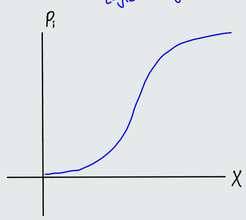
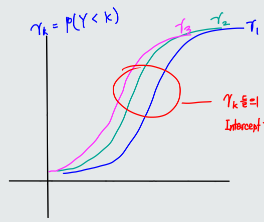
Logistic Regression



Cumulative Logistic



γ_k들의 기울기는 동일하여 통일된 β를 갖지만,
Intercept들이 다르기 때문에 이러한 차이들이 발생한다.

2. Statistical Modelling (4)

Statistical Modelling & Machine Learning

Jaejik Kim

Department of Statistics
Sungkyunkwan University

STA3036

- 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. normality assumption을 만족하지 못하면 통계적 추론을 할 수가 없다.
- 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). violates the normality assumption

Generalized Linear Model

=> normality assumption이 충족되지 않았을 때 사용하는 Model

- Generalized linear model (GLM): Extension of classical linear model. * GLM에서 MLE를 통해 parameters를 구하게 되면 closed form으로 나오지 않기 때문에 numerical method로 찾아야 한다.

* 범주형 자료분석에서 다룸

- 3 components of GLM:
  1. Systematic component: $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$. : i번째 observation 마다 분포의 parameter가 바뀜을 의미.
  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: GLM에서 필요한 조건이나 상황들이 exponential family에 대해서만 적용가능하다.

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

    - θ_i: Location parameter (usually our interest). : μ_i에 대한 함수로 표현될 수 있음
    - θ_i can be expressed as some function of $\mu_i = E(Y_i)$.
    - ϕ: Scale parameter (nuisance parameter). 분석에 필요한 parameter인데 분석의 관심사항이 아닐 때

  위 2개를 연결해주는 함수 3. Link function: The link between the systematic and random components.

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

μ_i가 θ_i에 대한 함수이므로, η_i와 θ_i를 연결시킨다.

where g is one-to-one and differentiable. * random component의 range를 systematic component에 맞춰주는 역할

Exponential Family

GLM 사용가능

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

nuisance parameter

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

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

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

$$
\begin{aligned}
p(y_i; p_i) &= \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i} \\
&= \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i} (1-p_i)^{-y_i} \\
&= \binom{n_i}{y_i} \left(\frac{p_i}{1-p_i}\right)^{y_i} (1-p_i)^{n_i} \\
&= \exp\left[y_i \log\left(\frac{p_i}{1-p_i}\right) + n_i \log(1-p_i) + \log\binom{n_i}{y_i} \right]
\end{aligned}
$$

$$
\begin{aligned}
\therefore\ \theta_i &= \log\left(\frac{p_i}{1-p_i}\right) \\
\phi &= 1 \\
a_i(\phi) &= \phi \\
b(\theta_i) &= n_i \log(1+e^{\theta_i}) \\
c(y_i, \phi) &= \log\binom{n_i}{y_i}
\end{aligned}
$$

Link Function

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

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

☆ 가장 유명

  - 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.
  주어진 분포에 대해서 많이 사용되는 link function
- Canonical Link: $\theta = \eta$. $\theta_i(\mu_i) = \eta_i$

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

Model parameter를 Estimate 하는 단계

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

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

  - $\theta_i = \theta_i(\mu_i)$.
  - $g(\mu_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$. "MLE의 함수도 MLE다"
  - 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: $g(\mu_i) = \underbrace{\log(\mu_i/(1-\mu_i))}_{=\theta_i}$, where $\mu_i = E(Y_i) = p_i$.

$$\log \frac{p_i}{1-p_i} = \underbrace{\boldsymbol{x}_i^\top \boldsymbol{\beta},}_{=\eta_i}$$

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 β

- Likelihood function:

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

$\Rightarrow$ 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).

4주차 1차시 끝

Multinomial Logistic Regression

Multinomial Distribution $Y \to$ nominal categorical

Mutinomial Logistic Regression을 사용하자

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

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

small "k"

capital "K"

for $\underbrace{k = 1, \ldots, K-1}_{k\text{가 변할 때마다 } \beta_k\text{, 식의 계수가 변한다.}}$ 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 $p_k = P(Y = k|\boldsymbol{X})$.

  Y가 k번째 category에 속할 확률

- Cumulative probability:

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

$\therefore \underset{\text{Capital case } K}{\gamma_K} = 1$ since it cumulates all K possibilities

- Cumulative logit: 일반적인 Logistic Regression에서는 P_i를 활용해서 log-odds를 만들었는데, Cumulative Logit 에서는 log-odds를 만드는 대상이 γ_k 이다.

