Linear and Generalized Linear Models

Week 7, Lecture 1

Logistic regression

Saskia le Cessie

Leiden University Medical Centre

Linear regression:

- Linear model: $\boldsymbol{y} = \boldsymbol{X\beta} + \epsilon$, where $\epsilon \sim N_n(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_n)$ and $\mathbf{X}_{n\times(k+1)}$ is the model matrix.
- Least-squares estimator: $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$
- Variance-covariance matrix for $\mathbf{b}$: $V(\mathbf{b}) = \sigma_\epsilon^2 (\mathbf{X}'\mathbf{X})^{-1}$.
- Distribution of $\mathbf{b}$: $\mathbf{b} \sim N_{k+1}(\boldsymbol{\beta}, \sigma_\epsilon^2 (\mathbf{X}'\mathbf{X})^{-1})$

Responses are not always continuous

- What if the response Y is a 0/1 variable (dichotomous variable, binary variable) ?
  - yes / no
  - ill / healthy
  - success / failure
- How does Y depend on one or more X-variables?

Example: Predicting Rheumatoid arthritis

- Early Arthritis Clinic, Department of Rheumatology, LUMC
- Patients in Leiden area with arthritis complaints (painfull joints, swollen joints) are referred to this clinic by their GP
- Some patients are diagnosed with undifferential arthritis
- Which of these patients will get Rheumatoid Arthritis (RA) within a year ?

Aim: a model to predict the risk of RA within 1 year

Patients with high risk can be treated with new treatments(expensive, side effects)

Notations

Y_i binary response variable for person *i* : (Y_i =1 , RA after 1 year, Y_i = 0 no RA)

Y_i follows a binomial distribution: Y_i ~ binomial (1, π_i).

Several predictors $(x_{i1}, \dots, x_{ik})$ are measured

目标 explaining

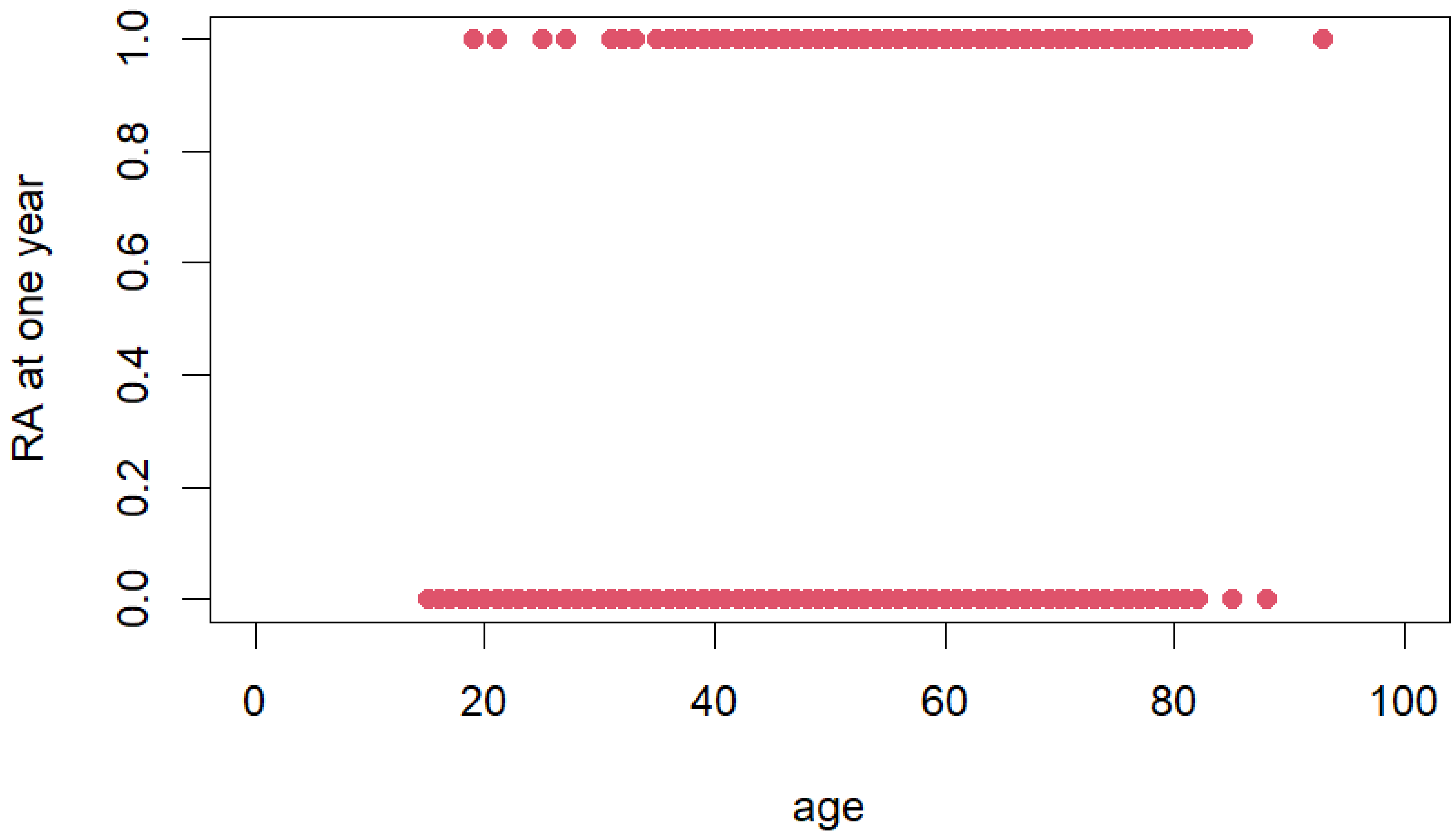
Aim: a model for $E(Y_i|x_{i1}, \dots, x_k) = P(Y_i = 1|x_{i1}, \dots, x_{ik}) = \pi_i$.

