Logistic Regression (part II)

Predictive Modeling & Statistical Learning

Gaston Sanchez

CC BY-SA 4.0

Logistic Regression Theory

Logistic Function

I'm afraid we have another issue. While the model:

$$E(Y|X = x_i) = p(x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

solves the issue about adequately approximating Y values in $[0, 1]$, it is NOT linear in its parameters.

Is there a way to linearize things?

Logit Function

To "recover" a linear model, we use the so-called **logit** function:

$$logit(p) = log\left(\frac{p}{1-p}\right)$$

invented by Joseph Berkson in 1944

By the inverse of the logistic function we have that:

$$log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$$

Question

What is this term?

$$\frac{p(x)}{1 - p(x)}$$

Question

What is this term?

$$\frac{p(x)}{1 - p(x)}$$

It is the **odds** of event $Y = 1$ for $X = x$

Logit and Odds

If we take the log of the odds, it turns out that we get the following expression:

$$log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$$

This is the so-called **logit** function.

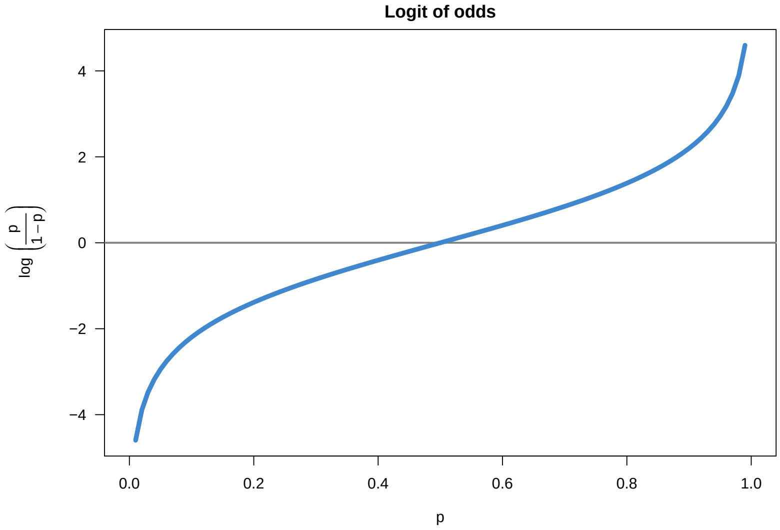
Logit and Odds

The logit function:

$$log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$$

- It is a particular case of the link functions in the framework of *generalized linear models*.
- It can range from $-\infty$ to ∞.
- There is no concern about the range of values that the linear predictors may produce.

Graph of Logit function



Odds

When we have multiple predictors $X_1, \ldots, X_p$ the logistic model becomes:

$$log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Logistic regression models the log-odds of the event as a linear function. In other words, we model the logit of the conditional expectation as a linear combination of the predictors.

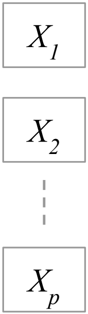
Interpretation of the Logistic Function

The linear predictor can be interpreted as the *propensity* to choose the $Y = 1$ "event"

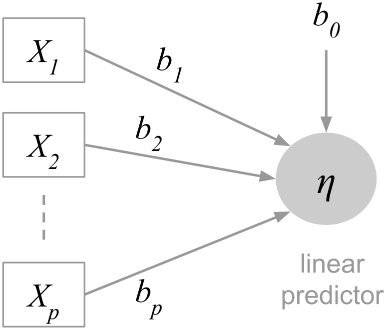
$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

If η_i is greater than a threshold (i.e. 0) then the individual chooses $Y = 1$, otherwise $Y = 0$

