AN INTRODUCTION TO LINEAR MODELS AND GENERALIZED LINEAR MODELS: CONCEPTS AND METHODS

David A. Binder[1]

Univariate statistical models, linear regression models and generalized linear models are briefly reviewed. Examples of a two-way analysis of variance, a three-way analysis of variance and logistic regression for a three way layout are given.

1. INTRODUCTION

The purpose of this presentation is to give a bird's-eye view of some of the concepts used in statistical applications for modelling data

The use of data sampled from a population to estimate means and proportions is now a common practice. In Section 2 we briefly review this concept and describe the interval estimates obtained from constructing confidence intervals.

Linear regression and analysis of variance models are often used to reduce multi-dimensional data to a model consisting of a few parameters. This tool is a valuable device for the analyst looking for a deeper understanding of a complex data set. These methods are reviewed in Section 3.

The concepts of linear regression methods can be extended to a much wider class of models through the generalized linear models described by Nelder and Wedderburn (1972). This is particularly useful when the dependent variable is categorical as opposed to continuous. In Section 4 we review the structure of these models.

Brief mention of appropriate diagnostics to guard against model failure and to detect multicollinearities is given in Section 5.

[1] David A. Binder, Institutional and Agriculture Survey Methods Division, Statistics Canada.

2. UNIVARIATE MODELS

2.1 Binomial Models

Suppose we have a large population from which we will select a sample and we take an observation from each selected unit. If the sample size is n, we denote the observations by $Y_1, Y_2, \ldots, Y_n$. The purpose of collecting this data is that we would like to make some inferences about the population based on this sample. For example, our population could be residents of Canada and our data are defined as

$$Y_j = \begin{cases} 1 & \text{if the person was born in Canada} \\ 0 & \text{if the person was born outside of Canada,} \end{cases}$$

for the j-th individual selected. Based on this sample we would like to make some inferences on the proportion of people in the population who were born in Canada.

If a simple random sample of n = 5000 residents is selected and the actual proportion of persons born in Canada is p = 0.85, then the number of persons in our sample who are born in Canada will be a random variable with a binomial distribution given by

$$f(y) = \binom{5000}{y}(.85)^y(.15)^{n-y}; \quad y = 0, 1, \ldots, 5000.$$

In this case, since we know p = .85, we can completely describe the properties of $Y = \Sigma Y_j$, the total in our sample who are born in Canada. For most statistical applications, though, we do not know all the characteristics of the population and we use our sample to make inferences about this population. For example, suppose we do not know the value of p in the previous example. Then we can say that the number of persons in our sample who were born in Canada will be a binomial random variable having a distribution given by

$$f(y) = \binom{5000}{y}p^y(1-p)^{5000-y}; \quad y = 0, 1, \ldots, 5000.$$

Now, the usual estimator for p, based on this data is $\hat{p} = \bar{Y} = \Sigma\ Y_j/5000$. We let $s(\hat{p}) = \{\hat{p}(1-\hat{p})/(5000)\}^{\frac{1}{2}}$. This is our estimate of the standard error of $\hat{p}$. Now, it turns out that $\hat{p} \pm 1.96\ s(\hat{p})$ is a random interval which has a 95% chance of including the true unknown value of p. This interval is called a 95% confidence interval. By changing the value of 1.96 we would either shorten or lengthen the confidence interval, thus changing the coefficient from 95% to some other value. These coefficients can be obtained from probabilities associated with the standard normal distribution.

We have described the binomial model via a simple random sample from a large population. Thus, all our inferences pertain to that population. However, in many contexts we would like our inferences to relate to other populations which we believe have been generated under similar conditions. For example, the number of deaths in Canada from a particular age-sex group in a given year may be thought of as a single realization from a binomial model, where each individual has the same probability of dying and the individual deaths are essentially independent. If this probability of dying is constant over a number of years then the number of deaths in one year can be used to make inferences for other years, even though the populations are different. (Life insurance companies and their actuaries rely on these types of assumptions in their calculations.) Providing that individual deaths are independent, assumptions about constancy of the probability of death are testable using these binomial models.

It should be pointed out that by using some generalized linear models to be described in Section 4, it may be possible to improve on the assumption of constant probabilities for all individuals, by allowing the probabilities to depend on other factors such as age, sex, health status, smoking habits, weight, etc.

2.2 Normal Models

An important distribution used in modeling data is the normal distribution given by

$$f(y) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y - \mu)^2\right\};\ -\infty < y < \infty.$$

The population mean is μ and is usually the parameter of interest. The population variance is σ^2.

If we observe data $Y_1, Y_2, \ldots, Y_n$ from this population, our usual estimator for μ is $\hat{\mu} = \bar{Y} = \Sigma Y_j/n$. Our estimator for the standard error of $\hat{\mu}$ is given by $s(\hat{\mu}) = s/n^{\frac{1}{2}}$, where

$$s^2 = \Sigma (Y_j - \hat{\mu})^2/(n - 1).$$

As in the case of the binomial model, for large samples the 95% confidence interval is given by $\hat{\mu} \pm 1.96s(\hat{\mu})$. This is a random interval which has a 95% chance of including the true value of μ. For small samples (e.g. $n < 60$), the value 1.96 may be replaced by the appropriate value from the t distribution for more accurate intervals. Other confidence coefficients may also be obtained by changing the value 1.96 to the appropriate percentile from the standard normal or t distribution.

In some applications, the assumption of constant variance is unrealistic, particularly in the linear models to be discussed in Section 3. A simple extension of this model is to assume that the variance of X_i is given by σ_i^2 where $\sigma_i^2 = \sigma^2/w_i$. Here we assume that $w_1, w_2, \ldots, w_n$ are known weights. In this case $\hat{\mu} = \Sigma w_j Y_j/\Sigma w_j$, a weighted average of the data. Also $s(\hat{\mu}) = s/(\Sigma w_j)^{\frac{1}{2}}$, where

$$s^2 = \Sigma w_j(Y_j - \hat{\mu})^2/(n - 1).$$

Confidence intervals for μ are obtained analogously. It should be pointed out here that the weights, $w_1, \ldots, w_n$ are based on the normal model specification and are usually unrelated to sampling weights which are derived from complex survey designs from finite populations. When fitting models to finite populations based on data from a complex survey design, the analyst may wish to incorporate both the model weights as well as the sampling weights in the estimation.

2.3 Exponential Family Models

The binomial and normal models just described can be viewed as special cases of a much wider class of models known as the exponential family. The general form which we will use for this model is given by:

$$f(y_j) = \exp[\kappa_j \{y_j\theta - b(\theta)\} + c(y_j, \kappa_j)],$$

where y_j takes values which do not depend on θ.

We assume $\kappa_j = \kappa w_j$ where $w_1, \ldots, w_n$ are known. In many cases κ will also be known.