Matrix notations:

$x_i{}' = (1, x_{i1}, \dots, x_{ik})$ X = $\begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & & x_{2k} \\ \vdots & & \ddots & \\ 1 & x_{n1} & & x_{nk} \end{pmatrix}$ Y = $\begin{pmatrix} Y_1 \\ Y_2 \\ \\ Y_n \end{pmatrix}$

We assume independent observations

Let's start with one x variable: age



Problems with linear regression model for π

>1 <0

- Predicted probabilities can be greater than 1 or less than 0.
- Variance of Y_i is $\pi_i(1 - \pi_i)$. However, the linear model assumes constant variance.

Therefore : develop a model directly suited for binomial data.

Transform $\pi = P(Y = 1|x)$

- Requirement: $0 \leq \pi \leq 1$
- Therefore $0 \leq \frac{\pi}{1-\pi} \leq \infty$ ($\frac{\pi}{1-\pi}$ is called the odds)
- And $-\infty \leq \log\left(\frac{\pi}{1-\pi}\right) \leq \infty$

→ Use a linear model for $\log\left(\frac{\pi}{1-\pi}\right)$ =logit(π)

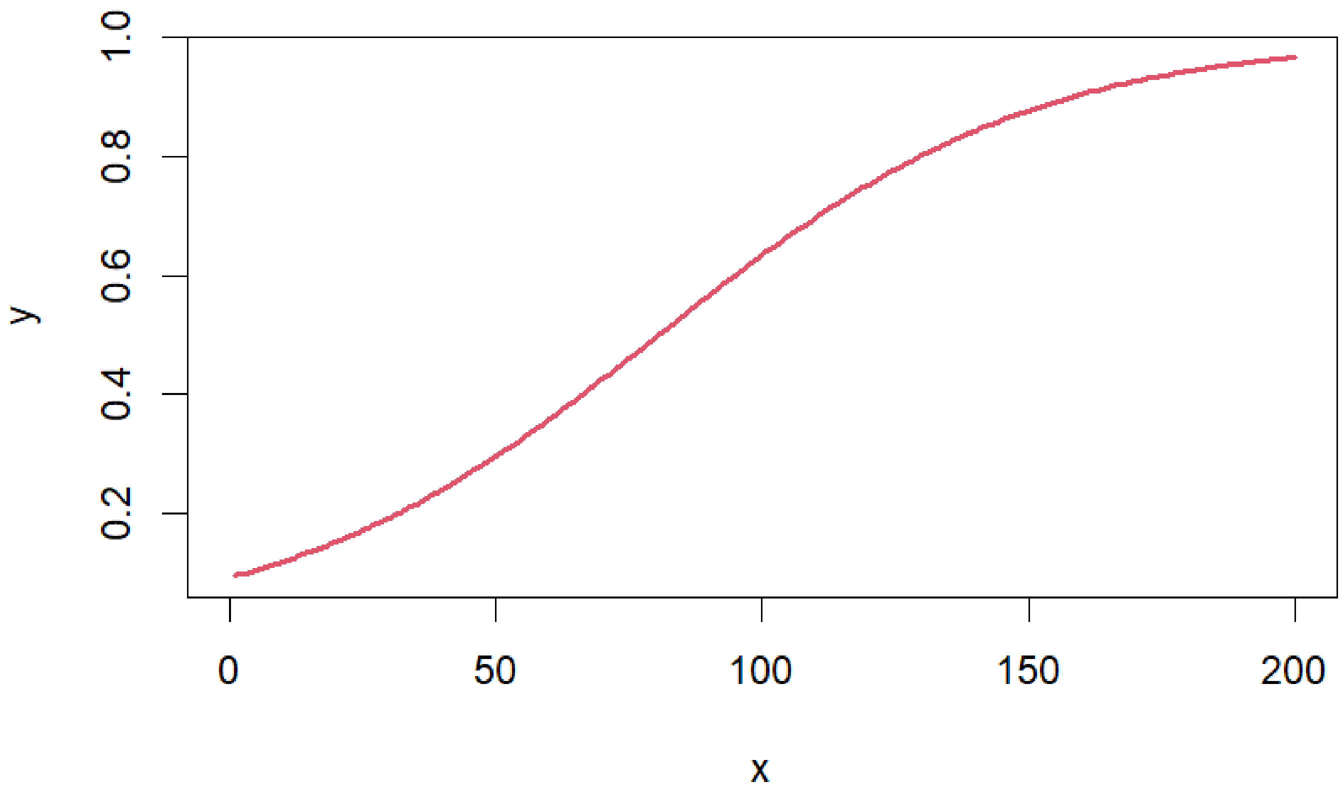
The logistic regression model

- $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + .. + \beta_k x_k$

- Or equivalent:

- $\pi = \frac{\exp(\beta_0+\beta_1 x_1+..+\beta_k x_k)}{1+\exp(\beta_0+\beta_1 x_1+..+\beta_k x_k)}$

(show it yourself)



