Linear Models in Statistics II

Lecture notes

Andrea Kraus

Department of Mathematics and Statistics
Faculty of Science
Masaryk University
Brno, Czech Republic

Spring 2017

Linear model: a reminder

- $Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k} + \varepsilon_i$, $i \in \{1, \ldots, n\}$
  - Y_i: outcome, response, output, dependent variable
    - random variable, we observe a realization y_i
    - (odezva, závisle proměnná, regresand)
  - $x_{i,1}, \ldots, x_{i,k}$: covariates, predictors, explanatory variables, input, independent variables
    - given, known
    - (nezávisle proměnné, regresory)
  - $\beta_0, \ldots, \beta_k$: coefficients
    - unknown
    - (regresní koeficienty)
  - ε_i: random error
    - random variable, unobserved
- $\varepsilon_i \overset{\text{iid}}{\sim} (0, \sigma^2)$, $i \in \{1, \ldots, n\}$
  - $\mathrm{E}\,\varepsilon_i = 0$: no systematic errors
  - $\mathrm{Var}\,\varepsilon_i = \sigma^2$: same precision
- we often assume that $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$, $i \in \{1, \ldots, n\}$

Example: `bloodpress` data

- from `sites.stat.psu.edu/~lsimon/stat501wc/sp05/data/`
- association between the mean arterial blood pressure[mmHg] and age[years], weight[kg], body surface area[m^2], duration of hypertension[years], basal pulse[beats/min], stress
- data:

BP	Age	Weight	BSA	DoH	Pulse	Stress
105	47	85.4	1.75	5.1	63	33
115	49	94.2	2.10	3.8	70	14
...	...	...	...	...	...	...
110	48	90.5	1.88	9.0	71	99
122	56	95.7	2.09	7.0	75	99

- model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

$$\begin{pmatrix} 105 \\ 115 \\ \dots \\ 110 \\ 122 \end{pmatrix} = \begin{pmatrix} 1 & 47 & 85.4 & 1.75 & 5.1 & 63 & 33 \\ 1 & 49 & 94.2 & 2.10 & 3.8 & 70 & 14 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 48 & 90.5 & 1.88 & 9.0 & 71 & 99 \\ 1 & 56 & 95.7 & 2.09 & 7.0 & 75 & 99 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \dots \\ \beta_6 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_{19} \\ \varepsilon_{20} \end{pmatrix}$$

Non-normal outcome

- linear model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
  - outcome $\mathbf{Y}$
    - random vector, we observe a realization $\mathbf{y}$
  - predictors $\mathbf{x}_{,1}, \ldots, \mathbf{x}_{,k}$
    - vector of given (known) constants
  - coefficients $\boldsymbol{\beta}$
    - vector of unknown constants
  - error $\boldsymbol{\varepsilon}$
    - unknown random vector, we do not observe its realization
  - assumptions: $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \mathbf{I})$
    - $\mathrm{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$: the expected value of $\mathbf{Y}$ is a linear function of $\boldsymbol{\beta}$
    - $\mathrm{E}\,\boldsymbol{\varepsilon} = \mathbf{0}$: no systematic errors
    - $\mathrm{Var}\,\boldsymbol{\varepsilon} = \sigma^2 \mathbf{I}$: independence and same precision
- normality not crucial with a large data set without influential observations BUT what if $\mathbf{Y}$ is nowhere close to normal?
- e.g. what if $Y_i \in \{0, 1\}\ \forall i$?

Example: `heart attack` data

- From: Simon Wood (2006). Generalized Additive models
- Question: Is the level of creatinine kinase (CK) in blood a marker of an on-going heart attack (HA)?
- Data:

CK level	HA (yes:1, no:0)
20	1
20	1
20	0
20	0
20	0
20	0
20	0
20	0
20	0
20	0
20	0
...	...
$\mathbf{X}$	$\mathbf{Y}$

- Structure:
  - outcome $\mathbf{Y}$
    - random vector, we observe a realization $\mathbf{y}$, $y_i \in \{0, 1\}\ \forall i$
  - predictors $\mathbf{x}_{,1}, \ldots, \mathbf{x}_{,k}$
    - vector of given (known) constants

Binary outcome

- outcome $\mathbf{Y}$
  - random vector, we observe a realization $\mathbf{y}$, $y_i \in \{0, 1\}\ \forall i$
- predictors $\mathbf{x}_{,1}, \ldots, \mathbf{x}_{,k}$
  - vector of given (known) constants
- coefficients $\boldsymbol{\beta}$
  - vector of unknown constants
- how to connect $\mathbf{Y}, \mathbf{X}$, and $\boldsymbol{\beta}$, so that we can describe the relationship between $\mathbf{Y}$ and $\mathbf{X}$ using $\boldsymbol{\beta}$?
  - clearly not $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$

Binary outcome

- outcome $\mathbf{Y}$
  - random vector, we observe a realization $\mathbf{y}$, $y_i \in \{0, 1\}\ \forall i$
- predictors $\mathbf{x}_{,1}, \ldots, \mathbf{x}_{,k}$
  - vector of given (known) constants
- coefficients $\boldsymbol{\beta}$
  - vector of unknown constants
- how to connect $\mathbf{Y}, \mathbf{X}$, and $\boldsymbol{\beta}$, so that we can describe the relationship between $\mathbf{Y}$ and $\mathbf{X}$ using $\boldsymbol{\beta}$?
  - clearly not $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
  - how about picking up on $\mathrm{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$?