Example 1 (Binomial Proportion)

We let $\bar{y}_j = y_j/n_j$ be the sample proportion from a binomial model based on n_j observations. Therefore we have:

$$f(\bar{y}_j) = \binom{n_j}{n_j\bar{y}_j} p^{n_j\bar{y}_j}(1 - p)^{n_j(1 - \bar{y}_j)} : \bar{y}_j = 0, \frac{1}{n_j}, \frac{2}{n_j}, \ldots, 1,$$

$$E(\bar{y}_j) = p, \quad Var(\bar{y}_j) = p(1 - p)/n_j,$$

$$\theta = \log[p/(1 - p)].$$

$$\kappa_j = n_j,$$

$$b(\theta) = \log(1 + e^{\theta}).$$

Example 2 (Normal)

Suppose y_j is normally distributed with mean μ and variance σ_j^2. We have:

$$f(y_j) = (2\pi\sigma_j^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\frac{y_j - \mu}{\sigma_j}\right)^2\right\} : -\infty < y_j < \infty$$

$$E(y_j) = \mu, \qquad Var(y_j) = \sigma_j^2,$$

$$\theta = \mu,$$

$$\kappa_j = 1/\sigma_j^2,$$

$$b(\theta) = \mu^2/2.$$

Example 3 (Poisson Mean)

Suppose y_j is Poisson with mean $n_j\lambda$. Letting $\bar{y}_j = y_j/n_j$, we have:

$$f(\bar{y}_j) = e^{-n_j\lambda}(n_j\lambda)^{n_j\bar{y}_j}/(n_j\bar{y}_j)!; \qquad \bar{y}_j = 0, \frac{1}{n_j}, \frac{2}{n_j}, \ldots,$$

$$E(\bar{y}_j) = \lambda, \qquad Var(\bar{y}_j) = \lambda/n_j,$$

$$\theta = \log \lambda,$$

$$\kappa_j = n_j,$$

$$b(\theta) = e^{\theta}.$$

Example 4 (χ^2)

Suppose y_j has a $\sigma^2\chi^2_{\nu_j}/\nu_j$ distribution. This is common for analysis of variance and variance components models, where y_j is the mean-square. Then, we have:

$$f(y_j) = y_j^{(\nu_j - 2)/2}\left(\frac{\nu_j}{2\sigma^2}\right)^{\nu_j/2}\exp\{-y_j\nu_j/(2\sigma^2)\}/\Gamma(\nu_j/2); \; y_j \geq 0,$$

$$E(y_j) = \sigma^2, \qquad Var(y_j) = 2\sigma^4/\nu_j,$$

$$\theta = -1/\sigma^2,$$

$$\kappa_j = \nu_j/2,$$

$$b(\theta) = -\log(-\theta).$$

As we can see from these examples, the exponential family includes a wide variety of common distributions. In general, we have

$$E(y_j) = b'(\theta) = \mu, \qquad Var(y_j) = b''(\theta)/\kappa_j = V_j$$

where $b'(\cdot)$ and $b''(\cdot)$ denote the first and second derivatives of $b(\cdot)$.

If $y_1, \ldots, y_n$ are independent, then the maximum likelihood estimate of θ is given by the solution to:

$$\hat{\mu} = \Sigma\, \kappa_j y_j / \Sigma\, \kappa_j = \Sigma\, w_j y_j / \Sigma\, w_j$$

where $\hat{\mu} = b'(\hat{\theta})$. This implies that there is a large family of models where a weighted sample mean provides an efficient estimator of the population mean. The estimated variance of $\hat{\mu}$ is given by

$$\hat{V}(\hat{\mu}) = (\Sigma\, \kappa_j^2 \hat{V}_j)/(\Sigma\, \kappa_j)^2$$

$$= b''(\hat{\theta})/(\Sigma\, \kappa_j).$$

For large samples, the 95% confidence interval for μ is given by $\mu \pm 1.96 \times \{\hat{V}(\hat{\mu})\}^{\frac{1}{2}}$, providing the model is true.

In cases where $\kappa_j = \kappa w_j$ is known only up to the constant of proportionality κ, (e.g. normal model), it will be necessary to estimate the value of κ. The maximum likelihood estimate is given by the solution to:

$$\Sigma\, w_j \left[y_j \theta - b(\theta) + \frac{\partial c(y_j, \kappa_j)}{\partial \kappa_j} \right] = 0.$$

Alternatively, an unbiased estimator for $\hat{V}(\hat{\mu})$ which is less model-dependent is given by

$$\hat{V}_1(\hat{\mu}) = \frac{\Sigma\, w_j (y_j - \hat{\mu})^2}{(n-1)(\Sigma\, w_j)}.$$

This may be used instead to create the confidence intervals for $\hat{\mu}$. The

main assumption required for the validity of this approach is that $Var(y_j) \propto 1/w_j$.

3. LINEAR MODELS

3.1 One Way Analysis of Variance

A simple extension of the univariate normal models, described in Section 2.2, is the one-way analysis of variance (ANOVA) model. Here, in addition to observing one characteristic from each individual sampled, we also have a sub-population identifier. Some such identifiers could be age-sex groups, industry/occupation groups, etc. Here the model could be written as

$$y_{ij} = \mu_i + \varepsilon_{ij};\ i = 1, \ldots, I;\ j = 1, \ldots, n_i,$$

where the μ's are population means, which differ among subpopulations and the ε's are assumed to be independent normal with variances $\sigma^2_{ij} = \sigma^2/w_{ij}$, where the w_{ij}'s are known weights. In most applications the weights are constant.

The usual estimator for μ_i in this model is

$$\hat{\mu}_i = \sum_j w_{ij}y_{ij}/\sum_j w_{ij}.$$

Under the model assumptions, the estimated means are independent normal with $E(\hat{\mu}_i) = \mu_i$ and $Var(\hat{\mu}_i) = \sigma^2/\sum_j w_{ij}$. From this, confidence intervals for the individual means may be derived.

An alternative but equivalent description of this model is

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

where $\Sigma\Sigma\ w_{ij}\alpha_i = 0$. Here we have

$$\mu = \Sigma\Sigma\ w_{ij}\mu_i/\Sigma\Sigma\ w_{ij}$$

$$\alpha_i = \mu_i - \mu.$$

An extension of this representation is particularly useful for two-way and higher order analysis of variance models, to be discussed in Sections 3.2 and 3.3. One of the main questions of interest for these models is whether all the means are equal. This is equivalent to $\mu_1 = \mu_2 = \ldots = \mu_I$ or $\alpha_1 = \alpha_2 = \ldots = \alpha_I = 0$. Standard ANOVA statistical packages (e.g. SAS, SPSS, etc.) are available to test these hypotheses. A related problem is: Which subpopulation means are equal, given that we have concluded already that not all means are equal? When we have no further structure (such as in a two-way ANOVA), this is known as the multiple comparison problems. Special treatments for this problem are available in many statistical packages.

3.2 Two-Way Analysis of Variance

The data of Table 1 has been taken from the 1975 Sri Lanka Fertility Survey (see Little, 1982). The cell means describe the average number of children ever born cross-classified by Marital Duration and Level of Education.

