Logistic Regression

Instructor: Junghye Lee

Department of Industrial Engineering

junghyelee@unist.ac.kr

Contents

Logistic Regression

- To explain a dichotomous response variable on numeric and/or categorical explanatory variable(s)
  - Goal: Model the probability of a particular as a function of the predictor variable(s)
  - Problem: Probabilities are bounded between 0 and 1

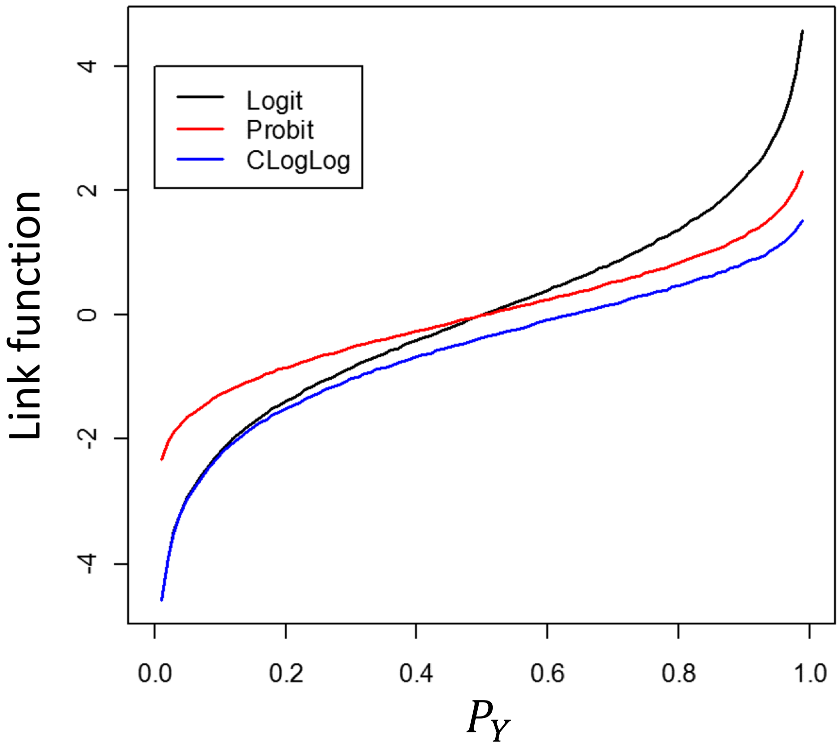
- Distribution of response variable: binomial distribution
- Link function
  - $\text{logit}(P(Y=1)) = \log\left(\frac{P(Y=1)}{1-P(Y=1)}\right)$
  - $\text{Probit}(P(Y=1)) = \Phi^{-1}(P(Y=1))$
  - $c\text{loglog}\big(P(Y=1)\big) = \log\big[-\log\big(1-P(Y=1)\big)\big]$