Binary outcome

- outcome $\mathbf{Y}$
  - random vector, we observe a realization $\mathbf{y}$, $y_i \in \{0, 1\}\ \forall i$
- predictors $\mathbf{x}_{,1}, \ldots, \mathbf{x}_{,k}$
  - vector of given (known) constants
- coefficients $\boldsymbol{\beta}$
  - vector of unknown constants
- how to connect $\mathbf{Y}, \mathbf{X}$, and $\boldsymbol{\beta}$, so that we can describe the relationship between $\mathbf{Y}$ and $\mathbf{X}$ using $\boldsymbol{\beta}$?
  - clearly not $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
  - how about picking up on $\mathrm{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$?
  - $Y_i \in \{0, 1\} \Rightarrow Y_i \sim \mathrm{Bernoulli}(p_i) = \mathrm{Bi}(1, p_i)$

Binary outcome

- outcome $\mathbf{Y}$
  - random vector, we observe a realization $\mathbf{y}$, $y_i \in \{0, 1\}\ \forall i$
- predictors $\mathbf{x}_{,1}, \ldots, \mathbf{x}_{,k}$
  - vector of given (known) constants
- coefficients $\boldsymbol{\beta}$
  - vector of unknown constants
- how to connect $\mathbf{Y}, \mathbf{X}$, and $\boldsymbol{\beta}$, so that we can describe the relationship between $\mathbf{Y}$ and $\mathbf{X}$ using $\boldsymbol{\beta}$?
  - clearly not $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
  - how about picking up on $\mathrm{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$?
  - $Y_i \in \{0, 1\} \Rightarrow Y_i \sim \mathrm{Bernoulli}(p_i) = \mathrm{Bi}(1, p_i)$
  - $\mathrm{E}\,Y_i = p_i = f(\mathbf{x}_{i,.}, \boldsymbol{\beta})$
  - how about $\mathrm{E}\,Y_i = \mathbf{x}_{i,.}^\top \boldsymbol{\beta}$? (i.e. $\mathrm{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$ again)

Binary outcome

- outcome $\mathbf{Y}$
  - random vector, we observe a realization $\mathbf{y}$, $y_i \in \{0, 1\}\ \forall i$
- predictors $\mathbf{x}_{,1}, \ldots, \mathbf{x}_{,k}$
  - vector of given (known) constants
- coefficients $\boldsymbol{\beta}$
  - vector of unknown constants
- how to connect $\mathbf{Y}, \mathbf{X}$, and $\boldsymbol{\beta}$, so that we can describe the relationship between $\mathbf{Y}$ and $\mathbf{X}$ using $\boldsymbol{\beta}$?
  - clearly not $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
  - how about picking up on $\mathrm{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$?
  - $Y_i \in \{0, 1\} \Rightarrow Y_i \sim \mathrm{Bernoulli}(p_i) = \mathrm{Bi}(1, p_i)$
  - $\mathrm{E}\,Y_i = p_i = f(\mathbf{x}_{i,.}, \boldsymbol{\beta})$
  - how about $\mathrm{E}\,Y_i = \mathbf{x}_{i,.}^\top \boldsymbol{\beta}$? (i.e. $\mathrm{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$ again)
  - $\mathrm{E}\,Y_i$ is a probability $\Rightarrow$ must be in $[0, 1]$ . . . unlike $\mathbf{x}_{i,.}^\top\,\boldsymbol{\beta}$

Binary outcome

- outcome $\mathbf{Y}$
  - random vector, we observe a realization $\mathbf{y}$, $y_i \in \{0, 1\}\ \forall i$
- predictors $\mathbf{x}_{,1}, \ldots, \mathbf{x}_{,k}$
  - vector of given (known) constants
- coefficients $\boldsymbol{\beta}$
  - vector of unknown constants
- how to connect $\mathbf{Y}, \mathbf{X}$, and $\boldsymbol{\beta}$, so that we can describe the relationship between $\mathbf{Y}$ and $\mathbf{X}$ using $\boldsymbol{\beta}$?
  - clearly not $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I})$
  - how about picking up on $\mathrm{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$?
  - $Y_i \in \{0, 1\} \Rightarrow Y_i \sim \mathrm{Bernoulli}(p_i) = \mathrm{Bi}(1, p_i)$
  - $\mathrm{E}\,Y_i = p_i = f(\mathbf{x}_{i,.}, \boldsymbol{\beta})$
  - how about $\mathrm{E}\,Y_i = \mathbf{x}_{i,.}^\top \boldsymbol{\beta}$? (i.e. $\mathrm{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$ again)
  - $\mathrm{E}\,Y_i$ is a probability $\Rightarrow$ must be in $[0, 1]$ ... unlike $\mathbf{x}_{i,.}^\top \boldsymbol{\beta}$
  - how about $\mathrm{E}\,Y_i = g^{-1}\left(\mathbf{x}_{i,.}^\top \boldsymbol{\beta}\right)$, where $g : (0, 1) \mapsto \mathbb{R}$?

Binary outcome

- outcome $\mathbf{Y}$
  - random vector, we observe a realization $\mathbf{y}$, $y_i \in \{0, 1\}\ \forall i$
- predictors $\mathbf{x}_{,1}, \ldots, \mathbf{x}_{,k}$
  - vector of given (known) constants
- coefficients $\boldsymbol{\beta}$
  - vector of unknown constants
- how to connect $\mathbf{Y}, \mathbf{X}$, and $\boldsymbol{\beta}$, so that we can describe the relationship between $\mathbf{Y}$ and $\mathbf{X}$ using $\boldsymbol{\beta}$?
  - clearly not $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\,\mathbf{I})$
  - how about picking up on $\mathrm{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$?
  - $Y_i \in \{0, 1\} \Rightarrow Y_i \sim \mathrm{Bernoulli}(p_i) = \mathrm{Bi}(1, p_i)$
  - $\mathrm{E}\,Y_i = p_i = f(\mathbf{x}_{i,.}, \boldsymbol{\beta})$
  - how about $\mathrm{E}\,Y_i = \mathbf{x}_{i,.}^\top \boldsymbol{\beta}$? (i.e. $\mathrm{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$ again)
  - $\mathrm{E}\,Y_i$ is a probability $\Rightarrow$ must be in $[0, 1]$ ... unlike $\mathbf{x}_{i,.}^\top\,\boldsymbol{\beta}$
  - how about $\mathrm{E}\,Y_i = g^{-1}\left(\mathbf{x}_{i,.}^\top\,\boldsymbol{\beta}\right)$, where $g : (0, 1) \mapsto \mathbb{R}$?
  - logistic regression: $\mathrm{E}\,Y_i = \exp\left\{\mathbf{x}_{i,.}^\top\,\boldsymbol{\beta}\right\} / \left(1 + \exp\left\{\mathbf{x}_{i,.}^\top\,\boldsymbol{\beta}\right\}\right)$

