REGRESSION MODELS

IN

MORBIDITY ESTIMATION

Gordon E. Willmot, M.Math, ASA
Actuarial Assistant
Mutual Life of Canada

September, 1982

Regression Models in Morbidity Estimates

Abstract

A statistical model is presented which analyzes group health and dental data. The purpose is to determine the cost of these benefits with special reference to the effects of various characteristics of individuals in the group.

The methodology involves a combination of a binary logistic regression and a normal multiple linear regression.

Regression Models In Morbidity Estimation

We are interested in the group insurance premium for any or all of the following benefits:

i) drug benefits
ii) hospital coverage (private or semi-private)
iii) vision care
iv) dental insurance
v) supplementary health insurance (e.g. ambulance, wheelchair, etc.)

In particular, we wish to determine the effects of an individual's characteristics on the cost of supplying these coverages.

We denote by

$$\underset{\sim}{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$$

a given set of covariates which may include

i) age
ii) sex
iii) occupation
iv) location of residence
v) marital status

among others.

For example, we might have

$$x_{i1} = 1 \quad \text{if } 18 \leq \text{age} < 35$$
$$= 0 \quad \text{otherwise}$$

$$x_{i2} = 1 \quad \text{if } 35 \leq \text{age} < 50$$
$$= 0 \quad \text{otherwise}$$

$x_{i3} = 1$ if $50 \leq$ age < 65
$= 0$ otherwise

$x_{i4} = 1$ if sex is male
$= 0$ if sex is female

etc.

If we denote by X the total claim amount for an individual for the premium paying period then we are interested in the random variable

$$X \mid \underset{\sim}{x}_i$$

or, more specifically, the first couple of moments of its associated distribution.

We assume we have the total amount of claims for each individual in the study for a period of time at least as great as the premium payment period (to eliminate the effects of seasonality on the claims). Furthermore, assume that all claim amounts are strictly positive so that X=0 implies no claims occurred.

Then,

$$\Pr\{X=x|\underset{\sim}{x}_i\} = \Pr\{\geq 1 \text{ claim}|\underset{\sim}{x}_i\} \cdot \Pr\{X=x \mid \underset{\sim}{x}_i, \geq 1 \text{ claim}\}$$

We will estimate the two quantities on the right hand side of this equation separately.

Let us divide the data into n groups such that all covariates in a given group are equal (or roughly so). Suppose in group i there

are n_i individuals, d_i of which have at least 1 claim. Let us assume that $\underset{\sim}{x}_i$ is the covariate vector of the ith group.

Part 1 - Logistic Regression

Let

$$p_i = \Pr \{ \geq 1 \text{ claim} \mid \underset{\sim}{x}_i \}$$

Then it is reasonable to assume that d_i is an observation from a binomial variate with parameters n_i and p_i.

Let

$$p_i = \frac{e^{\beta_1 x_{i1} + \ldots + \beta_p x_{ip}}}{1 + e^{\beta_1 x_{i1} + \ldots + \beta_p x_{ip}}} = \frac{e^{\underset{\sim}{x}_i \underset{\sim}{\beta}}}{1 + e^{\underset{\sim}{x}_i \underset{\sim}{\beta}}} = \frac{e^{\theta_i}}{1 + e^{\theta_i}}$$

where $\underset{\sim}{\beta}^T = \{ \beta_1, \beta_2, \ldots, \beta_p \}$, $\theta_i = \underset{\sim}{x}_i \underset{\sim}{\beta}$

Then we are interested in making inferences about $\underset{\sim}{\beta}$.

For notational simplicity, let

$$\underset{\sim}{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{pmatrix}_{n \times 1} \qquad \underset{\sim}{X} = \begin{pmatrix} \underset{\sim}{x}_1 \\ \underset{\sim}{x}_2 \\ \vdots \\ \underset{\sim}{x}_n \end{pmatrix}_{n \times p} \qquad \underset{\sim}{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}_{p \times 1}$$

To get estimates for $\underset{\sim}{\beta}$, we use likelihood methods.

