

EXERCISE 2

$n = 100$

$CHD_i = \begin{cases} 1 & \text{if individual } i \text{ has heart disease} \\ 0 & \text{otherwise} \end{cases}$

(a1) The response variable $CHD_i \in \{0, 1\}$ is binary, so I can fit a logistic regression (GLM for Bernoulli data based on the logit link)

- $CHD_i \sim \text{Bernoulli}(\pi_i)$ independent for $i = 1, \dots, 100$
 - linear predictor $\eta_i = \tilde{X}_i^T \beta = \beta_1 + \beta_2 AGE_i$
 - $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$ logit link function
- $$\Rightarrow \text{logit}(\pi_i) = \log \frac{\pi_i}{1-\pi_i} = \beta_1 + \beta_2 AGE_i$$

(a2) interpretation of AGE coefficient β_2

Consider two individuals i with covariate AGE_i

j with covariate $AGE_j = AGE_i + 1$

$$\text{logit}(\pi_i) = \beta_1 + \beta_2 AGE_i$$

$$\begin{aligned} \text{logit}(\pi_j) &= \beta_1 + \beta_2 AGE_j = \beta_1 + \beta_2 (AGE_i + 1) \\ &= \beta_1 + \beta_2 AGE_i + \beta_2 \end{aligned}$$

$$\begin{aligned} \beta_2 &= \text{logit}(\pi_j) - \text{logit}(\pi_i) \\ &= \log \frac{\pi_j}{1-\pi_j} - \log \frac{\pi_i}{1-\pi_i} \end{aligned}$$

odds for individual j odds for individual i

If I increase the age of 1 year, the log-odds of having a heart disease increase of 0.1109.

Equivalently

$$\begin{aligned} \beta_2 &= \log \frac{\frac{\pi_j}{1-\pi_j}}{\frac{\pi_i}{1-\pi_i}} \Rightarrow e^{\beta_2} = \frac{\frac{\pi_j}{1-\pi_j}}{\frac{\pi_i}{1-\pi_i}} \\ &\Rightarrow \frac{\pi_i}{1-\pi_i} \cdot e^{\beta_2} = \frac{\pi_j}{1-\pi_j} \end{aligned}$$

The odds change of a multiplicative factor $e^{0.1109}$ if I increase the age of 1 year

$$(a3) \begin{cases} H_0: \beta_2 = 0 \\ H_1: \beta_2 \neq 0 \end{cases}$$

the test is based on the statistic $z = \frac{\hat{\beta}_2}{\text{se}(\hat{\beta}_2)} = 4.61$ (in the table)

since the p-value is ~ 0 , I reject H_0 .

Hence, $\beta_2 \neq 0$, which means that age affects the probability of having a heart disease.

(b1) $CHD_i \sim \text{Bernoulli}(\pi_i)$ independent for $i = 1, \dots, 100$

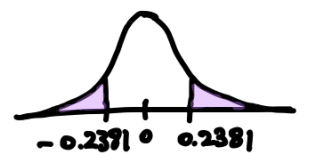
- linear predictor $\eta_i = \tilde{X}_i^T \beta = \beta_1 + \beta_2 AGE_i + \beta_3 AGE_i^2$
 - $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$ logit link function
- $$\Rightarrow \text{logit}(\pi_i) = \log \frac{\pi_i}{1-\pi_i} = \beta_1 + \beta_2 AGE_i + \beta_3 AGE_i^2$$

(b2) estimate of the intercept: $z_1 = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)} \Rightarrow \hat{\beta}_1 = z_1 \cdot \text{se}(\hat{\beta}_1) = -0.93 \cdot 4.2901 = -4.2472$

estimate of the age coeff.: $z_2 = \frac{\hat{\beta}_2}{\text{se}(\hat{\beta}_2)} \Rightarrow \hat{\beta}_2 = z_2 \cdot \text{se}(\hat{\beta}_2) = 0.315 \cdot 0.1947 = 0.0613$

p-value of age²: $z_3 = \frac{\hat{\beta}_3}{\text{se}(\hat{\beta}_3)} = 0.2381$

$$\begin{aligned} p\text{-value} &= P_{H_0}(|z_3| \geq |z_3^{\text{obs}}|) = 2 \cdot P_{H_0}(z_3 \geq |z_3^{\text{obs}}|) \\ &= 2 \cdot \underbrace{(1 - \Phi(0.2391))}_{> 0.10} > 0.20 \end{aligned}$$



