Logistic Regression:

- Until this point we have been talking about regression using continuous data.
- However in real life we often deal with situations when data is binary
- Especially true for biological scientist, economist, agriculture, social sciences, engineering etc
- In these situations we want to see how different predictors affect the binary response

Examples:

- Organism survived did not survive when exposed to different levels of Zinc
- Bank loan default or not based on level of income
- Dosage of medicine and disease remission
- Agree or disagree based on Income status, Education

One thing to keep in mind is that, all binary responses really have a continuous latent variable in the background.

Consider, the survival of the organism when exposed to Zinc. We can think that there is a continuous response, tolerance of the organism to Zinc, say Y*. When $Y^* \leq c$(some constant), we have the organism living, Y=1. Else, if Y*>c, the organism dies, Y=0. However, we do NOT have data on Y*, but only the binary end-point, Y.

Let us assume a Normal model for Y*, i.e.

$Y^* = \beta_0^* + \beta_1^* X + \varepsilon^*$

Now, $P(Y=1) = P(Y^* \leq c) = P(\beta_0^* + \beta_1^* X + \varepsilon^* \leq c) =$
$P(\varepsilon^* \leq c - (\beta_0^* + \beta_1^* X)) = P(Z \leq \beta_0 + \beta_1 X)$.

If we assume, Z follows a Normal distribution then, $P(Y=1)=\Phi(\beta_0+\beta_1 X)$, this leads to the Probit Formulation of the Problem.

If we assume that Z follows a Logistic distribution then, $P(Y=1)=F(\beta_0+\beta_1 X)$, where F represents the cdf of the logistic distribution i.e. $P_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1+\exp(\beta_0 + \beta_1 x_i)}$

If we assume that Z follows a Gumbel distribution then, $P(Y=1)=\mathcal{F}(\beta_0+\beta_1 X)$, where $\mathcal{F}$ represents the cdf of the Gumbel distribution.

$P_i = 1-\exp(-\exp(\beta_0 + \beta_1 x_i))$

Generally the Logistic Formulation is the most common due to computational advantages.

Another way of looking at this:

Model for ordinary regression:

$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

If Yi is binary then the model is very restrictive.

Here εi follows a Bernoulli distribution and hence

Var (εi) = E(Yi) [1-E(Yi)]

Which implies non-constant variance.

More importantly β0 and β1 can only take special values which depends upon xi.

Question is: How can we improve this?

One option is using the proportions of successes instead of successes. And modeling

$P_i = \beta_0 + \beta_1 x_i + \varepsilon_I$

Problem is that this will also be restrictive since it lies between 0 and 1.

In a nutshell we can think of this problem as:

How can we take something that is between 0 and 1 and make it between $-\infty$ and $+\infty$.

In one sense it's a transformation issue:

$$\ln[\frac{P_i}{(1-P_i)}] = \beta_0 + \beta_1 x_i$$

$$P_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1+\exp(\beta_0 + \beta_1 x_i)}$$

In the logistic regression model:

$$P(Y_i=1) = P_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1+\exp(\beta_0 + \beta_1 x_i)}$$

$E(Y_i) = 1.\ P(Y_i=1) + 0.\ P(Y_i=0) = P_i$

The more theoretical way of looking at this issue is to see what some of the key features were of GENERAL LINEAR MODEL (GLM).

Essentially in GLM we relate our response Y to the mean of the response and some error.

Now, $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
Is really of form:

$Y_i = E(Y_i) + \varepsilon_I$
And we make assumptions of normality, independence and equal variance on the εi and in turn Yi.
Note: we are also assuming that Yi is related to E(Yi) in a linear way, as a matter of fact as an identity.

We want to generalize from GLM to take into accounts situations were the data is clearly non-normal (like binary data).

We do this in two steps:

1. Generalize the parent distribution to accommodate Binomial, Poisson, Gamma distributions.
2. Allow for the LINK between Y and E(Y) to be NON-LINEAR. We call this relationship the LINK function.

This generalizes the General Linear Model into a GENERALIZED LINEAR MODEL (GliM).

Hence, Logistic Regression is a special case of GliM with Binomial distribution on the Y's and the link function:

$$\text{E(Yi)} = P_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

Which is non-linear in the X's.

Estimation of the parameters:

- Maximum likelihood methods used
- No closed form solutions, iterative methods have to be used
- SAS or R would do this

When we have multiple predictors we use multiple logistic regression

Parameter estimates are found using iterative methods

The applicability of logistic regression stems from the fact that it can easily be extended to multiple predictors.

Interpretation of the parameters:

In Logistic Regression:

$$E(Yi) = P_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)},$$

Hence, $\ln[\frac{P_i}{(1-P_i)}] = \beta_0 + \beta_1 x_i$

$[\frac{P_i}{(1-P_i)}]$ is also called the odds ratio. Hence β_1: is the change in the log odds-ratio for unit change in the predictor X.

For β_0: the interpretation is not as direct.