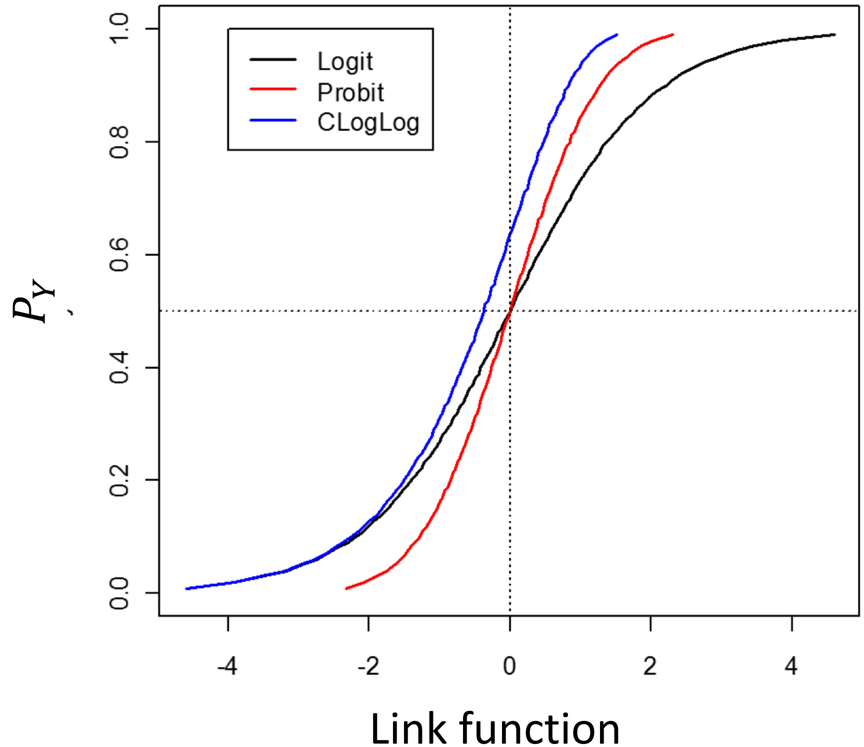
Properties of Link Functions

This is what you are interested in

- We denote $\boxed{P(Y=1)}$ by P_Y
- They can take any value on the real line for $0 \le P_Y \le 1$
- Consider logit:
  - If $P_Y = 0$, $\text{logit}(P_Y) = \log(0) = -Inf$
  - If $P_Y = 1$, $\text{logit}(P_Y) = \log(Inf) = Inf$

$$P(Y=1) = \frac{\exp(\boldsymbol{X\beta})}{1+\exp(\boldsymbol{X\beta})}$$





Logistic Regression with 1 Predictor

- Response: Presence/Absence of characteristic
- Predictor: Numeric variable observed for each case
- Model: $P(Y = 1|X) \equiv$ Probability of presence at predictor level X

$$\log\left(\frac{P(Y = 1)}{1 - P(Y = 1)}\right) = \beta_0 + \boxed{\beta X}$$

General form of logistic regression function

- β_0: constant
- $\beta = 0$: $P(Y = 1)$ is the same at each level of X
- $\beta > 0$: $P(Y = 1)$ increases as X increases
- $\beta < 0$: $P(Y = 1)$ decreases as X increases

Logistic Regression with 1 Predictor

- β_0, β are unknown parameters and must be estimated using maximum likelihood estimation
- Primary interest in estimating and testing hypotheses regarding β
- Large-sample test (Wald Test):
  - $H_0: \beta = 0$ vs. $H_A: \beta \neq 0$
  - Test statistic (TS): $X_{obs}^2 = \left(\frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}}\right)^2$
  - Rejected region (RR): $X_{obs}^2 \geq \chi_{\alpha,1}^2$
  - p value: $P(\chi^2 \geq X_{obs}^2)$

Maximum Likelihood Estimation of Logistic Regression Model

- $Y_i \sim \text{Bernoulli}(P_i)$, $\boxed{(x_1, y_1), \dots, (x_n, y_n)}$

This should be your training set. Those are given values.

Y_i	Probability
1	P_i
0	$1 - P_i$

$$f_i(y_i) = P(Y_i = y_i) = P_i^{y_i}(1 - P_i)^{1-y_i}, y_i \in \{0,1\}$$

- Likelihood Function

$$L = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} P_i^{y_i}(1 - P_i)^{1-y_i}$$

$$\log L = l = \sum_{i=1}^{n} y_i \log P_i + \sum_{i=1}^{n}(1 - y_i) \log(1 - P_i)$$

$$= \sum_{i=1}^{n} y_i \log \frac{P_i}{1-P_i} + \sum_{i=1}^{n} \log(1 - P_i)$$

$$= \sum_{i=1}^{n} y_i (\beta_0 + \beta x_i) - \sum_{i=1}^{n} \log(1 + e^{\beta_0 + \beta x_i})$$

Maximum Likelihood Estimation of Logistic Regression Model

- Partial derivatives w.r.t. β_0 and β

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \frac{e^{\beta_0+\beta x_i}}{1+e^{\beta_0+\beta x_i}}$$

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} x_i \frac{e^{\beta_0+\beta x_i}}{1+e^{\beta_0+\beta x_i}}$$

