Modelling Binary Data with Phi-divergence Measures

Leandro PARDO
Departamento de Estadística e Investigación Operativa
Facultad de Matemáticas
Universidad Complutense

Contents

Contents

Contents

Contents

Contents

Contents

- We consider a binary random variable

$$Z = \begin{cases} 1 & \text{if the outcome is a success} \quad \Pr(Z=1)=\pi \\ 0 & \text{if the outcome is a failure} \quad \Pr(Z=0)=1-\pi \end{cases}$$

Generalized Linear Model with Binary Data

- We consider a binary random variable

$$Z = \begin{cases} 1 & \text{if the outcome is a success} & \Pr(Z=1) = \pi \\ 0 & \text{if the outcome is a failure} & \Pr(Z=0) = 1-\pi \end{cases}$$

- $Z_1, \ldots, Z_n \equiv$ Independent binary random variables with $\Pr\left(Z_j = 1\right) = \pi$, $j = 1, \ldots, n$

$$Y \equiv \sum_{j=1}^{n} Z_j \equiv B(n, \pi)$$

Generalized Linear Model with Binary Data

- We consider a binary random variable

$$Z = \begin{cases} 1 & \text{if the outcome is a success} \quad \Pr(Z=1)=\pi \\ 0 & \text{if the outcome is a failure} \quad \Pr(Z=0)=1-\pi \end{cases}$$

- $Z_1, \ldots, Z_n \equiv$ Independent binary random variables with $\Pr\left(Z_j = 1\right) = \pi\,,\ j = 1, \ldots, n$

$$Y \equiv \sum_{j=1}^{n} Z_j \equiv B(n, \pi)$$

- $Y_1, \ldots, Y_I \equiv$ Independent Binomial random variables with $Y_j \equiv B\left(n_j, \pi_j\right)\,,\ j = 1, \ldots., I$

	1	2	.....	.....	I-1	I
Successes	n_{11}	n_{12}	.....	.....	n_{1I-1}	n_{1I}
Failures	$n_1 - n_{11}$	$n_2 - n_{12}$	.....	.....	$n_{I-1} - n_{1I-1}$	$n_I - n_{1I}$
Totals	n_1	n_2	.....	.....	n_{I-1}	n_I

- Example 1
  The data in the following table describes mortality of groups of the beetle *Tribolium Castaneum* to the insecticide γ-benzene hexachloride ($\gamma - BHC$)

	Concentration					
	1.08	1.16	1.21	1.26	1.31	1.35
(Success) Killed	15	24	26	24	29	29
(Failure) Not Killed	35	25	24	26	21	21
Totals	50	49	50	50	50	50

Generalized Linear Model

- Example 1
  The data in the following table describes mortality of groups of the beetle *Tribolium Castaneum* to the insecticide γ-benzene hexachloride $(\gamma - BHC)$

	Concentration					
	1.08	1.16	1.21	1.26	1.31	1.35
(Success) Killed	15	24	26	24	29	29
(Failure) Not Killed	35	25	24	26	21	21
Totals	50	49	50	50	50	50

  - Higher concentrations of the insecticide generally resulted in higher death rates, but notice that the empirical fraction killed is not monotone.

- Example 1
  The data in the following table describes mortality of groups of the beetle *Tribolium Castaneum* to the insecticide γ-benzene hexachloride $(\gamma - BHC)$

	Concentration					
	1.08	1.16	1.21	1.26	1.31	1.35
(Success) Killed	15	24	26	24	29	29
(Failure) Not Killed	35	25	24	26	21	21
Totals	50	49	50	50	50	50

  - Higher concentrations of the insecticide generally resulted in higher death rates, but notice that the empirical fraction killed is not monotone.
    - $Y_i \equiv$Number the dead insect where concentration is x_i, $B\left(n_i, \pi\left(x_i\right)\right)$

- Example 1
  The data in the following table describes mortality of groups of the beetle *Tribolium Castaneum* to the insecticide γ-benzene hexachloride $(\gamma - BHC)$

	Concentration					
	1.08	1.16	1.21	1.26	1.31	1.35
(Success) Killed	15	24	26	24	29	29
(Failure) Not Killed	35	25	24	26	21	21
Totals	50	49	50	50	50	50

  - Higher concentrations of the insecticide generally resulted in higher death rates, but notice that the empirical fraction killed is not monotone.
    - $Y_i \equiv$Number the dead insect where concentration is x_i, $B\left(n_i, \pi\left(x_i\right)\right)$
    - $X \equiv$ Explanatory variable (concentration insecticide)

- Example 1
  The data in the following table describes mortality of groups of the beetle *Tribolium Castaneum* to the insecticide γ-benzene hexachloride ($\gamma - BHC$)

	Concentration					
	1.08	1.16	1.21	1.26	1.31	1.35
(Success) Killed	15	24	26	24	29	29
(Failure) Not Killed	35	25	24	26	21	21
Totals	50	49	50	50	50	50

  - Higher concentrations of the insecticide generally resulted in higher death rates, but notice that the empirical fraction killed is not monotone.
    - $Y_i \equiv$Number the dead insect where concentration is x_i, $B\left(n_i, \pi\left(x_i\right)\right)$
    - $X \equiv$ Explanatory variable (concentration insecticide)
    - $\Pr\left(Z = 1/X = x\right) = \pi\left(x\right)$

- **Example 2.** The data, taken from Lee (1974), consist of patient characteristics and whether or not cancer remission accrued. The variable REMISS is the cancer remission indicator variable with a value of 1 for remission and a value of 0 for nonremission. The other six variables are the risk factors thought to be related to cancer remission.

  1 .8 .83 .66 1.9 1.1 .999

  1 .9 .36 .32 1.4 .74 .992 .

  0 .8 .88 .7 .8 .176 .982

  0 1 .87 .87 .7 1.053 .986

  1 .9 .75 .68 1.3 .519 .98

  0 1 .65 .65 .6 .519 .982

  1 .95 .97 .92 1 1.23 .992

  0 .95 .87 .83 1.9 1.354 1.02

  0 1 .45 .45 .8 .322 .999

  0 .95 .36 .34 .5 0 1.038

  0 .85 .39 .33 .7 .279 .988

Generalized Linear Model

- 1 1 .84 .84 1.9 2.064 1.02
  0 .65 .42 .27 .5 .114 1.014
  0 1 .75 .75 1 1.322 1.004
  0 .5 .44 .22 .6 .114 .99
  1 1 .63 .63 1.1 1.072 .986
  0 1 .33 .33 .4 .176 1.01
  0 .9 .93 .84 .6 1.591 1.02
  1 1 .58 .58 1 .531 1.002

- 1 1 .84 .84 1.9 2.064 1.02
  0 .65 .42 .27 .5 .114 1.014
  0 1 .75 .75 1 1.322 1.004
  0 .5 .44 .22 .6 .114 .99
  1 1 .63 .63 1.1 1.072 .986
  0 1 .33 .33 .4 .176 1.01
  0 .9 .93 .84 .6 1.591 1.02
  1 1 .58 .58 1 .531 1.002
  - $\underbrace{REMISS}_{Z} \equiv \begin{cases} 1 & \text{Remission of cancer} \\ 0 & \text{Nonremision} \end{cases}$

Generalized Linear Model

- 1 1 .84 .84 1.9 2.064 1.02
  0 .65 .42 .27 .5 .114 1.014
  0 1 .75 .75 1 1.322 1.004
  0 .5 .44 .22 .6 .114 .99
  1 1 .63 .63 1.1 1.072 .986
  0 1 .33 .33 .4 .176 1.01
  0 .9 .93 .84 .6 1.591 1.02
  1 1 .58 .58 1 .531 1.002
  - $\underbrace{REMISS}_{Z} \equiv \begin{cases} 1 & \text{Remission of cancer} \\ 0 & \text{Nonremision} \end{cases}$
  - $\underbrace{X_1, ..., X_6}_{\text{Explanatory Variables}} \equiv$ Risk factors

Generalized Linear Model

- 1 1 .84 .84 1.9 2.064 1.02
  0 .65 .42 .27 .5 .114 1.014
  0 1 .75 .75 1 1.322 1.004
  0 .5 .44 .22 .6 .114 .99
  1 1 .63 .63 1.1 1.072 .986
  0 1 .33 .33 .4 .176 1.01
  0 .9 .93 .84 .6 1.591 1.02
  1 1 .58 .58 1 .531 1.002
  - $\underbrace{REMISS}_{Z} \equiv \begin{cases} 1 & \text{Remission of cancer} \\ 0 & \text{Nonremision} \end{cases}$
  - $\underbrace{X_1, ..., X_6}_{\text{Explanatory Variables}} \equiv$ Risk factors
  - $\pi\left(x_1, ..., x_6\right) = \Pr\left(REMISS = 1/X_1 = x_1, ..., X_6 = x_6\right)$?

Generalized Linear Model

- Suppose Y_i $(i = 1, ..., I)$ are independent binomial random variables $(Y_i \equiv B(n_i, \pi_i))$

Generalized Linear Model

- Suppose Y_i $(i = 1, ..., I)$ are independent binomial random variables $(Y_i \equiv B(n_i, \pi_i))$
  - Probabilities $\pi_i = \pi(X_0 = x_{i0}, X_1 = x_{i1}, ..., X_k = x_{ik})$ are functions of $k+1$ explanatory variables and $k+1$ unknown parameters by the formula

$$\eta_j = g(\pi_j) = \sum_{l=0}^{K} \beta_l x_{jl},\ j = 1, ..., I,$$

Generalized Linear Model

- Suppose Y_i $(i = 1, ..., I)$ are independent binomial random variables $(Y_i \equiv B(n_i, \pi_i))$
  - Probabilities $\pi_i = \pi(X_0 = x_{i0}, X_1 = x_{i1}, ..., X_k = x_{ik})$ are functions of $k+1$ explanatory variables and $k+1$ unknown parameters by the formula

$$\eta_j = g(\pi_j) = \sum_{l=0}^{K} \beta_l x_{jl}, \ j = 1, ..., I,$$

    - $g : [0, 1] \to \mathbb{R}$ (monotonic and differentiable) LINK FUNCTION

Generalized Linear Model

- Suppose Y_i $(i = 1, ..., I)$ are independent binomial random variables $(Y_i \equiv B(n_i, \pi_i))$
  - Probabilities $\pi_i = \pi(X_0 = x_{i0}, X_1 = x_{i1}, ..., X_k = x_{ik})$ are functions of $k+1$ explanatory variables and $k+1$ unknown parameters by the formula

$$\eta_j = g(\pi_j) = \sum_{l=0}^{K} \beta_l x_{jl},\ j = 1, ..., I,$$

    - $g : [0,1] \to \mathbb{R}$ (monotonic and differentiable) LINK FUNCTION
    - $g^{-1}(\eta_j) = \pi_j$ (easy to compute)

Generalized Linear Model

- Suppose Y_i $(i = 1, ..., I)$ are independent binomial random variables $(Y_i \equiv B(n_i, \pi_i))$
  - Probabilities $\pi_i = \pi(X_0 = x_{i0}, X_1 = x_{i1}, ..., X_k = x_{ik})$ are functions of $k+1$ explanatory variables and $k+1$ unknown parameters by the formula
$$\eta_j = g(\pi_j) = \sum_{l=0}^{K} \beta_l x_{jl}, \; j = 1, ..., I,$$
    - $g : [0,1] \to \mathbb{R}$ (monotonic and differentiable) LINK FUNCTION
    - $g^{-1}(\eta_j) = \pi_j$ (easy to compute)
  - $\boldsymbol{\beta} = (\beta_0, ..., \beta_K)^T \in \Theta = \{(\beta_0, ..., \beta_K) : \beta_i \in (-\infty, \infty), \; i = 0, ..., K\}$

Generalized Linear Model

- Suppose Y_i $(i = 1, ..., I)$ are independent binomial random variables $(Y_i \equiv B(n_i, \pi_i))$
  - Probabilities $\pi_i = \pi(X_0 = x_{i0}, X_1 = x_{i1}, ..., X_k = x_{ik})$ are functions of $k+1$ explanatory variables and $k+1$ unknown parameters by the formula

$$\eta_j = g(\pi_j) = \sum_{l=0}^{K} \beta_l x_{jl}, \; j = 1, ..., I,$$

    - $g : [0,1] \to \mathbb{R}$ (monotonic and differentiable) LINK FUNCTION
    - $g^{-1}(\eta_j) = \pi_j$ (easy to compute)
  - $\boldsymbol{\beta} = (\beta_0, ..., \beta_K)^T \in \Theta = \{(\beta_0, ..., \beta_K) : \beta_i \in (-\infty, \infty), \; i = 0, ..., K\}$
  - $\mathbf{x}_i = (x_{i0}, x_{i1}, ..., x_{ik})$ and $\pi_i = \pi(\mathbf{x}_i \boldsymbol{\beta})$, $x_{i0} = 1$, $\forall i = 1, ..., I$.

