Generalized Linear Models

Let Y be a response variable, and $X_1, \cdots, X_p$, predictor variables.

For linear models, the conditional mean of Y given $\mathbf{X} = \mathbf{x}$ is given by

$$\mu_{Y|\mathbf{x}} = \beta_o + \beta_1 X_1 + \cdots + \beta_p X_p$$

Basic assumptions:

- $Var(Y \mid \mathbf{x})$ is constant, and
- Y is Guassian
- Error terms uncorrelated.

More generally, assume that there is a function g such that

$$g(\mu) = \beta_o + \beta_1 X_1 + \cdots + \beta_p X_p$$

and that

$$Var(Y \mid \mathbf{x}) = \phi \nu(\mu)$$

The function g is known as a *link function.*

ϕ is a dispersion parameter.

- *Linear regression*

  $g(\mu) = \mu$, the identity function,

  $\nu(\mu) = 1$.

- *Log linear model*

  Used to model count data, with Poisson family of distributions.

  Link function: $g(\mu) = \ln(\mu)$, and

  $\nu(\mu) = \mu$

- *Logistic regression*

  Used to model binomial data.

  The link function is given by

  $$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$$

  and is known as *logit* .

  $\nu(\mu) = \mu(1-\mu)/n.$

Interpretation of Coefficients

Suppose we are interested in determining the relationship between smoking status and occurrence of lung cancer.

$$X_1 = \begin{cases} 1, & \text{Smoker} \\ 0, & \text{Non-smoker} \end{cases}$$

$$Y = \begin{cases} 1, & \text{Cancer} \\ 0, & \text{No Cancer} \end{cases}$$

Let p_x denote the probability of cancer given the smoking status of the individual. Then

$$p_x = \frac{e^{\beta_0+\beta_1 X_1}}{1+e^{\beta_0+\beta_1 X_1}}$$

In terms of the link function

$$logit(p_x) = \beta_0 + \beta_1 x$$

Now, when $x = 0$, $logit(p_0) = \beta_0$, which gives

$$\exp\{\beta_0\} = \frac{p_0}{1 - p_0}$$

i.e., the odds of cancer for a smoker.

Similarly, when $x = 1$, we see that

$$\beta_1 = logit(p_1) - \beta_0$$

or

$$\beta_1 = \ln \frac{p_1}{1 - p_1} - \ln \frac{p_0}{1 - p_0}$$

β_1 is the *log odds ratio* of having cancer for a smoker relative to a non-smoker.

```
> kyphosis[1:4,]
  Kyphosis Age Number Start
1   absent  71      3     5
2   absent 158      3    14
3  present 128      4     5
4   absent   2      5     1
…….
```

```
>  Age60   <-    1*(Age > 60)
> Y   <-  1*(Kyphosis=="present")
```

```
> table(Age60,Y)
  ..      0     1 ..
     0   26  4
     1   38 13
```

Crude OR:
13*26/(38*4)
2.22

```
> fit1 <-glm(Y~ Age60,family="binomial")
> summary(fit1)
Coefficients:
                  Value          Std. Error      t value
(Intercept) -1.8718019  0.5369109  -3.486243
     Age60   0.7991651  0.6257093  1.277215

> exp(0.8)
[1] 2.225541
```

$$X_2 = \begin{cases} 1, & \text{Old} \\ 0, & \text{Young} \end{cases}$$

$$logit(p \mid x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

When $x_1 = 0$, $logit(p \mid x_1 = 0, x_2) = \beta_0 + \beta_2 x_2$.

Similarly, when $x_1 = 1$,

$$logit(p \mid x_1 = 1, x_2) = \beta_0 + \beta_1 + \beta_2 x_2$$

By subtraction, β_1 is the log odds ratio of having cancer for a smoker relative to a non-smoker, for *any age* group.

Suppose there is interaction between age and smoking status.