Generalized linear model

- outcome $\mathbf{Y}$
  - random vector, we observe a realization $\mathbf{y}$
- predictors $\mathbf{x}_{,1}, \ldots, \mathbf{x}_{,k}$
  - vector of given (known) constants
- coefficients $\boldsymbol{\beta}$
  - vector of unknown constants
- model: $Y_i \overset{iid}{\sim} \mathcal{L}$
  - $\mathcal{L}$ a probability distribution
    - from an exponential family of distributions
    - density/probability mass function satisfies that

$$f(y; \theta, \varphi) = \exp\left\{\frac{\theta y - b(\theta)}{\varphi} + c(y, \varphi)\right\}$$

      and there are some assumptions on b
    - $\mathrm{E}\, Y = b'(\theta)$, $\mathrm{Var}\, Y = \varphi\, b''(\theta)$
  - $g(\mathrm{E}\, Y_i) = \mathbf{x}_{i,.}^\top \boldsymbol{\beta}$
    - g is called link function

GLM example: linear regression

- exponential family of distributions

  $$f(y;\theta,\varphi) = \exp\left\{\frac{\theta y - b(\theta)}{\varphi} + c(y,\varphi)\right\}$$

  - Gaussian distribution $Y_i \overset{iid}{\sim} \mathrm{N}(\mu,\sigma^2)$

  $$\begin{aligned} f(y;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} \\ &= \exp\left\{\frac{\mu\, y - \mu^2/2}{\sigma^2} - \frac{1}{2\sigma^2}\, y^2 - \frac{1}{2}\, \log(2\pi\sigma^2)\right\} \end{aligned}$$

- link function $g(\mathrm{E}\, Y_i) = \mathbf{x}_{i,\cdot}^\top \boldsymbol{\beta}$
  - $\mathrm{E}\, Y = \mu$
  - canonical link: identity $g(x) = x \rightsquigarrow$ linear regression
- log-link may help address heteroskedasticity in linear regression

GLM example: Gamma regression

- exponential family of distributions

$$f(y;\theta,\varphi) = \exp\left\{\frac{\theta y - b(\theta)}{\varphi} + c(y,\varphi)\right\}$$

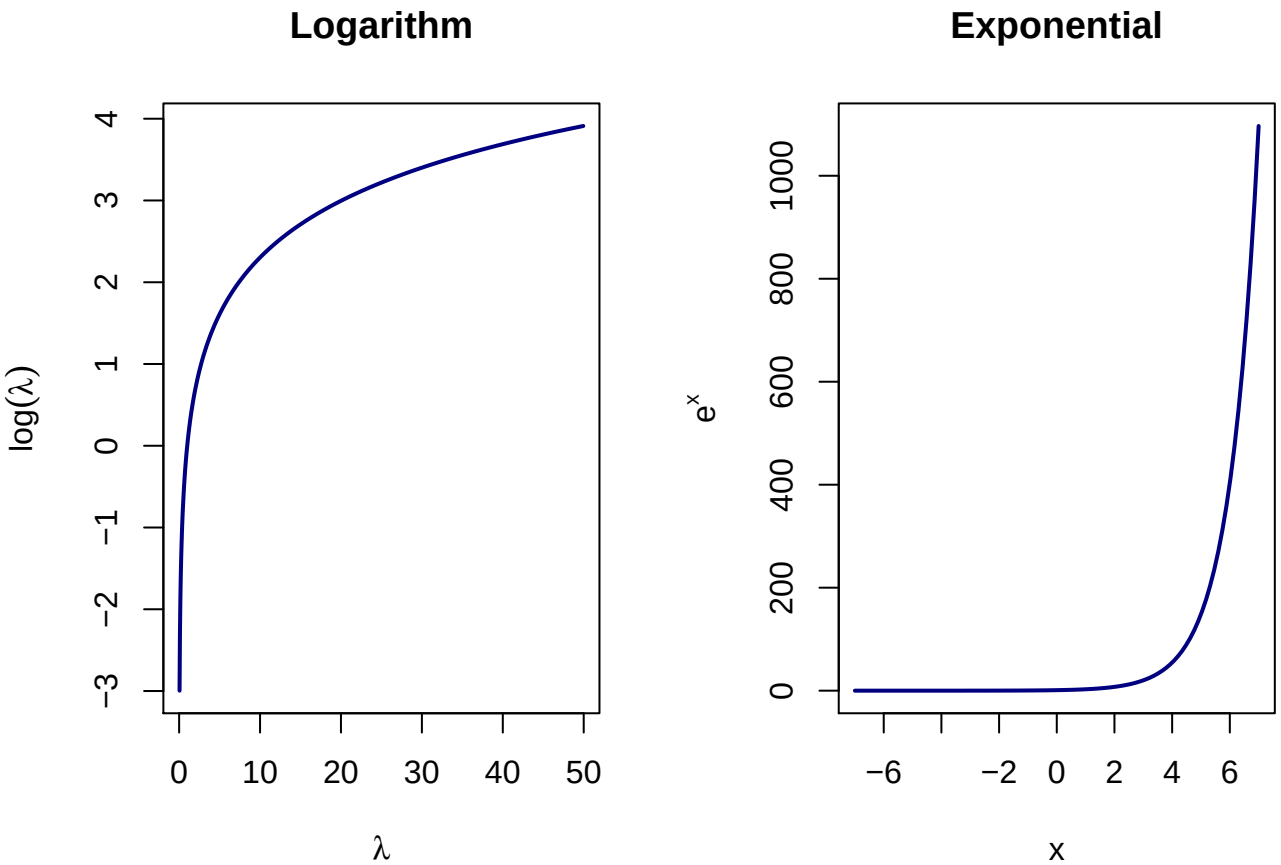
  - Gamma distribution $Y_i \overset{iid}{\sim} \mathrm{Gamma}(\alpha,\beta)$

$$\begin{aligned} f(y;\alpha,\beta) &= \frac{\beta^\alpha\, y^{\alpha-1}\, \mathrm{e}^{-\beta y}}{\Gamma(\alpha)} \\ &= \exp\left\{\alpha\, \log(\beta) + (\alpha-1)\log(y) - \beta y - \log(\Gamma(\alpha))\right\} \\ &= \exp\left\{\frac{(-\beta/\alpha)\, y + \log(\beta/\alpha)}{1/\alpha} + \alpha \log(\alpha\, y) - \log\left(\Gamma(\alpha)\, y\right)\right\} \end{aligned}$$