Generalized Linear Model

- Suppose Y_i $(i = 1, ..., I)$ are independent binomial random variables $(Y_i \equiv B(n_i, \pi_i))$
  - Probabilities $\pi_i = \pi(X_0 = x_{i0}, X_1 = x_{i1}, ..., X_k = x_{ik})$ are functions of $k+1$ explanatory variables and $k+1$ unknown parameters by the formula

$$\eta_j = g(\pi_j) = \sum_{l=0}^{K} \beta_l x_{jl}, \ j = 1, ..., I,$$

    - $g : [0,1] \to \mathbb{R}$ (monotonic and differentiable) LINK FUNCTION
    - $g^{-1}(\eta_j) = \pi_j$ (easy to compute)
  - $\boldsymbol{\beta} = (\beta_0, ..., \beta_K)^T \in \Theta = \{(\beta_0, ..., \beta_K) : \beta_i \in (-\infty, \infty), \ i = 0, ..., K\}$
  - $\mathbf{x}_i = (x_{i0}, x_{i1}, ..., x_{ik})$ and $\pi_i = \pi(\mathbf{x}_i \boldsymbol{\beta})$, $x_{i0} = 1$, $\forall i = 1, ..., I$.
  - $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_I)^T_{I \times (K+1)}$, rank$(\mathbf{X}) = K + 1$

Generalized Linear Model

- Logistic regression model

$$\pi\left(\mathbf{x}_i \boldsymbol{\beta}\right)=\frac{\exp\left(\sum_{j=0}^{k}\beta_j x_{ij}\right)}{1+\exp\left(\sum_{j=0}^{k}\beta_j x_{ij}\right)},\quad \eta_i=g\left(\pi_i\right)=\log\left(\pi_i/\left(1-\pi_i\right)\right)$$

Generalized Linear Model

- Logistic regression model

$$\pi\left(\mathbf{x}_i\boldsymbol{\beta}\right)=\frac{\exp\left(\sum_{j=0}^{k}\beta_j x_{ij}\right)}{1+\exp\left(\sum_{j=0}^{k}\beta_j x_{ij}\right)},\ \ \eta_i=g\left(\pi_i\right)=\log\left(\pi_i/\left(1-\pi_i\right)\right)$$

- Probit Model

$$\pi\left(\mathbf{x}_i\boldsymbol{\beta}\right)=\int_{-\infty}^{\sum_{j=0}^{k}\beta_j x_{ij}}\left(2\pi\right)^{-1/2}e^{-\frac{t^2}{2}}dt,\ \ \eta_i=g\left(\pi_i\right)=\Phi^{-1}\left(\pi_i\right)$$

Generalized Linear Model

- Logistic regression model

$$\pi\left(\mathbf{x}_i\boldsymbol{\beta}\right)=\frac{\exp\left(\sum_{j=0}^{k}\beta_j x_{ij}\right)}{1+\exp\left(\sum_{j=0}^{k}\beta_j x_{ij}\right)},\quad \eta_i=g\left(\pi_i\right)=\log\left(\pi_i/\left(1-\pi_i\right)\right)$$

- Probit Model

$$\pi\left(\mathbf{x}_i\boldsymbol{\beta}\right)=\int_{-\infty}^{\sum_{j=0}^{k}\beta_j x_{ij}}(2\pi)^{-1/2}e^{-\frac{t^2}{2}}dt,\quad \eta_i=g\left(\pi_i\right)=\Phi^{-1}\left(\pi_i\right)$$

- Log-Log Model

$$\pi\left(\mathbf{x}_i\boldsymbol{\beta}\right)=\exp\left\{-\exp\left(\sum_{j=0}^{k}\beta_j x_{ij}\right)\right\},\quad \eta_i=g\left(\pi_i\right)=-\log(-\log\pi_i)$$

Generalized Linear Model

- Complementary-log-log

$$\pi\ (\mathbf{x}_i \boldsymbol{\beta}) = 1 - \exp\left\{ - \exp(\sum_{j=0}^{k} \beta_j x_{ij}) \right\},\ \ \eta_i = g\ (\pi\ _i) = \log(-\log\ (1-\pi\ _i))$$

Generalized Linear Model

- Complementary-log-log

$$\pi\ (\mathbf{x}_i \boldsymbol{\beta}) = 1 - \exp\left\{ -\exp(\sum_{j=0}^{k} \beta_j x_{ij}) \right\},\ \ \eta_i = g\ (\pi_i) = \log(-\log\ (1-\pi_i))$$

- **Example 1**

$$\pi\ (\mathbf{x}\boldsymbol{\beta}) = \Pr\left(\underbrace{Z=1}_{\textit{Inset}\ \text{dies}} / \underbrace{X=x}_{\textit{Concentration}\ \text{insecticide}} \right) = \frac{\exp\ (\alpha+\beta x)}{1+\exp\ (\alpha+\beta x)}$$

Generalized Linear Model

- Complementary-log-log

$$\pi\ (\mathbf{x}_i\boldsymbol{\beta}) = 1 - \exp\left\{ - \exp(\sum_{j=0}^{k} \beta_j x_{ij}) \right\},\ \ \eta_i = g\ (\pi_i) = \log(-\log(1-\pi_i))$$

- **Example 1**

$$\pi\ (\mathbf{x}\boldsymbol{\beta}) = \Pr\left(\underbrace{Z=1}_{Inset \text{ dies}} / \underbrace{X=x}_{Concentration \text{ insecticide}} \right) = \frac{\exp\left(\alpha+\beta x\right)}{1+\exp\left(\alpha+\beta x\right)}$$

- **Example 2**

$$\begin{aligned} \pi\ (\mathbf{x}\boldsymbol{\beta}) &= \Pr\left(\underbrace{Z=1}_{\mathrm{Re}\, mission} / \underbrace{X_1 = x_1, \ldots, X_6 = x_6}_{Risk \text{ factors}} \right) \\ &= \frac{\exp\left(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots,+x_6\beta_6\right)}{1+\exp\left(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots,+x_6\beta_6\right)} \end{aligned}$$

Minimum phi-Divergence estimators in GLM with Binary Data

- Fix $\mathbf{x}_i = (x_{i0}, x_{i1}, ..., x_{ik})$ we obtain n_i observations of a binary random variable and we have a Binomial random variable $(Y_i \equiv B(n_i, \pi(\mathbf{x}_i))$ $)$ $Y_i, i = 1, ..., I$ $(N = \sum_{i=1}^{I} n_i)$.

	$\mathbf{x}_1$	$\mathbf{x}_2$	....	$\mathbf{x}_I$
Successes	n_{11}	n_{21}	....	n_{I1}
Failures	$n_{12} = n_1 - n_{11}$	$n_{22} = n_2 - n_{21}$	....	$n_{I2} = n_I - n_{I1}$
Totals	n_1	n_2	....	n_I

- The likelihood function for β is

$$L(\beta) = L(\alpha, \beta_1, ..., \beta_k) = \prod_{i=1}^{I} \binom{n_i}{n_{i1}} \pi(\mathbf{x}_i\beta)^{n_{i1}} (1 - \pi(\mathbf{x}_i\beta))^{n_i - n_{i1}}.$$

Maximum Likelihood estimator

- We consider the two following probability vectors

$$\left(\max_{\alpha,\beta_1,\ldots,\beta_k} \log L\left(\alpha,\beta_1,\ldots,\beta_k\right) \Longleftrightarrow \min_{\alpha,\beta_1,\ldots,\beta_k} D_K\left(\widehat{\mathbf{p}},\mathbf{p}\left(\boldsymbol{\beta}\right)\right)\right)$$

Maximum Likelihood estimator

- We consider the two following probability vectors
  - $\widehat{\mathbf{p}} = \left(\frac{n_{11}}{N}, \frac{n_{12}}{N}, \frac{n_{21}}{N}, \frac{n_{22}}{N}, ..., \frac{n_{I1}}{N}, \frac{n_{I2}}{N} \right)$

$$\left(\max_{\alpha,\beta_1,...,\beta_k} \log L\left(\alpha, \beta_1, ..., \beta_k\right) \Longleftrightarrow \min_{\alpha,\beta_1,...,\beta_k} D_K\left(\widehat{\mathbf{p}}, \mathbf{p}\left(\boldsymbol{\beta}\right)\right) \right)$$

Maximum Likelihood estimator

- We consider the two following probability vectors
  - $\widehat{\mathbf{p}} = \left(\frac{n_{11}}{N}, \frac{n_{12}}{N}, \frac{n_{21}}{N}, \frac{n_{22}}{N}, ..., \frac{n_{I1}}{N}, \frac{n_{I2}}{N} \right)$
  - $\mathbf{p}(\boldsymbol{\beta}) \equiv \left(\pi\left(\mathbf{x}_i \boldsymbol{\beta}\right) \frac{n_i}{N}, \left(1 - \pi\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right) \frac{n_i}{N} \right)_{i=1,...,I}$

$$\left(\max_{\alpha, \beta_1, ..., \beta_k} \log L\left(\alpha, \beta_1, ..., \beta_k\right) \Longleftrightarrow \min_{\alpha, \beta_1, ..., \beta_k} D_K\left(\widehat{\mathbf{p}}, \mathbf{p}\left(\boldsymbol{\beta}\right)\right) \right)$$

Maximum Likelihood estimator

- We consider the two following probability vectors
  - $\widehat{\mathbf{p}} = \left(\frac{n_{11}}{N}, \frac{n_{12}}{N}, \frac{n_{21}}{N}, \frac{n_{22}}{N}, \ldots, \frac{n_{I1}}{N}, \frac{n_{I2}}{N}\right)$
  - $\mathbf{p}(\boldsymbol{\beta}) \equiv \left(\pi(\mathbf{x}_i\boldsymbol{\beta})\frac{n_i}{N}, (1-\pi(\mathbf{x}_i\boldsymbol{\beta}))\frac{n_i}{N}\right)_{i=1,\ldots,I}$
- Kullback-Leibler Divergence between vectors $\widehat{\mathbf{p}}$ and $\mathbf{p}(\boldsymbol{\beta})$ is given by

$$
\begin{aligned}
D_K(\widehat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\beta})) &= \sum_{i=1}^{I}\left(\frac{n_{i1}}{N}\log\frac{n_{i1}/N}{\pi(\mathbf{x}_i\boldsymbol{\beta})\frac{n_i}{N}} + \frac{n_{i2}}{N}\log\frac{n_{i2}/N}{(1-\pi(\mathbf{x}_i\boldsymbol{\beta}))\frac{n_i}{N}}\right) \\
&= kte - \frac{1}{N}\log\prod_{i=1}^{I}\pi(\mathbf{x}_i\boldsymbol{\beta})^{n_{i1}}(1-\pi(\mathbf{x}_i\boldsymbol{\beta}))^{n_i-n_{i1}}.
\end{aligned}
$$

$$\left(\max_{\alpha,\beta_1,\ldots,\beta_k}\log L(\alpha,\beta_1,\ldots,\beta_k) \Longleftrightarrow \min_{\alpha,\beta_1,\ldots,\beta_k} D_K(\widehat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\beta}))\right)$$

- We consider the two following probability vectors
  - $\widehat{\mathbf{p}} = \left(\frac{n_{11}}{N}, \frac{n_{12}}{N}, \frac{n_{21}}{N}, \frac{n_{22}}{N}, ..., \frac{n_{I1}}{N}, \frac{n_{I2}}{N} \right)$
  - $\mathbf{p}(\boldsymbol{\beta}) \equiv \left(\pi\left(\mathbf{x}_i \boldsymbol{\beta}\right) \frac{n_i}{N}, \left(1 - \pi\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right) \frac{n_i}{N} \right)_{i=1,...,I}$
- Kullback-Leibler Divergence between vectors $\widehat{\mathbf{p}}$ and $\mathbf{p}(\boldsymbol{\beta})$ is given by

$$
\begin{aligned}
D_K\left(\widehat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\beta})\right) &= \sum_{i=1}^{I} \left(\frac{n_{i1}}{N} \log \frac{n_{i1}/N}{\pi\left(\mathbf{x}_i \boldsymbol{\beta}\right) \frac{n_i}{N}} + \frac{n_{i2}}{N} \log \frac{n_{i2}/N}{\left(1 - \pi\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right) \frac{n_i}{N}} \right) \\
&= kte - \frac{1}{N} \log \prod_{i=1}^{I} \pi\left(\mathbf{x}_i \boldsymbol{\beta}\right)^{n_{i1}} \left(1 - \pi\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right)^{n_i - n_{i1}}.
\end{aligned}
$$