Logistic regression in R

```
> model.lr <- glm(ra1year~Age, family=binomial, data=RAdata)
> summary(model.lr)

Call:
glm(formula = ra1year ~ Age, family = binomial, data = RAdata)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.281798   0.321459  -7.098 1.26e-12 ***
Age          0.028303   0.005717   4.951 7.39e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The model

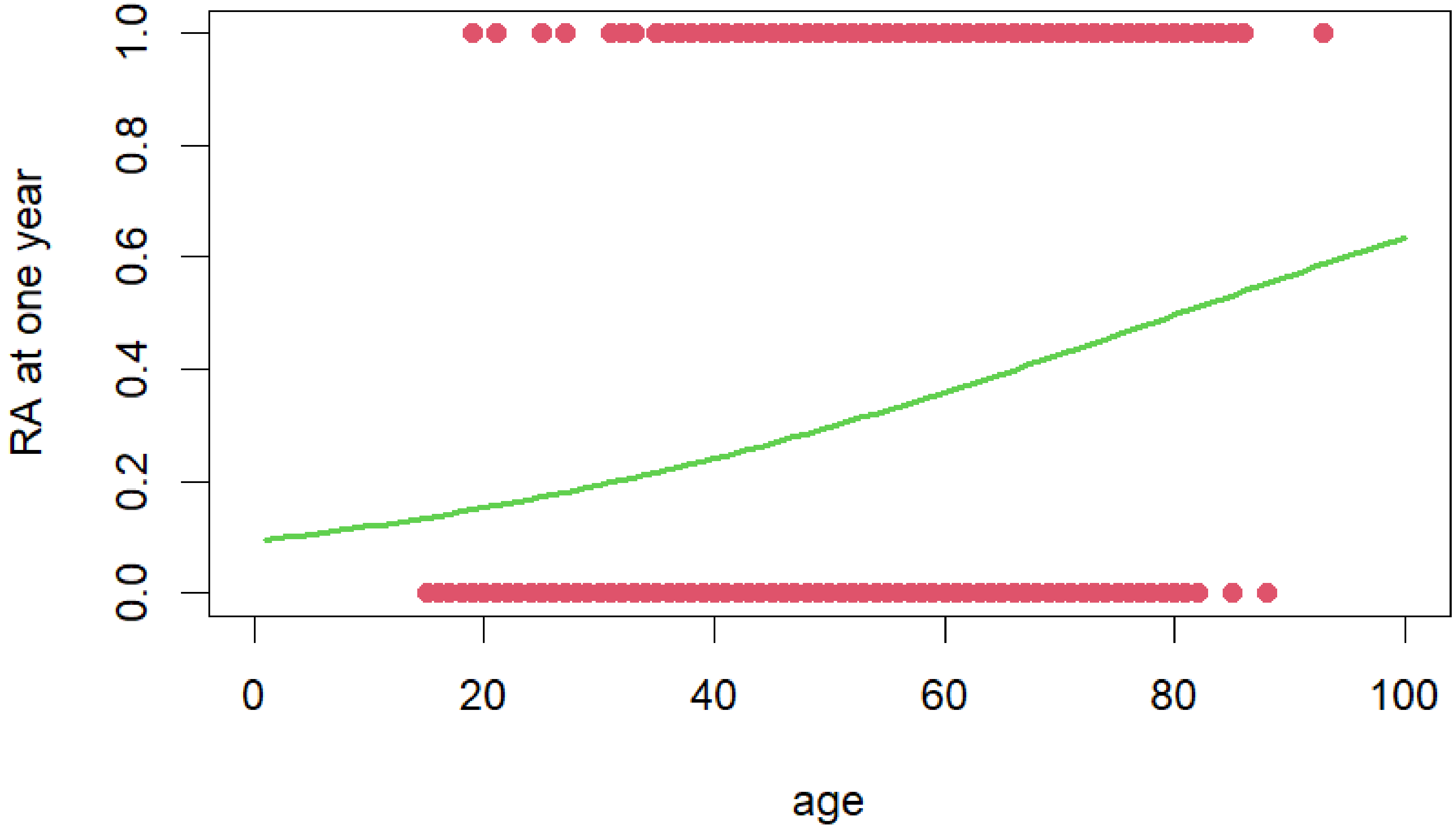
```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.281798   0.321459  -7.098 1.26e-12 ***
Age          0.028303   0.005717   4.951 7.39e-07 ***
```

Logistic model: $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1$

Estimated: $\widehat{\log\left(\frac{\pi}{1-\pi}\right)}$ = -2.281798 + 0.028303 age

$$\hat{\pi} = P(\text{RA withing one year}) = \frac{\exp(-2.281798 + 0.028303\, age)}{1+\exp(-2.281798 + 0.028303\, age)}$$

The logistic model



Interpretation of regression coefficients

additive

Interpretation coefficient for age, 0.028303:

- For two patients, with age difference of one year, the expected difference in log-odds, $\log\left(\frac{\pi}{1-\pi}\right)$ is 0.028303
- for a one year increase in age, the odds of RA becomes exp(0.028303) = 1.029 times larger.

Multiplicative

Logarithm converts multiplication and division to addition and subtraction. Exponentiation converts addition and subtraction back to multiplication and division.

- Interpretation intercept -2.281798
