Logistic Regression

Data Mining course

Master in Information Technologies
Enginyeria Informàtica

Tomàs Aluja
Logistic regression

Response variable: Binary
(or a proportion, ordinal variable, nominal variable)

Examples: To buy a product, to pass a course, to obtain acredit, to level the preference for a service, ...

\[
y_i = \begin{cases} 
1 & \text{if} \quad \text{with } p_i \\
0 & \text{if} \quad \text{with } (1 - p_i)
\end{cases}
\]

\[
E\left[ y_i / x_{i1}, \ldots, x_{ip} \right] = \hat{y} = b_0 + b_1 x_1 + \cdots + b_p x_p
\]

But with a linear fit we have: \( -\infty < \hat{y} < \infty \)

For each individual we know \( y_i \), but we would need an estimation of \( p_i \).

Violation of linear model hypothesis
How can we model binary responses?

The response is binary 0/1

\[ y_i = \begin{cases} 1 & \text{Prob}_i(1) = p_i, \\ 0 & \text{Prob}_i(0) = 1 - p_i. \end{cases} \]

\[ y_i = r(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) + \varepsilon_i = E\left[ y_i \mid x_{i1}, \ldots, x_{ip} \right] + \varepsilon_i \]

\[ E\left[ y_i \mid x_{i1}, \ldots, x_{ip} \right] = 1 \times p_i + 0 \times (1 - p_i) = p_i \]

Let’s do a transformation of the linear component: \( \beta'x_i \)

In such a way that \( r(\beta'x_i) \) mapping function in the interval 0:1 (=probability)

\[ p_i = r(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \]

\[ \varepsilon_i \text{ binomial } \quad \varepsilon \sim B(n_i, p_i) \]

**Generalized Linear Model**

\( n_i \) number of observations in individual \( i \)

\( p_i \) probability of \( y=1 \) for individual \( i \)
<table>
<thead>
<tr>
<th>Age Group</th>
<th>n</th>
<th>CHD absent</th>
<th>CHD present</th>
<th>Mean (Proportion)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20–29</td>
<td>10</td>
<td>9</td>
<td>1</td>
<td>0.10</td>
</tr>
<tr>
<td>30–34</td>
<td>15</td>
<td>13</td>
<td>2</td>
<td>0.13</td>
</tr>
<tr>
<td>35–39</td>
<td>12</td>
<td>9</td>
<td>3</td>
<td>0.25</td>
</tr>
<tr>
<td>40–44</td>
<td>15</td>
<td>10</td>
<td>5</td>
<td>0.33</td>
</tr>
<tr>
<td>45–49</td>
<td>13</td>
<td>7</td>
<td>6</td>
<td>0.46</td>
</tr>
<tr>
<td>50–54</td>
<td>8</td>
<td>3</td>
<td>5</td>
<td>0.63</td>
</tr>
<tr>
<td>55–59</td>
<td>17</td>
<td>4</td>
<td>13</td>
<td>0.76</td>
</tr>
<tr>
<td>60–69</td>
<td>10</td>
<td>2</td>
<td>8</td>
<td>0.80</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>57</td>
<td>43</td>
<td>0.43</td>
</tr>
</tbody>
</table>
Which $r$ function to choose?

**Logistic function**

The log odds (=logit) is a linear function of the predictors

$$\ln \frac{P(+/x_i)}{P(-/x_i)} = \beta'x_i$$

Logistic function is very close to the inverse of the normal distribution function (probit function)

$$p_i = \frac{1}{1 + \exp^{-\beta'x_i}} \approx \Phi^{-1}(\beta'x_i)$$
Interpretation of the logistic function

The linear predictor can be interpreted as the propensity to choose the + event

$$\eta_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

If $$\eta_i$$ is greater than a threshold (i.e. 0) then the individual choose +, otherwise -

In economy the propensity can be formulated as the difference of two utilities

$$y_i / x_1, \ldots, x_p = 1 \quad \text{if } u(1) > u(0)$$
$$y_i / x_1, \ldots, x_p = 0 \quad \text{if } u(0) > u(1)$$

Hence, supposing a linear model for the propensity

$$u(1) - u(0) = \beta' x_i + \varepsilon_i$$

$$P(y_i = 1 / x_{i1}, \ldots, x_{ip}) = P(u(1) - u(0) > 0) = P(\beta' x_i > -\varepsilon_i)$$
A graphical representation of the logistic regression

\[ \eta_i = b_0 + b_1 x_1 + \cdots + b_p x_p \]

\[ p_i = \frac{1}{1 + e^{-\eta_i}} \]

But, how to estimate the \( b_0, b_1, \ldots, b_p \)
Maximum Likelihood Estimation (MLE) remainder

Choose as estimates of the parameters those who minimize the probability of the observed data

\[
Max \ L(\theta) = \Pr(x_1, \ldots x_n / \theta) = \Pr(x_1 / \theta) \times \cdots \times \Pr(x_n / \theta)
\]