- link function $g(\mathrm{E}\, Y_i) = \mathbf{x}_{i,\cdot}^\top \boldsymbol{\beta}$
  - $\mathrm{E}\, Y = \alpha/\beta$
  - common links: $g(x) = 1/x$ (canonical), $g(x) = \log(x)$
- used for skewed non-negative data
- addresses heteroskedasticity and heavy tail (e.g. the size of an insurance claim) but other choices possible as well

Log link

- logarithm $g(\lambda) = \log(\lambda) : (0, \infty) \mapsto \mathbb{R}$
- exponential $g^{-1}(x) = \mathrm{e}^x : \mathbb{R} \mapsto (0, \infty)$



GLM example: log-linear model

- exponential family of distributions

$$f(y;\theta,\varphi) = \exp\left\{\frac{\theta y - b(\theta)}{\varphi} + c(y,\varphi)\right\}$$

  - Poisson distribution $Y_i \overset{iid}{\sim} \mathrm{Po}(\lambda)$

$$\begin{aligned} f(y;\lambda) &= \mathrm{e}^{-\lambda}\,\frac{\lambda^y}{y!} \\ &= \exp\left\{\log(\lambda)\,y - \lambda - \log(y!)\right\} \end{aligned}$$

- link function $g(\mathrm{E}\,Y_i) = \mathbf{x}_{i,\cdot}^\top \boldsymbol{\beta}$
  - $\mathrm{E}\,Y = \lambda$
  - canonical link: log link: $g(x) = \log(x)$
- convenient way of handling contingency tables
- used to model e.g. the number of insurance claims
- has connections to Cox PH model

GLM example: logistic regression

- exponential family of distributions

$$f(y;\theta,\varphi) = \exp\left\{\frac{\theta y - b(\theta)}{\varphi} + c(y,\varphi)\right\}$$

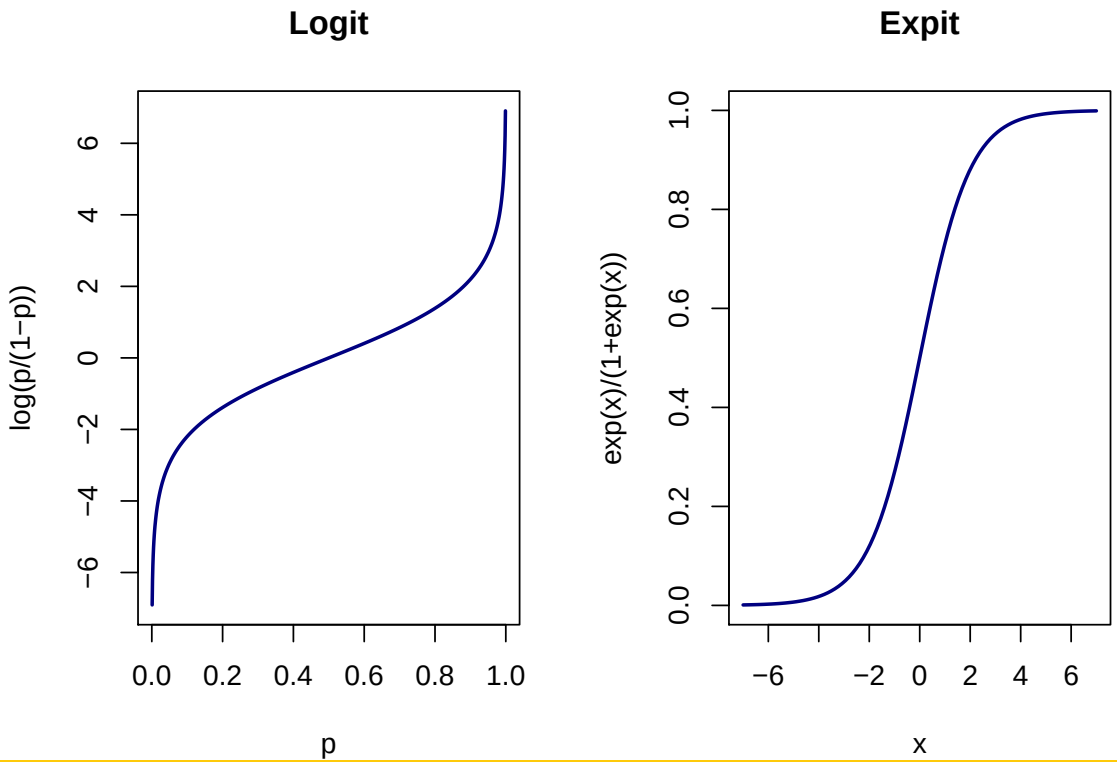
  - Bernoulli distribution: $Y_i \overset{iid}{\sim} \text{Bernoulli}(p)$

$$\begin{aligned} f(y;p) &= p^y(1-p)^{1-y} \\ &= \exp\left\{y\log(p) + (1-y)\log(1-p)\right\} \\ &= \exp\left\{\log\left(\frac{p}{1-p}\right) y + \log(1-p)\right\} \end{aligned}$$