The row and column means seem to indicate that the average number of children increases with longer marriage durations and decreases with more schooling. Now, the two-way analysis of variance model may be written as

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

where the ε's are assumed to be independent normal with variances $\sigma^2_{ijk} = \sigma^2/w_{ijk}$. The w's are known weights. In most applications the weights are constant. In order to estimate the parameters of this model, it is necessary to impose constraints on these parameters, otherwise they are not unique. The usual side conditions are:

$$\sum_i \sum_j \sum_k w_{ijk}\alpha_i = 0,$$

$$\sum_i \sum_j \sum_k w_{ijk}\beta_j = 0,$$

$$\sum_i \sum_k w_{ijk}\gamma_{ij} = 0,$$

$$\sum_j \sum_k w_{ijk}\gamma_{ij} = 0.$$

The estimators are defined by the equations:

$$\sum_i \sum_j \sum_k w_{ijk}(y_{ijk} - \hat{\mu}_{ij}) \frac{\partial \hat{\mu}_{ij}}{\partial \hat{\theta}_\ell} = 0$$

where $\hat{\theta}_1$, $\hat{\theta}_2$, ... correspondent to the parameter estimates $\hat{\mu}$, $\hat{\alpha}_i$, etc. The α's are β's are referred to as main effects and the γ's are the two-way interactions. This results in the following estimators:

$$\hat{\mu} = \bar{y}_{...},$$

$$\hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...} - \sum_j \sum_k w_{ijk}\hat{\beta}_j / \sum_j \sum_k w_{ijk},$$

$$\hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...} - \sum_i \sum_k w_{ijk}\hat{\alpha}_i / \sum_i \sum_k w_{ijk},$$

$$\gamma_{ij} = \bar{y}_{ij.} - \bar{y}_{...} - \hat{\alpha}_i - \hat{\beta}_j,$$

where $\bar{y}_{ij.}$, $\bar{y}_{i..}$, etc. are the appropriate weighted averages.

Now, the additive model specifies that $\mu_{ij} = \mu + \alpha_i + \beta_j$. We have plotted the cell means from Table 1 in Figure 1. The additive model would specify that all the lines are parallel. If the data of Table 1 are fitted to the additive model, we obtain the adjusted mean values in Table 2. These are plotted in Figure 2. As we can see, the effect of the level of education has been dramatically reduced after fitting this model. This is because the more educated women were not married for as long, so that the years since first marriage proves to be the important factor. However, as the analysis of variance in Table 3 shows, all the main effects and the interactions are significant. Hence the additive model is rejected. However, only 0.4% of the total variation is explained by the Education-Marital Durations interactions, whereas 49.7% of the variation is explained by the additive model. We may surmise from this that the additive model has led to a better understanding of the data and that the Education effect is not as dramatic as it first

seemed.

3.3 Regression Formulation

The above analysis of variance models can be considered as special cases of the multiple linear regression model, given by

$$y_j = \beta_0 X_{0j} + \beta_1 X_{1j} + \ldots + \beta_r X_{rj} + \varepsilon_j,$$

where X_{0j}, X_{1j}, ..., X_{rj} are known constants and β_0, β_1, ..., β_r are unknown coefficients. We assume that the ε's are independent normal with variances $\sigma_j^2 = \sigma^2/w_j$, where the w_j's are known weights. For example, in the one way analysis of variance, we could let

$X_{0j} = 1$ for all j

$X_{ij} = 1$ if the j-th individual is in the i-th sub-population

$= - a_i/a_I$ if the j-th individual is in the I-th sub-population

$= 0$ otherwise,

for $i = 1, \ldots, I - 1$, where a_i is the sum of the weights for individuals in the i-th sub-population. In this case we have

$$\mu_i = \beta_0 + \beta_i \qquad \text{for } i = 1, \ldots, I - 1,$$

$$\mu_I = \beta_0 - (a_1\beta_1 + \ldots + a_{I-1}\beta_{I-1})/a_I.$$

Therefore $\mu = \beta_0$ and $\alpha_i = \beta_i$ for $i = 1, \ldots, I - 1$.

A similar regression formulation is possible for two-way and higher order layouts as well.

Now, for the general regression model, the estimator for $\beta_0, \ldots, \beta_r$ is given by $\hat{\beta}_0, \ldots, \hat{\beta}_r$, the solution to

$$\Sigma\, w_j(y_j - \hat{y}_j)X_{ij}, \qquad i = 0, 1, \ldots, r$$

where $\hat{y}_j = \hat{\beta}_0 X_{0j} + \hat{\beta}_1 X_{ij} + \ldots + \hat{\beta}_r X_{rj}$.

In order to test hypotheses, perform model-building and develop confidence intervals for the β's, we need the covariance matrix of the $\hat{\beta}$'s. This is given by

$$\text{Var}(\underset{\sim}{\hat{\beta}}) = \sigma^2 A^{-1}$$

where A is the matrix with (k,ℓ)-th entry being $\sum_j w_j X_{kj} X_{\ell j}$. To estimate σ^2, we use $\hat{\sigma}^2 = \sum_j w_j (y_j - \hat{y}_j)^2/(n - r - 1)$.

Many statistical packages routinely perform various hypothesis tests on $\hat{\beta}$ using the estimated covariance matrix $\hat{\sigma}^2 A^{-1}$ and the critical values from the appropriate F-distribution (e.g. PROC REG, PROC ANOVA and PROC GLM in SAS).

For example, Koch, Gillings and Stokes (1980) give the data in Table 4 for the number of physician visits per person per year in 1973 in the U.S. cross-classified by size of city (SMSA = Standard Metropolitan Statistical Area vs. Non-SMSA), Income (3 groups) and Education (3 groups). This data is based on the 1973 Health Interview Survey, a survey using a complex probability sample. The data are illustrated in Figure 3.

By using a regression model and performing a number of statistical tests, the following reduced model was obtained:

$$E(Y_j) = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j},$$

where X_{1j} = 1 if the j-th person is in an SMSA
= 0 otherwise,

X_{2j} = 1 if the j-th person has less than \$5000 family income or more than 12 years education for the family head
= 0 otherwise.

The estimated parameters were $\hat{\beta}_0 = 4.18$ (standard error of 0.11), $\hat{\beta}_1 = 0.65$ (standard error of 0.11) and $\hat{\beta}_2 = 1.12$ (standard error of 0.09). The standard errors derived here were not those described above since the authors used the 18×18 estimated covariance matrix from the survey to obtain the standard errors. This approach removes the assumption of independent error terms in

the model-fitting and is a common approach for analysing data from complex surveys.

In Table 5 we summarize the results. These are illustrated in Figure 4. We see that the model fit is quite good. We have reduced the data from 18 values to 3 summary statistics and also have smaller standard errors (hence higher precision) of the estimated values.

4. GENERALIZED LINEAR MODELS

4.1 Regression with a Dichotomous Dependent Variable

One of the difficulties often encountered with the linear models discussed in Section 3 is that the error terms were assumed to be normally distributed. It is true that analyses similar to those in Section 3 may be performed with non-normal errors, providing the variances of the errors still satisfy $\sigma_j^2 = \sigma^2/w_j$ and the errors are uncorrelated. In this case the estimators we have described yield the minimum variance linear unbiased estimates of the model parameters, however better estimators (i.e. non-linear estimators) may be available. These considerations have led to generalized linear models (see Nelder and Wedderburn, 1972) and robust estimators (see Huber, 1973). We concentrate here on the generalized linear models.