No closed-form solution ☹

- You can easily expand to the case of $Y_i \sim \mathrm{Binomial}(r_i, P_i)$

$$f_i(y_i) = P(Y_i = y_i) = \binom{r_i}{y_i} P_i^{y_i} (1 - P_i)^{r_i - y_i}, \quad y_i = 0,1, \ldots, r_i$$

$$L = \prod_{i=1}^{n} = \binom{r_i}{y_i} P_i^{y_i} (1 - P_i)^{r_i - y_i}$$

Maximum Likelihood Estimation of Logistic Regression Model

- Newton-Raphson method (gradient descent method)
  - Iterative process

$$\boldsymbol{\beta}^{new} = \boldsymbol{\beta}^{old} - \left[l''\left(\boldsymbol{\beta}^{old}\right)\right]^{-1} l'\left(\boldsymbol{\beta}^{old}\right)$$

$$= \boldsymbol{\beta}^{old} + \left(\boldsymbol{X}^T \boldsymbol{W}^{old} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \left[\boldsymbol{y} - \boldsymbol{\mu}^{old}\right]$$ Matrix notation

$$\boldsymbol{W}^{old} = \mathrm{diag}\left(P\left(Y = 1 \middle| \boldsymbol{x}_i, \boldsymbol{\beta}^{old}\right)\left(1 - P\left(Y = 1 | \boldsymbol{x}_i, \boldsymbol{\beta}^{old}\right)\right)\right)$$

$\boldsymbol{\mu}^{old}$: n-dimensional vector with i-th element $P\left(Y = 1 \middle| \boldsymbol{x}_i, \boldsymbol{\beta}^{old}\right) = \frac{\exp({\boldsymbol{\beta}^{old}}^T \boldsymbol{x}_i)}{1+\exp({\boldsymbol{\beta}^{old}}^T \boldsymbol{x}_i)}$

It converges after 10~20 iterations!

c.f.) vector notation

$$l'\left(\boldsymbol{\beta}^{old}\right) = \sum_{i=1}^{n} \boldsymbol{x}_i (y_i - P(Y = 1 | \boldsymbol{x}_i, \boldsymbol{\beta})) = 0$$

$$l''\left(\boldsymbol{\beta}^{old}\right) = -\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T P(Y = 1 | \boldsymbol{x}_i, \boldsymbol{\beta})(1 - P(Y = 1 | \boldsymbol{x}_i, \boldsymbol{\beta})) = 0$$

Example - Rizatriptan for Migraine

- Response - Complete Pain Relief at 2 hours (Yes/No)
- Predictor - Dose (*mg*): Placebo (0),2.5,5,10

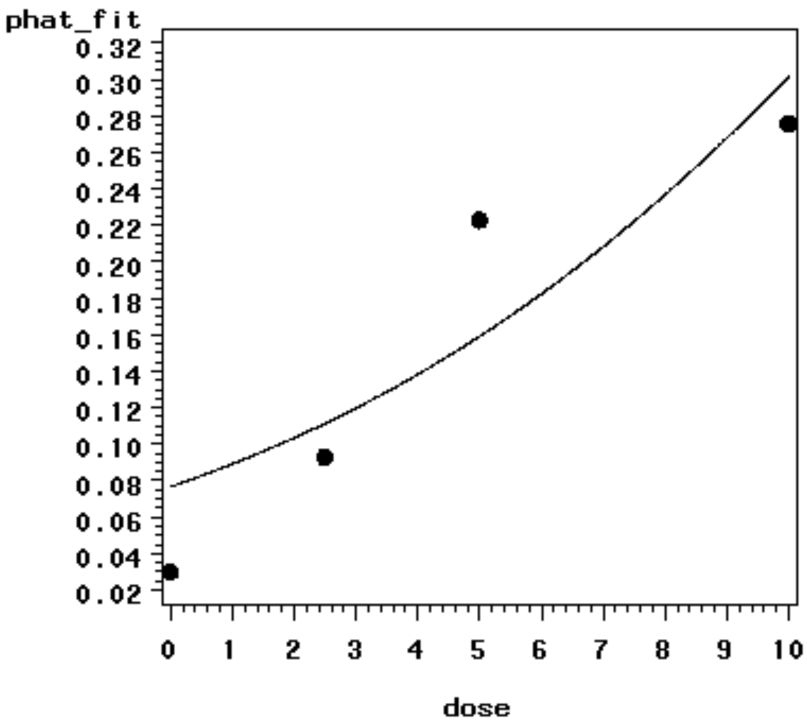
Dose	# Patients	# Relieved	% Relieved
0	67	2	3.0
2.5	75	7	9.3
5	130	29	22.3
10	145	40	27.6

