L 11. Logistic Regression

1. Model setup

- Suppose Y_i are independent r.v.s with <u>Binary outcomes</u>.

then Y_i has Bernoulli distribution $b(1, \pi_i)$, where π_i

is the unknown parameter value:

Y_i	Probability
0	$1-\pi_i$
1	π_i

$$\boxed{\begin{aligned} E(Y_i) &= \pi_i \\ Var(Y_i) &= \pi_i(1-\pi_i) \end{aligned}}$$

- Consider the simple regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad Y_i = 0, 1$$

Or mean response regression model

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

or

$$\boxed{\pi_i = \beta_0 + \beta_1 X_i} \leftarrow \text{Probability linear model}$$

- Special Problems when response is binary

(i) Error term ε_i is also binary:

when $Y_i = 1$: $\varepsilon_i = 1 - \beta_0 - \beta_1 X_i$

when $Y_i = 0$: $\varepsilon_i = -\beta_0 - \beta_1 X_i$

ε_i's are not normally distributed anymore

(2) $Var(Y_i) = Var(\varepsilon_i) = \pi_i(1-\pi_i) \overset{or}{=} (\beta_0 + \beta_1 X_i)(1 - \beta_0 - \beta_1 X_i)$

ε_i's don't have constant variance anymore, therefore,

OLS is not optimal anymore.

(3) Constraints on response value :

$$\pi_i = \beta_0 + \beta_1 X_i, \quad \underline{0 \le \pi_i \le 1}$$

where linear function don't keep this constraint.

If a linear function doesn't keep the range $[0, 1]$,

what functions do ?

- Probit Function* : (uncommon) Textbook p559.

$$\Phi^{-1}(\pi_i) = \beta_0^* + \beta_1^* X_i$$

where $\Phi(z)$ is the C.D.F. of $N(0,1)$

- logistic function

Motivation of logistic regression: imagine the binary outcomes are from a continuous scale; change the model to be similiar to SLR / MLR we have learned.

Ex: Suppose y_i^c is the time in weeks, at which mother i gives birth and X_i is the alcohol consumption.

then we can fit a SLR:

$$y_i^c = \beta_0^c + \beta_1^c X_i + \varepsilon_i^c$$

Now let's say we have a threshold week 38. then

$$Y_i = \begin{cases} 1 & \text{when } y_i^c \leq 38 \\ 0 & \text{when } y_i^c > 38 \end{cases}$$

$$\pi_i = P(Y_i = 1) = P(Y_i^c \leq 38)$$

$$= P(\beta_0^c + \beta_1^c X_i + \varepsilon_i^c \leq 38)$$

$$= P(\varepsilon_i^c \leq (38 - \beta_0^c) + (\beta_1^c) X_i)$$

or. $\pi_i = P(\varepsilon_i^c \le \beta_0^* + \beta_1^* X_i)$

* When we assum ε_i^c is from a normal distribution

AKA, $\varepsilon_i^c \sim N(0, \sigma^2)$

then $\pi_i = P(Z \le \beta_0 + \beta_1 X_i)$ after standardizing ε_i^c

$= \Phi(\beta_0 + \beta_1 X_i)$

so $\Phi^{-1}(\pi_i) = \beta_0 + \beta_1 X_i$ — Probit Function

When we assum $\varepsilon_i^* = \frac{\pi}{\sqrt{3}} \frac{\varepsilon_i^c}{\sigma}$ has logistic distribution:

A r.v. Y is said to have logistic distribution if

p.d.f. $f(y) = \frac{e^y}{(1+e^y)^2}$

c.d.f. $F(y) = \frac{e^y}{1+e^y}$

$E(Y) = 0$, $Var(Y) = \frac{\pi^2}{3}$.

