Journal of Modern Applied Statistical Methods

Volume 3 | Issue 1 Article 17

5-1-2004

A Generalized Quasi-likelihood Model Application To Modeling Poverty Of Asian American Women

Jeffrey R. Wilson
Arizona State University

Recommended Citation

Wilson, Jeffrey R. (2004) "A Generalized Quasi-likelihood Model Application To Modeling Poverty Of Asian American Women," *Journal of Modern Applied Statistical Methods*: Vol. 3 : Iss. 1 , Article 17.
DOI: 10.22237/jmasm/1083370620

Journal of Modern Applied Statistical Methods
May, 2004, Vol. 3, No. 1, 165-175

1538 – 9472/04/$30.00

A Generalized Quasi-likelihood Model Application To Modeling Poverty Of Asian American Women

Jeffrey R. Wilson
School of Health Management and Policy
W. P. Carey School of Business
Arizona State University

A generalized quasi-likelihood function that does not require the assumption of an underlying distribution when modeling jointly the mean and the variance, is introduced to examine poverty of Asian American women living in the West coast of the United States, using data from U.S. Census Bureau.

Key words: Logistic regression, extravariation, generalized linear models

Introduction

All systems of social inequality create poverty. In 1998, the U.S. Census Bureau (1998) states that 12.7% of the U.S. population is poor. Racial minorities are more likely to live in poverty than whites (U.S. Census Bureau, 1999). Previous studies on poverty have focused on whites and other racial minorities and few studies have modeled the poverty for Asian Americans. This research is useful since in recent years Asian Americans have increased significantly and are very diverse in socioeconomic status and country of origin. Poverty among Asian Americans has increased rapidly as a result of a large influx of Asian immigrants from many different countries, many of whom face difficulties in economic opportunities as a result of poor English fluency and low educational attainment. Data from the 1998 Current Population Survey were examined to study the effects of certain variables on the poverty level among Asian American women living in the Western region of the United States.

Because the use of ordinary least squares regression to predict binary response would violate the assumptions of a constant variance (homoscedasticity) and normal distribution (Allison 1999), it is common practice to model binary random variables using logistic regression models. As several variables of interest in social sciences and medical research are binary, logistic regression models have been used widely in these areas. Such models require a logistic transformation on the probability in such a way that the odds is modeled and thus the predicted probabilities are not outside the bounds for probability.

However, there may be times when the fitted logistic regression model does not adequately describe the observed proportions, because of the presence of extravariation or overdispersion as it is often referred to. The presence of overdispersion results in the assumption of binomial variability to be invalid (Collett, 1991). When overdispersion occurs, it may be necessary to consider other binary models. One such approach is to consider a quasi-likelihood model thus negating the need for the binomial variation assumptions. A quasi-likelihood model does not make any distributional assumption about the random variable in the mean modeling.

Modeling the mean of a binary response model consists of several approaches. Some approaches have been proposed where the

Jeffery R. Wilson is a Professor of Biostatistics and Director of the School of Health Management and Policy at Arizona State University, Tempe AZ 85287-4506.

parameters of the distribution are allowed to vary based on some known distribution (Williams, 1982; Crowder, 1978; Wilson, 1989; Wilson & Koehler, 1991). Other methods have made use of a mean-variance relation (Wedderburn, 1974; McCullagh, 1983; Firth, 1987; Moore & Tsiatis, 1991) and so the knowledge of an underlying distribution is not required.

These methods assume that the variance is related to the mean through the variance function, which is a function of the mean. Neither of these approaches considered modeling the variance of the distribution. Analyzing the poverty data among Asian Americans showed that through there is sufficient extravariation that needs to be modeled. A review of a binary logistic function is follows.

Generalized linear models (Nelder & Wedderburn, 1972) encompass a wide range of models. These models include linear regression, analysis of variance, logit and probit models for binary response data, and log-linear and multinomial response models for count data. A generalized linear model has three components. The random component specifies the distribution of the response variable from the exponential family of distributions. The systematic component defines a linear predictor based on some set of known covariates and the link component combines the random component and the systematic component. The link function is a monotonic twice-differentiable function that provides a relation between the mean of the response variable and the covariates.

Generalized linear models differ in their underlying distribution and in their link function. The systematic component of these models has a linear structure. Generalized linear models reduce the problem of scaling and do not require the assumption of normality and constancy of variance. For linear regression and analysis of variance models the distribution is normal with an identity link. For logit and probit models the distribution is binomial with logistic and cumulative distribution function of normal distribution as link functions, respectively. Log-linear models have a multinomial distribution with a log link. Estimation of these regression parameters in the systematic function can be done through maximum likelihood procedure (Finney, 1952). However, for exponential family distributions, the maximum likelihood estimation is equivalent to the weighted least squares method (Bradley, 1973). Thus, generalized linear models lead to a unified method for estimating the parameters for a wide range of models. They provide a method for modeling the mean of the distribution.

The modeling of the mean and the dispersion jointly through two sub models using a generalized linear model framework was first suggested by Pregibon (1984) and later addressed by Efron (1986), Aitkin (1987) and Smyth (1989). In the joint modeling of the mean and the variance, three components similar to the mean sub model are required for modeling the variance. The response variable for the dispersion submodel is the deviance obtained from the mean submodel. The extended quasi-likelihood function (Nelder & Pregibon, 1987; McCullagh & Nelder, 1989) and the pseudo likelihood function (Carroll & Ruppert, 1982) are useful for joint modeling of the mean and the dispersion, when only the relation between mean and variance has been specified for the mean submodel.