Example - Rizatriptan for Migraine (SPSS)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	DOSE	.165	.037	19.819	1	.000	1.180
	Constant	-2.490	.285	76.456	1	.000	.083

a. Variable(s) entered on step 1: DOSE.



- $\hat{P}(y = 1|X) = \frac{e^{-2.490+0.165X}}{1+e^{-2.490+0.165X}}$
- $H_0: \beta = 0$ vs. $H_A: \beta \neq 0$
- TS: $X^2_{obs} = \left(\frac{0.165}{0.037}\right)^2 = 19.819$
- RR: $X^2_{obs} \geq \chi^2_{0.05,1} = 3.84$
- p value: 0.000

Odds Ratio

- Interpretation of Regression Coefficient (β):
  - In linear regression, the slope coefficient is the change in the mean response as X increases by 1 unit
  - In logistic regression, we can show that:

$$\frac{odds(X+1)}{odds(X)} = e^{\beta} \quad \left(odds = \frac{P(Y=1)}{1-P(Y=1)}\right)$$

For example, β=1.1
We can introduce a specific case:
X → odds=1 (P=0.5)
X+1 → odds=3 (P=0.75)

- Thus e^{β} represents the change in the odds of the outcome (multiplicatively) by increasing X by 1 unit
  - If $\beta = 0$, the odds and probability are the same at all X levels ($e^{\beta} = 1$)
  - If $\beta > 0$, the odds and probability increase as X increases ($e^{\beta} > 1$)
  - If $\beta < 0$, the odds and probability decrease as X increases ($e^{\beta} < 1$)

Logistic Regression with p Predictors

- Extension to more than one predictor variable (either numeric or dummy variables).
- Model: $P(y = 1|\boldsymbol{X}) \equiv$ Probability of presence at predictor level $\boldsymbol{X}$ consisting of p random variables $(X_1, \ldots, X_p)$

$$\underbrace{\log\left(\underbrace{\boxed{\frac{P(Y=1)}{1-P(Y=1)}}}_{\text{Odd ratio}}\right)}_{\text{Logit}} = \boldsymbol{X\beta} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$P(Y=1) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

- Adjusted odds ratio for raising X_k by 1 unit, holding all other predictors constant:

$$e^{\beta_k}$$

- Many models have nominal/ordinal predictors, and widely make use of dummy variables

Testing Regression Coefficients

- Testing the overall model:

  $H_0: \beta_1 = \cdots = \beta_p = 0$ vs. $H_A: \text{Not all } \beta_k = 0 \; k = 1, \ldots, p$

  TS: $X^2_{obs} = 2\log\left(\frac{L_1}{L_0}\right) = -2\log(L_0) + 2\log(L_1)$

  RR: $X^2_{obs} \geq \chi^2_{\alpha,p}$

  $P = P\left(\chi^2 \geq X^2_{obs}\right)$

- L_0, L_1 are values of the maximized likelihood function based on the reduced model and full model.
- This logic can also be used to compare full and reduced models based on subsets of predictors.
- Testing for individual terms is done as in model with a single predictor.

Testing Regression Coefficients

- How can we make the reduced model?
  - For testing the entire model

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 \text{ (null, } L_0\text{)}$$

vs.

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \text{ (alternative, } L_1\text{)}$$

  - For testing a specific coefficient, for example β_1

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_2 X_2 + \beta_p X_p \text{ (null, } L_0\text{)}$$

vs.

