2. Statistical Modelling (4)

Statistical Modelling & Machine Learning

Jaejik Kim

Department of Statistics
Sungkyunkwan University

STA3036

Models for Discrete or Non-normal Y Variables

- Classical regression models assume continuous Y and normal error distribution.
- When Y is a discrete random variable or it does not have normal distribution, classical regression models do not work properly.
  - The range of $\mu = E(Y) = \boldsymbol{X}\boldsymbol{\beta}$.
  - Statistical inference due to normality assumption.
- E.g., suppose that Y has Bernoulli response $Y_i = 0$ or 1.
  $\mu_i = E(Y_i) = P(Y_i = 1) \in [0, 1]$
  $Var(Y_i) = \mu_i(1 - \mu_i)$ (not constant).

Generalized Linear Model

- Generalized linear model (GLM): Extension of classical linear model.
- 3 components of GLM:
  1. Systematic component: $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$.
  2. Random component: Y_i's are independent random variables with $E(Y_i) = \mu_i$ and pdf (pmf) in the exponential family as follows:

$$p(y_i; \theta_i, \phi) = \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\}, \qquad (1)$$

     - θ_i: Location parameter (usually our interest).
     - θ_i can be expressed as some function of $\mu_i = E(Y_i)$.
     - ϕ: Scale parameter (nuisance parameter).
  3. Link function: The link between the systematic and random components.

$$g(\mu_i) = \eta_i,$$

     where g is one-to-one and differentiable.

Exponential Family

- Exponential family: A set of distributions whose pdf (pmf) satisfies the format of (1).
  - Distributions in Exponential family: Normal, exponential, Bernoulli, binomial, Poisson, gamma, geometric, etc.

- E.g., Normal distribution: $Y_i \sim^{indep.} N(\mu_i, \sigma^2)$.

$$
\begin{aligned}
p(y_i; \mu_i, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \mu_i)^2}{2\sigma^2}\right\} \\
&= \exp\left\{\left[y_i\mu_i - \frac{\mu_i^2}{2}\right]\frac{1}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\}
\end{aligned}
$$

  - $\theta_i = \mu_i$, $\phi = \sigma^2$,
  - $a_i(\phi) = \phi$, $b(\theta_i) = \theta_i^2/2$, $c(y_i, \phi) = -[y_i^2/\phi + \log(2\pi\phi)]/2$.

Exponential Family

- E.g., Binomial distribution: $Y_i \sim^{indep.} Binom(n_i, p_i)$.

$$
\begin{aligned}
p(y_i; p_i) &= \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i - y_i} \\
&=
\end{aligned}
$$

Link Function

- For Y_i with a certain distribution in the exponential family, various link functions exist.
- E.g., for a binary random variable Y, we want to map $\mathbb{R}$ $(\eta = \boldsymbol{x}\boldsymbol{\beta})$ to $[0, 1]$ (the range of $\mu = E(Y)$).
  - All cdf can map $\mathbb{R}$ to $[0, 1]$.

$$\mu = F(\eta) \quad \Rightarrow \quad F^{-1}(\mu) = \eta.$$

  - Logit link: $\log \frac{\mu}{1-\mu}$ (inverse of unit logistic cdf).
  - Probit link: $\Phi^{-1}(\mu) = \eta$ (inverse of standard normal cdf).
  - log-log link: $\log(-\log(\mu)) = \eta$ (inverse of Gumbel cdf).

Canonical Link Function

- For each distribution, there is one link function that is mathematically convenient $\Rightarrow$ Canonical link function.

- Canonical Link: $\theta = \eta$.

Distribution	Canonical Link
Normal distribution	$g(\mu) = \mu$
Bernoulli distribution	$g(\mu) = \log(\mu/(1-\mu))$
Poisson distribution	$g(\mu) = \log \mu$
Gamma distribution	$g(\mu) = \mu^{-1}$

- Choice of link should be made on
  - model fit,
  - model interpretability,
  - mathematical convenience (canonical link or not).

Maximum Likelihood Estimation in GLM

- For independent $Y_1, \dots, Y_n$, the log-likelihood function is

$$l(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \log p(y_i; \theta_i),$$

  - $\theta_i = \theta_i(\mu_i)$.
  - $g(\mu_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$.
  - By the invariance property of MLE, MLE of μ_i and θ_i can be obtained by the MLE $\hat{\boldsymbol{\beta}}$ of the model parameters.

$$\hat{\boldsymbol{\beta}} = \max_{\boldsymbol{\beta}} l(\boldsymbol{\beta}; \boldsymbol{y}).$$

- No closed-form solution exists $\Rightarrow$ Numerical method (Newton-Raphson or Fisher scoring).

Logistic Regression

- Suppose that Y is a binary output variable.

- Canonical link function: $g(\mu_i) = \log(\mu_i/(1-\mu_i))$, where $\mu_i = E(Y_i) = p_i$.
$$\log \frac{p_i}{1-p_i} = \boldsymbol{x}_i^\top \boldsymbol{\beta},$$
where $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$ & $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$.

- $p_i = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})} \;\Rightarrow\; \frac{p_i}{1-p_i} = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$.