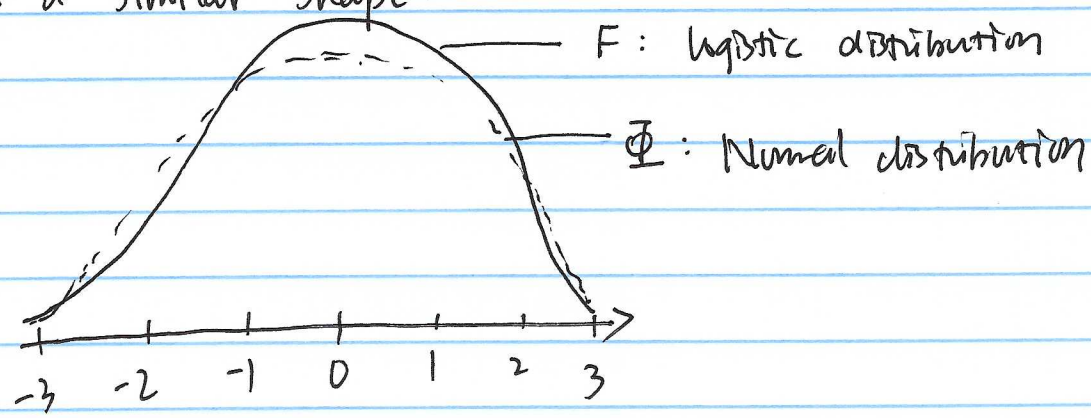
Then $\pi_i = P(\varepsilon_i^c \le \beta_0^* + \beta_1^* X_i)$

$= P(\varepsilon_i^* \le \beta_0 + \beta_1 X_i)$

$= F(\beta_0 + \beta_1 X_i)$

F and F^{-1} are much easier than Φ and Φ^{-1}

but with a similar shape:



$\boxed{\text{In Summary}}$:

The Model setup of Logistic Regression

- $Y_i \sim b(1, \pi_i)$ with binary outcomes.

  $Pr(Y_i = 1) = \pi_i$
  $Pr(Y_i = 0) = 1 - \pi_i$

- The logistic model is $\pi_i = \dfrac{e^{\beta_0 + \beta_1 X_i}}{1 + e^{\beta_0 + \beta_1 X_i}}$

or

$$\boxed{\ln \frac{\pi_i}{1-\pi_i} = \text{logit}(\pi_i) = \beta_0 + \beta_1 X_i}$$

- This is a model based on $E(Y_i) = \pi_i$, no error term.

II. Parameter Estimation

- OLS doesn't work anymore: we dont have error-residuals to be estimated.

- Use MLE:

$Y_i \sim b(1, \pi_i)$.

$$f(y_i) = \pi_i^{y_i} (1-\pi_i)^{1-y_i}$$

$$L(\pi) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i}$$

$$\ell(\pi) = \ln L(\pi) = \sum_{i=1}^{n} \left[y_i \cdot \ln(\pi_i / 1-\pi_i) + \ln(1-\pi_i) \right]$$

Plug in $\ln \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 X_i$

and $1-\pi_i = \frac{1}{1+e^{\beta_0+\beta_1 X_i}}$

$$\ell(\beta) = \sum_{i=1}^{n} y_i (\beta_0 + \beta_1 X_i) - \sum_{i=1}^{n} \ln(1+e^{\beta_0+\beta_1 X_i})$$

- we need to find $\hat{\beta}_0, \hat{\beta}_1 = \text{argmax}\ \ell(\beta)$

- In general, MLE are obtained by using iterative algorithms. (Supplemental material provided).

- After we obtain the parameter estimates $\hat{\beta}_0, \hat{\beta}_1$

then π_i can be estimated by

$$\hat{\pi}_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

- More commonly, we look at the estimate of "odds" of success.

$$\frac{\hat{\pi}_i}{1 - \hat{\pi}_i} = e^{\beta_0 + \beta_1 x_i}$$

The odds measures the comparison the chance of success would occur with the chance of success would not occur.

