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)

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

(a2) interpretation of AGE coefficient β_2

Consider two individuals i with covariate AGE_i

j with covariate $AGE_j = AGE_i + 1$

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

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

$$\beta_2 = \text{logit}(\pi_j) - \text{logit}(\pi_i) = \underbrace{\log \frac{\pi_j}{1-\pi_j}}_{\text{odds for individual } j} - \underbrace{\log \frac{\pi_i}{1-\pi_i}}_{\text{odds for individual } i}$$

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

Equivalently

$$\beta_2 = \log \frac{\frac{\pi_j}{1-\pi_j}}{\frac{\pi_i}{1-\pi_i}} \Rightarrow e^{\beta_2} = \frac{\frac{\pi_j}{1-\pi_j}}{\frac{\pi_i}{1-\pi_i}}$$

$$\Rightarrow \frac{\pi_i}{1-\pi_i} \cdot e^{\beta_2} = \frac{\pi_j}{1-\pi_j}$$

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 $z = \frac{\hat{\beta}_2}{\hat{se}(\hat{\beta}_2)} = 4.61$ (in the table)

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

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

(b1)
$$\left.\begin{array}{l} \bullet\ CHD_i \sim \text{Bernoulli}(\pi_i) \quad \text{independent for } i = 1, \dots, 100 \\ \bullet\ \text{linear predictor } \eta_i = \tilde{x}_i^T \beta = \beta_1 + \beta_2 AGE_i + \beta_3 AGE_i^2 \\ \bullet\ \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i \quad \text{logit link function} \end{array}\right\} \Rightarrow \text{logit}(\pi_i) = \log \frac{\pi_i}{1-\pi_i} = \beta_1 + \beta_2 AGE_i + \beta_3 AGE_i^2$$

(b2)
- estimate of the intercept: $z_1^{obs} = \frac{\hat{\beta}_1}{\hat{se}(\hat{\beta}_1)} \Rightarrow \hat{\beta}_1 = z_1^{obs} \cdot \hat{se}(\hat{\beta}_1) = -0.99 \cdot 4.2901 = -4.2472$
- estimate of the age coeff.: $z_2^{obs} = \frac{\hat{\beta}_2}{\hat{se}(\hat{\beta}_2)} \Rightarrow \hat{\beta}_2 = z_2^{obs} \cdot \hat{se}(\hat{\beta}_2) = 0.315 \cdot 0.1947 = 0.0613$
- z-value of age²: $z_3^{obs} = \frac{\hat{\beta}_3}{\hat{se}(\hat{\beta}_3)} = 0.2381$

$$\text{pvalue} = \mathbb{P}_{H_0}(|z_3| \geq |z_3^{obs}|) = 2 \cdot \mathbb{P}_{H_0}(z_3 \geq |z_3^{obs}|) = 2 \underbrace{\left(1 - \Phi(0.2391)\right)}_{>0.10} > 0.20$$

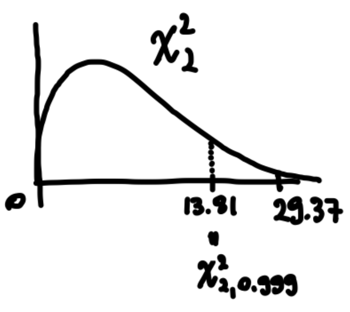
(b3) $\begin{cases} H_0: \beta_2 = \beta_3 = 0 \\ H_1: \text{at least one of } \beta_2, \beta_3 \text{ is } \neq 0 \end{cases}$

likelihood ratio test

$$W = 2 \log \frac{\hat{L}(\text{model})}{\tilde{L}(\text{null model})} = 2 \{ \hat{\ell}(\text{model}) - \tilde{\ell}(\text{null model}) \} \dot\sim \chi^2_2 \quad \text{under } H_0$$

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

$\mathbb{P}_{H_0}(W \geq w^{obs}) < 0.001$ I 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).

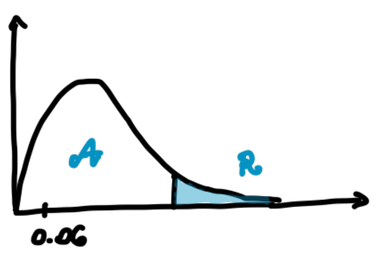
We can equivalently use the deviance to compare nested models

$$W = 2 \{ \hat{\ell}(\text{model } b) - \hat{\ell}(\text{model } a) \} \dot\sim \chi^2_1$$

$$w^{obs} = 2 \{ \tilde{\ell}(\text{saturated}) - \hat{\ell}(\text{model } a) - \tilde{\ell}(\text{saturated}) + \hat{\ell}(\text{model } b) \}$$

$$= D(\text{model } a) - D(\text{model } b) = 107.35 - 107.29 = 0.06$$

I do not reject H_0



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

$$\left.\begin{array}{l} \bullet\ CHD_i \sim \text{Bernoulli}(\pi_i) \quad \text{independent for } i = 1, \dots, 100 \\ \bullet\ \text{linear predictor } \eta_i = \tilde{x}_i^T \beta = \beta_1 + \beta_2 z_i \\ \bullet\ \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i \quad \text{logit link function} \end{array}\right\} \Rightarrow \text{logit}(\pi_i) = \log \frac{\pi_i}{1-\pi_i} = \beta_1 + \beta_2 z_i$$

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

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

$$\left.\begin{array}{l} \log \frac{\pi_i}{1-\pi_i} = \beta_1 \\ \log \frac{\pi_j}{1-\pi_j} = \beta_1 + \beta_2 \end{array}\right\} \Rightarrow \beta_2 = \log \frac{\pi_j}{1-\pi_j} - \log \frac{\pi_i}{1-\pi_i}$$

$$= \log \left\{ \frac{\frac{\pi_j}{1-\pi_j}}{\frac{\pi_i}{1-\pi_i}} \right\} = \log \left\{ \frac{\frac{\mathbb{P}(Y_i = 1 \mid AGE_i < 50)}{\mathbb{P}(Y_i = 0 \mid AGE_i < 50)}}{\frac{\mathbb{P}(Y_i = 1 \mid AGE_i \geq 50)}{\mathbb{P}(Y_i = 0 \mid AGE_i \geq 50)}} \right\} = -2.0989$$

The odds of having a coronary heart disease of an individual older than 50 y.o. are multiplied by -2.0989 to obtain the odds of an individual younger than 50 y.o.