Then the logit becomes:

$$logit(p \mid x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

When $x_1 = 0$,

$$logit(p \mid x_1 = 0, x_2) = \beta_0 + \beta_2 x_2$$

Similarly, when $x_1 = 1$, we note that

$$logit(p \mid x_1 = 1, x_2) = \beta_0 + \beta_1 + \beta_2 x_2 + \beta_3 x_2$$

Odds ratio of having cancer for a smoker relative to a nonsmoker is a function of the age group.

When X is *polytomous*, suitably defined *design variables* may be used.

Example. Suppose smoking has three categories:

Never Smoked, Current Smoker, and Smoked in the Past.

Take "Never Smoked" as the reference group and define the design variables as follows:

$$D_1 = \begin{cases} 1, & \text{Current Smoker} \\ 0, & \text{Otherwise} \end{cases}$$

$$D_2 = \begin{cases} 1, & \text{Past Smoker} \\ 0, & \text{Otherwise} \end{cases}$$

The logit is now given by

$$logit(p \mid D_1, D_2) = \beta_0 + \beta_1 D_1 + \beta_2 D_2$$

When $D_1 = 0$ and $D_2 = 0$, i.e., corresponding to Never Smoked,

$$logit(p) = \beta_o,$$

which is the log odds of cancer for someone who never smoked.

When $D_1 = 1$ and $D_2 = 0$,

$$logit(p) = \beta_o + \beta_1.$$

By subtraction, β_1 is the log odds ratio of cancer for a "Current Smoker" relative to someone who never smoked.

Similarly, β_2 is seen to be the log odds ratio of cancer for a Past Smoker relative to the reference group.

It is easy to see that the odds ratio of cancer for a Current Smoker relative to a Past Smoker is

$$e^{\beta_1-\beta_2} = \frac{e^{\beta_1}}{e^{\beta_2}}$$

When X is continuous, e.g., Age in Years.

e^{β} may be interpreted as the odds ratio of cancer for someone of a given age relative to another who is 1 year younger.

Likelihood Inference

Suppose $y \sim f(y;\theta)$. The log likelihood is given by

$$L(\theta) = \sum_{j=1}^{n} \ln f(y_j;\theta)$$

The *deviance* is defined as

$$D(y;\theta) = 2\phi[L(y) - L(\theta)]$$

where $L(y)$ is the saturated model.

In the Gaussian case, when $\hat{\theta}$ is the m.l.e., the deviance corresponds to the RSS.

Analysis of Deviance

Sums of squares for non-normal data are not appropriate measures of contributions of a sum to total variation.

Suppose θ_1 and θ_2 correspond to two competing models.

Difference in deviance is given by

$$D(\theta_1;\theta_2) = D(y;\theta_1) - D(y;\theta_2)$$

Under θ_1, $D(\theta_1;\theta_2)$ is approximately χ^2_ν, where $\nu = \nu_1 - \nu_2$, the difference in the corresponding model degrees of freedom.

For model selection, one would reject the θ_1 model, if the difference is too large, i.e., the model based on θ_2 fits better.

Residuals

- Deviance residuals

  Let d_j denote the contribution of the j'th observation to the deviance. Then

  $$r_j^D = sgn(y_j - \hat{\mu}_j)\sqrt{d_j}$$

  indicates the influence of the j'th observation to the fit.

- Working residuals

  Let

  $$r_j^W = (y_j - \hat{\mu}_j)\frac{\partial \hat{\eta}_j}{\partial \hat{\mu}_j}$$

- Pearson residuals

  $$r_j^P = \frac{(y_j - \hat{\mu}_j)}{\sqrt{Var(\hat{\mu}_j)}}$$

- Response residuals

  $$r_j^R = (y_j - \hat{\mu}_j)$$

Goodness-of-fit: Logistic Regression

Let $Y_1, \cdots, Y_n$ be the observed response, and $\hat{Y}_1, \cdots, \hat{Y}_n$ be the expected values under the model.