  - $D_K\left(\widehat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\beta})\right) = kte - \log L\left(\alpha, \beta_1, ..., \beta_k\right)$

$$\left(\max_{\alpha, \beta_1, ..., \beta_k} \log L\left(\alpha, \beta_1, ..., \beta_k\right) \Longleftrightarrow \min_{\alpha, \beta_1, ..., \beta_k} D_K\left(\widehat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\beta})\right) \right)$$

Minimum distance estimators

- $$\left\{ \begin{array}{l} \boxed{\max_{\alpha,\beta_1,\ldots,\beta_k} \log L\left(\alpha, \beta_1, \ldots, \beta_k\right)} \\ \\ \text{is equivalent to} \\ \\ \boxed{\min_{\alpha,\beta_1,\ldots,\beta_k} D_K\left(\widehat{\mathbf{p}}, \mathbf{p}\left(\alpha, \beta_1, \ldots, \beta_k\right)\right)} \end{array} \right.$$

$$D\left(\widehat{\mathbf{p}}, \mathbf{p}\left(\widehat{\alpha}^*, \widehat{\beta}_1^*, \ldots, \widehat{\beta}_k^*\right)\right) = \inf_{\alpha,\beta_1,\ldots,\beta_k} D\left(\widehat{\mathbf{p}}, \mathbf{p}\left(\alpha, \beta_1, \ldots, \beta_k\right)\right)$$

Minimum distance estimators

- $$\begin{cases} \boxed{\max_{\alpha,\beta_1,\ldots,\beta_k} \log L\left(\alpha,\beta_1,\ldots,\beta_k\right)} \\ \text{is equivalent to} \\ \boxed{\min_{\alpha,\beta_1,\ldots,\beta_k} D_K\left(\widehat{\mathbf{p}},\mathbf{p}\left(\alpha,\beta_1,\ldots,\beta_k\right)\right)} \end{cases}$$
- **Why do not use any distance between the unrestricted estimator $\widehat{\mathbf{p}}$ and the probability vector, $\mathbf{p}(\beta)$, that characterizes the model different to the Kullback-Leibler divergence?**

$$D\left(\widehat{\mathbf{p}},\mathbf{p}\left(\widehat{\alpha}^*,\widehat{\beta}_1^*,\ldots,\widehat{\beta}_k^*\right)\right)=\inf_{\alpha,\beta_1,\ldots,\beta_k} D\left(\widehat{\mathbf{p}},\mathbf{p}\left(\alpha,\beta_1,\ldots,\beta_k\right)\right)$$

Phi-divergence measures

- Given the probability vectors
  $\widehat{\mathbf{p}} = (p_1, ..., p_k)^T \qquad \mathbf{p}(\boldsymbol{\theta}) = (p_1(\boldsymbol{\theta}), ..., p_k(\boldsymbol{\theta}))^T$

$$D_\phi(\widehat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\theta})) = \sum_{j=1}^{k} p_j(\boldsymbol{\theta})\, \phi\left(\frac{\widehat{p}_j}{p_j(\boldsymbol{\theta})}\right), \qquad \phi \in \Phi^*$$

Phi-divergence measures

- Given the probability vectors
  $\widehat{\mathbf{p}} = (p_1, ..., p_k)^T \qquad \mathbf{p}(\boldsymbol{\theta}) = (p_1(\boldsymbol{\theta}), ..., p_k(\boldsymbol{\theta}))^T$

$$D_\phi(\widehat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\theta})) = \sum_{j=1}^{k} p_j(\boldsymbol{\theta}) \, \phi\left(\frac{\widehat{p}_j}{p_j(\boldsymbol{\theta})}\right), \qquad \phi \in \Phi^*$$

  - Where Φ^* is the class of all convex functions $\phi(x)$, $x \geq 0$, such that

Phi-divergence measures

- Given the probability vectors
  $\widehat{\mathbf{p}} = (p_1, ..., p_k)^T \qquad \mathbf{p}(\boldsymbol{\theta}) = (p_1(\boldsymbol{\theta}), ..., p_k(\boldsymbol{\theta}))^T$

$$D_\phi(\widehat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\theta})) = \sum_{j=1}^{k} p_j(\boldsymbol{\theta})\, \phi\left(\frac{\widehat{p}_j}{p_j(\boldsymbol{\theta})}\right), \qquad \phi \in \Phi^*$$

  - Where Φ^* is the class of all convex functions $\phi(x)$, $x \geq 0$, such that
    - at $x = 1, \phi(1) = 0$,

Phi-divergence measures

- Given the probability vectors
  $\widehat{\mathbf{p}} = (p_1, ..., p_k)^T \qquad \mathbf{p}(\boldsymbol{\theta}) = (p_1(\boldsymbol{\theta}), ..., p_k(\boldsymbol{\theta}))^T$

$$D_\phi(\widehat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\theta})) = \sum_{j=1}^{k} p_j(\boldsymbol{\theta}) \phi\left(\frac{\widehat{p}_j}{p_j(\boldsymbol{\theta})}\right), \qquad \phi \in \Phi^*$$

  - Where Φ^* is the class of all convex functions $\phi(x)$, $x \geq 0$, such that
    - at $x = 1, \phi(1) = 0$,
    - at $x = 0$, $0\phi\left(0/0\right) = 0$

Phi-divergence measures

- Given the probability vectors
  $\widehat{\mathbf{p}} = (p_1, ..., p_k)^T \qquad \mathbf{p}(\boldsymbol{\theta}) = (p_1(\boldsymbol{\theta}), ..., p_k(\boldsymbol{\theta}))^T$

$$D_\phi(\widehat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\theta})) = \sum_{j=1}^{k} p_j(\boldsymbol{\theta})\, \phi\left(\frac{\widehat{p}_j}{p_j(\boldsymbol{\theta})}\right), \qquad \phi \in \Phi^*$$

  - Where Φ^* is the class of all convex functions $\phi(x)$, $x \geq 0$, such that
    - at $x = 1, \phi(1) = 0$,
    - at $x = 0$, $0\phi\left(0/0\right) = 0$
    - $0\phi\left(p/0\right) = p \lim_{u\to\infty} \phi(u)/u$

Phi-divergence measures

- Given the probability vectors
  $\widehat{\mathbf{p}} = (p_1, ..., p_k)^T \qquad \mathbf{p}(\boldsymbol{\theta}) = (p_1(\boldsymbol{\theta}), ..., p_k(\boldsymbol{\theta}))^T$

$$D_\phi(\widehat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\theta})) = \sum_{j=1}^{k} p_j(\boldsymbol{\theta})\,\phi\left(\frac{\widehat{p}_j}{p_j(\boldsymbol{\theta})}\right), \qquad \phi \in \Phi^*$$

  - Where Φ^* is the class of all convex functions $\phi(x)$, $x \geq 0$, such that
    - at $x = 1, \phi(1) = 0$,
    - at $x = 0$, $0\phi(0/0) = 0$
    - $0\phi(p/0) = p \lim_{u \to \infty} \phi(u)/u$

- Power-Divergence family

$$I^\lambda(\widehat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\theta})) = \frac{1}{\lambda(\lambda+1)} \sum_{i=1}^{M} \widehat{p}_i \left[\left(\frac{\widehat{p}_i}{p_i(\boldsymbol{\theta})}\right)^\lambda - 1\right], \quad -\infty < \lambda < \infty$$

Phi-divergence measures

- Given the probability vectors
  $\widehat{\mathbf{p}} = (p_1, ..., p_k)^T \qquad \mathbf{p}(\boldsymbol{\theta}) = (p_1(\boldsymbol{\theta}), ..., p_k(\boldsymbol{\theta}))^T$

$$D_\phi(\widehat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\theta})) = \sum_{j=1}^{k} p_j(\boldsymbol{\theta})\, \phi\left(\frac{\widehat{p}_j}{p_j(\boldsymbol{\theta})}\right), \qquad \phi \in \Phi^*$$

  - Where Φ^* is the class of all convex functions $\phi(x)$, $x \geq 0$, such that
    - at $x = 1, \phi(1) = 0$,
    - at $x = 0$, $0\phi\left(0/0\right) = 0$
    - $0\phi\left(p/0\right) = p \lim_{u \to \infty} \phi(u)/u$
- Power-Divergence family

$$I^\lambda(\widehat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\theta})) = \frac{1}{\lambda(\lambda+1)} \sum_{i=1}^{M} \widehat{p}_i \left[\left(\frac{\widehat{p}_i}{p_i(\boldsymbol{\theta})}\right)^\lambda - 1\right], \quad -\infty < \lambda < \infty$$

  - $\phi(x) = x \log x - x + 1$ (Kullback-Leibler)

- Minimum *phi*-divergence estimator in the Generalized Linear Model are the values $\widehat{\boldsymbol{\beta}}_\phi = \left(\widehat{\beta}_0^\phi, \widehat{\beta}_1^\phi, \ldots, \widehat{\beta}_K^\phi\right)$ verifying

$$\widehat{\boldsymbol{\beta}}_\phi = \arg\min_{\boldsymbol{\beta}\in\Theta} D_\phi\left(\widehat{\mathbf{p}}, \mathbf{p}\left(\boldsymbol{\beta}\right)\right),$$

where

$$\begin{aligned} D_\phi\left(\widehat{\mathbf{p}}, \mathbf{p}\left(\boldsymbol{\beta}\right)\right) = & \sum_{i=1}^{I} \frac{n_i}{N} \left\{ \pi\left(\mathbf{x}_i^T\boldsymbol{\beta}\right) \phi\left(\frac{n_{i1}}{\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}\right) n_i} \right) \right. \\ & \left. + \left(1 - \pi\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)\right) \phi\left(\frac{n_{i2}}{\left(1 - \pi\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)\right) n_i} \right) \right\}. \end{aligned}$$

Minimum phi-divergence estimator in GLM

- Minimum *phi*-divergence estimator in the Generalized Linear Model are the values $\widehat{\boldsymbol{\beta}}_\phi = \left(\widehat{\beta}_0^\phi, \widehat{\beta}_1^\phi, \ldots, \widehat{\beta}_K^\phi\right)$ verifying

$$\widehat{\boldsymbol{\beta}}_\phi = \arg\min_{\boldsymbol{\beta}\in\Theta} D_\phi\left(\widehat{\mathbf{p}}, \mathbf{p}\left(\boldsymbol{\beta}\right)\right),$$

where

$$\begin{aligned} D_\phi\left(\widehat{\mathbf{p}}, \mathbf{p}\left(\boldsymbol{\beta}\right)\right) = & \sum_{i=1}^{I} \frac{n_i}{N} \left\{ \pi\left(\mathbf{x}_i^T\boldsymbol{\beta}\right) \phi\left(\frac{n_{i1}}{\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}\right) n_i}\right) \right. \\ & \left. + \left(1-\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)\right) \phi\left(\frac{n_{i2}}{\left(1-\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)\right) n_i}\right) \right\}. \end{aligned}$$

- Newton-Raphson

- Minimum *phi*-divergence estimator in the Generalized Linear Model are the values $\widehat{\boldsymbol{\beta}}_{\phi}=\left(\widehat{\beta}_0^{\phi},\widehat{\beta}_1^{\phi},...,\widehat{\beta}_K^{\phi}\right)$ verifying

$$\widehat{\boldsymbol{\beta}}_{\phi}=\arg\min_{\boldsymbol{\beta}\in\Theta} D_{\phi}\left(\widehat{\mathbf{p}},\mathbf{p}\left(\boldsymbol{\beta}\right)\right),$$

where

$$\begin{aligned} D_{\phi}\left(\widehat{\mathbf{p}},\mathbf{p}\left(\boldsymbol{\beta}\right)\right)= & \sum_{i=1}^{I}\frac{n_i}{N}\left\{\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)\phi\left(\frac{n_{i1}}{\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)n_i}\right)\right. \\ & \left.+\left(1-\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)\right)\phi\left(\frac{n_{i2}}{\left(1-\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)\right)n_i}\right)\right\}. \end{aligned}$$

- Newton-Raphson
  - $\boldsymbol{\beta}^{(t+1)}=\boldsymbol{\beta}^{(t)}-G\left(\boldsymbol{\beta}^{(t)}\right)^{-1}f\left(\boldsymbol{\beta}^{(t)}\right)$

$$G\left(\boldsymbol{\beta}^{(t)}\right)=\left(\frac{\partial^2 D_{\phi}\left(\widehat{\mathbf{p}},\mathbf{p}\left(\boldsymbol{\beta}\right)\right)}{\partial\beta_a\partial\beta_b}\right) \text{ and } f\left(\boldsymbol{\beta}^{(t)}\right)=\left(\frac{\partial D_{\phi}\left(\widehat{\mathbf{p}},\mathbf{p}\left(\boldsymbol{\beta}\right)\right)}{\partial\beta_a}\right)$$