- link function $g(\mathrm{E}\, Y_i) = \mathbf{x}_{i,\cdot}^\top \boldsymbol{\beta}$
  - $\mathrm{E}\, Y = p$
  - canonical link: "logit" $g(x) = \log\left(\frac{x}{1-x}\right)$
  - other common choices: "probit", "complementary log-log"
- used e.g. in credit risk analysis (probability of default, classification)

"Logit" link

- "logit" $g(p) = \log\left(\frac{p}{1-p}\right) : (0,1) \mapsto \mathbb{R}$
- "expit" $g^{-1}(x) = \frac{e^x}{1+e^x} : \mathbb{R} \mapsto (0,1)$



MLE for θ in exponential families

- exponential family of distributions

$$f(y;\theta,\varphi) = \exp\left\{\frac{\theta y - b(\theta)}{\varphi} + c(y,\varphi)\right\}$$

- likelihood

$$L(\mathbf{y};\theta,\varphi) = \prod_{i=1}^{n} f(y_i;\theta,\varphi) = \exp\left\{\frac{\theta}{\varphi}\sum_{i=1}^{n} y_i - n\,\frac{b(\theta)}{\varphi} + \sum_{i=1}^{n} c(y_i,\varphi)\right\}$$

- log-likelihood

$$\ell(\mathbf{y};\theta,\varphi) = \frac{\theta}{\varphi}\sum_{i=1}^{n} y_i - n\,\frac{b(\theta)}{\varphi} + \sum_{i=1}^{n} c(y_i,\varphi)$$

- score function (the θ-related part)

$$U_1(\mathbf{y};\theta,\varphi) = \frac{\partial}{\partial\theta}\,\ell(\mathbf{y};\theta,\varphi) = \frac{1}{\varphi}\sum_{i=1}^{n} y_i - n\,\frac{b'(\theta)}{\varphi}$$

- solution to the score equation (the θ-related part)

$$U_1(\mathbf{y};\theta,\varphi) = 0 \Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} y_i = b'(\theta)$$

From MLE for θ to MLE for β

- under some assumptions on the exponential family
  - $\exists$ unique MLE for θ:
$$\hat{\theta} = b'^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)$$
  - it can be shown that
$$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} \mathrm{N}\left(0, \varphi\left(b''(\theta)\right)^{-1}\right)$$
  - note that
    - $\hat{\theta}$ does not depend on φ
      $\Rightarrow$ we do not need φ for point estimation of θ
    - the (asymptotic) variance of $\hat{\theta}$ depends on φ $\Rightarrow$
      $\Rightarrow$ we do need $\hat{\varphi}$ for interval estimation of θ

From MLE for θ to MLE for β

- under some assumptions on the exponential family
  - $\exists$ unique MLE for θ:
$$\hat{\theta} = b'^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)$$
  - it can be shown that
$$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} \mathrm{N}\left(0, \varphi\left(b''(\theta)\right)^{-1}\right)$$
  - note that
    - $\hat{\theta}$ does not depend on φ
      $\Rightarrow$ we do not need φ for point estimation of θ
    - the (asymptotic) variance of $\hat{\theta}$ depends on φ $\Rightarrow$
      $\Rightarrow$ we do need $\hat{\varphi}$ for interval estimation of θ
- this is all very nice BUT in a GLM
  - $g(\mathrm{E}\, Y_i) = \mathbf{x}_{i,\cdot}^{\top}\boldsymbol{\beta} = b(\theta_i)$
    - so there is θ_i, not θ
    - and we want to estimate $\boldsymbol{\beta}$

Estimating β in GLMs

- log-likelihood for an exponential family:

$$\ell(\mathbf{y}; \theta, \varphi) = \frac{\theta}{\varphi} \sum_{i=1}^{n} y_i - n \frac{b(\theta)}{\varphi} + \sum_{i=1}^{n} c(y_i, \varphi)$$

- log-likelihood for a GLM:

$$\ell(\mathbf{y}; \boldsymbol{\beta}, \varphi) = \frac{1}{\varphi} \sum_{i=1}^{n} \left(\theta_i \, y_i - b(\theta_i)\right) + \sum_{i=1}^{n} c(y_i, \varphi)$$

- score function (the $\boldsymbol{\beta}$-related part)

$$U_{1:p}(\mathbf{y}; \boldsymbol{\beta}, \varphi) = \frac{\partial}{\partial \boldsymbol{\beta}} \ell(\mathbf{y}; \boldsymbol{\beta}, \varphi) = \frac{1}{\varphi} \sum_{i=1}^{n} \left(y_i - b'(\theta_i)\right) \frac{\partial}{\partial \boldsymbol{\beta}} \theta_i$$

- no closed-form solution $\Rightarrow$ numerical solution through the **I**teratively **R**e-weighted **L**east **S**quares algorithm
  - usually converges fast; if not, we might have a deeper problem
    - no guarantee that a solution exists
    - no guarantee that a solution is the MLE unless the link is canonical

Inference for $\widehat{\boldsymbol{\beta}}$

- $\widehat{\boldsymbol{\beta}}$ is a MLE $\Rightarrow$ we can use general theory on the asymptotic properties of the MLEs
  - the results are asymptotic (i.e. for large n)
  - hold under some assumptions but "if all is well"...
    1. $\widehat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$
    2. $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \mathbf{I}^{-1}(\boldsymbol{\beta}, \varphi))$
    3. $2\left(\ell(\mathbf{Y}; \widehat{\boldsymbol{\beta}}, \varphi) - \ell(\mathbf{Y}; \boldsymbol{\beta}, \varphi)\right) \xrightarrow{d} \chi^2_p$
- inference for $\widehat{\boldsymbol{\beta}}$
  1. tells us we are eventually getting what we want
  2. is a basis for Wald tests and CIs about $\boldsymbol{\beta}$
  3. and similar results basis for likelihood ratio tests and CIs for $\boldsymbol{\beta}$
     - CIs are based on profile likelihood
     - recommended over Wald (or Rao) tests and CIs
- we have treated φ as fixed so far but we need to estimate it in order to get the test statistics and CIs