For example, suppose the dependent variable, y_j, can take on only two values, 0 or 1. We now want to model $p_j = Pr(Y_j = 1)$ as a function of the linear expression $X_{0j}\beta_0 + X_{1j}\beta_1 + \ldots + X_{rj}\beta_r$. There are three popular approaches for this problem. One is to let $\hat{\beta}_0, \ldots, \hat{\beta}_r$ be the usual estimate from a stardard regression model. This is analogous to discriminant analysis where the variables $X_{0j}, \ldots, X_{rj}$ are not considered fixed known constants, but are themselves random variables (multivariate normal with constant covariance matrix) whose mean depends on the value of Y_j. The problem with this approach is that $\hat{Y}_j = X_{0j}\hat{\beta}_0 + \ldots + X_{rj}\hat{\beta}_r$ cannot be used directly to predict the value of p_j. Also, in many applications the X_{ij}'s are categorical, (e.g. province, occupation, etc.), thus violating the assumption of multivariate normality.

Two other popular approaches are known as probit analysis and logistic

regression. In probit analysis it is assumed that $p_j = \Phi(\sum_i X_{ij}\beta_i)$, where Φ is the cumulative distribution function of a standard normal random variable. In logistic regression, it is assumed that

$$\theta_j = \log[p_j/(1 - p_j)] = \sum_i X_{ij}\beta_i.$$

Both these approaches are valuable analytic tools, and are available in many statistical packages (e.g. SAS, BMDP). The two approaches may be viewed together by letting

$$\eta_j = g(p_j) = \sum_i X_{ij}\beta_i.$$

For probit analysis we have $\eta_j = \Phi^{-1}(p_j)$, whereas for logistic regression we have $\eta_j = \log [p_j/(1 - p_j)]$. The maximum likelihood estimate for $\beta_0, \ldots, \beta_r$ is the solution to

$$\sum_j \frac{(y_j - \hat{p}_j)X_{ij}}{\hat{p}_j(1 - \hat{p}_j)g'(\hat{p}_j)} = 0, \qquad \text{for } i = 0, \ldots, r,$$

where $g(\hat{p}_j) = \sum_i X_{ij}\hat{\beta}_i$. These equations often must be solved iteratively. For the probit analysis we have

$$g'(p_j) = \frac{1}{\phi[\Phi^{-1}(p_j)]}$$

where $\phi(\cdot)$ is the standard normal density function. For the logistic regression,

$$g'(p_j) = [p_j(1 - p_j)]^{-1}$$

so that the parameter estimate is given by the solution to

$$\sum_j (y_i - \hat{p}_j)X_{ij} = 0, \qquad \text{for } i = 0, \ldots, r.$$

The covariance matrix of $\hat{\beta}_0, \ldots, \hat{\beta}_r$ is A^{-1} where A is a matrix with (k, ℓ)-th entry given by

$$A_{k\ell} = \sum_j \frac{X_{kj}X_{\ell j}}{p_j(1 - p_j)\{q'(p_j)\}^2}$$

This can be used to construct confidence intervals and perform hypothesis tests and model-building.

For logistic regression, the covariance simplifies to

$$A_{k\ell} = \sum_j p_j(1 - p_j)X_{kj}X_{\ell j}.$$

As an example of the utility of these models, we consider an unpublished analysis performed by Dolson and Morin on the Canadian Health and Disability Survey. The dependent variable was whether or not a person would be screened in as potentially disabled using the Screening Test 2 of the January 1983 Labour Force Supplement on Disability. For details, see Dolson and Morin (1983). Analysis was restricted to males aged 15-64. Of the 13,897 respondents, 14.4% (unweighted) were screened in. The screened-in rates are cross-classified by age-groupings, labour force participation and a proxy/non-proxy variable (with 3 levels: non-proxy, proxy by male or proxy by female) in Table 6. (The fitted values from the model to be discussed below are also shown.) The data are illustrated in Figure 5.

The fitted model reduced the number of parameters from 30 to 11. The final model was given by

$$\log[p_{ijk}/(1 - p_{ijk})] = \mu + \alpha_i + \beta_j + \gamma_k + \delta_{ij},$$

where $\sum \alpha_i = \sum \beta_j = \sum \gamma_k = 0$, $\sum_j \delta_{ij} = 0$, $\sum_i \delta_{ij} = 0$, for the i-th age group, j-th labour force status and k-th proxy status (2 levels: non-proxy vs. proxy). The following were the estimated parameters.

Parameter	Subscript	Estimate
μ		-1.43
α	Age 15-24	-1.12
	Age 25-34	-0.571
	Age 35-44	0.0143
	Age 45-54	0.629
	Age 55-64	1.05
β	In Labour Force	-0.576
	Not in Labour Force	0.576
γ	Non-proxy	0.0859
	Proxy	-0.0859
δ	Age 15-24, in L.F.	0.385
	Age 25-34, in L.F.	0.0938
	Age 35-44, in L.F.	-0.175
	Age 45-54, in L.F.	-0.243
	Age 55-64, in L.F.	-0.0612
	Age 15-24, not in L.F.	-0.385
	Age 25-34, not in L.F.	-0.0938
	Age 35-44, not in L.F.	0.175
	Age 45-54, not in L.F.	0.243
	Age 55-64, not in L.F.	0.0612

The fitted values are illustrated in Figure 6.

We see that even after adjusting for age and labour force status, there is a proxy effect on the screening rates. This proxy effect does not seem to depend on the sex of the proxy respondent. Also, there is no interaction between the proxy and the age/labour force status variables. This model does not necessarily imply a proxy bias, but it indicates that a proxy bias may potentially be present. Without a special study such as a re-interview program for the proxy respondent, it is impossible to definitively conclude the existence of a proxy bias.

4.2 Generalized Linear Models

In the previous section we discussed a large class of linear models related to the binomial model, of which probit analysis and logistic regression were special cases. We now extend these to the exponential family as proposed by Nelder and Wedderburn (1972).

As in Section 2.3, we assume y_j has probability function given by

$$f(y_j) = \exp[\kappa_j \{y_j \theta_j - b(\theta_j)\} + c(y_j, \kappa_j)],$$

where $\mu_j = E[Y_j] = b'(\theta_j)$ and $V_j = \mathrm{Var}[Y_j] = b''(\theta_j)/\kappa_j$.

We let $\eta_j = q(\mu_j) = \sum_i X_{ij}\beta_i$ be the linear component of the model, where $q(\cdot)$ is a known function.