A graphical representation

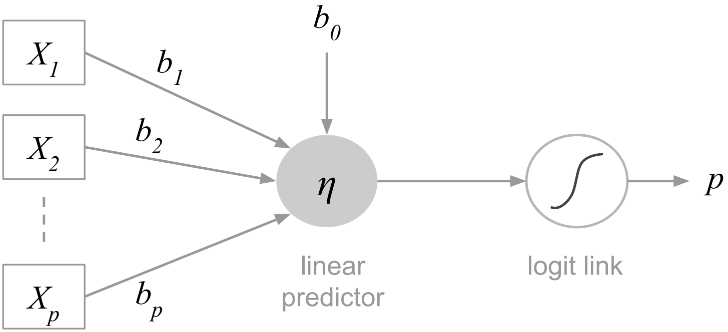


A graphical representation



$$\eta = b_0 + b_1 X_1 + \ldots + b_p X_p$$

A graphical representation



$$\eta = b_0 + b_1 X_1 + \ldots + b_p X_p \qquad p = \frac{e^{\eta}}{1 + e^{\eta}}$$

Estimation of Parameters

Estimation of parameters

For simplicty ...

- one binary response variable Y, coded 0 and 1
- one predictor variable X

The estimation of β_0 and β_1 is carried out by **Maximum Likelihood**

Likelihood Function

The probability of observing the data (independent observations)

$$[(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)]$$

is:

$$= \prod_{i=1}^{n} P(Y = y_i | X = x_i) = \prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}$$

$$= \prod_{i=1}^{n} \left(\frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right)^{y_i} \left(1 - \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right)^{1-y_i} = L(\beta_0, \beta_1)$$

ML Estimation of parameters

The estimation of β_0 and β_1 is carried out by **Maximum Likelihood**

We look for estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ that maximize the likelihood function $L(\beta_0, \beta_1) = L(\boldsymbol{\beta})$

As it is cutomary with ML estimation, it is more convenient to work with the log-likelihood $l(\boldsymbol{\beta}) = log(L(\boldsymbol{\beta}))$

Estimation of parameters

$$
\begin{aligned}
l(\boldsymbol{\beta}) &= log(L(\boldsymbol{\beta})) \\
&= \sum_{i=1}^{n} \{y_i log(p(x_i)) + (1-y_i) log(1-p(x_i))\} \\
&= \sum_{i=1}^{n} log(1-p(x_i)) + \sum_{i=1}^{n} y_i log\left(\frac{p(x_i)}{1-p(x_i)}\right) \\
&= \sum_{i=1}^{n} log(1-p(x_i)) + \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) \\
&= \sum_{i=1}^{n} -log\left(1+e^{\beta_0+\beta_1 x_i}\right) + \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i)
\end{aligned}
$$

ML Estimation of parameters

We look for $\hat{\beta}_0$ and $\hat{\beta}_1$ that maximize the log-likelihood $l(\hat{\beta}_0, \hat{\beta}_1)$. To do so, we set the first order partial derivatives of $l(\boldsymbol{\beta})$ to zero.

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_0} = \sum_{i=1}^{n} (y_i - p(x_i)) = 0$$

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_1} = \sum_{i=1}^{n} x_i (y_i - p(x_i)) = 0$$

There is no analytical solution to this problem.

Estimation of parameters

- We can use the Newton-Raphson method.
- Newton-Raphson requires second-derivatives or Hessian matrix

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\mathsf{T}} = -\sum_{i=1}^{n} x_i x_i^\mathsf{T} p(x_i)(1 - p(x_i))$$

Estimation of parameters

Starting with $\boldsymbol{\beta}^{old}$, a single Newton-Raphson update is:

$$\boldsymbol{\beta}^{new} = \boldsymbol{\beta}^{old} - \left(\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathsf{T}}} \right)^{-1} \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$$

where the derivatives are evaluated at $\boldsymbol{\beta}^{old}$

Estimation of parameters

The iteration can be compactly expressed in matrix form:

- Let $\mathbf{y}$ be the column vector of Y
- Let $\mathbf{X}$ be the $n \times (p+1)$ input (design) matrix
- Let $\mathbf{p}$ be the n-vector of fitted probabilities with the i-th element $p(x_i; \beta^{old})$
- Let $\mathbf{W}$ be an $n \times n$ diagonal matrix of weights with i-th element $p(x_i; \beta^{old})(1 - p(x_i; \beta^{old}))$
- Then

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{X}^\mathsf{T}(\mathbf{y} - \mathbf{p})$$

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\mathsf{T}} = -\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}$$