Deviance

- a model that fits $\widehat{\mathrm{E}\, Y_i} = Y_i$ is called saturated model in GLMs
  - has a parameter for each unique covariate combination
- unscaled deviance is $\varphi\times$ the difference between the maximized log-likelihood in the saturated and current model

$$D(\mathbf{y}, \widehat{\boldsymbol{\beta}}) = 2\varphi\big(\ell(\text{saturated model}) - \ell(\mathbf{y}; \widehat{\boldsymbol{\beta}}, \varphi)\big)$$

  - a goodness-of-fit measure
  - a generalization of the residual sum of squares from LM
- scaled deviance is the difference between the maximized log-likelihood in the saturated and current model

$$D^*(\mathbf{y}, \widehat{\boldsymbol{\beta}}) = 2\big(\ell(\text{saturated model}) - \ell(\mathbf{y}; \widehat{\boldsymbol{\beta}}, \varphi)\big)$$

  - difference between deviances of two nested models is the test statistic of the likelihood ratio test (with φ replaced by $\hat{\varphi}$)
- other goodness of fit and model selection tools
  - AIC, BIC, ...

Residuals and model diagnostics

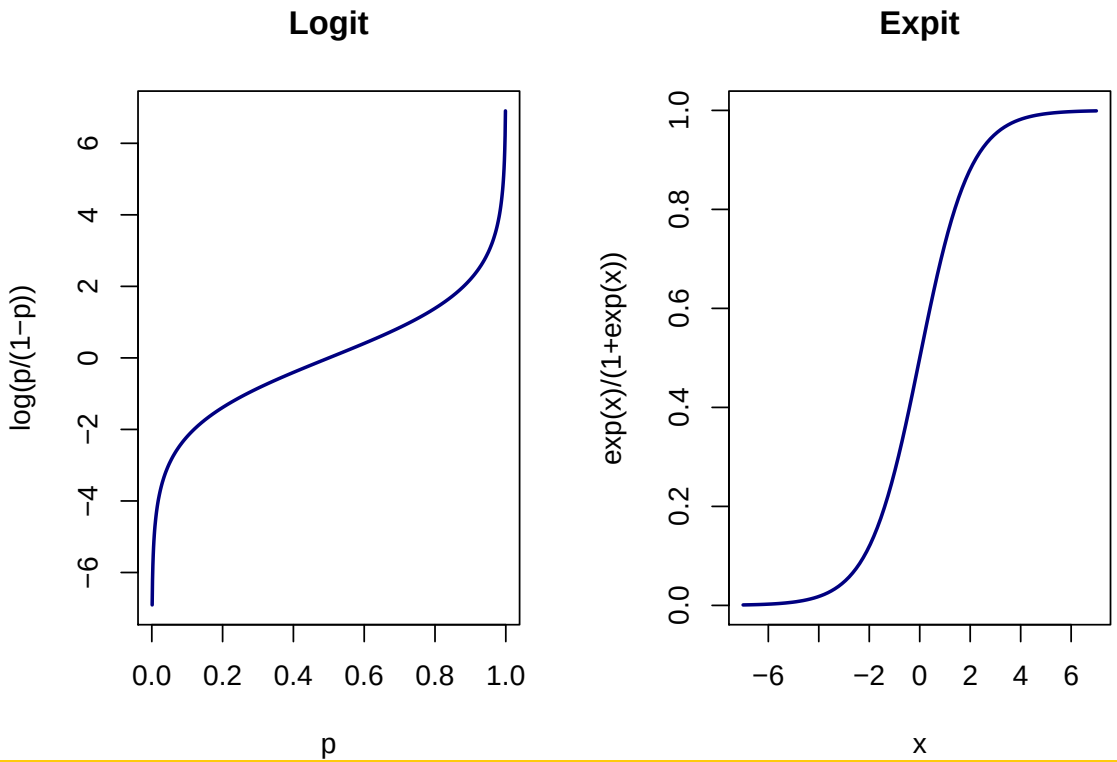
- Pearson residuals
$$r_i^P = \frac{Y_i - \widehat{\mathrm{E}\, Y_i}}{\sqrt{\widehat{\mathrm{Var}\, Y_i}}}$$
- $\mathrm{Var}\, r_i^P \approx \varphi(1 - h_{i,i})$ (**H** comes from the WLS in IRLS)
- standardized Pearson residuals
$$r_i^{SP} = \frac{Y_i - \widehat{\mathrm{E}\, Y_i}}{\sqrt{\hat{\varphi}\, \widehat{\mathrm{Var}\, Y_i}(1 - h_{i,i})}}$$
- deviance residuals
$$r_i^D = \mathrm{sgn}(Y_i - \widehat{\mathrm{E}\, Y_i})\, d_i$$
  - d_i^2 is the contribution of the i^{th} observation to the deviance
- standardized deviance residuals
$$r_i^{SD} = \frac{\mathrm{sgn}(Y_i - \widehat{\mathrm{E}\, Y_i})\, d_i}{\sqrt{\hat{\varphi}\,(1 - h_{i,i})}}$$
- residuals can be used for residual plots as in LM
- a generalization of leverage and Cook's distance from LM available for GLM

Logistic regression

- outcome $\mathbf{Y}$
  - random vector, we observe a realization $\mathbf{y}$, $y_i \in \{0, 1\}\ \forall i$
- predictors $\mathbf{x}_{,1}, \ldots, \mathbf{x}_{,k}$
  - vector of given (known) constants
- coefficients $\boldsymbol{\beta}$
  - vector of unknown constants
- model:
  - $Y_i \overset{iid}{\sim} \text{Bernoulli}(p_i)$
  - $p_i = \frac{\exp\left\{\mathbf{x}_{i,.}^\top \boldsymbol{\beta}\right\}}{1+\exp\left\{\mathbf{x}_{i,.}^\top \boldsymbol{\beta}\right\}}$ with "logit" link: $g(p) = \log\left(\frac{p}{1-p}\right)$
  - less common choices for the link function:
    - $p_i = \Phi\{\mathbf{x}_{i,.}^\top \boldsymbol{\beta}\}$ with Φ the distribution function of $\text{N}(0, 1)$ with "probit" link $g(p) = \Phi^{-1}(p)$
    - $p_i = 1 - \exp\left\{-\exp\{\mathbf{x}_{i,.}^\top \boldsymbol{\beta}\}\right\}$ with "complementary log-log" link $g(p) = \log\left(-\log(1-p)\right)$