Now the maximum likelihood estimates of $\underset{\sim}{\beta}$ are given by the solution to

$$\sum_j \frac{(y_j - \hat{\mu}_j)X_{ij}}{\hat{V}_j[q'(\hat{\mu}_j)]} = 0.$$

Nelder and Wedderburn (1972) have shown that a reasonable method for estimating $\underset{\sim}{\beta}$ is given by performing a number of weighted least-squares regressions, updating the weights and the dependent variables on successive iterations. This is called iteratively re-weighted least squares. In particular, the weights for the t-th iteration are given by

$$\hat{w}_j^{(t)} = \frac{1}{\hat{V}_j^{(t)}[q'(\hat{\mu}_j^{(t)})]^2}$$

and the dependent variables on the t-th iteration are given by

$$\hat{Z}_j^{(t)} = q(\hat{\mu}_j^{(t)}) + q'(\hat{\mu}_j^{(t)})(y_j - \hat{\mu}_j^{(t)}).$$

The (t + 1)-th iteration of $\hat{\underset{\sim}{\beta}}$ is then the solution to

$$\sum_j \hat{w}_j^{(t)}[\hat{Z}_j^{(t)} - \sum_\ell X_{\ell j}\hat{\beta}_\ell^{(t+1)}] X_{kj} = 0.$$

The estimated covariance matrix of $\hat{\underset{\sim}{\beta}}$ is given by A^{-1} where the (k, ℓ)-th entry for A is

$$A_{k\ell} = \sum_j \hat{w}_j X_{kj} X_{\ell j}.$$

This implies that many standard weighted least-squares packages could be invoked to perform analysis of these generalized linear models.

For example, a common analysis of contingency tables, called log-linear models assumes a basic Poisson model with $\log \mu_j = \sum_i X_{ij}\beta_i$. Here we have

$$V_j = \mu_j,$$

$$q(\mu_j) = \log \mu_j,$$

so that the iteratively reweighted solution is given by assigning

$$\hat{w}_j^{(t)} = \hat{\mu}_j^{(t)},$$

$$\hat{Z}_j^{(t)} = \log \hat{\mu}_j^{(t)} + \frac{y_j - \hat{\mu}_j^{(t)}}{\hat{\mu}_j^{(t)}}.$$

Hence, models similar to those described in Section 3 can be analyzed analogously using the generalized linear model formulation.

5. DIAGNOSTICS

Linear regression methods have been known now for over a century; see Hocking (1983) for a review of developments over the last 25 years. In more recent years attention has been focused on difficulties encountered when there is multicollinearity in the variables (leading to large variances of the parameter estimates) and when the models may fail. Some of these diagnostics are now available in SAS and SPSS-X.

The methods discussed in this paper extend linear regression to a much wider class of problems. Newer diagnostic techniques for models of this sort

are discussed in Landwehr, Pregibon and Shoemaker (1984).

In many statistical applications, the proposed model is only used as an approximation to reality. Therefore, the user of these models should employ these diagnostic tools in the course of the analysis.

REFERENCES

[1] Dolson, D. and Morin, J.-P. (1983). Disability data development project: Analysis of screening questionnaires. Technical Report, Health Division, Statistics Canada.

[2] Hocking, R.R. (1983). Developments in linear regression methodology: 1959-1982 (with discussion). Technometrics, 25, pp. 219-249.

[3] Huber, P.J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. Ann. Statist., 1, pp. 799-821.

[4] Koch, G.G., Gillings, D.B., and Stokes, M.E. (1980). Biostatistical implications of design, sampling and measurement to health science data analysis. Ann. Rev. Public Health, 1, pp. 163-225.

[5] Landwehr, J.M., Pregibon, D., and Shoemaker, A. (1984). Graphical methods for assessing logistic regression models (with discussion). J. Amer. Statist. Assoc., 79, pp. 61-83.

[6] Little, R.J.A. (1982). Direct standardization: A tool for teaching linear models for unbalanced data. Amer. Statist., 36, pp. 38-43.

[7] Nelder, J.A., and Wedderburn, R.W.M., (1972). Generalized linear models. J. Roy. Statist. Soc., Ser. A, 135, pp. 370-384.

Table 1: Mean Number of Children Ever Born, by Marital Duration and Education Level. Sri Lanka 1975 (from Little, 1982)

Years since First Marriage		Level of Education				
		No School	1 - 5 Years	6 - 9 Years	10+ Years	Row
0 - 4	Mean	0.96	0.88	0.95	0.92	0.92
	Count	112	376	442	351	1281
5 - 9	Mean	2.54	2.46	2.39	2.39	2.44
	Count	172	442	362	255	1231
10 - 14	Mean	3.87	3.91	3.73	3.14	3.76
	Count	197	482	293	145	1117
15 - 19	Mean	5.13	4.97	4.61	4.13	4.84
	Count	239	461	262	95	1057
20 - 24	Mean	6.22	5.87	5.22	4.47	5.79
	Count	292	377	184	40	893
25+	Mean	6.92	6.55	6.23	5.97	6.65
	Count	501	548	161	22	1232
Column	Mean	5.17	4.24	3.26	2.30	3.94
	Count	1513	2686	1704	908	6811

Table 2: Interactions for Mean Number of Children from Table 1

Years Since First Marriage		Level of Education				
		No School	1 - 5 Years	6 - 9 Years	10+ Years	Row
0 - 4	Raw Mean	0.96	0.88	0.95	0.92	0.92
	Adjusted Mean	1.31	1.07	0.86	0.71	1.02
	Interaction	-0.35	-0.19	0.09	0.21	
5 - 9	Raw Mean	2.54	2.46	2.39	2.39	2.44
	Adjusted Mean	2.78	2.54	2.33	2.18	2.49
	Interaction	-0.24	-0.08	0.06	0.21	
10 - 14	Raw Mean	3.87	3.91	3.73	3.14	3.76
	Adjusted Mean	4.06	3.82	3.61	3.46	3.77
	Interaction	-0.19	0.09	0.12	-0.32	
15 - 19	Raw Mean	5.13	4.97	4.61	4.13	4.84
	Adjusted Mean	5.11	4.87	4.66	4.51	4.82
	Interaction	0.02	0.10	-0.05	-0.38	
20 - 24	Raw Mean	6.22	5.87	5.22	4.47	5.79
	Adjusted Mean	6.01	5.77	5.56	5.41	5.72
	Interaction	0.21	0.10	-0.34	-0.94	
25+	Raw Mean	6.92	6.55	6.23	5.97	6.65
	Adjusted Mean	6.82	6.58	6.37	6.22	6.53
	Interaction	0.10	-0.03	-0.14	-0.25	
Column		5.17	4.24	3.26	2.30	3.94
		4.23	3.99	3.78	3.63	3.94

Table 3: Analysis of Variance of Data from Table 1

Source	Sum of Squares	Proportion of Total SS	DF	Mean Square	F	Signif. of F
Main Effects						
Marital Duration	27402.684	0.493	5	5480.537	1340.990	.000
Education/Duration	225.535	0.004	3	75.178	18.395	.000
Interactions						
Duration×Education	206.965	0.004	15	13.798	3.376	.000
Residual	27729.848	0.499	6787	4.986		
Total	55565.031		6810			

Table 4. Physician Visits per Person per Year by Residence Size, Family Income and Education of Family Head, U.S. 1973

Education in Years	Family Income		
	0 - 4999	5000 - 14999	15000 or more
SMSA			
Less than 12	6.15 (0.18)	4.73 (0.13)	4.82 (0.25)
12	6.17 (0.41)	4.98 (0.17)	4.70 (0.18)
More than 12	6.31 (0.49)	6.08 (0.19)	5.66 (0.16)
Non-SMSA			
Less than 12	5.08 (0.26)	4.14 (0.15)	4.42 (0.37)
12	5.36 (0.44)	4.32 (0.19)	4.49 (0.33)
More than 12	4.58 (0.58)	5.06 (0.29)	4.48 (0.31)

Note: Bracketed figures indicate standard errors of estimate.