Estimation of parameters

The Newton-Raphson step is:

$$\begin{aligned}
\boldsymbol{\beta}^{new} &= \boldsymbol{\beta}^{old} + (\mathbf{X}^\intercal \mathbf{W} \mathbf{X})^{-\mathbf{1}} \mathbf{X}^\intercal (\mathbf{y} - \mathbf{p}) \\
&= (\mathbf{X}^\intercal \mathbf{W} \mathbf{X})^{-\mathbf{1}} \mathbf{X}^\intercal \mathbf{W} (\mathbf{X} \boldsymbol{\beta}^{old} + \mathbf{W}^{-\mathbf{1}} (\mathbf{y} - \mathbf{p})) \\
&= (\mathbf{X}^\intercal \mathbf{W} \mathbf{X})^{-\mathbf{1}} \mathbf{X}^\intercal \mathbf{W} \mathbf{z}
\end{aligned}$$

where $\mathbf{z} = \mathbf{X} \boldsymbol{\beta}^{old} + \mathbf{W}^{-\mathbf{1}} (\mathbf{y} - \mathbf{p})$

MLE of the Logistic Regression

Newton-Raphson:

$$\boldsymbol{\beta}^{new} = \boldsymbol{\beta}^{old} - \left(\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta'}\right)^{-1}\left(\frac{\partial l(\beta)}{\partial\beta}\right)$$

$$\boldsymbol{\beta}^{new} = \boldsymbol{\beta}^{old} + (\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\mathsf{T}(\mathbf{y} - \mathbf{p})$$

Estimation of parameters

If $\mathbf{z}$ is viewed as a response and $\mathbf{X}$ is the input matrix, $\boldsymbol{\beta}^{new}$ is the solution to a weighted least squares problem:

$$\boldsymbol{\beta}^{new} \longleftarrow \underset{\beta}{argmin} \left\{ (\mathbf{z} - \mathbf{X}\boldsymbol{\beta})^{\mathsf{T}} \mathbf{W} (\mathbf{z} - \mathbf{X}\boldsymbol{\beta}) \right\}$$

$\mathbf{z}$ is referred to as the adjusted response; and the algorithm is referred to as iteratively reweighted least squares.

IRLS Pseudo Code

1. $\mathbf{b}^{\mathbf{old}} \longleftarrow \mathbf{0}$
2. Compute $\mathbf{p}$ by setting its elements to:

$$p(x_i) = \frac{e^{\mathbf{x_i}^\top \mathbf{b}^{\mathbf{old}}}}{1 + e^{\mathbf{x_i}^\top \mathbf{b}^{\mathbf{old}}}}$$

3. Compute the diagonal matrix $\mathbf{W}$ with the i-th diagonal element: $p(x_i)(1 - p(x_i)), \quad i = 1, \ldots, n$
4. $\mathbf{z} \longleftarrow \mathbf{X}\mathbf{b}^{\mathbf{old}} + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{p})$
5. $\mathbf{b}^{\mathbf{new}} \longleftarrow (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{z}$
6. Check whether $\mathbf{b}^{\mathbf{old}}$ and $\mathbf{b}^{\mathbf{new}}$ are close "enough", otherwise update $\mathbf{b}^{\mathbf{old}} \leftarrow \mathbf{b}^{\mathbf{new}}$, and go back to step 2.

Computational Efficiency

Since $\mathbf{W}$ is an $n \times n$ diagonal matrix, direct matrix operations with it may be very inefficient.

A modified pseudo code is provided next.