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \text{ (alternative, , } L_1\text{)}$$

Example - Erectile dysfunction (ED)

- Response: Presence/Absence of ED in older Dutch men ($n = 1688$)
- Predictors ($p = 12$):
  - Age stratum (50-54*, 55-59, 60-64, 65-69, 70-78)
  - Smoking status (Nonsmoker*, Smoker)
  - BMI stratum (<25*, 25-30, >30)
  - Lower urinary tract symptoms (None*, Mild, Moderate, Severe)
  - Under treatment for cardiac symptoms (No*, Yes)
  - Under treatment for COPD (No*, Yes)

    * Baseline group for dummy variables

Example - Erectile dysfunction (ED)

Predictor	b	s_b	Adjusted OR (95% CI)
Age 55-59 (vs 50-54)	0.83	0.42	2.3 (1.0 – 5.2)
Age 60-64 (vs 50-54)	1.53	0.40	4.6 (2.1 – 10.1)
Age 65-69 (vs 50-54)	2.19	0.40	8.9 (4.1 – 19.5)
Age 70-78 (vs 50-54)	2.66	0.41	14.3 (6.4 – 32.1)
Smoker (vs nonsmoker)	0.47	0.19	1.6 (1.1 – 2.3)
BMI 25-30 (vs <25)	0.41	0.21	1.5 (1.0 – 2.3)
BMI >30 (vs <25)	1.10	0.29	3.0 (1.7 – 5.4)
LUTS Mild (vs None)	0.59	0.41	1.8 (0.8 – 4.3)
LUTS Moderate (vs None)	1.22	0.45	3.4 (1.4 – 8.4)
LUTS Severe (vs None)	2.01	0.56	7.5 (2.5 – 22.5)
Cardiac symptoms (Yes vs No)	0.92	0.26	2.5 (1.5 – 4.3)
COPD (Yes vs No)	0.64	0.28	1.9 (1.1 – 3.6)

- Interpretations: Risk of ED appears to be:
  - Increasing with age, BMI, and LUTS strata
  - Higher among smokers

Nominal Predictors

- Create dummy variables according to the number of categories (c), i.e., generate $c - 1$ variables

	$\boldsymbol{D_1}$	$\boldsymbol{D_2}$	$\boldsymbol{D_3}$
Spring	0	0	0
Summer	1	0	0
Fall	0	1	0
Winter	0	0	1

- You can treat an ordinal variable as a continuous variable

Additional Material

Logistic Regression for Nominal Response

- When the dependent variable has three or more nominal type category (no natural ordering): Baseline-category logit model
- The odds of falling in category j or below:

$$\frac{P(Y_i = j)}{P(Y_i = c)}, j = 1, \dots, c-1$$

$$\sum_{j=1}^{c} P(Y_i = j) = 1$$

- Logit (log odds) of cumulative probabilities are modeled as linear functions of predictor variable(s) $\boldsymbol{x_i}$:

$$\text{logit}[P(Y_i = j)] = \log\left[\frac{P(Y_i=j)}{P(Y_i=c)}\right] = \boldsymbol{\beta}_j^{\boldsymbol{T}} \boldsymbol{x_i}, j = 1, \dots, c-1$$

Logistic Regression for Nominal Response

- **As a set of independent binary regressions**

$$\ln\left[\frac{P(Y_i = 1)}{P(Y_i = c)}\right] = \boldsymbol{\beta}_1^{\boldsymbol{T}}\boldsymbol{x_i}$$

$$\vdots$$

$$\ln\left[\frac{P(Y_i = c-1)}{P(Y_i = c)}\right] = \boldsymbol{\beta}_{c-1}^{\boldsymbol{T}}\boldsymbol{x_i}$$

- **Using (1)** $P(Y_i = c) = 1 - \sum_{j=1}^{c-1} P(Y_i = j)$

  (2) $P(Y_i = j) = P(Y_i = c)\exp(\boldsymbol{\beta}_j^{\boldsymbol{T}}\boldsymbol{x_i})$

$$P(Y_i = c), P(Y_i = j)?$$

Logistic Regression for Nominal Response

- $P_j = P(Y_i = j) = \frac{\exp(\boldsymbol{\beta}_j^T \boldsymbol{x}_i)}{1+\sum_{j=1}^{c-1} \exp(\boldsymbol{\beta}_j^T \boldsymbol{x}_i)}, j = 1, \dots, c-1$
- $P_c = P(Y_i = c) = \frac{1}{1+\sum_{j=1}^{c-1} \exp(\boldsymbol{\beta}_j^T \boldsymbol{x}_i)}$