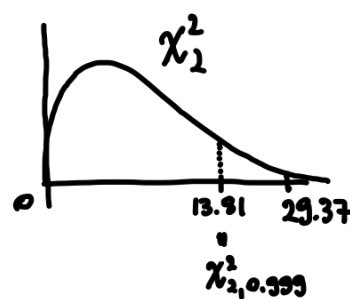
$$(b3) \begin{cases} H_0: \beta_2 = \beta_3 = 0 \\ H_1: \text{at least one of } \beta_2, \beta_3 \text{ is } \neq 0 \end{cases}$$

likelihood ratio test

$$W = 2 \log \frac{\hat{L}(\text{model})}{\hat{L}(\text{null model})} = 2 \{ \hat{E}(\text{model}) - \hat{E}(\text{null model}) \} \sim \chi_2^2 \text{ under } H_0$$

$$w^{\text{obs}} = D(\text{null}) - D(\text{model}) = 136.66 - 107.29 = 29.37$$

$$P_{H_0}(W \geq w^{\text{obs}}) < 0.001 \quad | \text{ reject } H_0$$



$$(b4) \begin{cases} H_0: \beta_3 = 0 \\ H_1: \beta_3 \neq 0 \end{cases}$$

We already performed the test in (b2). We do not reject H_0 .

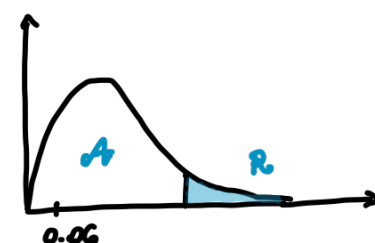
Hence we prefer model (a).

We can equivalently use the deviance to compare nested models

$$W = 2 \{ \hat{E}(\text{model b}) - \hat{E}(\text{model a}) \} \sim \chi_1^2$$

$$\begin{aligned} w^{\text{obs}} &= 2 \{ \hat{E}(\text{saturated}) - \hat{E}(\text{model a}) - \hat{E}(\text{saturated}) + \hat{E}(\text{model b}) \} \\ &= D(\text{model a}) - D(\text{model b}) = 104.35 - 104.29 = 0.06 \end{aligned}$$

I do not reject H_0



(c1) the new variable $z_i = AGE_i < 50$ is $= \begin{cases} 1 & \text{if } AGE_i < 50 \\ 0 & \text{if } AGE_i \geq 50 \end{cases}$

$CHD_i \sim \text{Bernoulli}(\pi_i)$ independent for $i = 1, \dots, 100$

- linear predictor $\eta_i = \tilde{X}_i^T \beta = \beta_1 + \beta_2 z_i$
 - $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$ logit link function
- $$\Rightarrow \text{logit}(\pi_i) = \log \frac{\pi_i}{1-\pi_i} = \beta_1 + \beta_2 z_i$$

(c2) I consider individual i with $AGE_i \geq 50$ ($\Rightarrow z_i = 0$)

individual j with $AGE_j < 50$ ($\Rightarrow z_j = 1$)

$$\begin{aligned} \log \frac{\pi_i}{1-\pi_i} &= \beta_1 \\ \log \frac{\pi_j}{1-\pi_j} &= \beta_1 + \beta_2 \end{aligned} \quad \Rightarrow \quad \beta_2 = \log \frac{\pi_j}{1-\pi_j} - \log \frac{\pi_i}{1-\pi_i}$$

$$= \log \left\{ \frac{\frac{\pi_j}{1-\pi_j}}{\frac{\pi_i}{1-\pi_i}} \right\} = \log \left\{ \frac{\frac{P(Y_i = 1 | AGE_i < 50)}{P(Y_i = 0 | AGE_i < 50)}}{\frac{P(Y_i = 1 | AGE_i \geq 50)}{P(Y_i = 0 | AGE_i \geq 50)}} \right\} = -2.0969$$

The odds of having a coronary heart disease of an individual older than 50 y.o. are multiplied by -2.0969 to obtain the odds of an individual younger than 50 y.o.