Estimation of $\boldsymbol{\beta}$

- Likelihood function:

$$\begin{aligned}\max_{\boldsymbol{\beta}} L(\boldsymbol{\beta}; \boldsymbol{y}) &= \max_{\boldsymbol{\beta}} \left[\prod_{i:y_i=1} p_i \prod_{i':y_{i'}=0} (1 - p_{i'}) \right] \\ &= \max_{\boldsymbol{\beta}} \prod_{i:y_i=1} \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}} \prod_{i':y_{i'}=0} \frac{1}{1 + e^{\boldsymbol{x}_{i'}^\top \boldsymbol{\beta}}}.\end{aligned}$$

  $\Rightarrow$ Numerical method (Iteratively Reweighted Least Squares)
  $\Rightarrow \hat{\boldsymbol{\beta}}$ (MLE).

- $\hat{\boldsymbol{\beta}} > 0$, $p(x) \uparrow$,
  $\hat{\boldsymbol{\beta}} < 0$, $p(x) \downarrow$. (Not linear relationship).

Multinomial Logistic Regression

- Logistic regression with K classes ($K > 2$) $\Rightarrow$ Multinomial logistic regression:

$$\ln \frac{P(Y = k|X = x)}{P(Y = K|X = x)} = x^\top \boldsymbol{\beta}_k$$

  for $k = 1, \ldots, K - 1$ with $\sum_{k=1}^{K} P(Y = k|X = x) = 1$.

- $x = (1, x_1, \ldots, x_p)^\top$ & $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \ldots, \beta_{kp})^\top$.

- The choice of denominator, the Kth class, is arbitrary.

Multinomial Logistic Regression

- By solving for $P(Y = k|X = x)$, we have

$$P(Y = k|X = x) = \frac{\exp\left(x^\top \boldsymbol{\beta}_k\right)}{1 + \sum_{j=1}^{K-1} \exp\left(x^\top \boldsymbol{\beta}_j\right)}$$

for $k = 1, \ldots, K-1$ and

$$P(Y = K|X = x) = \frac{1}{1 + \sum_{j=1}^{K-1} \exp\left(x^\top \boldsymbol{\beta}_j\right)}.$$

$\Rightarrow$ ML estimation (numerical method)
$\Rightarrow \hat{\boldsymbol{\beta}}_k,\ k = 1, \ldots, K-1$ (MLE).

Ordinal Response: Cumulative Logit Model

- Ordinal data: Categories are ordered (e.g., good, medium, bad).
- Suppose that response Y takes ordered category values $k = 1, \ldots, K$, let $p_k = P(Y = k|\boldsymbol{X})$.
- Cumulative probability:

$$\gamma_k = \sum_{j=1}^{k} p_j = P(Y \leq k|\boldsymbol{X}), \; k = 1, \ldots, K.$$

- Cumulative logit:

$$\log \frac{\gamma_k}{1-\gamma_k} = \log \frac{p_1 + \cdots + p_k}{p_{k+1} + \cdots + p_K}, \; k = 1, \ldots, K-1.$$

Ordinal Response: Cumulative Logit Model

- Cumulative Logit Model (Proportional odds model):

$$\log \frac{\gamma_{ik}}{1-\gamma_{ik}} = \alpha_k + \boldsymbol{x}_i^\top \boldsymbol{\beta}, \ i = 1, \ldots, n, \ k = 1, \ldots, K-1.$$

  - α_k is increasing in k because γ_k is increasing in k for fixed $\boldsymbol{x}$.
  - This model has the same effects $\boldsymbol{\beta}$ for each logit model (The $K-1$ logistic curves have the same shape).

- Two observations with input vector $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively.

$$\log \frac{\gamma_{1k}}{1-\gamma_{1k}} - \log \frac{\gamma_{2k}}{1-\gamma_{2k}} = \log \frac{\gamma_{1k}/(1-\gamma_{1k})}{\gamma_{2k}/(1-\gamma_{2k})} = (\boldsymbol{x}_1 - \boldsymbol{x}_2)^\top \boldsymbol{\beta}.$$

  - Log cumulative odds ratio does not depend on k, only on $(\boldsymbol{x}_1 - \boldsymbol{x}_2)$.
  - The ratio of odds of being in the kth or smaller category under two different inputs is the same for all categories. $\Rightarrow$ Proportional odds model.

Maximum Likelihood Estimation

- Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iK})^\top$, $i = 1, \ldots, n$, where $y_{ik} = 1$ if the ith obs. is in the kth ordered category. Otherwise, $y_{ik} = 0$.