IRLS Simplified Pseudo Code

1. $\mathbf{b}^{\mathbf{old}} \longleftarrow \mathbf{0}$
2. Compute $\mathbf{p}$ by setting its elements to:

$$p(x_i) = \frac{e^{\mathbf{x_i}^\top \mathbf{b}^{\mathbf{old}}}}{1 + e^{\mathbf{x_i}^\top \mathbf{b}^{\mathbf{old}}}}$$

3. Compute the $n \times (p+1)$ matrix $\tilde{\mathbf{X}}$ by multiplying the *i*-th row of matrix $\mathbf{X}$ by $p(x_i)(1 - p(x_i)), \quad i = 1, \ldots, n$
4. $\mathbf{b}^{\mathbf{new}} \longleftarrow \mathbf{b}^{\mathbf{old}} + (\mathbf{X}^\top \tilde{\mathbf{X}})^{-\mathbf{1}} \mathbf{X}^\top (\mathbf{y} - \mathbf{p})$
5. Check whether $\mathbf{b}^{\mathbf{old}}$ and $\mathbf{b}^{\mathbf{new}}$ are close "enough", otherwise update $\mathbf{b}^{\mathbf{old}} \leftarrow \mathbf{b}^{\mathbf{new}}$, and go back to step 2.

More about the Parameters

Variance of estimators

$$\hat{V}(\hat{\beta}) = \left[-\frac{\partial l(\beta)}{\partial \beta}\right]^{-1}_{\beta=\hat{\beta}} = (\mathbf{X}^{\mathsf{T}}\hat{\mathbf{V}}\mathbf{X})^{-1}$$

where:

$$\mathbf{X} = \begin{bmatrix} 1 & \cdots & x_1 \\ 1 & \cdots & x_2 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_n \end{bmatrix} \quad \text{and} \quad \hat{\mathbf{V}} = \begin{bmatrix} \hat{p}_1(1-\hat{p}_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{p}_n(1-\hat{p}_n) \end{bmatrix}$$

Interpreting β_1

- Interpreting what β_1 means is not straightforward because we are predicting $P(Y|X)$ not Y.
- If $\beta_1 = 0$, this means that there is no relationship between Y and X.
- If $\beta_1 > 0$, this means that when X gets larger, the probability that $Y = 1$ gets larger too.
- If $\beta_1 < 0$, this means that when X gets larger, the probability that $Y = 1$ gets smaller.
- But how much bigger or smaller depends on where we are on the slope.

Are coefficients significant?

To see whether β_0 and β_1 are significant, we use a Z-test instead of a t-test.

```
log_reg <- glm(chd ~ age, data = dat, family = "binomial")

summary(log_reg)
```

Are coefficients significant?

```
Call:
glm(formula = chd ~ age, family = "binomial", data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9718  -0.8456  -0.4576   0.8253   2.2859

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.30945    1.13365  -4.683 2.82e-06 ***
age          0.11092    0.02406   4.610 4.02e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 136.66  on 99  degrees of freedom
Residual deviance: 107.35  on 98  degrees of freedom
AIC: 111.35

Number of Fisher Scoring iterations: 4
```

Making Predictions

Suppose an individual has an age of 27. What is the probability of having CHD?

$$\hat{p}(27) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}} = \frac{e^{-5.309 + 0.11 \times 27}}{1 + e^{-5.309 + 0.11 \times 27}} = 0.0899$$

The predicted probability of CHD for an 27yr individual is less than 1%

References

- **Extending the Linear Model with R** by Julian Faraway (2006) *Chapter 2: Binomial Data*. CRC Press.
- **The origins and development of the logit model** by J.S. Cramer (2003) http://www.cambridge.org/resources/0521815886/1208_default.pdf
- **Applied Logistic Regression** by Hosmer and Lemeshow (2000).
- **Data Mining and Statistics for Decision Making** by Stephane Tuffery (2011). *Chapter 11: Classification and prediction methods*. Wiley.

References (French Literature)

- **Modeles Statistiques pour Donnees Qualitatives** by Droesbeke et al (2005). *Chapter 6: Modele a reponse dichotomique* by P.L. Gonzalez. Editions Technip, Paris.
- **Statistique Explicative Appliquee** by Nakache and Confais (2003). *Chapter 4: Modele logistique binaire*. Editions Technip, Paris.
- **Probabilites, analyse des donnees et statistique** by Gilbert Saporta (2011). *Chapter 18: Analyse discriminante et regression logistique*. Editions Technip, Paris.
- **Statistique: Methodes pour decrire, expliquer et prevoir** by Michel Tenenhaus (2008). *Chapter 11: La regression logistique binaire*. Dunod, Paris.