- For someone of age 0, the estimated log-odds is -2.281798 and $\hat{\pi} = \frac{\exp(-2.281798)}{1+\exp(-2.281798)}$ =0.09

Odds

INDEPENDENT BETTING powered by gambling.com

Betting > Football

England vs Netherlands predictions: Women’s Nations League tips and odds

Our tipster offers three betting predictions for England Women’s crunch Nations League encounter with Netherlands Women

Last Updated: 30th of November 2023
Joe Short • Football Writer

England vs Netherlands betting tips:

- England to win both halves – 13/8 with Unibet
- Beth Mead to score first – 9/2 with 10Bet
- Over 2.5 goals – 8/13 with BetVictor

Most Popular

CRICKET
Big Bash League 13 predictions: Cricket betting…

FOOTBALL

A binary risk factor: rheumafactor (RF)

	Y	
X	RA	No RA
Rheumafactor	84	56
No rheumafactor	93	336

Let π_1=P(Y=1|X=1) , $\hat{\pi}_1$ =

Let π_0=P(Y=1|X=0) , $\hat{\pi}_0$ =

Measures of association in 2 by 2 table:

- Risk difference π_1- π_0
- Risk ratio π_1/ π_0
- Odds ratio $(\pi_1/(1-\pi_1))/(\pi_0/(1-\pi_0))$

The odds ratio

$$\frac{\frac{p^*}{1-p^*}}{\frac{p}{1-p}}$$

OR = $\frac{\pi_1/(1-\pi_1)}{\pi_0/\,(1-\pi_0)}$

OR >1 : $\pi_1 > \pi_0$

OR <1 : $\pi_1 < \pi_0$

OR = 1: $\pi_1 = \pi_0$

- If the outcome Y=1 is rare (π_1 and $\pi_0$close to 0), we have:
- $\pi_1/(1-\pi_1) \approx \pi_1$
- $\pi_0/(1-\pi_0) \approx \pi_0$
- and the odds ratio ≈ π_1/π_0, the relative risk

Logistic regression with one binary predictor