Table 5. Estimated Physician Visits from Table 4, Original and Fitted Values

Education (in Years)		Family Income		
		0 - 4999	5000 - 14999	15000 or more
SMA				
Less than 12	Original	6.15 (0.18)	4.73 (0.13)	4.82 (0.25)
	Fitted	5.95 (0.07)	4.83 (0.07)	4.83 (0.07)
	Difference	0.20	-0.10	-0.01
12	Original	6.17 (0.41)	4.98 (0.17)	4.70 (0.18)
	Fitted	5.95 (0.07)	4.83 (0.07)	4.83 (0.07)
	Difference	0.22	0.15	-0.13
More than 12	Original	6.31 (0.49)	6.08 (0.19)	5.66 (0.16)
	Fitted	5.95 (0.07)	5.95 (0.07)	5.95 (0.07)
	Difference	0.36	0.13	-0.29
Non-SMSA				
Less than 12	Original	5.08 (0.26)	4.14 (0.15)	4.42 (0.37)
	Fitted	5.30 (0.11)	4.18 (0.11)	4.18 (0.11)
	Difference	-0.22	-0.04	0.24
12	Original	5.36 (0.44)	4.32 (0.19)	4.49 (0.33)
	Fitted	5.30 (0.11)	4.18 (0.11)	4.18 (0.11)
	Difference	0.06	0.14	0.31
More than 12	Original	4.58 (0.58)	5.06 (0.29)	4.48 (0.31)
	Fitted	5.30 (0.11)	5.30 (0.11)	5.30 (0.11)
	Difference	-0.72	-0.24	-0.82

Table 6· Unadjusted and Fitted Screened-in Rates from Test 2, Canadian Health and Disability Survey, Males Aged 15-64, by Labour Force Participation and Proxy Status, Canada January 1983 (Unweighted)

Age		Non-Proxy	Male Proxy	Female Proxy
In Labour Force				
15 - 24	Unadjusted	.065(.0067)	.055(.0143)	.056(.0069)
	Fitted	.065(.0051)	.056(.0044)	.056(.0044)
	Difference	.000	-.001	.000
25 - 34	Unadjusted	.085(.0058)	.058(.0252)	.069(.0069)
	Fitted	.085(.0048)	.071(.0046)	.071(.0046)
	Difference	.000	-.013	-.002
35 - 44	Unadjusted	.113(.0079)	.029(.0290)	.094(.0086)
	Fitted	.111(.0064)	.093(.0059)	.093(.0059)
	Difference	.002	-.064	.001
45 - 54	Unadjusted	.180(.0109)	.082(.0351)	.154(.0120)
	Fitted	.177(.0088)	.153(.0083)	.153(.0083)
	Difference	.003	-.071	.001
55 - 64	Unadjusted	.284(.0150)	.207(.0752)	.250(.0183)
	Fitted	.283(.0124)	.249(.0124)	.249(.0124)
	Difference	.001	-.042	.001
Not in Labour Force				
15 - 24	Unadjusted	.104(.0127)	.071(.0190)	.074(.0084)
	Fitted	.104(.0078)	.079(.0065)	.079(.0065)
	Difference	.000	-.008	-.005
25 - 34	Unadjusted	.146(.0239)	.367(.1450)	.227(.0365)
	Fitted	.192(.0213)	.167(.0194)	.167(.0194)
	Difference	-.046	.200	.060
35 - 44	Unadjusted	.348(.0372)	.455(.1501)	.324(.0544)
	Fitted	.359(.0309)	.320(.0299)	.320(.0299)
	Difference	-.011	.135	.004
45 - 54	Unadjusted	.534(.0361)	.625(.1712)	.454(.0505)
	Fitted	.525(.0293)	.483(.0301)	.483(.0301)
	Difference	.009	.142	-.029
55 - 64	Unadjusted	.571(.0220)	.563(.1240)	.591(.0420)
	Fitted	.585(.0194)	.543(.0217)	.543(.0217)
	Difference	-.014	.020	.048

NOTE: Bracketed figures are Standard Errors

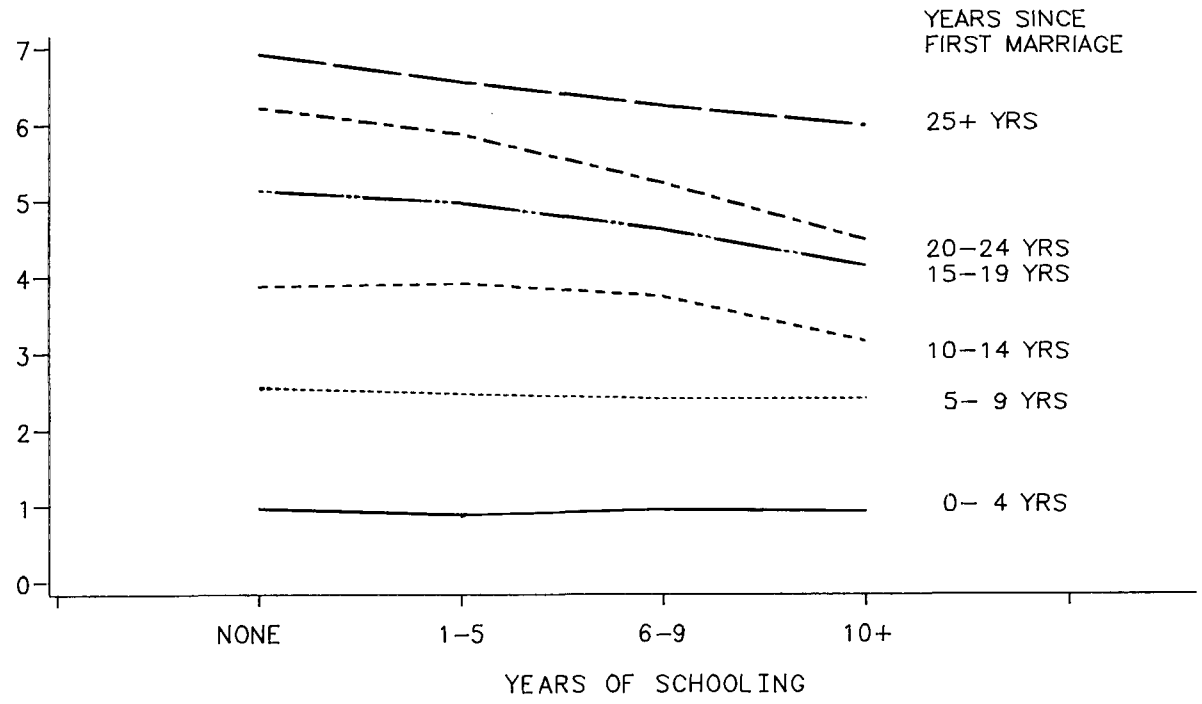

Figure 1: Observed Means from Sri Lanka Fertility Survey, 1975. Data source: Little (1982).