  ※ $P_i + (1 - P_i) = \gamma_k + (1 - \gamma_k) = 1$

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

k^{th} order 보다 큰 observations 들의 합

Ordinal Response: Cumulative Logit Model

- Cumulative Logit Model (Proportional odds model):

dependent on k

동일된 β를 가정하지 않으면 order가 역전되지 않게 하기 위한 constraints들이 많이 생기고, 이러한 constraints들에 맞춰서 optimize하기 상당히 까다롭다.

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

K개의 intercept가 개별존재한다.

모든 K개의 categories들이 똑같은 β coefficients 값을 가진다.

☆ 동일한 β를 갖는 이유 : 정해진 order가 역전되지 않게 하기 위해서 기울기를 고정하고 α_k만 움직이게 해야 했다.

  - 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).

형태는 모두 동일하되 위치만 다르다.

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

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

x_1, x_2 모두 k번째 category에 속하기 때문에 동일한 α_k값을 가지므로 소거된다.

  - Log cumulative odds ratio does not depend on k, only on $(\boldsymbol{x}_1 - \boldsymbol{x}_2)$.

$= \log \frac{\gamma_{1k}/(1-\gamma_{1k})}{\gamma_{2k}/(1-\gamma_{2k})}$

  - 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.

~ one-hot encoding

- 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:

$$\underbrace{L(\boldsymbol{\alpha}, \boldsymbol{\beta}; \boldsymbol{y})}_{\alpha,\ \beta\text{ 에 대한 Likelihood Function}} = \prod_{i=1}^{n} \left[\prod_{k=1}^{K} p_k^{y_{ik}} \right] = \prod_{i=1}^{n} \left[\prod_{k=1}^{K} (\gamma_k - \gamma_{k-1})^{y_{ik}} \right]$$

$$= \prod_{i=1}^{n} \left[\prod_{k=1}^{K} \left\{ \frac{\exp(\alpha_k + \boldsymbol{x}_i^T \beta)}{1 + \exp(\alpha_k + \boldsymbol{x}_i^T \beta)} - \frac{\exp(\alpha_{k-1} + \boldsymbol{x}_i^T \beta)}{1 + \exp(\alpha_{k-1} + \boldsymbol{x}_i^T \beta)} \right\}^{y_{ik}} \right].$$

☆ $\log \frac{\gamma_{ik}}{1-\gamma_{ik}} = \alpha_k + x_i^T \beta$

$\Rightarrow \gamma_{ik} = \frac{\exp(\alpha_k + x_i^T \beta)}{1 + \exp(\alpha_k + x_i^T \beta)}$

Poisson Regression

- Y: Count data $\Rightarrow Y_1, \ldots, Y_n \overset{indep.}{\sim} Poisson(\mu_i)$.
- 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:

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

포아송의 단점!

- Overdispersion: Data variation is higher than model's expectation.

- Overdispersion in Poisson regression is typical because $E(Y_i) = Var(Y_i) = \mu_i$.

  거의 모든 상황에서 분산은 평균보다 큰데 포아송은 평균과 분산이 동일해서

- 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: $\underbrace{T}_{\text{살아있는 시간, 기간}}$ 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.

  => 피실험자의 중도포기 또는 한정된 관측기간 때문에 실험이 끝날 시점에도 잘 알 수 없음.

- 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). t 에서의 순간사망률 (사력...?)
  - $h(t)\cancel{d(t)}$ dt: Prob. of dying in the next small time interval $\cancel{d(t)}$ dt given survival to time t. t까지 살다가 t가 되자마자 사망할 확률

$$h(t)dt = P(T \in \underbrace{[t, t+dt]}_{\text{이 작은 틈에 사망}} | \underbrace{T > t}_{\text{생존시간 자체는 t보다 크다.}}) = \frac{f(t)dt}{1 - F(t)}$$

$$\Rightarrow \quad h(t) = \frac{f(t)}{1 - F(t)}.$$

Proportional Hazard Model

☆ $X^\top \beta = 0$ 이면 $h(t|x) = \lambda(t)$ 가 되므로,
$\lambda(t)$를 Baseline Hazard라고 칭한다.

- 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.

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

=> Independent from t. 위험율의 비율은 시간에 상관없이 일정하다.

  - 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,

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

t 가 없으므로 상수취급이 되면서 그대로 나옴

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,

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

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

- Log-likelihood function:

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

β에 대한 추정은 parametric method인데,
$\lambda(t)$에 대한 추정은 non-parametric 해서
$l(\beta ; t, w)$를 semi-parametric이라 한다.

$\lambda(t_i)$를 추정해야 하는데, $\lambda(t_i)$는 non-parametric 하게 추정된다.

=> 모델의 해석 : X가 치료여부에 대한 binary data인 경우, 만약 $\frac{h(t \mid X_1)}{h(t \mid X_2)} = \exp[(X_1 - X_2)^T \beta]$ 에서 $X_1 = 1$, $X_2 = 0$, $\beta > 0$ 이라면 치료를 안 받았을 때 위험률이 e^β 만큼 증가한다고 할 수 있다.