```
> model.lr2 <- glm(ra1year~rfactor, family=binomial, data=RAdata)
> summary(model.lr2)

Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.2845     0.1172 -10.963  < 2e-16 ***
rfactor       1.6900     0.2085   8.104 5.33e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Exp(1.6900) = 5.419 → the oddsratio in the 2 by 2 table
- e^{β_1}=odds ratio

Logistic regression with multiple x-variables

- $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + .. + \beta_k x_k \;\; = x'\boldsymbol{\beta}$

- Or equivalent:

- $\pi = \frac{\exp(\beta_0+\beta_1 x_1+..+\beta_k x_k)}{1+\exp(\beta_0+\beta_1 x_1+..+\beta_k x_k)} = \frac{\exp(x'\boldsymbol{\beta})}{1+\exp(x'\boldsymbol{\beta})}$

Logistic regression with multiple x- variables

```
Call:
glm(formula = ra1year ~ rfactor + Age, family = binomial, data = RAdata)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.708391   0.353936  -7.652 1.98e-14 ***
rfactor      1.658540   0.212768   7.795 6.44e-15 ***
Age          0.027257   0.006126   4.450 8.61e-06 ***
```

$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2\, x_2$ = -2.282+1.659 rfactor+ 0.0278 age

Logistic regression with multiple x- variables

```
Call:
glm(formula = ra1year ~ rfactor + Age, family = binomial, data = RAdata)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.708391   0.353936  -7.652 1.98e-14 ***
rfactor      1.658540   0.212768   7.795 6.44e-15 ***
Age          0.027257   0.006126   4.450 8.61e-06 ***
```

- For two patients with the same age, the odds on RA is exp(1.658540)= 5.252 times larger for those with rheumafactor present, compared to those with rheumafactor absent.

- For two patients with the same value for rfactor, differing by 1 year in age, the odds on RA is exp(0.027257)=1.0276 times larger for the older patient.

Estimation of parameters β = ($\boldsymbol{\beta_0}$,… $\boldsymbol{\beta_p}$)’

Maximum likelihood:

- Choose values for β_i’s that are most likely to have produced the data.
- Suppose we have n observations (x_1, y_1) … (x_n, y_n)
- What is P(data | β) = $\mathrm{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n)$?

- Contribution of person *i* : $\mathrm{P}(Y_i = y_i)$
    - if y_i =1 : π_i
      if y_i =0 : (1- π_i)
- $\mathrm{P}(Y_i = y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$

The likelihood function

$$\pi_i = \frac{e^{x_i \vec{\beta}}}{1+e^{x_i \vec{\beta}}} = 1 - \frac{1}{1+e^{x_i \vec{\beta}}}$$

The likelihood function: $L(\beta) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$

- The larger the likelihood, the better the model fits the data.
- Select those values for b, for which L(β) attains its maximum.

- Easier to maximize the log of the likelihood
- $\log L(\beta) = \sum_{i=1}^{n} y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)$

- Log likelihood will take its maximum at the same values of b .