Classify the sample to the class t

$$t = \mathrm{argmax}_j\, P_j$$

This applies to the others as well

- (Reference) Likelihood of $y_i \sim$ Multinomial distribution

$$L = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \prod_{j=1}^{c} P_j^{y_{ij}}$$

where $y_{ij} = \begin{cases} 1 & y_i = j \\ 0 & y_i \neq j \end{cases}$

$$\log L = \sum_{i=1}^{n} \sum_{j=1}^{c} y_{ij} \log P_j$$

Logistic Regression for Ordinal Response

- When the dependent variable has three or more ordinal type category (natural ordering): cumulative logit model
- The probability falling in category j or below:

$$P(Y_i \leq j) = P_1 + P_2 + \cdots + P_j, j = 1, \ldots, c-1$$

- Logit (log odds) of cumulative probabilities are modeled as linear functions of predictor variable(s) $\boldsymbol{X}$:

$$\text{logit}[P(Y_i \leq j)] = \log\left[\frac{P(Y_i \leq j)}{1 - P(Y_i \leq j)}\right] = \alpha_j + \boxed{\boldsymbol{\beta}^{\boldsymbol{T}}}\boldsymbol{x}_{\boldsymbol{i}}, j = 1, \ldots, c-1$$

- This is called the proportional odds model, and assumes the effect of X_k is the same for each cumulative probability.
  - able to reduce the number of parameters to be estimated

Logistic Regression for Ordinal Response

- $P_1 = P(Y_i \leq 1) = \frac{\exp(\alpha_1 + \boldsymbol{\beta}^T \boldsymbol{x_i})}{1+\exp(\alpha_1 + \boldsymbol{\beta}^T \boldsymbol{x_i})}$
- $P_j = P(Y_i \leq j) - P(Y_i \leq j-1) = \frac{\exp(\alpha_j + \boldsymbol{\beta}^T \boldsymbol{x_i})}{1+\exp(\alpha_j + \boldsymbol{\beta}^T \boldsymbol{x_i})} - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}^T \boldsymbol{x_i})}{1+\exp(\alpha_{j-1} + \boldsymbol{\beta}^T \boldsymbol{x_i})}$

  , $j = 2, \dots, c-1$
- $P_c = 1 - \frac{\exp(\alpha_{c-1} + \boldsymbol{\beta}^T \boldsymbol{x_i})}{1+\exp(\alpha_{c-1} + \boldsymbol{\beta}^T \boldsymbol{x_i})}$

Logistic Regression for Ordinal Response

- When the dependent variable has three or more ordinal type category (natural ordering): adjacent-categories logit model
- Logs of adjacent categories are modeled as linear functions of predictor variable(s) $\boldsymbol{X}$:

$$\log\left(\frac{P_{j+1}}{P_j}\right) = \boldsymbol{\beta}_j^T \boldsymbol{x_i}, j = 1, \dots, c-1$$

where $P_j = P_1 \exp\left(\sum_{k=2}^{j-1} \boldsymbol{\beta}_k^T \boldsymbol{x_i}\right), j = 2, \dots, c$

- Then, $P_1 = \left\{1 + \exp(\boldsymbol{\beta}_1^T \boldsymbol{x_i}) + \cdots + \exp\left(\sum_{k=1}^{c-1} \boldsymbol{\beta}_k^T \boldsymbol{x_i}\right)\right\}^{-1}$

⇨ We can represent in a new way, $\log\left(\frac{P_j}{\boxed{P_1}}\right) = \boldsymbol{\gamma}_j^T \boldsymbol{x_i}, j = 2, \dots, c$

(Similar to nominal response except for forcing P_1 as baseline)

Logistic Regression for Ordinal Response

- $P_1 = \frac{1}{1+\sum_{j=2}^{c} \exp(\boldsymbol{\gamma}_j^T \boldsymbol{x}_i)}$
- $P_j = P_1 \exp(\boldsymbol{\gamma}_j^T \boldsymbol{x}_i) = \frac{\exp(\boldsymbol{\gamma}_j^T \boldsymbol{x}_i)}{1+\sum_{j=2}^{c} \exp(\boldsymbol{\gamma}_j^T \boldsymbol{x}_i)}, j = 2, \dots, \mathrm{c}$