$$L(\underset{\sim}{\beta}) = \prod_{i=1}^{n} \frac{e^{\underset{\sim}{x}_i \underset{\sim}{\beta} \cdot d_i}}{(1 + e^{\underset{\sim}{x}_i \underset{\sim}{\beta}})^{n_i}} \qquad \text{(since } d_i \text{ is binomial)}$$

$$= \prod_{i=1}^{n} \frac{e^{\theta_i d_i}}{(1 + e^{\theta_i})^{n_i}}$$

Calculate the score function

$$S(\underset{\sim}{\beta}) = \left\{ \frac{\partial}{\partial \beta_i} \log L(\underset{\sim}{\beta}) \right\}_{p \times 1}$$

$$= \left\{ \sum_{i=1}^{n} \left[d_i x_{ij} - \frac{n_i x_{ij} e^{\underset{\sim}{x}_i \underset{\sim}{\beta}}}{1 + e^{\underset{\sim}{x}_i \underset{\sim}{\beta}}} \right] \right\}_{p \times 1}$$

and the information matrix

$$I(\underset{\sim}{\beta}) = \left\{ - \frac{\partial^2 \log L(\underset{\sim}{\beta})}{\partial \beta_j \partial \beta_k} \right\}_{p \times p}$$

$$= \left\{ \sum_{i=1}^{n} \frac{n_i x_{ij} x_{ik} e^{\underset{\sim}{x}_i \underset{\sim}{\beta}}}{(1 + e^{\underset{\sim}{x}_i \underset{\sim}{\beta}})^2} \right\}_{p \times p}$$

then the estimates $\hat{\underset{\sim}{\beta}}$ may be calculated interatively using Newton-Raphson techniques as follows ($\hat{\underset{\sim}{\beta}}_i$ is the value of $\hat{\underset{\sim}{\beta}}$ from the ith interation) until $\hat{\underset{\sim}{\beta}}_i$ is sufficiently close to $\hat{\underset{\sim}{\beta}}_{i-1}$.

$$\hat{\underset{\sim}{\beta}}_i = \hat{\underset{\sim}{\beta}}_{i-1} + I^{-1}(\hat{\underset{\sim}{\beta}}_{i-1}) \cdot S(\hat{\underset{\sim}{\beta}}_{i-1})$$

To get a starting value $\hat{\underset{\sim}{\beta}}_0$, we go back to the likelihood as a function of $\underset{\sim}{\theta}$. We have

$$L(\underset{\sim}{\theta}) = \prod_{i=1}^{n} \frac{e^{\theta_i d_i}}{(1+e^{\theta_i})^{n_i}}$$

and so

$$\hat{\underset{\sim}{\theta}}_j = \log \frac{d_j}{n_j - d_j}$$

It can be shown using large sample theory techniques that (asymptotically)

$$\hat{\theta}_j \sim N\left\{\theta_j, \frac{n_j}{d_j(n_j - d_j)}\right\}$$

or $$\log \frac{d_j}{n_j - d_j} \sim N\left\{x_{i1}\beta_1 + \ldots + x_{ip}\beta_p, \frac{n_j}{d_j(n_j - d_j)}\right\}$$

which is the usual weighted least squares multiple regression model. Thus, if this were the true result, we would have

$$\hat{\underset{\sim}{\beta}} = (V^T V)^{-1} V^T Y$$

where

$$V = \left\{\sqrt{\frac{d_j(n_j-d_j)}{n_j}}\, x_{jk}\right\} \quad n \times p$$

$$Y = \left\{\sqrt{\frac{d_j(n_j-d_j)}{n_j}}\, \log \frac{d_j}{n_j-d_j}\right\}$$

This can thus serve as a good starting value for $\hat{\underset{\sim}{\beta}}$ (ie. as $\hat{\underset{\sim}{\beta}}_0$). Note that if $d_j = 0$, it is recommended to replace

$$\log \frac{d_j}{n_j - d_j} \quad \text{by} \quad \log \frac{0.5}{n_j + 0.5}$$

and if $d_j = n_j$, replace

$$\log \frac{d_j}{n_j - d_j} \quad \text{by} \quad \log \frac{n_j + 0.5}{0.5}$$

in the Y matrix above so that $\hat{\theta}_j \neq \infty$. This does not affect the final answer.

From large sample theory, it can be shown that $\hat{\underset{\sim}{\beta}}$ is approximately normal with mean $\hat{\underset{\sim}{\beta}}$ and variance covariance matrix $I^{-1}(\hat{\underset{\sim}{\beta}})$. Thus, the usual tests can be applied to the solutions $\hat{\underset{\sim}{\beta}}$ for inference purposes (ie. to test whether $\hat{\underset{\sim}{\beta}}_i = 0$ so that the corresponding covariates may be dropped).