"Logit" link

- "logit" $g(p) = \log\left(\frac{p}{1-p}\right) : (0,1) \mapsto \mathbb{R}$
- "expit" $g^{-1}(x) = \frac{e^x}{1+e^x} : \mathbb{R} \mapsto (0,1)$



Example: `heart attack` data

- Is the level of creatinine kinase (CK) in blood a marker of an on-going heart attack (HA)?
- Data:

CK level	HA (yes:1, no:0)
20	1
20	1
20	0
20	0
20	0
20	0
20	0
20	0
20	0
20	0
20	0
...	...
...	...
...	...

- Data (equivalent form):

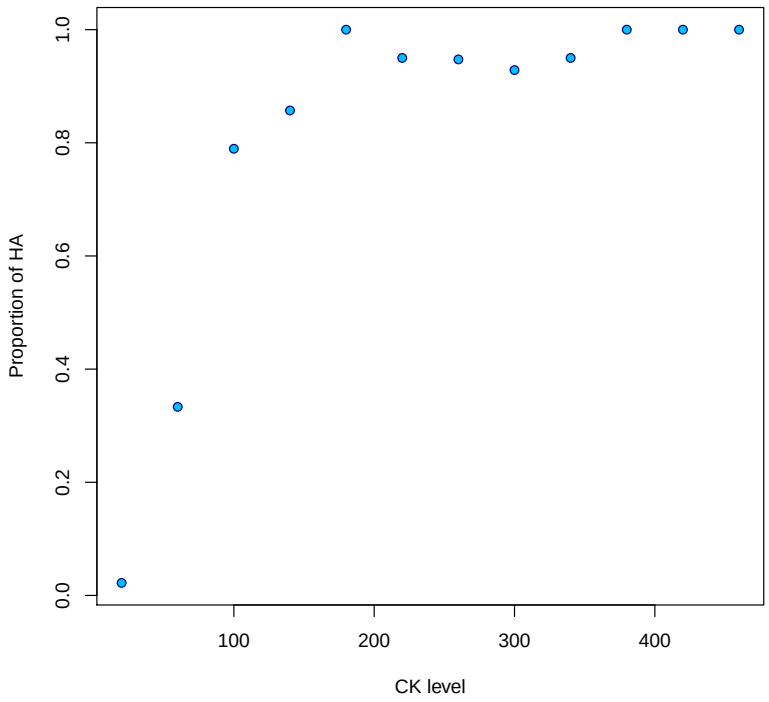
CK level	Nr. of HAs	Nr. of no HAs
20	2	88
60	13	26
100	30	8
140	30	5
180	21	0
220	19	1
260	18	1
300	13	1
340	19	1
380	15	0
420	7	0
460	8	0

Binomial form for the `heart attack` data

Data

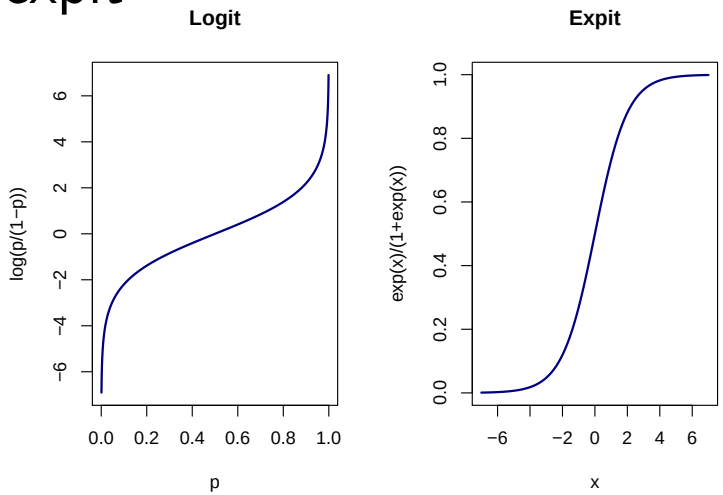
CK level	Nr. of HAs	Nr. of no HAs
20	2	88
60	13	26
100	30	8
140	30	5
180	21	0
220	19	1
260	18	1
300	13	1
340	19	1
380	15	0
420	7	0
460	8	0

Observed proportions

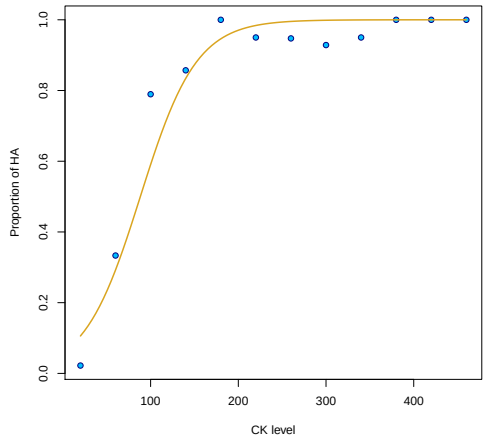


Model fit: a logistic curve

- "logit" and "expit"