- (Reference) Likelihood of y_i ~ Multinomial distribution (same in ordinal response)

$$L = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \prod_{j=1}^{c} P_j^{y_{ij}}$$

where $y_{ij} = \begin{cases} 1 & y_i = 1 \\ 0 & y_i \neq j \end{cases}$

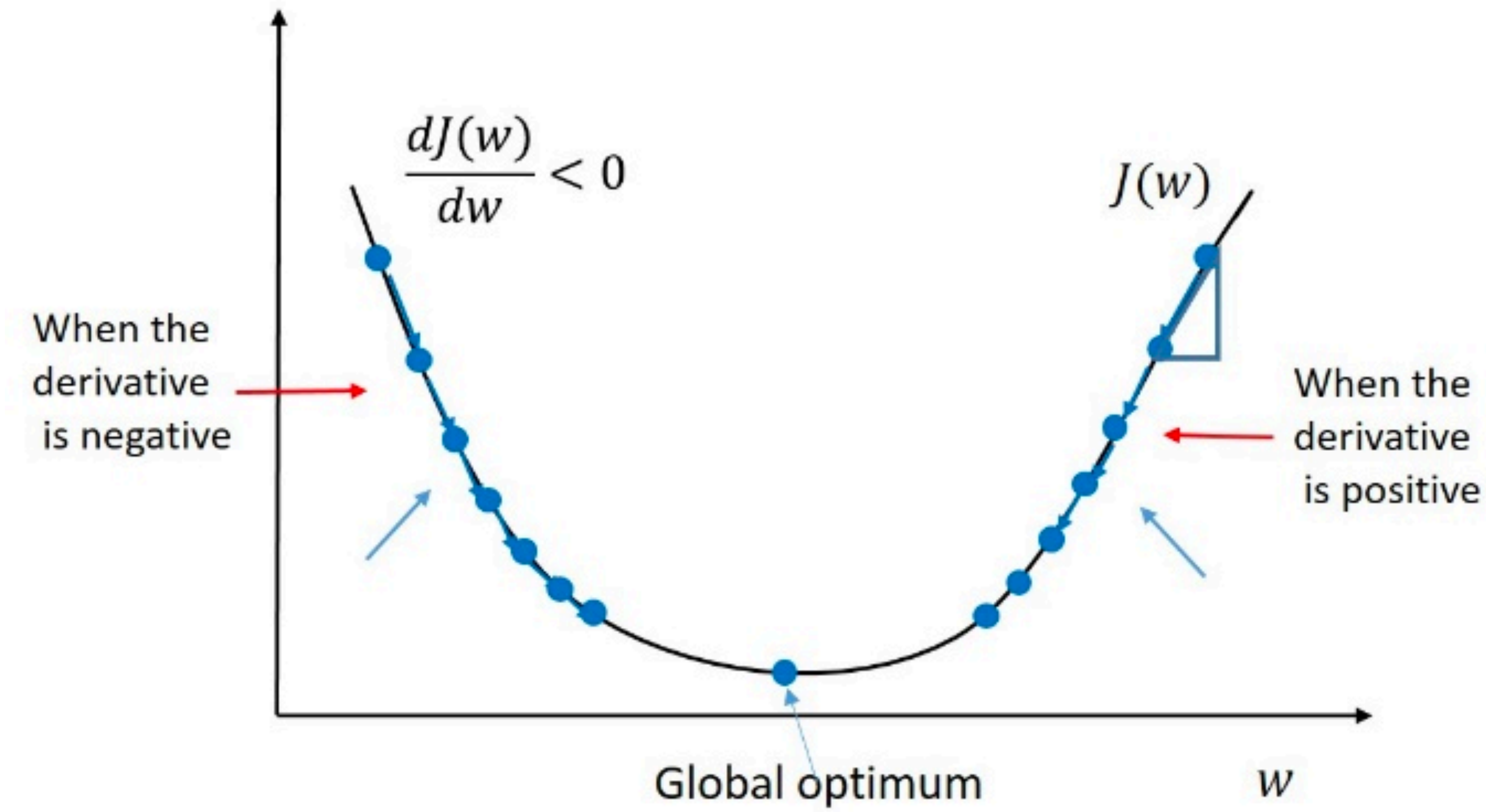
$$\log L = \sum_{i=1}^{n} \sum_{j=1}^{c} y_{ij} \log P_j$$

Questions?