Maximize $\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log(\pi_i) + (1 - y_i)\log(1 - \pi_i)$

- Differentiate w.r.t. $\beta_0, \beta_1 \ldots, \beta_p$ and set the first derivatives equal to 0.

$y_i\left(1+\frac{1}{e^{x_i\beta}}\right)\left[\frac{1}{(1+e^{x_i\beta})^2}\right]\cdot x_i + (-y_i)\frac{-1}{1-\pi_i}\frac{1}{(1+e^{x_i\beta})^2}x_i$

- You can show that:

$$\begin{bmatrix} \frac{\partial l(\beta)}{\partial \beta_0} \\ \frac{\partial l(\beta)}{\partial \beta_1} \\ \vdots \\ \frac{\partial l(\beta)}{\partial \beta_k} \end{bmatrix} = \begin{bmatrix} \sum_i (y_i - \pi_i) \\ \sum_i x_{i1}(y_i - \pi_i) \\ \vdots \\ \sum_i x_{ik}(y_i - \pi_i) \end{bmatrix} = X'(Y - \pi)$$

- So we need to solve : $X'(Y - \pi)$ =0
- In general, this can not be solved exactly , numerical techniques are needed.

Standard errors of $\hat{\beta}$

- Calculate the second derivatives of the log likelihood:

- $\frac{\partial^2 \log L(\beta)}{\partial\beta\partial\beta\prime} = -X'V\,X$ with V $= \begin{pmatrix} \pi_1(1-\pi_1) & 0 & \cdots & 0 \\ 0 & \pi_2(1-\pi_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & & \pi_n(1-\pi_n) \end{pmatrix}$

- Asymptotic variance-covariance $(k+1)\times(k+1))$ matrix of MLE is then:

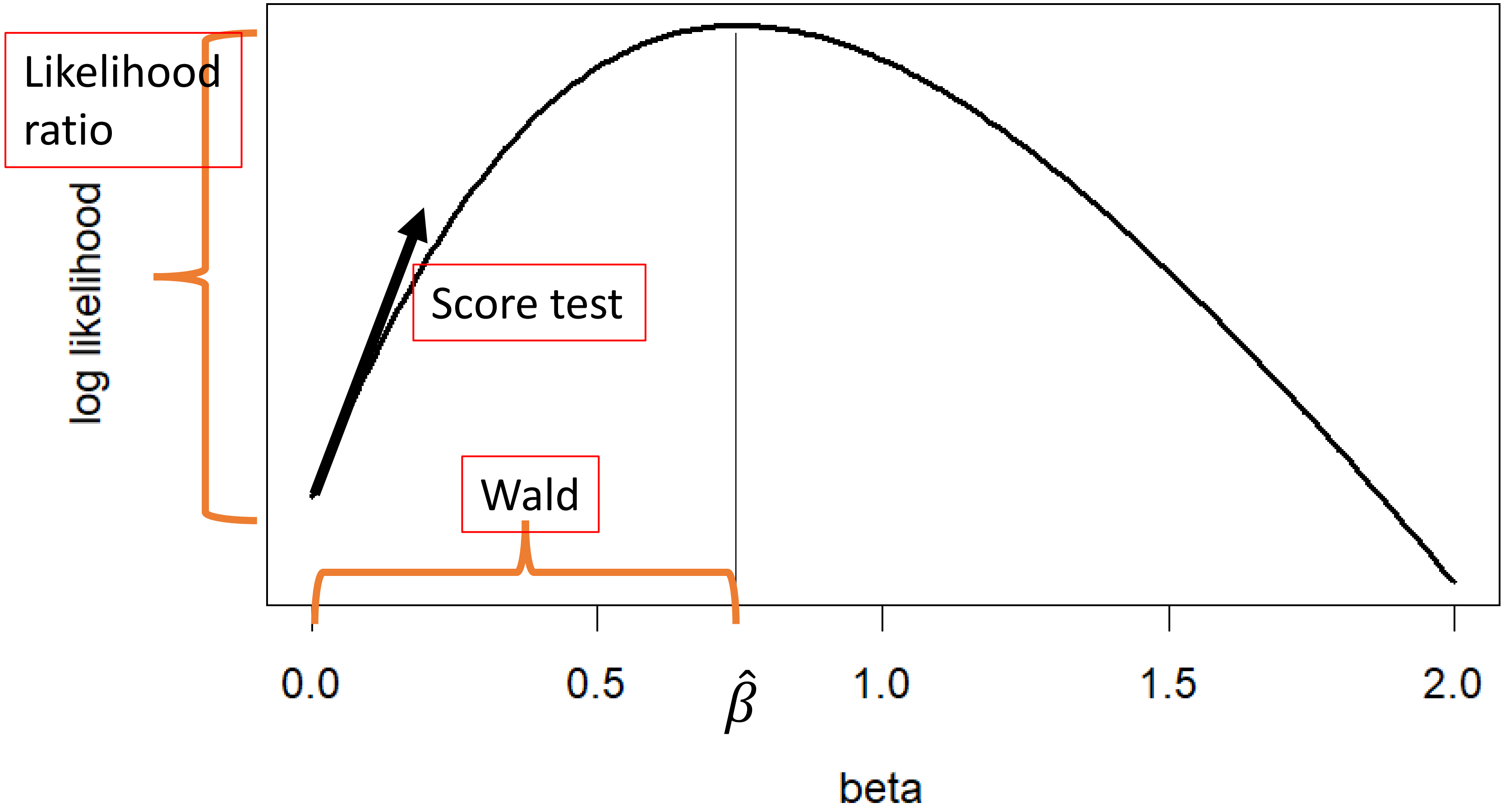
•$V(\hat{\beta}) = \left\{-E\left[\frac{\partial^2 \log L(\boldsymbol{\beta})}{\partial\beta\partial\boldsymbol{\beta}'}\right]\right\}^{-1}$ $=(X'V\,X)^{-1}$