Multiple Logistic Regressions:

The popularity of Logistic Regression stems from the fact that we can easily extend it to multiple predictors like Multiple Regression

Here, we can write this as:

$$X'\beta = \beta_0 + \beta_1 X_1 + ... + \beta_{p-1} X_{p-1}$$

And,

$$E(Y) = \frac{\exp(X'\beta)}{1 + \exp(X'\beta)}$$

Model Fitting follows exactly the same methods as Simple Logistic Regression.

It is possible to use Polynomial Logistic Regressions, i.e. higher order terms in the model. However, it is strongly advised to center the terms (X- $\overline{X}$) first.

Model Building:

- Most computer packages are equipped to deal with model building with logistic regression as with linear regression.
- Stepwise selection is a standard method

Options include SLENTRY, SLSTAY, INCLUDE

As with linear regression.

Often used Criterion:

$AIC_p = -2\ln L(b) + 2p$
$SBC_p = -2\ln L(b) + p\ln(n)$

Good models small numbers for these measures. These are analogous to a small SSE.

Best Subsets and Stepwise Models are available in most Commercial Software.

Significance of the parameters:

Test on Individual Parameters:

Wald Tests are employed. These are available is most computer Packages.

Can do Confidence interval on the parameters as well as the odds ratios.

Can use Likelihood techniques or Deviance Likelihoods for testing several parameters equal to 0. (Similar to General Linear Testing).

Model Deviance:
Deviance of the model compares the log likelihood of the fitted model with the log likelihood of the saturated model.

Saturated Model: A model with n parameters

i.e. # of parameters=#of observation

For the saturated model:

$$\hat{P}_{is} = Y_i$$

$$\log_e L(\hat{P}_{1s},...,\hat{P}_{ns}) = \sum_{i=1}^{n} \{Y_i \log_e (Y_i) + (1-Y)_i \log_e (1-Y_i)\}$$

Log likelihood for the fitted model when MLE techniques are used to estimate $b_0,\ldots b_{p-1}$ is:

$$\ln L(b_0,...,b_{p-1}) = \sum_{i=1}^{n} Y_i (b_0 + b_1 x_{1i} + ... + b_{p-1} x_{p-1i}) -$$

$$\sum_{i=1}^{n} \ln [1 + \exp(b_0 + b_1 x_{1i} + ... + b_{p-1} x_{p-1i})]$$

DEVIANCE$(X_0,\ldots,X_{p-1})$ =

$$2 \ln L(\hat{P}_{is},...,\hat{P}_{ns}) - 2 \ln L(b_0,...,b_{p-1})$$

This simplifies to:

DEVIANCE$(X_0,\ldots,X_{p-1})$ =

$$-2 \sum_{i=1}^{n} [Y_i \ln(\hat{P}_i) + (1 - Y_i) \ln(1 - \hat{P}_i)]$$

With: $\hat{P}_i = \dfrac{\exp(b_0 + b_1 x_{1i} + ... + b_{p-1} x_{p-1i})}{1 + \exp(b_0 + b_1 x_{1i} + ... + b_{p-1} x_{p-1i})}$

Smaller Deviance indicates better fit.

Deviance is analogous to SSE in Linear Regression.

Partial Deviance:

- For each fitted model we can calculate deviance
- Between two fitted models we can calculate partial deviance

Same idea as general linear testing and partial sums of squares

To test:

H_0: $\beta_q=\beta_{q+1}=\ldots=\beta_{p-1}=0$

H_A: not all equal to 0

$Dev(X_q,\ldots,X_{p-1}| X_0,\ldots X_{q-1}) = Dev(X_q,\ldots,X_{p-1}) Dev(X_0,\ldots X_{q-1})$

If $Dev(X_q,\ldots,X_{p-1}| X_0,\ldots X_{q-1}) > \chi^2(1-\alpha,p-q)$ then Reject H_0.

Comment:

- Idea of deviance stems from the idea of the likelihood ratio test
- Using partial deviance to test hypothesis is identical to the corresponding likelihood ratio tests

MODEL DIAGNOSTICS

- **Goodness of Fit**

Assessing Fit:

-2 log likelihood:

Let $\hat{p}_i$ estimate $p_j = P(Y_i=y_i \mid X_1,\ldots, X_{p-1})$ which is obtained by replacing the b's in the equation.

$$P_i = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_{p-1} x_{p-1i})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_{p-1} x_{p-1i})}$$

$$-2\ln L = -2\sum_{i=1}^{n} \ln(\hat{p}_j)$$

Under Ho this follows a χ^2 distribution hence p-values are printed in most computer programs.

