EXERCISE 2

y_i = diabetes$_i$ = $\begin{cases} 1 & \text{if woman } i \text{ has diabetes} \\ 0 & \text{if woman } i \text{ does not have diabetes} \end{cases}$

$i = 1, \dots, n$ with $n = 724$

Moreover $\sum_{i=1}^{n} y_i = 249$ $\quad\quad$ $\sum_{i=1}^{n} (1 - y_i) = 475$

a)
- $Y_i \sim$ Bernoulli (π_i) $\quad$ $i = 1, \dots, n$ independent $\quad\quad$ (distribution)
- $\eta_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}$

  with x_{i1} = pregnant$_i$ $\quad$ (dummy) $\quad\quad$ (linear prediction)
  x_{i3} = glucose$_i$
  x_{i4} = pressure$_i$
  x_{i5} = BMI$_i$
  x_{i6} = age$_i$
- $g(\pi_i)$ = logit(π_i) = $\log \frac{\pi_i}{1-\pi_i} = \eta_i$ $\quad\quad$ (link function)

b)
- Interpretation of the "age" coefficient (β_6)

We study the odds of having diabetes for two women i and k with $\underbrace{age_k}_{x_{k6}} = \underbrace{age_i}_{x_{i6}} + 1$ and all other covariates equal.

odds woman i: $\frac{\pi_i}{1-\pi_i} = \frac{e^{x_i^T\beta}}{1 + e^{x_i^T\beta}} \cdot \left(1 + e^{x_i^T\beta}\right) = e^{x_i^T\beta}$

$= \exp\{\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}\}$

odds woman k: $\frac{\pi_k}{1-\pi_k} = e^{x_k^T\beta} = \exp\{\beta_1 + \beta_2 \underbrace{x_{k2}}_{x_{i2}} + \beta_3 \underbrace{x_{k3}}_{x_{i3}} + \beta_4 \underbrace{x_{k4}}_{x_{i4}} + \beta_5 \underbrace{x_{k5}}_{x_{i5}} + \beta_6 \underbrace{x_{k6}}_{x_{i6}+1}\}$

if we look at the ODDS RATIO

$\frac{\frac{\pi_k}{1-\pi_k}}{\frac{\pi_i}{1-\pi_i}} = e^{\beta_6} \Rightarrow \frac{\pi_k}{1-\pi_k} = e^{\beta_6} \cdot \frac{\pi_i}{1-\pi_i}$ $\quad$ and $e^{\hat{\beta}_6} = 1.033 > 1$

$\frac{\pi_k}{1-\pi_k}$: odds with age increased by 1

The odds of having diabetes increase by a multiplicative factor 1.033 for every additional year, keeping all other covariates fixed.

- Interpretation of the "pregnant" coefficient (β_2)

We study the odds of two women i and k, where woman i had no pregnancies, and woman k had at least one pregnancy, and all other covariates are equal.

odds woman i: $\frac{\pi_i}{1-\pi_i} = \frac{e^{x_i^T\beta}}{1 + e^{x_i^T\beta}} \cdot \left(1 + e^{x_i^T\beta}\right) = e^{x_i^T\beta}$

$= \exp\{\beta_1 + \beta_2 \overbrace{x_{i2}}^{0} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}\}$

odds woman k: $\frac{\pi_k}{1-\pi_k} = e^{x_k^T\beta} = \exp\{\beta_1 + \beta_2 \underbrace{x_{k2}}_{1} + \beta_3 \underbrace{x_{k3}}_{x_{i3}} + \beta_4 \underbrace{x_{k4}}_{x_{i4}} + \beta_5 \underbrace{x_{k5}}_{x_{i5}} + \beta_6 \underbrace{x_{k6}}_{x_{i6}}\}$

if we look at the ODDS RATIO

$\frac{\frac{\pi_k}{1-\pi_k}}{\frac{\pi_i}{1-\pi_i}} = e^{\beta_2} \Rightarrow \frac{\pi_k}{1-\pi_k} = e^{\beta_2} \cdot \frac{\pi_i}{1-\pi_i}$ $\quad$ and $e^{\hat{\beta}_2} = 1.279 > 1$

$\frac{\pi_k}{1-\pi_k}$: odds with at least 1 pregnancy

The odds of having diabetes increase by a multiplicative factor 1.279 when we consider a woman who had at least one pregnancy, compared to a woman with no pregnancies and all other covariates equal.

c) If we look at the p-value in the table, it is 0.40.

It is the p-value α^{obs} of the test for testing

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

Hence $0.40 = P_{H_0}(|Z| > |Z^{obs}|)$ $\quad$ probability of observing "more extreme" values.

Since $\alpha^{obs} > \alpha$ for all usual α (0.01, 0.05, 0.1), we do not reject H_0.

That is, β_2 is not significant and we can remove the corresponding covariate.

d) We want to compare model A and model B.

First of all, notice that model B is nested w.r.t. model A and can be obtained by constraining $\beta_2 = \beta_3 = 0$.

Hence, we want to test

$\begin{cases} H_0: \beta_2 = \beta_3 = 0 \\ H_1: \exists\, r \in \{2,3\} : \beta_r \neq 0 \end{cases}$

We use the likelihood ratio test

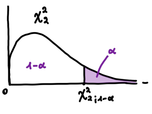
$W = 2\left(\hat{\ell}(\text{model A}) - \hat{\ell}(\text{model B})\right) \overset{H_0}{\sim} \chi^2_{p-p_0}$ $\quad$ Here $p = 6$, $p_0 = 4$

$\hat{\ell}(\text{model A})$: max of the log-likelihood in the FULL model
$\hat{\ell}(\text{model B})$: max of the log-likelihood in the RESTRICTED model

observed value of the test

$w^{obs} = 2\left(\hat{\ell}(\text{model A}) - \hat{\ell}(\text{model B})\right)$
$= 2\left(\hat{\ell}(\text{model A}) - \hat{\ell}(\text{model B}) + \tilde{\ell}(\text{saturated}) - \tilde{\ell}(\text{saturated})\right)$
$= 2\left[\left(\tilde{\ell}(\text{saturated}) - \hat{\ell}(\text{model B})\right) - \left(\tilde{\ell}(\text{saturated}) - \hat{\ell}(\text{model A})\right)\right]$
$= D(\text{model B}) - D(\text{model A})$
$= 636.15 - 634.45 = 1.70$

Reject region is for large values of the test using a fixed significance level α



$R_1 = (\chi^2_{2;1-\alpha}; +\infty)$

if $\alpha = 0.05$ $\quad R_1 = (\chi^2_{2;0.95}; +\infty) = (5.99; +\infty)$

$w^{obs} \notin R_1$: do not reject H_0
I prefer model B

e) NULL model

is the model with only the intercept: it assumes that none of the covariates has an effect.

$\begin{cases} Y_i \sim \text{Bern}(\pi_i) \\ \eta_i = \beta_1 \\ \text{logit}\,\pi_i = \eta_i \end{cases}$

π_i does not depend on i: it is constant

$\pi_i = \pi = \frac{e^{\beta_1}}{1 + e^{\beta_1}}$ $\quad$ for all $i = 1, \dots, n$

In the null model, the MLE of π is $\hat{\pi} = \bar{y} = \frac{249}{724} = 0.344$

Thus, $\hat{\beta}_1 = \log \frac{\bar{y}}{1-\bar{y}} = -0.646$