Hypothesis testing for a regression coefficient

- H_0: β_1=0

Three different methods:

- Wald test
- Likelihood ratio test
- Score test



Method 1: Wald test

- Consider Z = $\hat{\beta}_1/se(\hat{\beta}_1)$. Under H_0 Z follows approximately a standard normal distribution. Reject H_0 if | Z | >z_α
- Sometimes Z^2 instead of Z is used (compare Z^2 to a $\chi_{(1)}^2$ distribution)

- Example:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.281798   0.321459  -7.098 1.26e-12 ***
Age          0.028303   0.005717   4.951 7.39e-07 ***
```

Method 2: likelihood ratio test

- The (log) likelihood measures how well the model fits the data.

- Compare $\log L(\beta_0,0)$, the log-likelihood of model with β_1=0: with $\log L(\beta_0, \beta_1)$, the log-likelihood of complete model:

- Under H_0 : 2($\log L(\beta_0, \beta_1)$ - $\log L(\beta_0,0)$) has (approximately) a $\chi_{(1)}^2$ distribution.

Likelihood ratio tests

内嵌的

- Can be used more generally to compare nested models
- Can sometimes be used as goodness of fit test (to compare a model to a perfectly fitted model)
- More tomorrow

Method 3: score test

- Use only the null model and consider how steep the likelihood function is in $\beta = 0$ is.

- The three methods are asymptotically equivalent; for small numbers the likelihood ratio method is usually better.

Confidence intervals

Wald method:

- 100(1-α) % confidence interval for β_i : $(\hat{\beta}_i - z_{\alpha/2}\,se(\hat{\beta}_i)\ ,\ \hat{\beta}_i + z_{\alpha/2}\,se(\hat{\beta}_i))$
- Confidence interval for odds ratio e^{β_i}: $(\ e^{\text{lower limit}}\ ,\ e^{\text{upper limit}})$

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.281798   0.321459  -7.098 1.26e-12 ***
Age          0.028303   0.005717   4.951 7.39e-07 ***
```

95% CI for coefficient for age:

95% CI for odds ratio for age:

Confidence intervals-Wald method

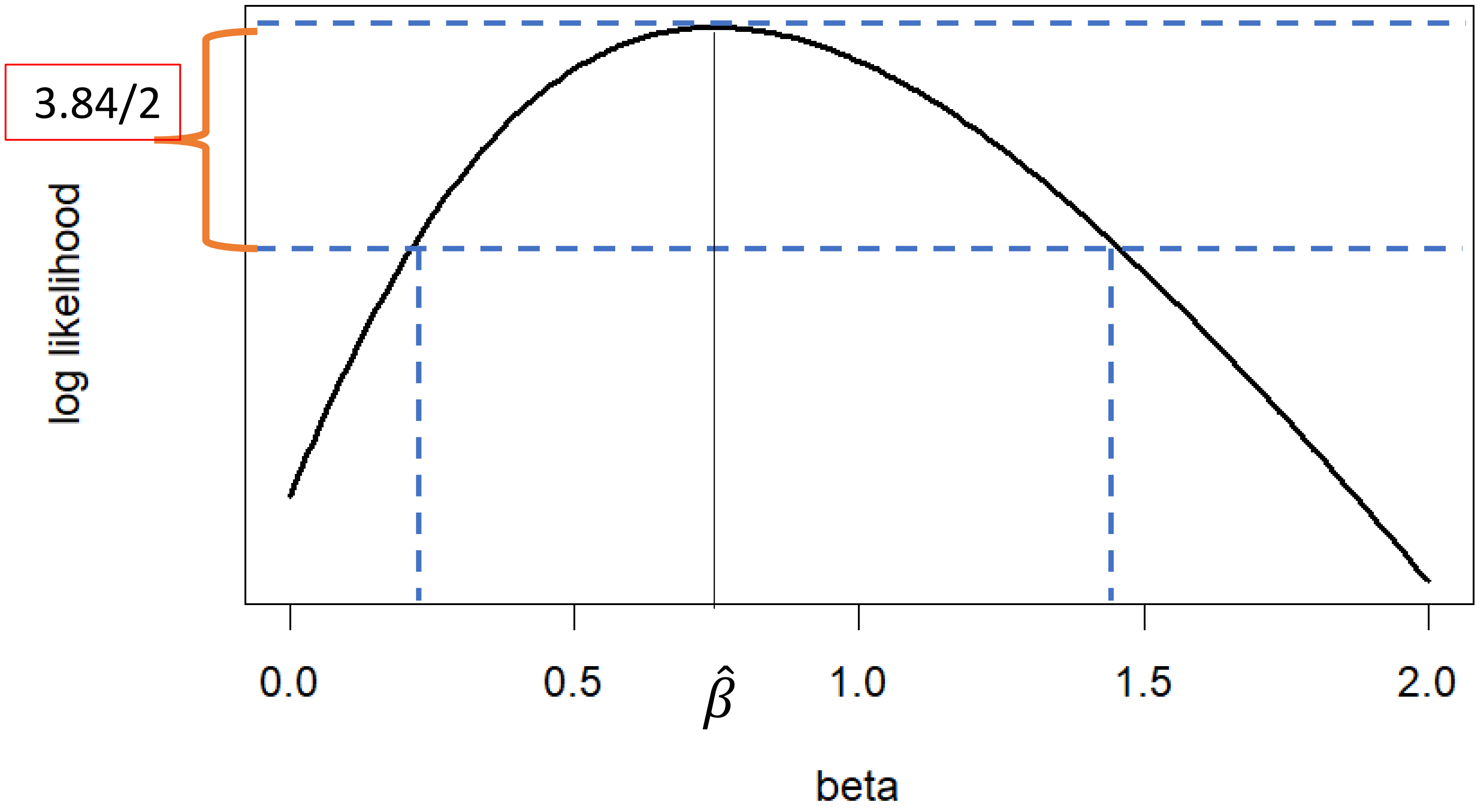
```
cbind(exp(coefficients(model.lr)), exp(confint.default(model.lr)))
                           2.5 %    97.5 %
(Intercept) 0.1021004 0.05437525 0.1917141
Age         1.0287074 1.01724528 1.0402987
```

Conclusion on relation between age and RA after 1 year based on the 95% CI?

Confidence intervals based on the (profile) likelihood

- All values of β_i for which the 2(profile) log likelihood is less then the critical value of the $\chi_{(1)}^2$ distribution.

- For $\alpha = 0.05$, this value is 3.84 ($=1.96^2$)



Profile likelihood confidence intervals

```
cbind(exp(coefficients(model.lr)), exp(confint(model.lr)))
Waiting for profiling to be done...
                          2.5 %    97.5 %
(Intercept) 0.1021004 0.05354001 0.1891023
Age         1.0287074 1.01739913 1.0404893
```