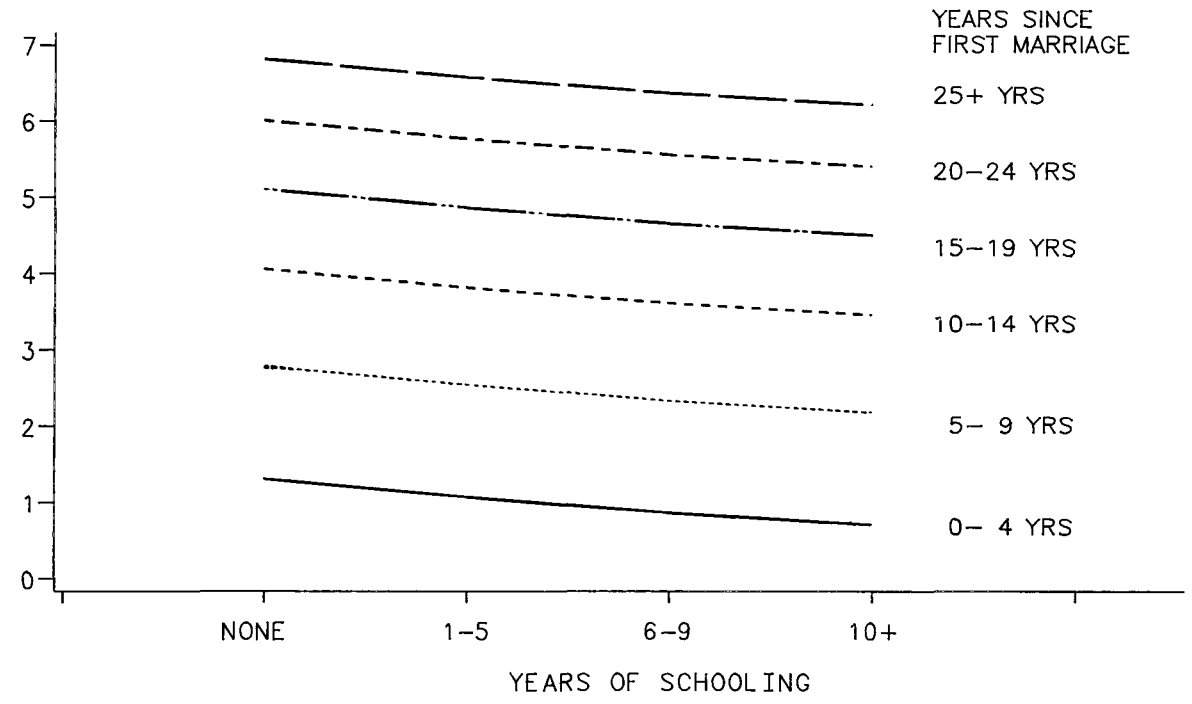

Figure 2: Adjusted Means from Sri Lanka Fertility Survey, 1975. Data source: Little (1982)

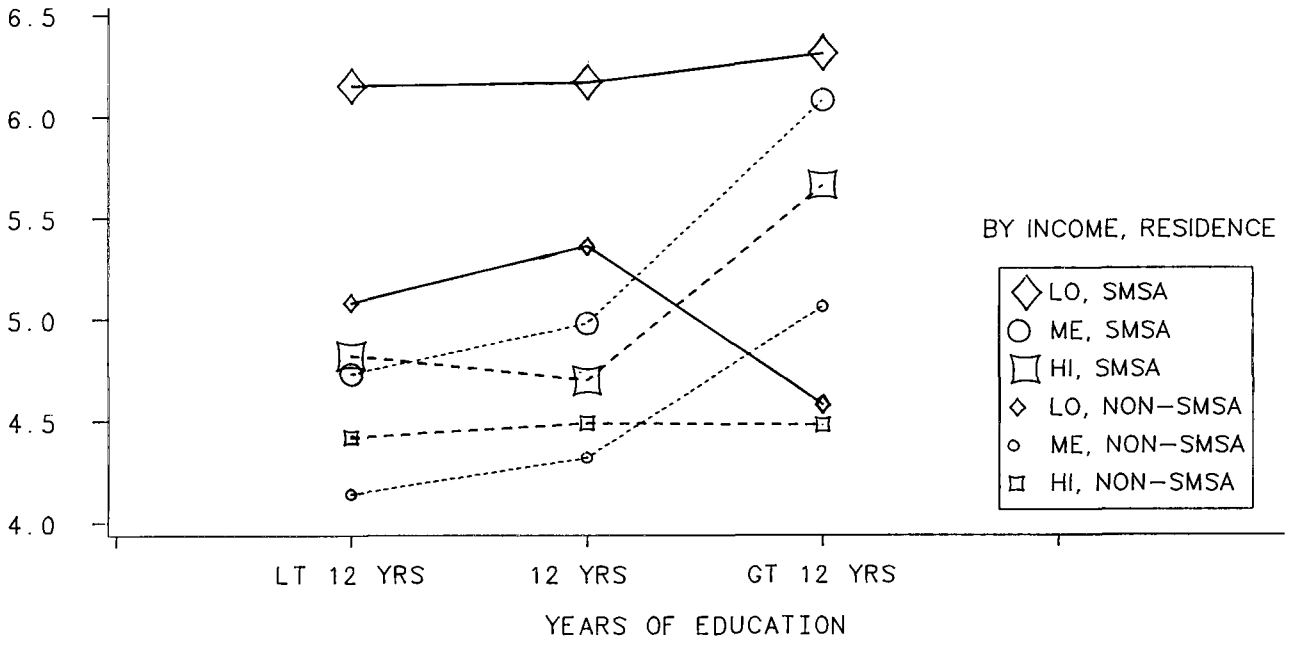

Figure 3: Observed Mean Number of Physician Visits per Person per Year. U.S.A., 1973.