- fitted logistic curve for the `heart attack` data



Fitted model for the `heart attack` data

```
> summary(glm.ha)

Call:
glm(formula = cbind(ha.ha, ha.ok) ~ ck, family = "binomial",
    data = heart.attack)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.08184  -1.93008   0.01652   0.41772   2.60362

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.758358   0.336696  -8.192 2.56e-16 ***
ck           0.031244   0.003619   8.633  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 271.712  on 11  degrees of freedom
Residual deviance:  36.929  on 10  degrees of freedom
AIC: 62.334

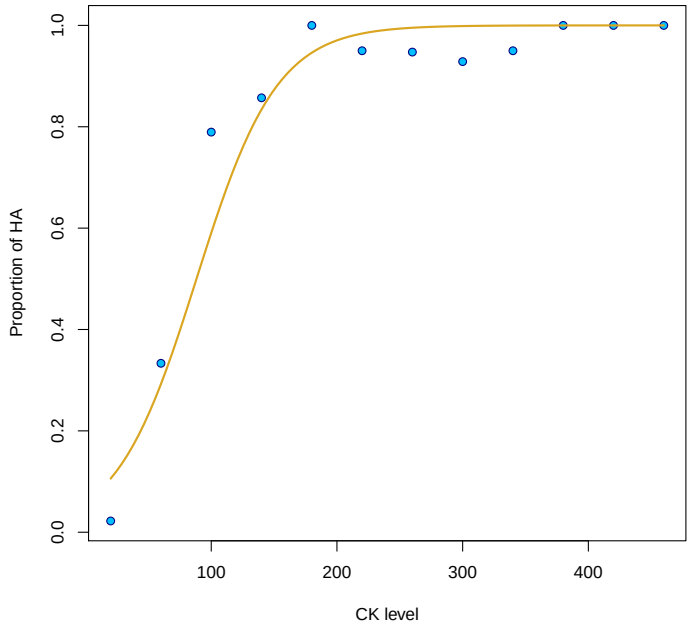
Number of Fisher Scoring iterations: 6
```

Fitted logistic curve for the `heart attack` data

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.758358   0.336696  -8.192 2.56e-16 ***
ck           0.031244   0.003619   8.633  < 2e-16 ***
```

- fitted probability

$$p(\text{ck}) = \frac{\exp\{-2.76 + 0.03\text{ck}\}}{1 + \exp\{-2.76 + 0.03\text{ck}\}}$$



Interpretation of the parameters

- fitted probability
$$p(\text{ck}) = \frac{\exp\{-2.76 + 0.03\,\text{ck}\}}{1 + \exp\{-2.76 + 0.03\,\text{ck}\}}$$
- is there a nice way to see $\hat{\beta}_1 = 0.03$?
- odds
$$\frac{p}{1-p}$$
- odds ratio
$$\left(\frac{p}{1-p}\right) / \left(\frac{\tilde{p}}{1-\tilde{p}}\right)$$
- $e^{\hat{\beta}_1}$ is the estimated odds ratio for two patients whose difference in CK level is one unit
- estimated odds for heart attack become $e^{\hat{\beta}_1} = 1.03$ times higher when the CK level increases by one unit
- with more covariates the interpretation remains the same *when the values of all other covariates are kept fixed*

Fitted model for the `heart attack` data

```
> summary(glm.ha)

Call:
glm(formula = cbind(ha.ha, ha.ok) ~ ck, family = "binomial",
    data = heart.attack)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.08184  -1.93008   0.01652   0.41772   2.60362

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.758358   0.336696  -8.192 2.56e-16 ***
ck           0.031244   0.003619   8.633  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 271.712  on 11  degrees of freedom
Residual deviance:  36.929  on 10  degrees of freedom
AIC: 62.334

Number of Fisher Scoring iterations: 6
```

Inference for the `heart attack` data

- Wald test statistics (and confidence intervals)

```
> summary(glm.ha)
...
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.758358   0.336696  -8.192 2.56e-16 ***
ck           0.031244   0.003619   8.633  < 2e-16 ***
```

- likelihood ratio confidence intervals (preferred)

```
> confint(glm.ha)
Waiting for profiling to be done...
                  2.5 %      97.5 %
(Intercept) -3.46305890 -2.13705606
ck           0.02467179  0.03889618
```

- likelihood ratio test (preferred)

```
> anova(glm.ha.null, glm.ha, test="Chisq")
Analysis of Deviance Table

Model 1: cbind(ha.ha, ha.ok) ~ 1
Model 2: cbind(ha.ha, ha.ok) ~ ck
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        11    271.712
2        10     36.929  1   234.78 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fitted model for the `heart attack` data

```
> summary(glm.ha)

Call:
glm(formula = cbind(ha.ha, ha.ok) ~ ck, family = "binomial",
    data = heart.attack)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.08184  -1.93008   0.01652   0.41772   2.60362

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.758358   0.336696  -8.192 2.56e-16 ***
ck           0.031244   0.003619   8.633  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 271.712  on 11  degrees of freedom
Residual deviance:  36.929  on 10  degrees of freedom
AIC: 62.334

Number of Fisher Scoring iterations: 6
```

Goodness of fit for the `heart attack` data

- ```
  > summary(glm.ha)
  ...
      Null deviance: 271.712  on 11  degrees of freedom
  Residual deviance:  36.929  on 10  degrees of freedom
  AIC: 62.334
  ```
- null deviance: deviance of the null model (only intercept)
- residual deviance: deviance of the current model
- a generalization of the proportion explained

  ```
  > (271.712 - 36.929)/271.712
  [1] 0.8640877
  ```
- residual variance should be $\approx \chi^2_{10}$ if the model is OK: deviance sometimes used for goodness of fit (caution...) but primary use is for model comparison

  ```
  > 1-pchisq(36.929, df=10)
  [1] 5.821642e-05
  ```
- other measures of goodness of fit/model comparison/selection

  ```
  > AIC(glm.ha)
  [1] 62.3339
  > BIC(glm.ha)
  [1] 63.30371
  ```
  ```

Fitted model for the `heart attack` data

```
> summary(glm.ha)

Call:
glm(formula = cbind(ha.ha, ha.ok) ~ ck, family = "binomial",
    data = heart.attack)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.08184  -1.93008   0.01652   0.41772   2.60362

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.758358   0.336696  -8.192 2.56e-16 ***
ck           0.031244   0.003619   8.633  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 271.712  on 11  degrees of freedom
Residual deviance:  36.929  on 10  degrees of freedom
AIC: 62.334

Number of Fisher Scoring iterations: 6
```

Example: diagnostic plots for the `heart attack` data