Example: Logistic Regression (Pneumonia Data)

A study carried out on several cohorts of coal miners sought to determine the relationship between duration of coal dust exposure (measured in years) and incidence of severe pneumonia. We use logistic regression to analyze these data.

(a) What is the overall probability of severe pneumonia?

n_i : number of miners

y_i : number of Pneumonia incidence

x_i : duration (year)

overall probability: $\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} n_i} = 0.121$

(b) Write down a plausible logistic regression model based on the logit link.

Logistic regression model:

$$logit(\pi_i) = \beta_0 + \beta_1 x_i$$

$$y_i \sim Binomial(n_i, \pi_i)$$

(c) Fit the model listed in (b) using PROC LOGISTIC.

$\widehat{\beta}_0 = -4.5383, \quad \exp\{\widehat{\beta}_0\} = 0.0107$

$\widehat{\beta}_1 = 0.0869, \quad \exp\{\widehat{\beta}_1\} = 1.091$

See the SAS code

(d) Interpret $\widehat{\beta}_1$ and $\exp\{\widehat{\beta}_1\}$

$\widehat{\beta}_1$: expected increase in logit risk (log odds) of Pneumonia for a one-unit increase in exposure.

$\exp\{\widehat{\beta}_1\}$: expected odds ratio of Pneumonia for a one-unit increase in exposure.

(e) Should $\exp\{\beta_1\}$ be an accurate approximation to the relative risk?

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

It can be shown that

$$OR = RR\frac{1-\pi_0}{1-RR\pi_0}. \tag{1}$$

From (1)

$$RR = \frac{OR}{1-\pi_0+OR\pi_0} \tag{2}$$

The above equation shows that OR is close to RR when π_0 is small or OR is close to 1. Since the OR estimate ($\exp\{\widehat{\beta}_1\}$) is 1.09 and the overall risk is small, we can conclude that the OR estimate is an accurate approximation to RR.

(f) Interpret $\widehat{\beta}_0$ and $\exp\{\widehat{\beta}_0\}$

$\widehat{\beta}_0$: expected logit risk (log odds) of Pneumonia for a subject with zero exposure.

$\exp\{\widehat{\beta}_0\}$: expected odds of Pneumonia for a subject with zero exposure.

(g) Test $H_0 : \beta_1 = 0$ using the Wald statistic.

$$X_w = \frac{\widehat{\beta}_1^2}{\widehat{SE}(\widehat{\beta}_1)^2} = 34.9410 >> 3.84 = \chi^2_{1,0.95}$$

Reject the null hypothesis.

(h) Repeat the hypothesis test, but this time using the LRT.

Full model: $logit(\pi_i) = \beta_0 + \beta_1 x_i$

Reduced model: $logit(\pi_i) = \beta_0$

$X_L = -2\{l(\widehat{\beta}^0) - l(\widehat{\beta})\} = 46.554 >> 3.84 = \chi^2_{1,0.95}$

Reject the null hypothesis.

(i) Again, using the score test.

$$X_S = U(\widehat{\beta}^0)' I^{-1}(\widehat{\beta}^0) U(\widehat{\beta}^0) = 44.13 >> 3.84 = \chi^2_{1,0.95}$$

(j) For the score test just computed, what does $\widehat{\boldsymbol{\beta}}^0$ equal?

$$\widehat{\beta}^0 = (\widehat{\beta}^0_0, 0)'$$

$$\widehat{\beta}^0_0 = logit(\widehat{\pi}_0) = logit(0.121) = -1.983$$

(k) Verify your calculation in (j) using PROC LOGISTIC.

See the SAS code

(l) Calculate pseudo R^2 and generalized R^2.

Minimal model: $logit(\pi_i) = \beta_0 \Rightarrow \quad \widehat{\pi}^{intercept}$

Fitted model: $logit(\pi_i) = \beta_0 + \beta_1 x_i \Rightarrow \quad \widehat{\pi}$

- Pseudo R^2 (Cox & Snell)

$$R^2 = 1 - \left\{ \frac{L(\widehat{\pi}^{intercept})}{L(\widehat{\pi})} \right\}^{2/N}$$

(N is the number of subjects. So in our data, $N = 371$)

- From -2 Log L in SAS output

$$-2l_0 = 274.165; -2l_1 = 227.611$$

$L_0 = exp(l_0) = 2.92e{-}60; L_1 = exp(l_1) = 3.76e{-}50$

$$R^2 = 1 - \left\{ \frac{2.92e - 60}{3.76e - 50} \right\}^{2/371} = 0.1179$$

You can obtain generalized (Cox & Snell) and max-adjusted (NagelKerke) R^2 in SAS by using RSQ option in the model statement. See the SAS code.

- Maximum of the Cox & Snell R^2 can be smaller than 1. So NagelKerke proposed a max adjusted Cox & Snell R^2. From the SAS output (with RSQ option):

$$\text{max-adjusted } R^2 = 0.6478$$

(m) Write out the model equation for a saturated logistic model; include an intercept.

$$logit(\pi_i) = \beta_0 + \beta_1 I(x_i = 15.0) + \cdots + \beta_7 I(x_i = 51.5)$$

(n) Interpret the parameters from the saturated model.

β_0: log odds when $x_i = 5.8$

β_1: log odds ratio between $x_i = 15.0$ vs $x_i = 5.8$

$\ldots$

(o) Write out another saturated model, this time *not including* an intercept.

$$logit(\pi_i) = \beta_1 I(x_i = 5.8) + \cdots + \beta_8 I(x_i = 51.5)$$

(p) Interpret the parameters from this version of the saturated model.

β_1: log odds when $x_i = 5.8$

β_2: log odds when $x_i = 15.0$

$\ldots$

(q) Fit the no-intercept saturated model without using PROC LOGISTIC or GENMOD.

$\widehat{\beta}_i = logit(\widehat{\pi}_i)$, where $\widehat{\pi}_i = y_i/n_i$

See the SAS code

(r) Fit the no-intercept saturated model using PROC LOGISTIC. Compare the parameter estimates with your data-step calculations.

See the SAS code

(s) Fit the saturated model with an intercept using PROC LOGISTIC.

- See the SAS code

(t) Plot the fitted logits from the saturated and linear logistic model. Does the linear model appear to fit the data based on this plot?

See the SAS code

- To compare these two models, LRT can be used.
- H_0: linear logistic model fits data well

- Full model ($q = 8$):

  $$logit(\pi_i) = \beta_0 + \beta_1 I(x_i = 15.0) + \cdots + \beta_7 I(x_i = 51.5)$$

- Reduced model ($q = 2$):

  $$logit(\pi_i) = \beta_0 + \beta_1 x_i$$

- LRT test statistic:

  $$X_L = 2\{-112.22 + 113.81\} = 3.18 < 12.59 = \chi^2_{6,0.95}$$

- At $\alpha = 0.05$, we cannot reject H$_0$. There is no enough evidence that the saturated model fits better than the linear logistic model.

(u) This time, plot the fitted probabilities.

See the SAS code

(v) Change the group level data format to the individual level data format and fit the linear logistic model. Are there any difference?

- Parameter estimates and log likelihood are the same. Full log likelihood are different.

- In the individual level format, deviance and Pearson χ^2 GOF statistics are not presented.