Asymptotic Properties of minimum Phi-divergence estimator

- We have

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_{\phi}-\boldsymbol{\beta}_{0}\right)\underset{N\to\infty}{\overset{L}{\to}}\mathcal{N}\left(\mathbf{0},\left(\mathbf{X}^{T}\mathbf{W}^{*}\mathbf{X}\right)^{-1}\right)$$

Asymptotic Properties of minimum Phi-divergence estimator

- We have

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_{\phi}-\boldsymbol{\beta}_0\right) \underset{N\to\infty}{\overset{L}{\longrightarrow}} \mathcal{N}\left(\mathbf{0},\left(\mathbf{X}^T\mathbf{W}^*\mathbf{X}\right)^{-1}\right)$$

  - $\mathbf{W}^* = \lim_N \mathbf{W}_N$

Asymptotic Properties of minimum Phi-divergence estimator

- We have

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_{\phi}-\boldsymbol{\beta}_0\right) \underset{N\to\infty}{\overset{L}{\to}} \mathcal{N}\left(\mathbf{0}, \left(\mathbf{X}^T\mathbf{W}^*\mathbf{X}\right)^{-1}\right)$$

  - $\mathbf{W}^* = \lim_N \mathbf{W}_N$
  - $\mathbf{W}_N = Diag\left(\left(\frac{n_i}{N\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)\left(1-\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)\right)}\left(\frac{\partial\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)}{\partial\eta_i}\right)^2\right)_{i=1,\ldots,I}\right)$

Asymptotic Properties of minimum Phi-divergence estimator

- We have

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_{\phi}-\boldsymbol{\beta}_0\right)\underset{N\to\infty}{\overset{L}{\to}}\mathcal{N}\left(\mathbf{0},\left(\mathbf{X}^T\mathbf{W}^*\mathbf{X}\right)^{-1}\right)$$

  - $\mathbf{W}^*=\lim_N \mathbf{W}_N$
  - $\mathbf{W}_N=Diag\left(\left(\frac{n_i}{N\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)\left(1-\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)\right)}\left(\frac{\partial\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)}{\partial\eta_i}\right)^2\right)_{i=1,\ldots,I}\right)$
  - $\mathbf{X}=\begin{pmatrix} x_{10} & \ldots\ldots\ldots\ldots & x_{1k}\\ . & \ldots\ldots\ldots\ldots & .\\ x_{I0} & \ldots\ldots\ldots\ldots & x_{Ik}\end{pmatrix}$

Asymptotic Properties of minimum Phi-divergence estimator

- We have

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_{\phi}-\boldsymbol{\beta}_0\right) \underset{N\to\infty}{\overset{L}{\to}} \mathcal{N}\left(\mathbf{0},\left(\mathbf{X}^T\mathbf{W}^*\mathbf{X}\right)^{-1}\right)$$

  - $\mathbf{W}^* = \lim_N \mathbf{W}_N$
  - $\mathbf{W}_N = Diag\left(\left(\frac{n_i}{N\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)\left(1-\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)\right)}\left(\frac{\partial\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)}{\partial\eta_i}\right)^2\right)_{i=1,\ldots,I}\right)$
  - $\mathbf{X} = \begin{pmatrix} x_{10} & \ldots\ldots\ldots\ldots & x_{1k} \\ . & \ldots\ldots\ldots\ldots & . \\ x_{I0} & \ldots\ldots\ldots\ldots & x_{Ik} \end{pmatrix}$

- BAN (Best asymptotically normal)

Asymptotic Properties of minimum Phi-divergence estimator

- We have

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_{\phi}-\boldsymbol{\beta}_0\right) \underset{N\to\infty}{\overset{L}{\longrightarrow}} \mathcal{N}\left(\mathbf{0},\left(\mathbf{X}^T\mathbf{W}^*\mathbf{X}\right)^{-1}\right)$$

  - $\mathbf{W}^* = \lim_N \mathbf{W}_N$
  - $\mathbf{W}_N = Diag\left(\left(\frac{n_i}{N\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)\left(1-\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)\right)}\left(\frac{\partial\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)}{\partial\eta_i}\right)^2\right)_{i=1,\ldots,I}\right)$
  - $\mathbf{X} = \begin{pmatrix} x_{10} & \cdots\cdots\cdots & x_{1k} \\ . & \cdots\cdots\cdots & . \\ x_{I0} & \cdots\cdots\cdots & x_{Ik} \end{pmatrix}$
- BAN (Best asymptotically normal)
  - They are consistent, i.e., the estimator converges to the true value of the parameter as $n \to \infty$.

Asymptotic Properties of minimum Phi-divergence estimator

- We have

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_{\phi}-\boldsymbol{\beta}_0\right) \underset{N\to\infty}{\overset{L}{\to}} \mathcal{N}\left(\mathbf{0},\left(\mathbf{X}^T\mathbf{W}^*\mathbf{X}\right)^{-1}\right)$$

  - $\mathbf{W}^* = \lim_N \mathbf{W}_N$
  - $\mathbf{W}_N = Diag\left(\left(\frac{n_i}{N\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)\left(1-\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)\right)}\left(\frac{\partial\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)}{\partial\eta_i}\right)^2\right)_{i=1,\ldots,I}\right)$
  - $\mathbf{X} = \begin{pmatrix} x_{10} & \ldots\ldots\ldots\ldots & x_{1k} \\ . & \ldots\ldots\ldots\ldots & . \\ x_{I0} & \ldots\ldots\ldots\ldots & x_{Ik} \end{pmatrix}$

- BAN (Best asymptotically normal)
  - They are consistent, i.e., the estimator converges to the true value of the parameter as $n \to \infty$.
  - They are asymptotically normally distributed.

Asymptotic Properties of minimum Phi-divergence estimator

- We have

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_{\phi}-\boldsymbol{\beta}_0\right) \underset{N\to\infty}{\overset{L}{\longrightarrow}} \mathcal{N}\left(\mathbf{0},\left(\mathbf{X}^T\mathbf{W}^*\mathbf{X}\right)^{-1}\right)$$

  - $\mathbf{W}^* = \lim_N \mathbf{W}_N$
  - $\mathbf{W}_N = Diag\left(\left(\frac{n_i}{N\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)\left(1-\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)\right)}\left(\frac{\partial\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}_0\right)}{\partial\eta_i}\right)^2\right)_{i=1,\dots,I}\right)$
  - $\mathbf{X} = \begin{pmatrix} x_{10} & \dots\dots\dots\dots & x_{1k} \\ . & \dots\dots\dots\dots & . \\ x_{I0} & \dots\dots\dots\dots & x_{Ik} \end{pmatrix}$
- BAN (Best asymptotically normal)
  - They are consistent, i.e., the estimator converges to the true value of the parameter as $n \to \infty$.
  - They are asymptotically normally distributed.
  - They are asymptotically efficient, in that no other estimator can have smaller variance, as $n \to \infty$.

Simulation study

- Minimum power-divergence estimator

$$\widehat{\boldsymbol{\beta}}^{(\lambda)} = \arg\min_{\alpha,\beta_1} I^{\lambda}\left(\widehat{\mathbf{p}}, \mathbf{p}\left(\boldsymbol{\beta}\right)\right)$$

where $I^{\lambda}\left(\widehat{\mathbf{p}}, \mathbf{p}\left(\boldsymbol{\beta}\right)\right)$ is given by

$$\frac{1}{\lambda\left(\lambda+1\right)}\left(\sum_{i=1}^{I}\left(\frac{n_{i1}^{\lambda+1}}{\left(\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)n_i\right)^{\lambda}}+\frac{n_{i2}^{\lambda+1}}{\left(1-\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)n_i\right)^{\lambda}}\right)-1\right).$$

- Minimum power-divergence estimator

$$\widehat{\boldsymbol{\beta}}^{(\lambda)} = \arg\min_{\alpha,\beta_1} I^{\lambda}\left(\widehat{\mathbf{p}}, \mathbf{p}\left(\boldsymbol{\beta}\right)\right)$$

where $I^{\lambda}\left(\widehat{\mathbf{p}}, \mathbf{p}\left(\boldsymbol{\beta}\right)\right)$ is given by

$$\frac{1}{\lambda\left(\lambda+1\right)}\left(\sum_{i=1}^{I}\left(\frac{n_{i1}^{\lambda+1}}{\left(\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)n_i\right)^{\lambda}} + \frac{n_{i2}^{\lambda+1}}{\left(1-\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)n_i\right)^{\lambda}}\right) - 1\right).$$

- Models

Simulation study

- Minimum power-divergence estimator

$$\widehat{\boldsymbol{\beta}}^{(\lambda)} = \arg\min_{\alpha,\beta_1} I^{\lambda}\left(\widehat{\mathbf{p}}, \mathbf{p}\left(\boldsymbol{\beta}\right)\right)$$

where $I^{\lambda}\left(\widehat{\mathbf{p}}, \mathbf{p}\left(\boldsymbol{\beta}\right)\right)$ is given by

$$\frac{1}{\lambda\left(\lambda+1\right)}\left(\sum_{i=1}^{I}\left(\frac{n_{i1}^{\lambda+1}}{\left(\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)n_i\right)^{\lambda}}+\frac{n_{i2}^{\lambda+1}}{\left(1-\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)n_i\right)^{\lambda}}\right)-1\right).$$

- Models
  - cloglog $\equiv \log\left(-\log\left(1-\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)\right)\right) = \beta_0 + \beta_1\, x_i$

Simulation study

- Minimum power-divergence estimator

$$\widehat{\beta}^{(\lambda)} = \arg\min_{\alpha,\beta_1} I^{\lambda}\left(\widehat{\mathbf{p}}, \mathbf{p}\left(\beta\right)\right)$$

where $I^{\lambda}\left(\widehat{\mathbf{p}}, \mathbf{p}\left(\beta\right)\right)$ is given by

$$\frac{1}{\lambda\left(\lambda+1\right)}\left(\sum_{i=1}^{I}\left(\frac{n_{i1}^{\lambda+1}}{\left(\pi\left(\mathbf{x}_i^T\beta\right)n_i\right)^{\lambda}}+\frac{n_{i2}^{\lambda+1}}{\left(1-\pi\left(\mathbf{x}_i^T\beta\right)n_i\right)^{\lambda}}\right)-1\right).$$

- Models
  - cloglog $\equiv \log\left(-\log\left(1-\pi\left(\mathbf{x}_i^T\beta\right)\right)\right) = \beta_0+\beta_1\ x_i$
  - Probit $\equiv \Phi_{N(0,1)}^{-1}\log\left(1-\pi\left(\mathbf{x}_i^T\beta\right)\right) = \beta_0+\beta_1\ x_i$

- Minimum power-divergence estimator

$$\widehat{\boldsymbol{\beta}}^{(\lambda)} = \arg\min_{\alpha,\beta_1} I^{\lambda}\left(\widehat{\mathbf{p}}, \mathbf{p}\left(\boldsymbol{\beta}\right)\right)$$

where $I^{\lambda}\left(\widehat{\mathbf{p}}, \mathbf{p}\left(\boldsymbol{\beta}\right)\right)$ is given by

$$\frac{1}{\lambda\left(\lambda+1\right)}\left(\sum_{i=1}^{I}\left(\frac{n_{i1}^{\lambda+1}}{\left(\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)n_i\right)^{\lambda}}+\frac{n_{i2}^{\lambda+1}}{\left(1-\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)n_i\right)^{\lambda}}\right)-1\right).$$

- Models
  - cloglog $\equiv \log\left(-\log\left(1-\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)\right)\right) = \beta_0+\beta_1\ x_i$
  - Probit $\equiv \Phi_{N(0,1)}^{-1}\log\left(1-\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)\right) = \beta_0+\beta_1\ x_i$
  - logit $\equiv \log\dfrac{\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)}{1-\pi\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)} = \beta_0+\beta_1\ x_i$

Simulation study

- Description of Simulation study

Simulation study

- Description of Simulation study
  - $\boldsymbol{\beta} = \left(\beta_0, \beta_1\right)^T$, $\beta_0 = -1$ and $\beta_1 = 2$, and $x_i = -1(0.2)1, i = 1, ..., 11$

Simulation study

- Description of Simulation study
  - $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, $\beta_0 = -1$ and $\beta_1 = 2$, and $x_i = -1(0.2)1, i = 1, ..., 11$
  - $M = 10000$