Given the vector of covariates $\mathbf{X} = (X_1, \cdots X_p)'$, let m_j be the number of subjects with $\mathbf{X} = \mathbf{x_j}$, $j = 1, \cdots, J < n$, and $\Sigma_j m_j = n$. Let

$$\hat{p}_j = \frac{e^{\mathbf{x_j}'\hat{\beta}}}{1 + e^{\mathbf{x_j}'\hat{\beta}}}$$

so that $\hat{Y}_j = m_j \hat{p}_j$

- *Pearson residuals*

  Let
  $$r_j^P = \frac{(Y_j - m_j\hat{p}_j)}{\sqrt{m_j\hat{p}_j(1-\hat{p}_j)}}$$
  Then $X^2 = \sum_j^J (r_j^P)^2$ has an approximate χ^2_{J-p} distribution under the model.

- *Deviance residuals*

  Let
  $$r_j^D = \pm\left\{2(y_j \ln(\frac{Y_j}{m_j\hat{p}_j}) + (m_j - Y_j)\ln\frac{m_j - Y_j}{m_j(1-\hat{p}_j)})\right\}^2$$
  Then $D = \sum_j^J r_j^D$ has an approximate $\chi^2_{J-(p+1)}$ distribution under the model. The approximation may not be reliable if $J \approx n$.

- *Hosmer-Lemshow Tests*

  These tests require grouping the data based on estimated probabilities.

  - Group data based on percentiles of estimated probabilities

    For n subjects, form **10** groups, each of size $m \approx n/10$. The lowest group then contains those observations having the smallest ten $\hat{p}_j$'s, etc.

  - Collapse the data based on fixed values of estimated probabilities.

    Use as cutpoints, the probabilities $\frac{k}{g}$, $k = \mathbf{1}, 2, \cdots, g - \mathbf{1}$, where g is a suitably defined number of groups.

  Having determined the g classes, let n_k be the number of covariate patterns in the k'th group,

  Let o_k be the number of successes among the n_k covariate patterns of /Hosmer the k'th group.

  Denote the average estimated probability for the k'th group by $\tilde{p}_k$.

Then the Hosmer-Lemshow goodness-of-fit test is given by

$$T_{HL} = \sum_{k=1}^{g} \frac{(o_k - n_k \tilde{p}_k)^2}{n_k \tilde{p}_k (1 - \tilde{p}_k)}$$

and has an approximate χ^2_{g-2} distribution under the model.

Model Selection in GLM

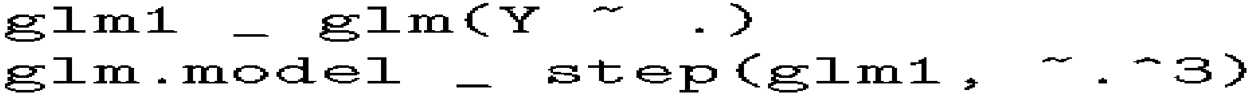
The S-PLUS function *glm()* provides parameter estimates and other inferential results for generalized linear models. S-PLUS also provides several routines for computing the various residuals, e.g., *resid()*.

The *step.glm()* and *step()* functions allow model selection in glm.

- Start with a *glm* object
- Compute selection criterion for entering or removing variables, e.g., C_p.
- Compute the model criterion: e.g., Akaike Information Criterion (AIC)

  $$AIC = Deviance + 2 * scale * df.resid.$$

  Choose model with the smallest AIC.
- Stop when no step will decrease the criterion or a model boundary is reached.



```
glm1 _ glm(Y ~ .)
glm.model _ step(glm1, ~.^3)
```

In R: step()

Library(MASS)

stepAIC

Problem Set 8

Read Chapter 20: Ramsey & Schafer

Consider the Muscular Dystrophy, Exercise 12, page 604:

1) Define “High CK” to have a value of 1 if the value of CK > 60, and 0, otherwise. Fit a logistic regression of “Carrier” on “High CK”.
   a) Estimate the parameters of the regression, and give the associated 95% confidence intervals.
   b) Interpret what the estimated parameters denote.

2) Do Problem 12.