Logistic Regression with R

- How to fit a logistic model in R
- How to interpret the coefficients?

Increasing X1 by one unit
= changing the odds by exp(beta1)

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

Logistic Regression with R

- What's null deviance and residual deviance

Deviance = 2*[loglik_saturated_model - loglik_current_model]

Loglik = $$\sum_{i:y_i=1} \log \hat{p}_i + \sum_{i:y_i=0} \log(1-\hat{p}_i)$$

Saturated Model

$$(x_i, y_{i,1}, \ldots, y_{i,n_i})$$

$$\hat{p}_i = \frac{\sum_j y_{ij}}{n_i}$$

Null Model

$$\hat{p}_i^0 = \frac{\sum_i y_i}{n}$$

If x_i's are **unique**, p-hat=0/1,

Deviance = -2*loglik_current model

Logistic Regression with R

- How to do model selection with AIC/BIC
  Stepwise/backward/forward
- How to do model selection with Lasso

More on Logistic Regression

- Convergence issue with logistic regression when data are well-separated
- Multinomial logistic regression
- Move beyond linear decision boundary: add quadratic terms to logistic regression
- Retrospect sampling (both LDA and Logistic can handle this)

We usually assume that data (x_i, y_i)'s are collected as iid samples from a population of interest. However, in some applications, we may have data collected retrospectively. For example, in a study on cancer, instead of taking a random sample of 100 people from the whole population (then the number of cancer patients will be too small), we may draw 50 samples from the cancer group and 50 from the control group.

Test Samples

Training Samples

Let Z indicate whether an individual is sampled or not. Using retrospective samples, we are estimating $P(Y = 1|Z = 1, X = x)$, while what we care is $P(Y = 1|X = x)$. Assume the logit of the latter follows a linear model, that is,

$$P(Y = 1|X = x) = \frac{\exp(\alpha + x^t\beta)}{1 + \exp(\alpha + x^t\beta)}, \tag{2}$$

where we isolate the intercept α from other coefficients β. Then the question is whether we can estimate α or β in (2) based on a retrospective sample.

It turns out that for logistic models, the coefficients estimated from a retrospective sample is roughly the same as the one from an ordinary random sample. Again, let Z indicate whether an individual is sampled or not and denote the sample proportions (in each class) by

$$r_0 = \mathbb{P}(Z=1|Y=0), \quad r_1 = \mathbb{P}(Z=1|Y=1).$$

Cancer patients are more likely to be sampled

Then

$$\begin{aligned}
\mathbb{P}(Y=1|Z=1, X=x) &= \frac{\mathbb{P}(Z=1|Y=1,x)\mathbb{P}(Y=1|x)}{\mathbb{P}(Z=1,Y=0|x)+\mathbb{P}(Z=1,Y=1|x)} \\
&= \frac{r_1 \exp(\alpha + x^t\beta)}{r_0 + r_1 \exp(\alpha + x^t\beta)} \\
&= \frac{\exp(\alpha^* + x^t\beta)}{1+\exp(\alpha^* + x^t\beta)}, \qquad \alpha^* = \alpha + \log\frac{r_1}{r_0}
\end{aligned}$$

P(Y=1, Z=1| X=x)

P(Z=1| X=x)
= P(Y=1, Z=1|X=x)
+ P(Y=0, Z=1|X=x)

Although the data have been sampled retrospectively, the logistic model continues to apply with the same coefficients β but a different intercept.