps of MLE for GLM

1

- Define the likelihood function: $L = \prod_{i=1}^{n} f(y_i)$ (each observation is independent, and there are n observations in total).
- Log-likelihood: $l = \ln(L) = \ln(\prod_{i=1}^{n} f(y_i)) = \sum_{i=1}^{n} \ln(f(y_i))$.
- To maximize L is equivalent to maximize $l = \ln(L)$. Therefore, $\frac{\partial l}{\partial \widehat{\beta_0}} = 0, ..., \frac{\partial l}{\partial \widehat{\beta_k}} = 0$.
- Solve for $\widehat{\beta_0}, ..., \widehat{\beta_k}$.

In R's MLE calculation, it uses IRLS (Iterative Re-Weighted Least Squares) method instead, which is a method that will obtain equivalent estimated coefficients as MLE does. The detail for IRLS is not required in this course.

No need to calculate bi's manually
We will use R to generate those bi's

Week6 Recap:

- L.R: $\begin{cases} Y_i \mid X_i = x \sim ND \\ \mu_{Y_i} = \Sigma \beta X \end{cases}$

  (SL relation)

  special case:
  $g(\mu_Y) = \mu_Y = \Sigma\beta X$

- GLM: $Y_i \mid X_i = x \sim EFD$

  $g(\mu_{Y_i}) = \Sigma\beta X$ (link fun) $\rightarrow \mu_{Y_i} = \cdots$ (inverse link fun)

  (non-SL relation)

  notations $\begin{cases} Y_i \rightarrow \text{non-aggregate data's response} \\ Z_i \rightarrow \text{aggregate data's response} \end{cases}$

  Math proof of $Y_i \sim EFD$
  - $\rightarrow$ general density ...
  - $\rightarrow$ $a(\phi)$ given, θ, $b(\theta)$, $c(y_i;\phi)$
    - $b'(\theta) = E(Y)$, $b''(\theta)$ = variance fun
    - $a(\phi) \cdot b''(\theta) = V(Y)$
  - $\rightarrow$ $\underbrace{g(\mu_Y)}_{\text{link fun}} = \Sigma\beta X \underbrace{= \theta}_{\text{canonical link fun}}$

  obtain coefficients $(\hat{\beta}_0 \sim \hat{\beta}_k)$: MLE