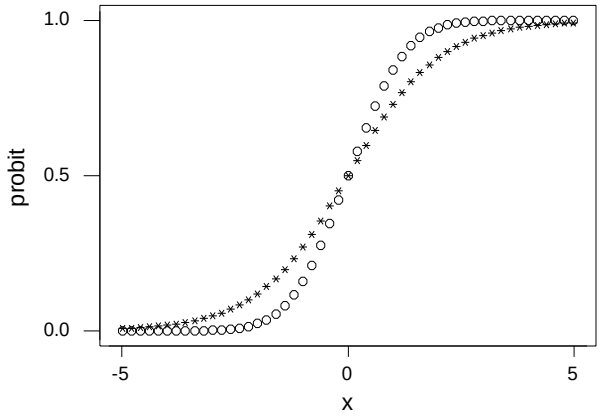
Deviance Goodness of Fit

If $\text{Dev}(X_q,\ldots,X_{p-1}) > \chi^2(1-\alpha,p-q)$ then Reject H_0.

H_0: model is correctly specified.

Goodness of Fit (informal)

- Divide data into classes (about equal number in each class)
- For each class
  - find the mid-point,x
  - find the proportion of successes,y
- Plot y vs x and look for the S shape.



Chi-square Goodness of Fit

Ho: $E(Y) = [1+\exp(-\beta' X)]^{-1}$
Ha: $\neq$

- Group data into classes as before, j=1,…,c.
- For the jth class O_{j1} number of outcomes 1, and O_{j0} as the number of outcomes 0.
- $E(Yi)= \pi_i =[1+\exp(-\beta' X)]^{-1}$
- Estimated by: $\hat{\pi}_i =[1+\exp(-b' X)]^{-1}$
- Hence, $E_{j1}=\sum \hat{\pi}_i$ and $E_{j0}=\sum (1-\hat{\pi}_i) = n_j - E_{j1}$
- And $X^2 = \sum_{j=1}^{c} \sum_{k=0}^{1} \frac{(O_{jk} - E_{jk})^2}{E_{jk}} \rightarrow \chi^2_{(c-2)}$

Hosmer and Lemeshow Goodness of Fit

Similar to Chi-square Goodness of Fit. Group the data based on Estimated Probabilities, i.e. group data points with similar $\hat{\pi}_i$ in the same group. Calculate the Pearson Goodness of Fit using c-2 degrees of freedom.

Residual analysis

- Residuals in logistic regression can only be 0 or 1
- Hence different methods are used to calculate these, deviance residuals, Pearson correlation residuals

$$e_i = \begin{cases} 1-\hat{\pi}_i, \text{if } \mathrm{Y_i} = 1 \\ -\hat{\pi}_i, \text{if } \mathrm{Y_i} = 0 \end{cases}$$

Pearson Residuals

$$r_{Pi} = \frac{Y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}$$

Arising from the Pearson Goodness of Fit

Deviance Residuals

- $dev_i = \pm\sqrt{\{-2[Y_i \log e(\hat{\pi}_i) + (1-Y_i)\log e(1-\hat{\pi}_i)]\}}$

If $Y_i \geq \hat{\pi}_i$ sign ‘+’ else, ‘-‘

$$\sum_{i=1}^{n} dev_i = DEV(X_0,...,X_{p-1})$$

Like individual Residual squares add up to SSE.

Diagnostics:

Fitted versus residual plot with Lowess Smooth is used to assess if the model is correct.

Under the Logistic Regression Model E(ei)=0, hence a LOWESS smooth of the residual versus predicted values should be horizontal. Any departure, may indicate that the model is not appropriate.

Influence Diagnostics are also available using Influence Plots. Here Pearson Residuals, the deviance residual are used.

If you have multiple categories, one may use POLYTOMOUS logistic regression. If the categories are ordered, one uses ORDINAL logistic Regression.

Poisson Regression:

Poisson distribution is used to model DISCRETE random variables.

Examples:

1. Now of telephone calls coming to a switchboard, as a function of the network size.
2. Number of trips to the grocery store, based upon number of children in the family and income level
3. Count of cars in a busy intersection based upon time of day, population of area, nearness to Malls, etc.

These examples show a need for modeling Discrete Random Variables as a function of Numerical or Categorical Predictors.

The Poisson Random Variable:

The mass function is given by

$$P(Y = y) = f(y) = \frac{\mu^y \exp(-\mu)}{y!}, y = 0,1,2...$$

Properties:
Has a range from 0 to positive infinity
Mean, $E(Y)=\mu$
Variance, $\sigma^2(Y)=\mu$
Hence, mean and Variance are same.

Commonly Y refers to units of time or space, for instance the time in minutes within which so many telephone calls come.

Regression Model:

$$Y_i = E(Y_i) + \varepsilon_i$$

We can model the mean function, E(Yi) or mi as:

$$\mu_i = X'\beta$$
$$\mu_i = \exp(X'\beta)$$
$$\mu_i = \ln(X'\beta)$$

In any case, we assume X'β to be nonnegative. The exponential form is the most commonly used LINK.

Here too, we estimate parameters using Maximum Likelihood Estimates. Use Model Deviance and Deviance Residuals.

Testing, Diagnostics are done similar to Logistic Regression.