Simulation study

- Description of Simulation study
  - $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, $\beta_0 = -1$ and $\beta_1 = 2$, and $x_i = -1(0.2)1, i = 1, ..., 11$
  - $M = 10000$
  - $n = (n_1, ..., n_I)^T \in \mathcal{N} = \{n^1, n^2, n^3, n^4, n^5, n^6, n^7, n^8\}$,

$$
\begin{aligned}
n^1 &= (15, 15, 15, 15, 30, 30, 30, 30, 40, 40, 40)\,, \\
n^2 &= (5, 5, 5, 5, 15, 15, 15, 15, 40, 40, 40)\,, \\
n^3 &= (10, 10, 10, 10, 20, 20, 20, 20, 15, 15, 15)\,, \\
n^4 &= (5, 5, 5, 5, 30, 30, 30, 30, 15, 15, 15)\,, \\
n^5 &= (10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10)\,, \\
n^6 &= (20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20)\,, \\
n^7 &= (40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40)\,, \\
n^8 &= (60, 60, 60, 60, 60, 60, 60, 60, 60, 60, 60)\,.
\end{aligned}
$$

Simulation study

- Criteria to choose the best estimators

λ	$-1/2$	0	$2/3$	1	2	3
n^1	.0569	.0356	.0333	.0352	.0458	.0585
n^2	.1171	.0693	.0649	.0703	.0962	.1249
n^3	.1400	.0669	.0594	.0637	.0844	.1065
n^4	.1201	.0663	.0630	.0695	.1009	.1357
n^5	.2798	.1018	.0828	.0875	.1119	.1373
n^6	.0843	.0452	.0403	.0425	.0542	.0678
n^7	.0294	.0215	.0203	.0210	.0256	.0316
n^8	.0168	.0141	.0136	.0139	.0164	.0199
	.1086	.0601	**.0539**	.0577	.0765	.0975

Table 2: Mean Values $\sum_{j=0}^{1} mse\left(\widehat{\beta}_j^{\lambda}\right)/2$ *for the c* $\log\log$ mod *el*

Simulation study

- Criteria to choose the best estimators
  - $mse\left(\widehat{\beta}_j^{\lambda}\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\beta}_{j(i)}^{\lambda} - \beta_j\right)^2, \quad j = 0, 1$ (Mean squared error)

λ	$-1/2$	0	$2/3$	1	2	3
n^1	.0569	.0356	.0333	.0352	.0458	.0585
n^2	.1171	.0693	.0649	.0703	.0962	.1249
n^3	.1400	.0669	.0594	.0637	.0844	.1065
n^4	.1201	.0663	.0630	.0695	.1009	.1357
n^5	.2798	.1018	.0828	.0875	.1119	.1373
n^6	.0843	.0452	.0403	.0425	.0542	.0678
n^7	.0294	.0215	.0203	.0210	.0256	.0316
n^8	.0168	.0141	.0136	.0139	.0164	.0199
	.1086	.0601	**.0539**	.0577	.0765	.0975

Table 2: Mean Values $\sum_{j=0}^{1} mse\left(\widehat{\beta}_j^{\lambda}\right)/2$ *for the* $c \log \log$ $\mathrm{mod}\, el$

- Criteria to choose the best estimators
  - $mse\left(\widehat{\beta}_j^{\lambda}\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\beta}_{j(i)}^{\lambda} - \beta_j\right)^2, \quad j = 0, 1$ (Mean squared error)
  - ($\widehat{\beta}_{j(i)}^{\lambda} \equiv \lambda$-minimum power-divergence estimator of β_j for the *ith* simulation)

λ	$-1/2$	0	2/3	1	2	3
n^1	.0569	.0356	.0333	.0352	.0458	.0585
n^2	.1171	.0693	.0649	.0703	.0962	.1249
n^3	.1400	.0669	.0594	.0637	.0844	.1065
n^4	.1201	.0663	.0630	.0695	.1009	.1357
n^5	.2798	.1018	.0828	.0875	.1119	.1373
n^6	.0843	.0452	.0403	.0425	.0542	.0678
n^7	.0294	.0215	.0203	.0210	.0256	.0316
n^8	.0168	.0141	.0136	.0139	.0164	.0199
	.1086	.0601	**.0539**	.0577	.0765	.0975

Table 2: Mean Values $\sum_{j=0}^{1} mse\left(\widehat{\beta}_j^{\lambda}\right)/2$ *for the c* log log *model*

Fitting the Generalized Linear Model

- $H_0: \ \mathbf{p} = \ \mathbf{p}(\beta) \in \{\ \mathbf{p}(\beta) : \beta \in \Theta$ with a given link function $g\ \}$

Fitting the Generalized Linear Model

- $H_0: \mathbf{p} = \mathbf{p}(\boldsymbol{\beta}) \in \{ \mathbf{p}(\boldsymbol{\beta}) : \boldsymbol{\beta} \in \Theta$ with a given link function g $\}$
  - We denote by $\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_k$ the maximum Likelihood estimators of $\beta_0, \beta_1, ..., \beta_k$.

Fitting the Generalized Linear Model

- $H_0 : \ \mathbf{p} = \ \mathbf{p}(\boldsymbol{\beta}) \in \{ \ \mathbf{p}(\boldsymbol{\beta}) : \boldsymbol{\beta} \in \Theta$ with a given link function g $\}$
  - We denote by $\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_k$ the maximum Likelihood estimators of $\beta_0, \beta_1, ..., \beta_k$.
  - Pearson chi-square statistic

$$X^2 = \sum_{i=1}^{I} \frac{\left(n_{i1} - n_i \widehat{\pi}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right)^2}{n_i \widehat{\pi}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\left(1 - \widehat{\pi}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right)}$$

Fitting the Generalized Linear Model

- H_0 : $\mathbf{p} = \mathbf{p}(\boldsymbol{\beta}) \in \{\mathbf{p}(\boldsymbol{\beta}) : \boldsymbol{\beta} \in \Theta$ with a given link function g $\}$
  - We denote by $\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_k$ the maximum Likelihood estimators of $\beta_0, \beta_1, ..., \beta_k$.
  - Pearson chi-square statistic

$$X^2 = \sum_{i=1}^{I} \frac{\left(n_{i1} - n_i \widehat{\pi}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right)^2}{n_i \widehat{\pi}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\left(1 - \widehat{\pi}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right)}$$

  - Likelihood ratio statistic

$$D^2 = \sum_{i=1}^{I} 2\left\{ n_{i1} \log \frac{n_{i1}}{n_i \widehat{\pi}\left(\mathbf{x}_i \boldsymbol{\beta}\right)} + \left(n_i - n_{i1}\right) \log \frac{n_i - n_{i1}}{n_i\left(1 - \widehat{\pi}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right)} \right\}$$

- Phi-divergence statistics

$$
\begin{aligned}
T_{\phi_1,\phi_2} &= \tfrac{2N}{\phi_1''(1)} D_{\phi_1}\left(\widehat{\mathbf{p}}, \mathbf{p}\left(\widehat{\boldsymbol{\beta}}^{\phi_2}\right)\right) \\
&= \begin{aligned}[t] &\sum_{i=1}^{I} \frac{n_i}{N} \left(\widehat{\pi}^{\phi_2}\left(\mathbf{x}_i \boldsymbol{\beta}\right) \phi_1\left(\frac{n_{i1}}{\widehat{\pi}^{\phi_2}\left(\mathbf{x}_i \boldsymbol{\beta}\right) n_i}\right)\right. \\
&\left. + \left(1 - \widehat{\pi}^{\phi_2}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right) \phi_1\left(\frac{n_{i2}}{\left(1 - \widehat{\pi}^{\phi_2}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right) n_i}\right)\right)\end{aligned}
\end{aligned}
$$

$$
\widehat{\mathbf{p}} = \left(\frac{n_{11}}{N}, \frac{n_{12}}{N}, \frac{n_{21}}{N}, \frac{n_{22}}{N}, \ldots, \frac{n_{I1}}{N}, \frac{n_{I2}}{N}\right)
$$

$$
\mathbf{p}\left(\boldsymbol{\beta}\right) \equiv \left(\pi\left(\mathbf{x}_i \boldsymbol{\beta}\right) \frac{n_i}{N}, \left(1 - \pi\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right) \frac{n_i}{N}\right)_{i=1,\ldots,I}
$$

Fitting the Generalized Linear Model

- Phi-divergence statistics

$$
\begin{aligned}
T_{\phi_1,\phi_2} &= \frac{2N}{\phi_1''(1)} D_{\phi_1}\left(\widehat{\mathbf{p}}, \mathbf{p}\left(\widehat{\boldsymbol{\beta}}^{\phi_2}\right)\right) \\
&= \sum_{i=1}^{I} \frac{n_i}{N}\left(\widehat{\pi}^{\phi_2}\left(\mathbf{x}_i \boldsymbol{\beta}\right) \phi_1\left(\frac{n_{i1}}{\widehat{\pi}^{\phi_2}\left(\mathbf{x}_i \boldsymbol{\beta}\right) n_i}\right)\right. \\
&\quad \left. + \left(1-\widehat{\pi}^{\phi_2}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right) \phi_1\left(\frac{n_{i2}}{\left(1-\widehat{\pi}^{\phi_2}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right) n_i}\right)\right)
\end{aligned}
$$

$$
\widehat{\mathbf{p}} = \left(\frac{n_{11}}{N}, \frac{n_{12}}{N}, \frac{n_{21}}{N}, \frac{n_{22}}{N}, \ldots, \frac{n_{I1}}{N}, \frac{n_{I2}}{N}\right)
$$

$$
\mathbf{p}\left(\boldsymbol{\beta}\right) \equiv \left(\pi\left(\mathbf{x}_i \boldsymbol{\beta}\right) \frac{n_i}{N}, \left(1-\pi\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right) \frac{n_i}{N}\right)_{i=1,\ldots,I}
$$

  - $\phi_2(x) = x \log x - x + 1,\ \phi_1(x) = \frac{1}{2}(x-1)^2 \Rightarrow T_{\phi_1,\phi_2} \equiv X^2$

Fitting the Generalized Linear Model

- Phi-divergence statistics

$$
\begin{aligned}
T_{\phi_1,\phi_2} &= \frac{2N}{\phi_1''(1)} D_{\phi_1}\left(\widehat{\mathbf{p}}, \mathbf{p}\left(\widehat{\boldsymbol{\beta}}^{\phi_2}\right)\right) \\
&= \sum_{i=1}^{I} \frac{n_i}{N} \left(\widehat{\pi}^{\phi_2}\left(\mathbf{x}_i \boldsymbol{\beta}\right) \phi_1 \left(\frac{n_{i1}}{\widehat{\pi}^{\phi_2}\left(\mathbf{x}_i \boldsymbol{\beta}\right) n_i} \right) \right. \\
&\quad \left. + \left(1 - \widehat{\pi}^{\phi_2}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right) \phi_1 \left(\frac{n_{i2}}{\left(1 - \widehat{\pi}^{\phi_2}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right) n_i} \right) \right)
\end{aligned}
$$

$$
\widehat{\mathbf{p}} = \left(\frac{n_{11}}{N}, \frac{n_{12}}{N}, \frac{n_{21}}{N}, \frac{n_{22}}{N}, \ldots, \frac{n_{I1}}{N}, \frac{n_{I2}}{N} \right)
$$

$$
\mathbf{p}\left(\boldsymbol{\beta}\right) \equiv \left(\pi\left(\mathbf{x}_i \boldsymbol{\beta}\right) \frac{n_i}{N}, \left(1 - \pi\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right) \frac{n_i}{N} \right)_{i=1,\ldots,I}
$$

  - $\phi_2(x) = x \log x - x + 1,\ \phi_1(x) = \frac{1}{2}(x-1)^2 \Rightarrow T_{\phi_1,\phi_2} \equiv X^2$
  - $\phi_2(x) = x \log x - x + 1,\ \phi_1(x) = x \log x - x + 1 \Rightarrow T_{\phi_1,\phi_2} \equiv D^2$

Goodness of fit

Theorem

Suppose that the data Y_i, $i = 1, \ldots, I$ are binomially distributed with parameters n_i and $\pi(\mathbf{x}_i\boldsymbol{\beta})$. Choose functions ϕ_1 and $\phi_2 \in \Phi$. Then, under the null hypothesis

$$H_0 : \ \mathbf{p} = \ \mathbf{p}(\boldsymbol{\beta}) \in \{ \ \mathbf{p}(\boldsymbol{\beta}) : \boldsymbol{\beta} \in \Theta \ \textit{with a given link function } g \ \},$$

the phi-divergence statistic

$$\begin{aligned} T_{\phi_1,\phi_2} &= \sum_{i=1}^{I} \frac{n_i}{N} \widehat{\pi}^{\phi_2}(\mathbf{x}_i\boldsymbol{\beta})\, \phi_1\left(\frac{n_{i1}}{\widehat{\pi}^{\phi_2}(\mathbf{x}_i\boldsymbol{\beta})\, n_i} \right) \\ &\quad + \left(1 - \widehat{\pi}^{\phi_2}(\mathbf{x}_i\boldsymbol{\beta})\right) \phi_1\left(\frac{n_{i2}}{\left(1 - \widehat{\pi}^{\phi_2}(\mathbf{x}_i\boldsymbol{\beta})\right) n_i} \right) \end{aligned}$$

has an asymptotic chi-square distribution with $I - (k+1)$ degrees of freedom.