A silly example, estimate the probability of heads in 10 coin tosses if we get 7 heads

```r
> n = 10
> n1 = 7
> n0 = n - n1
> p = seq(from=0, to=1, by=0.1)
> p
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
> fv = p^n1*(1-p)^n0
> fv
0.0000000000 0.0000000729 0.0000065536 0.0000750141 0.0003538944 0.0009765625 0.0017915904 0.0022235661 0.0016777216 0.0004782969 0.0000000000
> plot(p,fv,type="l")
```
MLE of the Logistic Regression

\[ L(\beta) = \Pr((y_1, x_1), \ldots, (y_n, x_n)) = \prod_{i=1}^{n} \Pr(y_i / x_i) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i} \]

\[ \log L(\beta) = l(\beta) = \sum_{i} \log p_i = \sum_{i} (y_i \log p_i + (1 - y_i) \log(1 - p_i)) \]

\[ p_i^{y_i} (1 - p_i)^{1-y_i} = \left( \frac{p_i}{1 - p_i} \right)^{y_i} (1 - p_i) = \left( e^{-\beta x_i} \right)^{y_i} (1 + e^{\beta x_i}) \]

\[ l(\beta) = \sum_{i} (y_i \beta' x_i + \log(1 + e^{\beta' x_i})) \]

\[ \frac{\partial l(\beta)}{\partial \beta} = X' (y - p) \]

\[ \frac{\partial^2 l(\beta)}{\partial \beta \partial \beta'} = -X'WX \]

\[ W = \begin{bmatrix} \vdots & \vdots & \vdots \\ \end{bmatrix} \]

\[ \begin{bmatrix} p_i (1 - p_i) \\ \end{bmatrix} \]

DM Logistic Regression, T. Aluja 9
MLE of the Logistic Regression

**Newton-Raphson**

\[
\beta^{t+1} = \beta^t - \left( \frac{\partial^2 l(\beta)}{\partial \beta \partial \beta'} \right)^{-1} \left( \frac{\partial l(\beta)}{\partial \beta} \right)
\]

\[
\beta^{t+1} = \beta^t + (X'WX)^{-1} X'(y - p) = (X'WX)^{-1} X'Wz
\]

\[
z = X \beta^t + W^{-1}(y - p)
\]

**Iterated Reweighted Least Squares (IRLS algorithm)**

Initialize \( \beta_0 = \log(n_+/n_-) \) \( \beta_j = 0, \ j=1,...,p \) (null model)

Iterate till convergence

- Estimate \( p \) and \( W \)
- Calculate \( z \)
- Update \( \beta \) by weighted regression
How can we assess the quality of the logistic regression?

Deviance
Likelihood ratio test of the proposed model respect to the saturated model (the one providing a perfect fit $p_i = y_i$). It can be interpreted as a proximity measure of the model fit respect to the data, is a similar fashion as the sum of residual squares in linear models.