To test the fit of the model, a one-tailed test involving the chi-squared statistic

$$\chi^2_{(n-p)} = \sum_{i=1}^{n} \frac{(d_i - n_i \hat{p}_i)^2}{n_i \hat{p}_i (1 - \hat{p}_i)} \qquad \text{where}$$

$$\hat{p}_i = \frac{e^{\underset{\sim}{x}_i \hat{\underset{\sim}{\beta}}}}{1 + e^{\underset{\sim}{x}_i \hat{\underset{\sim}{\beta}}}}$$

(large values give evidence against the model).

Part 2 - Multiple Linear Regression

In order to consider

$$X \mid \underset{\sim}{x}_i \geq 1 \text{ claim}$$

we consider only the d_i individuals who had claims. Let T_{ij} be the total claim amount for person j, group i.

Our model is

$$f(T_{ij}) \sim N\{\tau_1 x_{i1} + \ldots + \tau_p x_{ip}, \sigma^2\}$$

where f is chosen so that the data appears to be roughly normal. Appropriate choices of f include the following

$$f(x) = x$$
$$f(x) = \log x$$
$$f(x) = \sqrt{x}$$

among others.

Let

$$y_{ij} = f(t_{ij}), \quad \bar{y}_i = \sum_{j=1}^{d_j} y_{ij}/d_j$$

Then

$$\bar{Y}_i \sim N\{\tau_1 x_{i1} + \ldots + \tau_p x_{ip}, \sigma^2/d_i\}$$

The least squares estimates of $\underset{\sim}{\tau}$ are then

$$\hat{\underset{\sim}{\tau}} = (V^T V)^{-1} V^T Z$$

where

$$V = \{\sqrt{d_i}\, x_{ij}\}_{n \times p}, \quad Z = \{\sqrt{d_i}\, \bar{y}_i\}_{n \times 1}$$

and

$$\hat{\underset{\sim}{\tau}} \sim N \{ \tau, \sigma^2 (V^T V)^{-1} \}$$

An estimate of σ^2 is

$$S_1^2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{d_i} (y_{ij} - \bar{y}_i)^2}{(\sum_{i=1}^{n} d_i) - n}$$

To test the fit of the model, compute the F statistic

$$F_{n-p,(\sum_{i=1}^{n} d_i)-n} = \frac{\sum_{i=1}^{n} d_i (\bar{y}_i - \hat{\tau}_1 x_{i1} - \ldots - \hat{\tau}_p x_{ip})^2 \div (n-p)}{S_1^2}$$

This is the ratio of the "lack of fit" sum of squares to the "pure error" sum of squares. A one-tailed test is appropriate. If the model is good, a better estimate of σ^2 is

$$S_2^2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{d_i} (y_{ij} - \hat{\tau}_1 x_{i1} - \ldots - \hat{\tau}_p x_{ip})^2}{(\sum_{i=1}^{n} d_i) - p}$$

This problem could have been formulated as a weighted least squares problem with identical results. The use of the mean rather than the original data is primarily to cut down on the computing time.

Part 3 The Distribution

We may now compute the distribution of $X|\underset{\sim}{x}_i$.

The density for $x > 0$ is

$$g_i(x) = \hat{P}_i \cdot \phi\left\{\frac{f(x) - \underset{\sim}{x}_i\hat{\tau}}{\hat{\sigma}}\right\} \cdot \frac{f'(x)}{\hat{\sigma}}$$

where

$$\phi(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2}}$$

Also

$$g_i(0) = Pr\{X = 0|\underset{\sim}{x}_i\} = 1 - \int_{0^+}^{\infty} g_i(x)dx$$

and

$$E\{X|\underset{\sim}{x}_i\} = \int_{0^+}^{\infty} x\, g_i(x)dx$$

For a deductible of amount c, we are interest in the random variable

$$W|\underset{\sim}{x}_i = \max\{X-c, 0\}|\underset{\sim}{x}_i$$

and $Pr\{W = 0|\underset{\sim}{x}_i\} = Pr\{X_i \leq c|\underset{\sim}{x}_i\}$

and the density for $x > 0$ is

$$g_i(x + c)$$

The expected value is

$$\int_c^{\infty} (x - c)\, g_i(x)dx$$

I wish to thank Dr. R. J. Mackay and Dr. J. G. Kalbfleisch of the University of Waterloo for their extremely helpful advice and support in this venture.

References

1) Theoretical Statistics, Cox and Hinkley, 1974, Chapman and Hall, London

2) Applied Regression Analysis, Draper and Smith, 1966, John Wiley and Sons, New York

3) Discrete Data Analysis Lecture Notes, J. G. Kalbfleisch, University of Waterloo, Waterloo