- In Multiple logistic regression, similarly,

$$\ln \frac{\pi_i}{1 - \pi_i} = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_{p-1} X_{i,p-1}$$

After obtain MLE of $\hat{\beta}_0, \cdots, \hat{\beta}_{p-1}$

$$\frac{\hat{\pi}_i}{1 - \hat{\pi}_i} = e^{\hat{\beta}_0 + \hat{\beta}_1 X_{i,1} + \cdots + \hat{\beta}_{p-1} X_{i,p-1}}$$

III. Coefficient Interpretation. — odds ratio.

we still take the example with one predictor.

$$\frac{\hat{\pi}_i}{1-\hat{\pi}_i} = e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}$$

when $X_i = x$. the odds of success $\Big|_{x_i = x} = e^{\hat{\beta}_0 + \hat{\beta}_1 x}$

when $X_i = x+1$ the odds of success $\Big|_{x_i = x+1} = e^{\hat{\beta}_0 + \hat{\beta}_1 (x+1)}$

The $\boxed{\text{odds ratio}}$ when the predictor X ~~chang~~ increases by 1.

is given by

$$\frac{e^{\hat{\beta}_0 + \hat{\beta}_1 (x+1)}}{e^{\hat{\beta}_0 + \hat{\beta}_1 x}} = e^{\hat{\beta}_1}$$

Therefore.

- when $\hat{\beta}_1 < 0$: odds ratio < 1, the odds of success decreases when X value increases.
- when $\hat{\beta}_1 > 0$: odds ratio > 1, the odds of success increases when X value increases

- when $\hat{\beta}_1 = 0$, odds ratio $= 1$. the odds of success doesn't change with X values.

- In the multiple predictor case, the interpretation of $\hat{\beta}_K$ is similar with all other predictor values stay the same.

- The odds of success $= e^{\hat{\beta}_0}$ when all predictor values $= 0$.

Note: The odds of success changes when X increases by 1, by a multiplicative factor of $e^{\hat{\beta}_1}$, or by $100(e^{\hat{\beta}_1} - 1)\%$.

Ex. if $\hat{\beta}_1 = 2$. odds ratio $= e^2 = \dfrac{\text{odds}|_{x+1}}{\text{odds}|_x} = 7.3890$.

The odds of success ~~chang~~ increases 638.9% when X increases 1 unit.

Summary and Extra Notes.

I. Model Set-up.

$$\pi_i = \frac{e^{\vec{x}\vec{\beta}}}{1+e^{\vec{x}_i\vec{\beta}}}$$

II. Estimation of $\vec{\beta}$: Maximum likelihood Estimation using the distribution of

$$Y_i \sim b(1, \pi_i)$$

III. Interpretation of parameters

(i) Coefficient $\hat{\beta}_K$ of predictor X_K:

- when $\hat{\beta}_K = 0$, odds of success does not change with X_K
- when $\hat{\beta}_K < 0$: odds of success decreases when X_K values increases.
- when $\hat{\beta}_K > 0$: odds of success increases when X_K values increases.
- The change in odds is $100(e^{\hat{\beta}_K} - 1)\%$.

(ii) The odds of success $= e^{\hat{\beta}_0}$

when all predictors in the model is zero.

Notes on prediction:

- logistic regression predicts the chance or probability of $Y_i = 1$: $\hat{\pi}_i$

- More often, people prefer a classification result:

  $Y_{new} = 1$ or 0 ?

(i) Choose a threshold π^*:

If $\hat{\pi}_{new} \geq \pi^* \rightarrow Y_{new} = 1$

If $\hat{\pi}_{new} < \pi^* \rightarrow Y_{new} = 0$

OPTIONAL → Not included.

(ii) To decide π^* : define a confusion matrix

		Model 0	Model 1
Truth	0	☑ TN	FP
	1	FN	☑ TP

$$\text{Precision} = \frac{TP}{TP + FP}$$

proportion of the ~~actual~~ accurated identified successes among all classified as successes by the model

$$\text{Recall} = \frac{TP}{TP + FN}$$

proportion of the accurated identified successes among all actual successes.

Then you can use the train data / test data / cross-validation to choose a π^* that optimize a function

$$\varphi(\text{result}) = c_1 * \text{Precision} + c_2 * \text{Recall}.$$