$$D = -2 \sum_{i=1}^{n} \left( y_i \log p_i + (1 - y_i) \log(1 - p_i) \right)$$

Null deviance: Deviance of the null model (just with constant term)
Residual deviance: Deviance of the proposed model
AIC: Deviance with complexity penalization ($+2p$)
```r
> learn <- sample(1:n, round(0.67*n))
> l3 = glm(dict ~ edat+ratfin+tiptreb, family = binomial, data = dd[learn,])
> summary(l3)

glm(formula = dict ~ edat + ratfin + tiptreb, family = binomial(link = logit),
    data = dd[learn, ])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1157  -1.0444   0.4602   1.0010   1.9476

Coefficients:               Estimate Std. Error  z value Pr(>|z|)
(Intercept)   -0.515779    0.875162  -0.589 0.555625    
edat           0.033935    0.010838   3.131 0.001742 **
ratfin        -0.033892    0.006085  -5.569 2.56e-08 ***
tiptrebfixe   1.619291    0.662626   2.444 0.014536 *
tiptrebtemp   2.231853    0.657498   3.394 0.000688 ***
tiptrebtemp    0.562770    0.766715   0.734 0.462948
---

Null deviance: 563.92  on 406  degrees of freedom
Residual deviance: 489.35  on 401  degrees of freedom
AIC: 501.35

Number of Fisher Scoring iterations: 4
```
Could we simplify the model?

```r
> step(l3)
Start:  AIC= 501.35
  dict ~ edat + ratfin + tiptreb
           Df Deviance    AIC
<none>         489.35 501.35- edat     1   499.60 509.60- tiptreb  3   520.95 526.95- ratfin   1   525.10 535.10

Call:  glm(formula = dict ~ edat + ratfin + tiptreb, family = binomial(link = logit),
  data = dd[learn, ])

Coefficients:

(Intercept) edat ratfin tiptrebaution tiptrebfixe tiptrebtemp
   -0.51578   0.03394  -0.03389    1.61929    2.23185    0.56277

Degrees of Freedom: 406 Total (i.e. Null); 401 Residual

Null Deviance: 563.9 Residual Deviance: 489.3 AIC: 501.3
Interpreting the coefficients of a logistic regression

Lets take one predictor $x=0,1$ (i.e. $0=$single, $1=$married)

The odds for a married person, express how much likely is having a positive dictamen rather than a negative dictament being married

Likewise, ...

The ODDS RATIO

The exponential of the $\beta_1$ coefficient measures the change in the odds of $+$ against $-$, when passing from $x=0$ to $x=1$
The obtained model:
\[
\log \frac{p_i}{1 - p_i} = -0.51578 + 0.03394 \cdot edat - 0.03389 \cdot ratfin + 1.61929 \cdot auton + 2.23185 \cdot fixe + 0.56277 \cdot temp
\]

\[i: \text{ edat}=25, \text{ ratfin}=40, \text{ temp}=1\]
\[
\log \frac{p_i}{1 - p_i} = -0.51578 + 0.03394 \times 25 - 0.03389 \times 40 + 0.56277 = -0.46011 \quad p_i = 0.387
\]

\[i': \text{ edat}=26, \text{ ratfin}=40, \text{ temp}=1\]
\[
\log \frac{p_i}{1 - p_i} = -0.51578 + 0.03394 \times 26 - 0.03389 \times 40 + 0.56277 = -0.42617 \quad p_{i'} = 0.395
\]

efecto de la edat: \[\log \frac{p_{i'}}{1 - p_{i'}} - \log \frac{p_i}{1 - p_i} = 0.03394 \quad \frac{p_{i'}}{1 - p_{i'}} / p_i / 1 - p_i = e^{0.03394} = 1.0345\]

---

Interpret the coefficients

\[
> \exp(13\text{coefficients})
\]

<table>
<thead>
<tr>
<th>(Intercept)</th>
<th>edat</th>
<th>ratfin</th>
<th>tiptrebauton</th>
<th>tiptrebfixe</th>
<th>tiptrebttemp</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5970355</td>
<td>1.0345176</td>
<td>0.9666757</td>
<td>5.0495067</td>
<td>9.3171141</td>
<td>1.7555279</td>
</tr>
</tbody>
</table>
Plot of the linear predictor and the estimated probabilities

```r
> plot(l3$linear.predictors, l3$fitted.values)
```
Importance of the variables

Descomposition of the Deviance

```r
> anova(l3)
Analysis of Deviance Table
Model: binomial, link: logit

Response: dict
Terms added sequentially (first to last)

<table>
<thead>
<tr>
<th>Term</th>
<th>Df</th>
<th>Deviance</th>
<th>Resid. Df</th>
<th>Resid. Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td></td>
<td></td>
<td>406</td>
<td>563.92</td>
</tr>
<tr>
<td>edat</td>
<td>1</td>
<td>9.38</td>
<td>405</td>
<td>554.54</td>
</tr>
<tr>
<td>ratfin</td>
<td>3</td>
<td>31.60</td>
<td>401</td>
<td>489.35</td>
</tr>
</tbody>
</table>
```

\[ \text{Deviance}_1 - \text{Deviance}_2 \sim \chi^2_{\nu_1 - \nu_2} \]

\[ E[\chi^2_{\nu}] = \nu \]
Selecting the model

Estimate of the Generalization Error in a test sample:

<table>
<thead>
<tr>
<th>Error rate in learn</th>
<th>GE in test</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; l3pred=NULL</td>
<td>&gt; l3t = predict(l3, dd[-learn,])</td>
</tr>
<tr>
<td>&gt; l3pred[l3$fitted.values&lt;0.5]=0</td>
<td>&gt; pt = 1/(1+exp(-l3t))</td>
</tr>
<tr>
<td>&gt; l3pred[l3$fitted.values&gt;=0.5]=1</td>
<td>&gt; l3predt = NULL</td>
</tr>
<tr>
<td>&gt; table(dict[learn],l3pred)</td>
<td>&gt; l3predt[pt&lt;0.5]=0</td>
</tr>
<tr>
<td>13pred</td>
<td>&gt; l3predt[pt&gt;=0.5]=1</td>
</tr>
<tr>
<td>0</td>
<td>&gt; table(dict[-learn],l3predt)</td>
</tr>
<tr>
<td>1</td>
<td>13predt</td>
</tr>
<tr>
<td>0 118 80</td>
<td>0 1</td>
</tr>
<tr>
<td>1 67 142</td>
<td>0 65 40</td>
</tr>
</tbody>
</table>

\[P_{acierto} = 63.9\%\]

\[P_{acierto} = 62.0\%\]
Graphical comparison of the real response respect to the predicted in the test sample

Actual response (*test*)

Prediction (*test*)
ROC and Concentration curves

Concentration Curve

ROC Curve

DM Logistic Regression, T. Aluja