- The statistics

  provide a single number which summarizes the agreement of observed and fitted values. The advantage (as well as the disadvantage) of these statistics is that a single number is used to summarize considerable information

Diagnostic for the GLM based on phi-divergence measures

- The statistics
  - $X^2 = \sum_{i=1}^{I} \frac{(n_{i1} - n_i \widehat{\pi}(\mathbf{x}_i \beta))^2}{n_i \widehat{\pi}(\mathbf{x}_i \beta)(1 - \widehat{\pi}(\mathbf{x}_i \beta))}$

provide a single number which summarizes the agreement of observed and fitted values. The advantage (as well as the disadvantage) of these statistics is that a single number is used to summarize considerable information

Diagnostic for the GLM based on phi-divergence measures

- The statistics
  - $X^2 = \sum_{i=1}^{I} \frac{(n_{i1} - n_i \widehat{\pi}(\mathbf{x}_i \beta))^2}{n_i \widehat{\pi}(\mathbf{x}_i \beta)(1 - \widehat{\pi}(\mathbf{x}_i \beta))}$
  - $D^2 = \sum_{i=1}^{I} 2 \left\{ n_{i1} \log \frac{n_{i1}}{n_i \widehat{\pi}(\mathbf{x}_i \beta)} + (n_i - n_{i1}) \log \frac{n_i - n_{i1}}{n_i (1 - \widehat{\pi}(\mathbf{x}_i \beta))} \right\}$

  provide a single number which summarizes the agreement of observed and fitted values. The advantage (as well as the disadvantage) of these statistics is that a single number is used to summarize considerable information

Diagnostic for the GLM based on phi-divergence measures

- The statistics
  - $X^2 = \sum_{i=1}^{I} \frac{(n_{i1} - n_i \widehat{\pi}(\mathbf{x}_i \boldsymbol{\beta}))^2}{n_i \widehat{\pi}(\mathbf{x}_i \boldsymbol{\beta})(1 - \widehat{\pi}(\mathbf{x}_i \boldsymbol{\beta}))}$
  - $D^2 = \sum_{i=1}^{I} 2 \left\{ n_{i1} \log \frac{n_{i1}}{n_i \widehat{\pi}(\mathbf{x}_i \boldsymbol{\beta})} + (n_i - n_{i1}) \log \frac{n_i - n_{i1}}{n_i (1 - \widehat{\pi}(\mathbf{x}_i \boldsymbol{\beta}))} \right\}$
  - $T_{\phi_1, \phi_2} = \frac{2N}{\phi_1''(1)} D_{\phi_1} \left(\widehat{\mathbf{p}}, \mathbf{p} \left(\widehat{\boldsymbol{\beta}}^{\phi_2} \right) \right)$

  provide a single number which summarizes the agreement of observed and fitted values. The advantage (as well as the disadvantage) of these statistics is that a single number is used to summarize considerable information

- The statistics
  - $X^2 = \sum_{i=1}^{I} \frac{(n_{i1} - n_i \widehat{\pi}(\mathbf{x}_i \boldsymbol{\beta}))^2}{n_i \widehat{\pi}(\mathbf{x}_i \boldsymbol{\beta})(1 - \widehat{\pi}(\mathbf{x}_i \boldsymbol{\beta}))}$
  - $D^2 = \sum_{i=1}^{I} 2 \left\{ n_{i1} \log \frac{n_{i1}}{n_i \widehat{\pi}(\mathbf{x}_i \boldsymbol{\beta})} + (n_i - n_{i1}) \log \frac{n_i - n_{i1}}{n_i (1 - \widehat{\pi}(\mathbf{x}_i \boldsymbol{\beta}))} \right\}$
  - $T_{\phi_1, \phi_2} = \frac{2N}{\phi_1''(1)} D_{\phi_1} \left(\widehat{\mathbf{p}}, \mathbf{p} \left(\widehat{\boldsymbol{\beta}}^{\phi_2} \right) \right)$

  provide a single number which summarizes the agreement of observed and fitted values. The advantage (as well as the disadvantage) of these statistics is that a single number is used to summarize considerable information
- Additional diagnostic analysis are necessary to describe the nature of one lack of fit.

Diagnostic for the GLM based on phi-divergence measures

- The statistics
  - $X^2 = \sum_{i=1}^{I} \frac{(n_{i1} - n_i \widehat{\pi}(\mathbf{x}_i \boldsymbol{\beta}))^2}{n_i \widehat{\pi}(\mathbf{x}_i \boldsymbol{\beta})(1 - \widehat{\pi}(\mathbf{x}_i \boldsymbol{\beta}))}$
  - $D^2 = \sum_{i=1}^{I} 2 \left\{ n_{i1} \log \frac{n_{i1}}{n_i \widehat{\pi}(\mathbf{x}_i \boldsymbol{\beta})} + (n_i - n_{i1}) \log \frac{n_i - n_{i1}}{n_i (1 - \widehat{\pi}(\mathbf{x}_i \boldsymbol{\beta}))} \right\}$
  - $T_{\phi_1, \phi_2} = \frac{2N}{\phi_1''(1)} D_{\phi_1} \left(\widehat{\mathbf{p}}, \mathbf{p} \left(\widehat{\boldsymbol{\beta}}^{\phi_2} \right) \right)$

  provide a single number which summarizes the agreement of observed and fitted values. The advantage (as well as the disadvantage) of these statistics is that a single number is used to summarize considerable information
- Additional diagnostic analysis are necessary to describe the nature of one lack of fit.
  - Detection of "outliers"

Diagnostic for the GLM based on phi-divergence measures

- The statistics
  - $X^2 = \sum_{i=1}^{I} \frac{\left(n_{i1} - n_i \widehat{\pi}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right)^2}{n_i \widehat{\pi}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\left(1 - \widehat{\pi}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right)}$
  - $D^2 = \sum_{i=1}^{I} 2 \left\{ n_{i1} \log \frac{n_{i1}}{n_i \widehat{\pi}\left(\mathbf{x}_i \boldsymbol{\beta}\right)} + (n_i - n_{i1}) \log \frac{n_i - n_{i1}}{n_i \left(1 - \widehat{\pi}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right)} \right\}$
  - $T_{\phi_1, \phi_2} = \frac{2N}{\phi_1''(1)} D_{\phi_1} \left(\widehat{\mathbf{p}}, \mathbf{p}\left(\widehat{\boldsymbol{\beta}}^{\phi_2} \right) \right)$

  provide a single number which summarizes the agreement of observed and fitted values. The advantage (as well as the disadvantage) of these statistics is that a single number is used to summarize considerable information
- Additional diagnostic analysis are necessary to describe the nature of one lack of fit.
  - Detection of "outliers"
  - Identification of " influents observations".

Residuals

- Residuals based on the Pearson's statistic X^2

$$e_i = \frac{n_{i1} - n_i \widehat{\pi}\left(\mathbf{x}_i \boldsymbol{\beta}\right)}{\sqrt{n_i \widehat{\pi}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\left(1 - \widehat{\pi}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right)}}$$

Residual based on the ϕ-divergence statistic T_{ϕ_1,ϕ_2}

$$\begin{aligned} c_i^{\phi_1,\phi_2} = \quad & I \times \sqrt{\frac{2n_i}{\phi_1''(1)}} \left\{ \widehat{\pi}^{\phi_2}\left(\mathbf{x}_i \boldsymbol{\beta}\right) \phi_1\left(\tfrac{n_{i1}}{\widehat{\pi}^{\phi_2}\left(\mathbf{x}_i \boldsymbol{\beta}\right) n_i} \right) + \left(1 - \widehat{\pi}^{\phi_2}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right) \right. \\ & \left. \phi_1\left(\tfrac{n_{i2}}{\left(1 - \widehat{\pi}^{\phi_2}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right) n_i} \right) \right\}^{1/2} \end{aligned}$$

Residuals

- Residuals based on the Pearson's statistic X^2

$$e_i = \frac{n_{i1} - n_i \widehat{\pi}\left(\mathbf{x}_i \boldsymbol{\beta}\right)}{\sqrt{n_i \widehat{\pi}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\left(1 - \widehat{\pi}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right)}}$$

- Residuals based on the Likelihood ratio statistic D^2

$$d_i = l \times \sqrt{2\left\{n_{i1} \log \frac{n_{i1}}{n_i \widehat{\pi}\left(\mathbf{x}_i \boldsymbol{\beta}\right)} + \left(n_i - n_{i1}\right) \log \frac{n_i - n_{i1}}{n_i\left(1 - \widehat{\pi}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right)}\right\}},$$

$\left(l = sig\left(n_{i1} - n_i \widehat{\pi}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right)\right)$

Residual based on the ϕ-divergence statistic T_{ϕ_1,ϕ_2}

$$\begin{aligned} c_i^{\phi_1,\phi_2} = \; & l \times \sqrt{\frac{2n_i}{\phi_1''(1)}} \left\{\widehat{\pi}^{\phi_2}\left(\mathbf{x}_i \boldsymbol{\beta}\right) \phi_1\left(\frac{n_{i1}}{\widehat{\pi}^{\phi_2}\left(\mathbf{x}_i \boldsymbol{\beta}\right) n_i}\right) + \left(1 - \widehat{\pi}^{\phi_2}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right)\right. \\ & \left.\phi_1\left(\frac{n_{i2}}{\left(1 - \widehat{\pi}^{\phi_2}\left(\mathbf{x}_i \boldsymbol{\beta}\right)\right) n_i}\right)\right\}^{1/2} \end{aligned}$$

Residuals

Theorem

Suppose that the data Y_i, $i = 1, ..., I$ are binomially distributed with parameters n_i and $\pi(\mathbf{x}_i\beta)$. Choose functions ϕ_1 and $\phi_2 \in \Phi$. Then

$$c_j^{\phi_1,\phi_2} \underset{N\to\infty}{\overset{L}{\to}} N\left(0, 1 - h_{jj}\right)$$

where h_{jj} are the diagonal elements of the matrix

$$\mathbf{H} = (\mathbf{W}^*)^{1/2}\mathbf{X}\left(\mathbf{X}^T\mathbf{W}^*\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{W}^*)^{1/2} \text{ (“hat matrix”).}$$

Definition

The standardized phi-residuals are given by

$$\left(c_j^{\phi_1,\phi_2}\right)^* = c_j^{\phi_1,\phi_2}/\sqrt{1 - \widehat{h}_{jj}^{\phi_2}}.$$

Residuals

- The estimated values $\widehat{h}_{jj}^{\phi_2}$ of the elements h_{jj} in the hat matrix are given by

$$\widehat{h}_{jj}^{\phi_2} = \frac{n_j}{N\pi\left(\mathbf{x}_j^T\widehat{\boldsymbol{\beta}}_{\phi_2}\right)\left(1-\pi\left(\mathbf{x}_j^T\widehat{\boldsymbol{\beta}}_{\phi_2}\right)\right)}\left(\frac{\partial\pi\left(\mathbf{x}_j^T\widehat{\boldsymbol{\beta}}_{\phi_2}\right)}{\partial\eta_j}\right)^2 \mathbf{x}_j\mathbf{I}_F\left(\widehat{\boldsymbol{\beta}}_{\phi_2}\right)^{-1}\mathbf{x}_j^T.$$

Residuals

- The estimated values $\widehat{h}_{jj}^{\phi_2}$ of the elements h_{jj} in the hat matrix are given by

$$\widehat{h}_{jj}^{\phi_2} = \frac{n_j}{N\pi\left(\mathbf{x}_j^T\widehat{\boldsymbol{\beta}}_{\phi_2}\right)\left(1-\pi\left(\mathbf{x}_j^T\widehat{\boldsymbol{\beta}}_{\phi_2}\right)\right)}\left(\frac{\partial\pi\left(\mathbf{x}_j^T\widehat{\boldsymbol{\beta}}_{\phi_2}\right)}{\partial\eta_j}\right)^2 \mathbf{x}_j\mathbf{I}_F\left(\widehat{\boldsymbol{\beta}}_{\phi_2}\right)^{-1}\mathbf{x}_j^T.$$

- A large value of $\widehat{h}_{jj}^{\phi_2}$ indicates a data point where the variance of the residual deviate more from the binomial variance than for other data point. Such a data point can, therefore, influence the evaluation of the fit of the model in an uncharacteristic way.