- Likelihood function:

$$
\begin{aligned}
L(\boldsymbol{\alpha}, \boldsymbol{\beta}; \boldsymbol{y}) &= \prod_{i=1}^{n} \left[\prod_{k=1}^{K} p_k^{y_{ik}} \right] = \prod_{i=1}^{n} \left[\prod_{k=1}^{K} (\gamma_k - \gamma_{k-1})^{y_{ik}} \right] \\
&= \prod_{i=1}^{n} \left[\prod_{k=1}^{K} \left\{ \frac{\exp(\alpha_k + \boldsymbol{x}_i^T \boldsymbol{\beta})}{1 + \exp(\alpha_k + \boldsymbol{x}_i^T \boldsymbol{\beta})} - \frac{\exp(\alpha_{k-1} + \boldsymbol{x}_i^T \boldsymbol{\beta})}{1 + \exp(\alpha_{k-1} + \boldsymbol{x}_i^T \boldsymbol{\beta})} \right\}^{y_{ik}} \right].
\end{aligned}
$$

Poisson Regression

- Y: Count data $\Rightarrow Y_1, \ldots, Y_n \sim^{indep.} Poisson(\mu_i)$.
- Canonical link function: $g(\mu_i) = \log(\mu_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$.
- Model: $\log \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$, $i = 1, \ldots, n$.
- Log-likelihood function:

$$
\begin{aligned}
l(\boldsymbol{\beta}; \boldsymbol{y}) &= \sum_{i=1}^{n} [y_i \log(\mu_i) - \mu_i - \log(y_i!)] \\
&= \sum_{i=1}^{n} \left[y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} - \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) - \log(y_i!) \right].
\end{aligned}
$$

Poisson Regression

- Overdispersion: Data variation is higher than model's expectation.
- Overdispersion in Poisson regression is typical because $E(Y_i) = Var(Y_i) = \mu_i$.
- To solve the overdispersion problem, the negative binomial model can be considered. $Y \sim NB(\mu, \alpha)$

$$p(y) = \frac{\Gamma(y + \alpha^{-1})}{y!\Gamma(\alpha^{-1})} \left(\frac{\alpha\mu}{1 + \alpha\mu} \right)^{y} \left(\frac{1}{1 + \alpha\mu} \right)^{\alpha^{-1}}, \ y = 0, 1, 2, \ldots.$$

- $E(Y) = \mu$ and $Var(Y) = \mu + \alpha\mu^2$.
- Negative binomial model: $\log \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \ i = 1, \ldots, n$.

Survival Model

- Survival data: T is the survival time until death or failure.
- Censoring: Property of survival data
  - It occurs when the outcome of a particular patient or component is unknown at the end of the study.
- Let the survival time T have a pdf $f(t)$ and the cdf $F(t)$.
  - $F(t)$: The fraction of the population dying by time t.
  - $1 - F(t)$: Survival function (fraction still surviving at time t).
  - $h(t)$: Hazard function (instantaneous risk).
  - $h(t)d(t)$: Prob. of dying in the next small time interval $d(t)$ given survival to time t.

$$
\begin{aligned}
h(t)dt &= P(T \in [t, t+dt] | T > t) = \frac{f(t)dt}{1 - F(t)} \\
&\Rightarrow \quad h(t) = \frac{f(t)}{1 - F(t)}.
\end{aligned}
$$

Proportional Hazard Model

- Proportional hazard model:

$$h(t|\boldsymbol{x}) = \lambda(t)\exp(\boldsymbol{x}^\top\boldsymbol{\beta}).$$

- Under this model, consider two observations with $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively.

$$\frac{h(t|\boldsymbol{x}_1)}{h(t|\boldsymbol{x}_2)} = \exp[(\boldsymbol{x}_1 - \boldsymbol{x}_2)^\top\boldsymbol{\beta}].$$

  - Proportional hazard: This ratio does not depend on t.
- From the proportional hazard model,

$$h(t) = f(t)/[1 - F(t)] = \lambda(t)\exp(\boldsymbol{x}^\top\boldsymbol{\beta}),$$

  by taking integral on both sides,

$$-\log[1 - F(t)] = \Lambda(t)\exp(\boldsymbol{x}^\top\boldsymbol{\beta}),$$

  where $\Lambda(t) = \int_{-\infty}^{t}\lambda(u)du$ (Cumulative hazard).

ML Estimation of PH Model

- Survival function:

$$S(t) = 1 - F(t) = \exp\{-\Lambda(t)\exp(\boldsymbol{x}^\top \boldsymbol{\beta})\}.$$

- By minus derivative w.r.t. t,

$$f(t) = \lambda(t)\exp\{\boldsymbol{x}^\top \boldsymbol{\beta} - \Lambda(t)\exp(\boldsymbol{x}^\top \boldsymbol{\beta})\}.$$

- Likelihood function:
  - An object who died at time t contributes a factor $f(t)$ to the likelihood.
  - An object who censored at time t contributes $S(t)$.
  - If the ith observation is died at time t, $w_i = 1$. Otherwise, $w_i = 0$.

ML Estimation of PH Model

- Log-likelihood function:

$$
\begin{aligned}
l(\beta; \boldsymbol{t}, \boldsymbol{w}) &= \sum_{i=1}^{n} [w_i \log f(t_i) + (1 - w_i) \log S(t_i)] \\
&= \sum_{i=1}^{n} \left[w_i \{\log \lambda(t_i) + \boldsymbol{x}_i^\top \beta\} - \Lambda(t_i) \exp(\boldsymbol{x}_i^\top \beta) \right].
\end{aligned}
$$