Appendix

Gradient Descent Algorithm

- f(x), J(w), F(a) are notations of objective functions in terms of x, w, a respectively.



Gradient Descent Algorithm

- Gradient descent is a first-order iterative optimization algorithm for finding a local minimum of a differentiable function (minimization problem).

- Gradient descent is based on the observation that if the multivariate function f is defined and differentiable in a neighborhood of a point $\boldsymbol{x}$, then $f(\boldsymbol{x})$ decreases fastest if one goes from $\boldsymbol{x}$ in the direction of the negative gradient of f at $\boldsymbol{x}$, $-\nabla f(\boldsymbol{x})$. It follows that, if

$$\boldsymbol{x}_{\boldsymbol{t}+\mathbf{1}} = \boldsymbol{x}_{\boldsymbol{t}} - \eta \nabla f(\boldsymbol{x}_{\boldsymbol{t}})$$

for $\eta \in \mathbb{R}_{+}$ small enough, then $f(\boldsymbol{x}_{\boldsymbol{t}}) \geq f(\boldsymbol{x}_{\boldsymbol{t}+\mathbf{1}})$

Newton-Raphson Method

- An iterative method for finding the roots of a twice differentiable function *f*. Given *f*, we seek to solve the optimization problem

$$\min_{x \in \mathbb{R}} f(x) \qquad \text{vs.} \qquad \min_{\boldsymbol{\theta} \in \mathbb{R}^p} f(\boldsymbol{\theta})$$

- Newton's method performs the iteration

$$x_{t+1} = x_t - \frac{f'(x_t)}{f''(x_t)} \qquad \text{vs.} \qquad \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - [f''(\boldsymbol{\theta})]^{-1} f'(\boldsymbol{\theta})$$

Standard Error $\hat{\sigma}_{\widehat{\beta}_i}$

- You can find the second derivatives of the p+1 parameters

$$\boldsymbol{H} = \left[h_{ij}\right], i, j = 0,1, \ldots, p$$

  where $h_{ii} = \frac{\partial^2 \log L}{\partial \beta_i^2}, i = 0,1, \ldots, p$ and $h_{ij} = \frac{\partial^2 \log L}{\partial \beta_i \partial \beta_j}, i \neq j$

- The approximate variance-covariance matrix of the estimated parameters

$$\boldsymbol{S} = \left[\left(-h_{ij}\right)_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\right]^{-1}$$

- Therefore, the standard error $s_{\widehat{\beta}_i} (= \hat{\sigma}_{\widehat{\beta}_i})$ is the square-root of the (i+1)-th diagonal component (since there is no 0-th component).

Standard Error $\hat{\sigma}_{\hat{\beta}_i}$

- For example, we have

$$logit(P) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

then,

$$\boldsymbol{H} = \begin{bmatrix} -\sum \hat{P}_i(1-\hat{P}_i) & -\sum x_{1i}\hat{P}_i(1-\hat{P}_i) & -\sum x_{2i}\hat{P}_i(1-\hat{P}_i) \\ -\sum x_{1i}\hat{P}_i(1-\hat{P}_i) & -\sum x_{1i}^2\hat{P}_i(1-\hat{P}_i) & -\sum x_{1i}x_{2i}\hat{P}_i(1-\hat{P}_i) \\ -\sum x_{2i}\hat{P}_i(1-\hat{P}_i) & -\sum x_{1i}x_{2i}\hat{P}_i(1-\hat{P}_i) & -\sum x_{2i}^2\hat{P}_i(1-\hat{P}_i) \end{bmatrix}$$

where $\hat{P}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i})}{1+\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i})}$.