Cook's distance

- **Cook's distance** (Influence of the ith-observation)

Cook's distance

- **Cook's distance** (Influence of the ith-observation)
  - Consists of comparing the MϕE of $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$ obtained using the original data and the MϕE obtained after deleting the ith-observation

Cook's distance

- **Cook's distance** (Influence of the ith-observation)
  - Consists of comparing the MϕE of $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$ obtained using the original data and the MϕE obtained after deleting the ith-observation
  - If we consider the minimum ϕ_2-divergence estimator, we can consider ,

$$D_i = \frac{1}{k+1}\left(\widehat{\boldsymbol{\beta}}^{\phi_2} - \widehat{\boldsymbol{\beta}}^{\phi_2}_{(i)}\right) \widehat{Cov\left(\widehat{\boldsymbol{\beta}}^{\phi_2}\right)}^{-1} \left(\widehat{\boldsymbol{\beta}}^{\phi_2} - \widehat{\boldsymbol{\beta}}^{\phi_2}_{(i)}\right)^T$$

  where

- **Cook's distance** (Influence of the ith-observation)
  - Consists of comparing the MϕE of $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$ obtained using the original data and the MϕE obtained after deleting the ith-observation
  - If we consider the minimum ϕ_2-divergence estimator, we can consider ,

$$D_i = \frac{1}{k+1}\left(\widehat{\boldsymbol{\beta}}^{\phi_2} - \widehat{\boldsymbol{\beta}}^{\phi_2}_{(i)}\right) \widehat{Cov\left(\widehat{\boldsymbol{\beta}}^{\phi_2}\right)}^{-1}\left(\widehat{\boldsymbol{\beta}}^{\phi_2} - \widehat{\boldsymbol{\beta}}^{\phi_2}_{(i)}\right)^T$$

where

- $\widehat{\boldsymbol{\beta}}^{\phi_2} = \left(\widehat{\beta}^{\phi_2}_0, \widehat{\beta}^{\phi_2}_1, ..., \widehat{\beta}^{\phi_2}_k\right) \equiv M\phi_2 E$ of $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$

Cook's distance

- **Cook's distance** (Influence of the ith-observation)
  - Consists of comparing the MϕE of $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$ obtained using the original data and the MϕE obtained after deleting the ith-observation
  - If we consider the minimum ϕ_2-divergence estimator, we can consider ,

$$D_i = \frac{1}{k+1} \left(\widehat{\boldsymbol{\beta}}^{\phi_2} - \widehat{\boldsymbol{\beta}}^{\phi_2}_{(i)} \right) \widehat{Cov\left(\widehat{\boldsymbol{\beta}}^{\phi_2} \right)}^{-1} \left(\widehat{\boldsymbol{\beta}}^{\phi_2} - \widehat{\boldsymbol{\beta}}^{\phi_2}_{(i)} \right)^T$$

where

- $\widehat{\boldsymbol{\beta}}^{\phi_2} = \left(\widehat{\beta}^{\phi_2}_0, \widehat{\beta}^{\phi_2}_1, ..., \widehat{\beta}^{\phi_2}_k \right) \equiv M\phi_2E$ of $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$
- $\widehat{\boldsymbol{\beta}}^{\phi_2}_{(i)} = \left(\widehat{\beta}^{\phi_2}_{0(i)}, \widehat{\beta}^{\phi_2}_{1(i)}, ..., \widehat{\beta}^{\phi_2}_{k(i)} \right) \equiv M\phi_2E$ of $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$ without the i-th observation

- **Cook's distance** (Influence of the ith-observation)
  - Consists of comparing the MϕE of $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$ obtained using the original data and the MϕE obtained after deleting the ith-observation
  - If we consider the minimum ϕ_2-divergence estimator, we can consider ,

$$D_i = \frac{1}{k+1}\left(\widehat{\boldsymbol{\beta}}^{\phi_2} - \widehat{\boldsymbol{\beta}}^{\phi_2}_{(i)}\right) \widehat{Cov\left(\widehat{\boldsymbol{\beta}}^{\phi_2}\right)}^{-1} \left(\widehat{\boldsymbol{\beta}}^{\phi_2} - \widehat{\boldsymbol{\beta}}^{\phi_2}_{(i)}\right)^T$$

where

- $\widehat{\boldsymbol{\beta}}^{\phi_2} = \left(\widehat{\beta}^{\phi_2}_0, \widehat{\beta}^{\phi_2}_1, ..., \widehat{\beta}^{\phi_2}_k\right) \equiv M\phi_2 E$ of $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$
- $\widehat{\boldsymbol{\beta}}^{\phi_2}_{(i)} = \left(\widehat{\beta}^{\phi_2}_{0(i)}, \widehat{\beta}^{\phi_2}_{1(i)}, ..., \widehat{\beta}^{\phi_2}_{k(i)}\right) \equiv M\phi_2 E$ of $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$ without the i-th observation
- $\widehat{Cov\left(\widehat{\boldsymbol{\beta}}^{\phi_2}\right)} = \left(\mathbf{X}^T \mathbf{W}^* \mathbf{X}\right)^{-1}$

Dimensionality reduction: Phi-divergence statistics

- Determine if the explanatory variables in the model are "significantly" related to the outcome variable.

- Determine if the explanatory variables in the model are "significantly" related to the outcome variable.
    - **Stage 1**

Dimensionality reduction: Phi-divergence statistics

- Determine if the explanatory variables in the model are "significantly" related to the outcome variable.
  - **Stage 1**
    - We consider the Generalized Linear Model with all available explanatory variables; i.e., with all the parameters

$$\beta_1, ..., \beta_k$$

and we test

$$H_{Null} : \beta_j = 0 \text{ versus } H_{Alt} : \beta_j \neq 0 \;\; (j = 1, ..., k) \tag{1}$$

Dimensionality reduction: Phi-divergence statistics

- Determine if the explanatory variables in the model are "significantly" related to the outcome variable.
  - **Stage 1**
    - We consider the Generalized Linear Model with all available explanatory variables; i.e., with all the parameters

$$\beta_1, ..., \beta_k$$

and we test

$$H_{Null} : \beta_j = 0 \text{ versus } H_{Alt} : \beta_j \neq 0 \;\; (j = 1, ..., k) \qquad (1)$$

    - We delete the explanatory variable associated with the regression parameter β_{k_1} if the $p-$value associated with test (1) for $j = k_1$, p_{k_1}, is the highest.

- **Stage 2**

- **Stage 2**
  - We consider the Generalized Linear Model with the parameters,

$$\beta_1, ..., \beta_{k_1-1}, \beta_{k_1+1}, ..., \beta_k$$

and we consider the tests

$$H_{Null} : \beta_j = 0 \text{ against } H_{Alt} : \beta_j \neq 0 \ \ (j = 1, ..., k_{1-1}, k_{1+1}, ..., k)$$

- **Stage 2**
  - We consider the Generalized Linear Model with the parameters,

$$\beta_1, ..., \beta_{k_1-1}, \beta_{k_1+1}, ..., \beta_k$$

and we consider the tests

$$H_{Null} : \beta_j = 0 \text{ against } H_{Alt} : \beta_j \neq 0 \;\; (j = 1, ..., k_{1-1}, k_{1+1}, ..., k)$$

  - We delete the explanatory variable with the associated highest $p-$value

- **Stage 2**
  - We consider the Generalized Linear Model with the parameters,

$$\beta_1, ..., \beta_{k_1-1}, \beta_{k_1+1}, ..., \beta_k$$

and we consider the tests

$$H_{Null} : \beta_j = 0 \text{ against } H_{Alt} : \beta_j \neq 0 \ \ (j = 1, ..., k_{1-1}, k_{1+1}, ..., k)$$

  - We delete the explanatory variable with the associated highest $p-$value
  - We stop the procedure when the maximum $p-$value associated with the Generalized Linear Model with the remaining explanatory variables is sufficiently small.

- How to test in the stage t, $H_{Null} : \beta_j = 0$ versus $H_{Alt} : \beta_j \neq 0$?

- How to test in the stage t, $H_{Null} : \beta_j = 0$ versus $H_{Alt} : \beta_j \neq 0$?

  - First procedure

$$T_{\phi_2}^{\beta_j,t} = \frac{\widehat{\beta}_j^{\phi_2,t}}{\sqrt{\widehat{Var\left(\widehat{\beta}_j^{\phi_2,t}\right)}}}$$

- How to test in the stage t, $H_{Null} : \beta_j = 0$ versus $H_{Alt} : \beta_j \neq 0$?
  - First procedure

$$T_{\phi_2}^{\beta_j,t} = \frac{\widehat{\beta}_j^{\phi_2,t}}{\sqrt{\widehat{Var\left(\widehat{\beta}_j^{\phi_2,t}\right)}}}$$

    - $\widehat{Var\left(\widehat{\beta}_j^{\phi_2,t}\right)}$ *jth* diagonal element of the asymptotic covariance matrix)

$$\left(\mathbf{X}^T\mathbf{W}^*\mathbf{X}\right)^{-1}$$

- How to test in the stage t, $H_{Null} : \beta_j = 0$ versus $H_{Alt} : \beta_j \neq 0$?
    - First procedure

$$T_{\phi_2}^{\beta_j, t} = \frac{\widehat{\beta}_j^{\phi_2, t}}{\sqrt{\widehat{Var\left(\widehat{\beta}_j^{\phi_2, t}\right)}}}$$

        - $\widehat{Var\left(\widehat{\beta}_j^{\phi_2, t}\right)}$ *jth* diagonal element of the asymptotic covariance matrix)

$$\left(\mathbf{x}^T \mathbf{w}^* \mathbf{x}\right)^{-1}$$

        - $\widehat{\beta}_j^{\phi_2, t} \equiv$ Minimum ϕ_2-divergence estimator of parameter β_j in the stage t

- Second procedure

$$S_{\phi_1,\phi_2}^{\beta_j,t} = \frac{2N}{\phi_1''(1)} D_{\phi_1}\left(\mathbf{p}\left(\widehat{\boldsymbol{\beta}}^{\phi_2,t}\right), \mathbf{p}\left({}^{j}\widehat{\boldsymbol{\beta}}^{\phi_2,t}\right)\right)$$

Theorem

Suppose that the data Y_i, $i = 1, ..., I$ are binomially distributed with parameters n_i and $\pi(\mathbf{x}_i)$. Choose functions ϕ_1 and $\phi_2 \in \Phi$. Then for testing $H_{Null} : \beta_j = 0$ versus $H_{Alt} : \beta_j \neq 0$ the ϕ-divergence statistic

$$S_{\phi_1,\phi_2}^{\beta_j,t} = \frac{2N}{\phi_1''(1)} D_{\phi_1}\left(\mathbf{p}\left(\widehat{\boldsymbol{\beta}}^{\phi_2,t}\right), \mathbf{p}\left({}^{j}\widehat{\boldsymbol{\beta}}^{\phi_2,t}\right)\right)$$

has an asymptotic chi-square distribution with 1 degree of freedom.

- Second procedure

$$S_{\phi_1,\phi_2}^{\beta_j,t} = \frac{2N}{\phi_1''(1)} D_{\phi_1}\left(\mathbf{p}\left(\widehat{\boldsymbol{\beta}}^{\phi_2,t}\right), \mathbf{p}\left({}^{j}\widehat{\boldsymbol{\beta}}^{\phi_2,t}\right)\right)$$

  - $\widehat{\boldsymbol{\beta}}^{\phi_2,t} \equiv$ Minimum ϕ_2-divergence estimator of $\beta_1, ..., \beta_{k+1-t}$

Theorem

Suppose that the data Y_i, $i = 1, ..., I$ are binomially distributed with parameters n_i and $\pi(\mathbf{x}_i)$. Choose functions ϕ_1 and $\phi_2 \in \Phi$. Then for testing $H_{Null} : \beta_j = 0$ versus $H_{Alt} : \beta_j \neq 0$ the ϕ-divergence statistic

$$S_{\phi_1,\phi_2}^{\beta_j,t} = \frac{2N}{\phi_1''(1)} D_{\phi_1}\left(\mathbf{p}\left(\widehat{\boldsymbol{\beta}}^{\phi_2,t}\right), \mathbf{p}\left({}^{j}\widehat{\boldsymbol{\beta}}^{\phi_2,t}\right)\right)$$

has an asymptotic chi-square distribution with 1 degree of freedom.

- Second procedure

$$S_{\phi_1,\phi_2}^{\beta_j,t} = \frac{2N}{\phi_1''(1)} D_{\phi_1}\left(\mathbf{p}\left(\widehat{\boldsymbol{\beta}}^{\phi_2,t}\right), \mathbf{p}\left({}^{j}\widehat{\boldsymbol{\beta}}^{\phi_2,t}\right)\right)$$

  - $\widehat{\boldsymbol{\beta}}^{\phi_2,t} \equiv$ Minimum ϕ_2-divergence estimator of $\beta_1, ..., \beta_{k+1-t}$
  - ${}^{j}\widehat{\boldsymbol{\beta}}^{\phi_2} \equiv$ Minimum ϕ_2-divergence estimator of $\left(\beta_{0,} ..., \beta_{j-1}, 0, \beta_{j+1}, ..., \beta_{k+1-t}\right)$ with all the observations but without the explanatory variable x_j in the stage t.