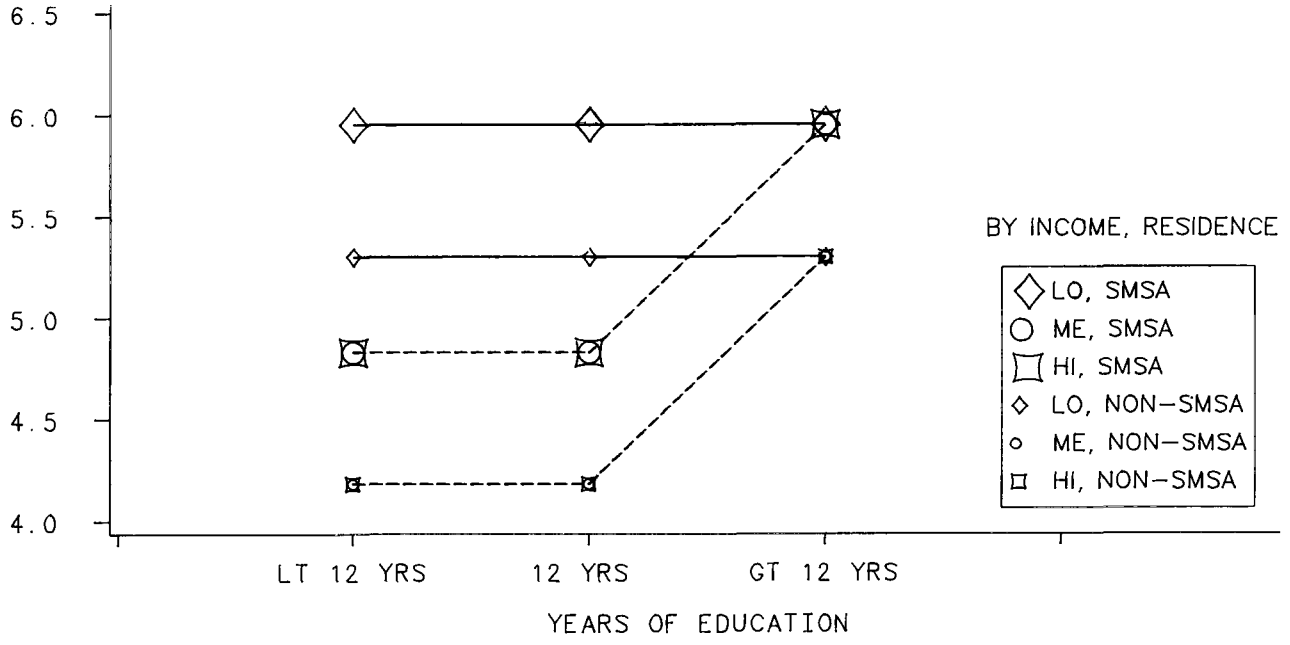

Figure 4: Model Predicted Mean Number of Physician Visits per Person per Year. U.S.A., 1973.

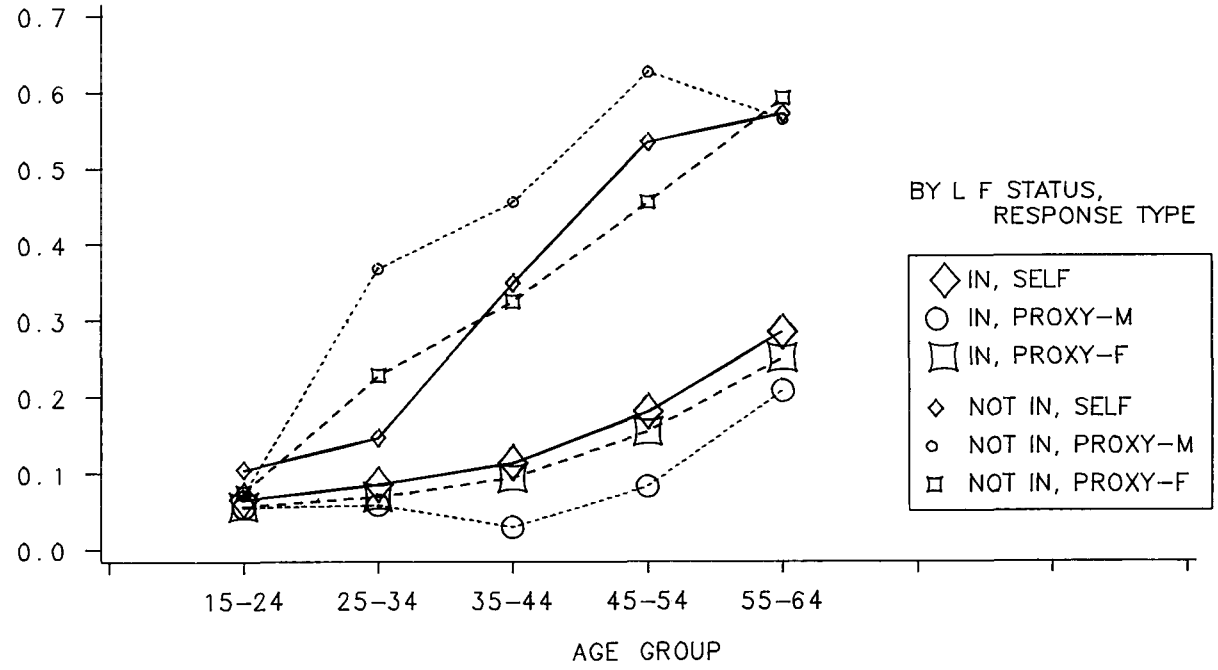

Figure 5: Observed Screening Rates, Disability Survey, January 1983, Males 15-64.

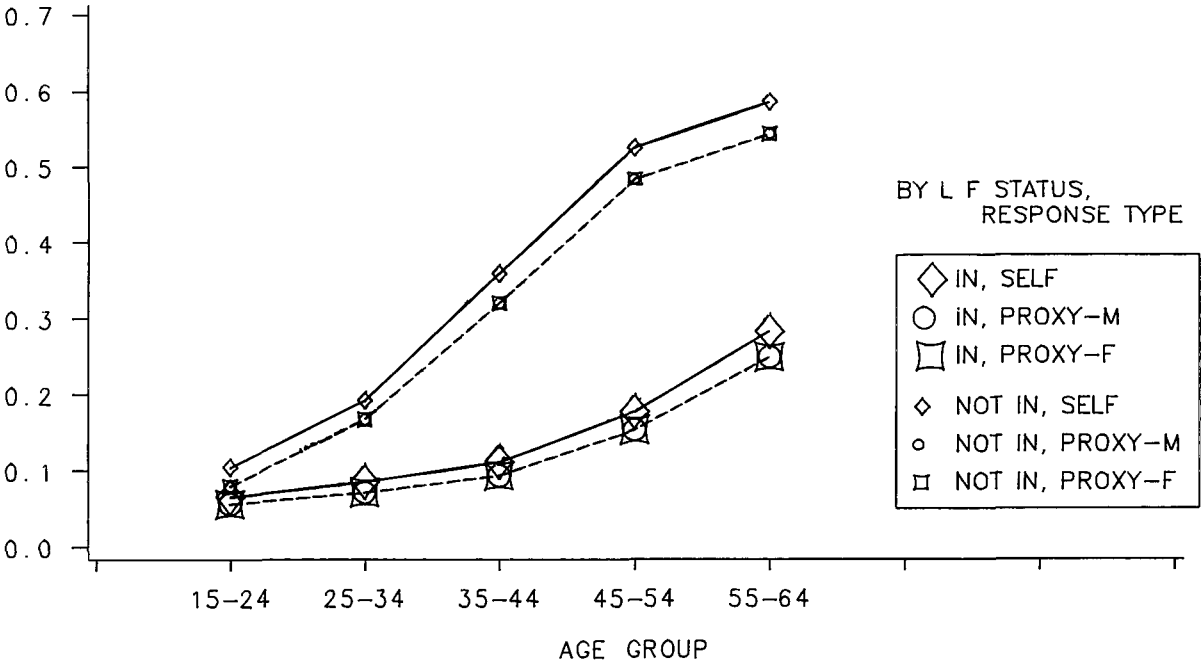

Figure 6: Predicted Screening Rates, Disability Survey. January 1983. Males 15-64.