Theorem

Suppose that the data Y_i, $i = 1, ..., I$ are binomially distributed with parameters n_i and $\pi(\mathbf{x}_i)$. Choose functions ϕ_1 and $\phi_2 \in \Phi$. Then for testing $H_{Null} : \beta_j = 0$ versus $H_{Alt} : \beta_j \neq 0$ the ϕ-divergence statistic

$$S_{\phi_1,\phi_2}^{\beta_j,t} = \frac{2N}{\phi_1''(1)} D_{\phi_1}\left(\mathbf{p}\left(\widehat{\boldsymbol{\beta}}^{\phi_2,t}\right), \mathbf{p}\left({}^{j}\widehat{\boldsymbol{\beta}}^{\phi_2,t}\right)\right)$$

has an asymptotic chi-square distribution with 1 degree of freedom.

Restricted Minimum phi-divergence estimator

- $H_{Null} : \beta_j = 0$ is equivalent to testing $H_0 : \mathbf{K}^T \boldsymbol{\beta} = \mathbf{m}$ with

Restricted Minimum phi-divergence estimator

- $H_{Null} : \beta_j = 0$ is equivalent to testing $H_0 : \mathbf{K}^T \boldsymbol{\beta} = \mathbf{m}$ with
  - $K^T = (0, ..., \underbrace{1}_{j}, ...0)$

Restricted Minimum phi-divergence estimator

- $H_{Null} : \beta_j = 0$ is equivalent to testing $H_0 : \mathbf{K}^T \boldsymbol{\beta} = \mathbf{m}$ with
  - $K^T = (0, ..., \underbrace{1}_{j}, ...0)$
  - $\mathbf{m} = (0, ..., 0)^T$

Restricted Minimum phi-divergence estimator

- $H_{Null} : \beta_j = 0$ is equivalent to testing $H_0 : \mathbf{K}^T \boldsymbol{\beta} = \mathbf{m}$ with
    - $K^T = (0, ..., \underbrace{1}_{j}, ...0)$
    - $\mathbf{m} = (0, ..., 0)^T$
- We are going to give a general procedure for testing

$$H_0 : \mathbf{K}^T \boldsymbol{\beta} = \mathbf{m} \text{ against } H_1 : \mathbf{K}^T \boldsymbol{\beta} \neq \mathbf{m}$$

Restricted Minimum phi-divergence estimator

- $H_{Null} : \beta_j = 0$ is equivalent to testing $H_0 : \mathbf{K}^T \boldsymbol{\beta} = \mathbf{m}$ with
  - $K^T = (0, ..., \underbrace{1}_{j}, ...0)$
  - $\mathbf{m} = (0, ..., 0)^T$
- We are going to give a general procedure for testing

$$H_0 : \mathbf{K}^T \boldsymbol{\beta} = \mathbf{m} \text{ against } H_1 : \mathbf{K}^T \boldsymbol{\beta} \neq \mathbf{m}$$

- We consider the phi-divergence statistics

$$\begin{aligned} T_N^{\phi_1,\phi_2} &= \frac{2N}{\phi_1''(1)} D_{\phi_1}\left(\mathbf{p}\left(\widehat{\boldsymbol{\beta}}_{\phi_2}\right), \mathbf{p}\left(\widehat{\boldsymbol{\beta}}_{\phi_2}^{H_0}\right)\right) \\ &= \frac{2}{\phi_1''(1)} \sum_{i=1}^{I} n_i \sum_{j=1}^{2} \pi_j\left(\mathbf{x}_i^T \widehat{\boldsymbol{\beta}}_{\phi_2}^{H_0}\right) \phi_1\left(\frac{\pi_j\left(\mathbf{x}_i^T \widehat{\boldsymbol{\beta}}_{\phi_2}\right)}{\pi_j\left(\mathbf{x}_i^T \widehat{\boldsymbol{\beta}}_{\phi_2}^{H_0}\right)}\right) \end{aligned}$$

Restricted Minimum phi-divergence estimator

- $H_{Null} : \beta_j = 0$ is equivalent to testing $H_0 : \mathbf{K}^T\boldsymbol{\beta} = \mathbf{m}$ with
  - $K^T = (0, ..., \underbrace{1}_{j}, ...0)$
  - $\mathbf{m} = (0, ..., 0)^T$
- We are going to give a general procedure for testing

$$H_0 : \mathbf{K}^T\boldsymbol{\beta} = \mathbf{m} \text{ against } H_1 : \mathbf{K}^T\boldsymbol{\beta} \neq \mathbf{m}$$

- We consider the phi-divergence statistics

$$\begin{aligned} T_N^{\phi_1,\phi_2} &= \frac{2N}{\phi_1''(1)} D_{\phi_1}\left(\mathbf{p}\left(\widehat{\boldsymbol{\beta}}_{\phi_2}\right), \mathbf{p}\left(\widehat{\boldsymbol{\beta}}_{\phi_2}^{H_0}\right)\right) \\ &= \frac{2}{\phi_1''(1)} \sum_{i=1}^{I} n_i \sum_{j=1}^{2} \pi_j\left(\mathbf{x}_i^T \widehat{\boldsymbol{\beta}}_{\phi_2}^{H_0}\right) \phi_1\left(\frac{\pi_j\left(\mathbf{x}_i^T \widehat{\boldsymbol{\beta}}_{\phi_2}\right)}{\pi_j\left(\mathbf{x}_i^T \widehat{\boldsymbol{\beta}}_{\phi_2}^{H_0}\right)}\right) \end{aligned}$$

- $\widehat{\boldsymbol{\beta}}_{\phi_2}^{H_0}$ ≡Restricted minimum phi-divergence estimator

Definition

The restricted minimum phi-divergence estimator is defined by

$$\widehat{\boldsymbol{\beta}}_{\phi}^{H_0} \equiv \arg\min_{\boldsymbol{\beta}\in\Theta_0} D_{\phi}\left(\widehat{\mathbf{p}}, \mathbf{p}\left(\boldsymbol{\beta}\right)\right),$$

where $\Theta_0 = \left\{\boldsymbol{\beta} \in \Theta : \mathbf{K}^T\boldsymbol{\beta} = \mathbf{m}\right\}$ and

$$\Theta = \left\{\boldsymbol{\beta} = \left(\beta_0, ..., \beta_k\right) : \beta_i \in (-\infty, \infty), \;\; i = 0, ..., k\right\}.$$

- We have,

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_{\phi}^{H_0} - \boldsymbol{\beta}_0\right) \underset{N\to\infty}{\overset{L}{\to}} \mathcal{N}\left(\mathbf{0}, \mathbf{S}\right)$$

where

$$\mathbf{S} = \left(\mathbf{I} - \left(\mathbf{X}^T\mathbf{W}^*\mathbf{X}\right)^{-1}\mathbf{K}\left(\mathbf{K}^T\left(\mathbf{X}^T\mathbf{W}^*\mathbf{X}\right)^{-1}\mathbf{K}\right)\right)\left(\mathbf{X}^T\mathbf{W}^*\mathbf{X}\right)^{-1}$$

Theorem

Consider the Generalized Linear Model with Binary Data

$$\eta_j = g\left(\pi_j\right) = \sum_{l=0}^{K} \beta_l x_{jl},\ j = 1, ..., I,$$

Under the null hypothesis

$$H_0 : \mathbf{K}^T \beta = \mathbf{m},\ \left(\mathbf{K}^T \equiv r \textit{ rows and } k+1 \textit{ columns}\right)$$

the asymptotic distribution of the phi-divergence test statistics

$$T_N^{\phi_1,\phi_2} = \frac{2N}{\phi_1''(1)} D_{\phi_1}\left(\mathbf{p}\left(\widehat{\boldsymbol{\beta}}_{\phi_2}\right), \mathbf{p}\left(\widehat{\boldsymbol{\beta}}_{\phi_2}^{H_0}\right)\right)$$

is chi-squared with r degrees of freedom.

Example

- The data, taken from Cox and Snell (1989, pp. 10-11), consist of the number, $n_{(1)} = \left(n_{11}, ..., n_{19,1}\right)^T$, of ingots not ready for rolling, out of $n = (n_1, ..., n_{19})^T$, tested, for a number of heating time $(\mathbf{x}_1)$ and soaking time $(\mathbf{x}_2)$.

x_{i1}	x_{i2}	n_{i1}	n_i	x_{i1}	x_{i2}	n_{1i}	n_i	x_{i1}	x_{i2}	n_{i1}	n_i	x_{i1}
7	1	0	10	14	1	0	31	27	1	1	56	51
7	1.7	0	17	14	1.7	0	43	27	1.7	4	44	51
7	2.2	0	7	14	2.2	2	33	27	2.2	0	21	51
7	2.8	0	12	14	2.8	0	31	27	2.8	1	22	51
7	4	0	9	14	4	0	19	27	4	1	16	

- We consider the following Logistic regression Model

$$\Pr\left(Y = 1/X_1 = x_{i1}, X_2 = x_{i2}\right) = \pi\left(\mathbf{x}_i \boldsymbol{\beta}\right) = \frac{\exp\left(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2}\right)}{1 + \exp\left(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2}\right)}$$

- **Stage 1**
  We consider the Logistic Regressing Models with the three parameters β_0, β_1 and β_2,

$$\pi\left(\mathbf{x}\right)=\frac{\exp\left(\beta_0+x_1\beta_1+x_2\beta_2\right)}{1+\exp\left(\beta_0+x_1\beta_1+x_2\beta_2\right)}$$

  and we consider the following test statistics

  - $H_{Null}:\beta_1=0$ versus $H_{Alter}:\beta_1\neq 0$

$$S_{\phi_{(\lambda)},\phi_{(0)}}^{\beta_1,1}=\frac{2N}{\phi''_{(\lambda)}\left(1\right)}D_{\phi_{(\lambda)}}\left(\mathbf{p}\left(\widehat{\beta}_0,\widehat{\beta}_1,\widehat{\beta}_2\right),\mathbf{p}\left(\widehat{\beta}_0,\widehat{\beta}_2\right)\right)$$

  - $H_{Null}:\beta_2=0$ versus $H_{Alter}:\beta_2\neq 0$

$$S_{\phi_{(\lambda)},\phi_{(0)}}^{\beta_2,1}=\frac{2N}{\phi''_{(\lambda)}\left(1\right)}D_{\phi_{(\lambda)}}\left(\mathbf{p}\left(\widehat{\beta}_0,\widehat{\beta}_1,\widehat{\beta}_2\right),\mathbf{p}\left(\widehat{\beta}_0,\widehat{\beta}_1\right)\right)$$

Example

- Values of test statistics and *p*-values

λ	-2	-1	-1/2	0	2/3	1	2
$S^{\beta_2,1}_{\phi_{(\lambda)},\phi_{(0)}}$	2.0918	1.1733	0.9233	0.7486	0.5894	0.5310	0.479
p-values	0.1481	0.2787	0.3366	0.3891	0.4426	0.4662	0.523

λ	-2	-1	-1/2	0	2/3
$S^{\beta_1,1}_{\phi_{(\lambda)},\phi_{(0)}}$	11.157	9.5246	9.8289	11.033	14.921
p-values	8.4×10^{-4}	2.1×10^{-3}	1.7×10^{-3}	8.9×10^{-4}	1.1×10^{-4}

- **First stage the "soaking time variable" (associated to the parameter β_2) can be omitted**

Example

- **Stage 2**

We consider the Logistic Regression Model with two parameters

$$\pi\left(\mathbf{x}\boldsymbol{\beta}\right)=\frac{\exp\left(\beta_0+x_1\beta_1\right)}{1+\exp\left(\beta_0+x_1\beta_1\right)}$$

and we consider the following tests

- $H_{Nula}:\beta_1=0$ versus $H_{Alter}:\beta_1\neq 0$

$$S_{\phi_{(\lambda)},\phi_{(0)}}^{\beta_1,2}=\frac{2N}{\phi_{(\lambda)}''(1)}D_{\phi_{(\lambda)}}\left(\mathbf{p}\left(\widehat{\beta}_0,\widehat{\beta}_1\right),\mathbf{p}\left(\widehat{\beta}_0\right)\right)$$

  - Values of test statistics and *p*-values

λ	-2	-1	-0.5	0	2/3
$S_{\phi_{(\lambda)},\phi_{(0)}}^{\beta_1,\beta_2}$	10.81	13.389	14.929	17.902	16.312
	1.0×10^{-2}	2.5×10^{-4}	1.1×10^{-4}	2.3×10^{-5}	5.4×10^{-5}

  - The p-values are too small, we reject the null hypothesis and the stepwise procedure finishes.