Extended quasi-likelihood and pseudo likelihood functions can be used for comparison of the link and the variance function. Further generalizations and modifications of the extended quasi-likelihood functions have been presented by Yanez and Wilson (1995).

Binary logistic function

Consider Y_i for $i = 1,........n;$ to denote the i th observation for each of the Asian women with mean p_i where p_i is the probability that an Asian woman falls below the poverty level. A linear logistic model for poverty level based on martial status, educational attainment, residence, employment status, and number of children for each of these women is

$$\begin{aligned}\text{logit}\ (p_i) &= \log(\frac{p_i}{1-p_i}) \\ &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki},\end{aligned} \quad (1)$$

where x_{ki} is the value of the k th variable on the i th woman. Thus the probability of an event is:

$$p_i = \frac{e^{(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki})}}{1 + e^{(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki})}} \quad (2)$$

and the variance function is defined by $\text{var}(y) = \Phi \frac{p_i}{1 - p_i}$. In most cases Φ is one. When $\Phi \neq 1$, it is usually common to use quasi-likelihood models (McCullagh & Nelder, 1989). For modeling the poverty data pertaining to Asian Americans, both the mean and variance parameters are modeled using a quasi-likelihood function.

Methodology

A generalized quasi-likelihood model (GQL) for poverty among Asian American women is now proposed. The model is simple and less restricted in that it does not require the assumption of an underlying distribution, when modeling either the mean or the variance jointly. The generalized quasi-likelihood function assumes that the distributional form for both the mean and the dispersion submodels are not known and relies on a mean-variance relation. In the dispersion submodel the mean and the variance of the response variable are ϕ_i^α and $2\phi_i^{2\alpha}$ respectively, where α is a nonlinear parameter.

Thus, the variance function is assumed to be a squared function of the mean in the dispersion submodel, with a dispersion parameter of value two. In the analysis of these data the link and variance functions used for the mean submodel is quasi and log-root, respectively, whereas the link and variance for the variance submodel is quasi and square root, respectively.

For a single observation y_i with mean μ_i i = 1, 2, . . ., n; a generalized quasi-likelihood function is defined as

$$Q_1^*(y_i, \mu_i, \phi_i, \alpha, \tau)$$
$$= -\frac{1}{2}\left[\frac{d(y_i; \mu_i)}{\phi_i^\alpha} + \ln(\phi_i^\alpha) + \ln(2\pi V_\tau(y_i))\right],$$

where $d(y_i, \mu_i) = -2\int_{y_i}^{\mu_i} \frac{y_i - u_i}{V_\tau(u_i)} du_i$, ϕ_i is the dispersion parameter for the mean submodel, $V_\tau\left(y_i\right)$ is the variance function evaluated at y_i, and α and τ are nonlinear parameters. The generalized quasi-likelihood model has a mean submodel with random, systematic, and link components as $Y_i \sim \left(\mu_i, \phi_i^\alpha V_\tau(\mu_i)\right)$, $\eta_i = \mathbf{x}_i' \vec{\beta}$, and $\eta_i = g\left(\mu_i\right)$, respectively.

Its dispersion submodel has response variable $d_i \sim \left(\phi_i^\alpha, 2\phi_i^{2\alpha}\right)$, $\eta_i^* = \mathbf{v}' \vec{\gamma}_i$, and $\eta_i^* = h\left(\phi_i, \alpha\right)$ as the random, systematic, and link function component, respectively. The estimating equations for the linear parameters $\overline{\beta} = (\beta_1, \beta_2, \ldots \beta_p)$, in the mean submodel based on the GQL function are

$$\frac{\partial Q_1^*}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi_i^\alpha V_\tau(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j}.$$

Similarly, the estimating equations for the linear parameter $\vec{\gamma} = (\gamma_1, \gamma_2, \gamma_3, \ldots \gamma_p)$ in the dispersion submodel are

$$\frac{\partial Q_1^*}{\partial \gamma_r} = \frac{\alpha}{2} \sum_{i=1}^{n} \frac{d_i - \phi_i^\alpha}{\phi_i^{\alpha+1}} \frac{\partial \phi_i}{\partial \gamma_r}.$$

A simultaneous iterative weighted least squares procedure is used to solve these estimating equations as $\vec{\beta}$ and $\vec{\gamma}$ are orthogonal. The orthogonality of μ_i and ϕ_i, leads to the orthogonality between $\vec{\beta}$ and $\vec{\gamma}$ which follows since the expected partial derivatives,

$$\mathrm{E}\left[\frac{\sum \partial^2 Q_1^*}{\sum \partial \mu_i \sum \partial \phi_i}\right] = \mathrm{E}\left[-\frac{\alpha\left(y_i - \mu_i\right)}{\mu_i^\tau \phi_i^{\alpha+1}}\right] = 0.$$

Thus holding α, τ and ϕ_i fixed, the maximum quasi-likelihood estimator, $\vec{\hat{\beta}}$ are obtained for the function Q_1^* through $\mathbf{X}'\mathbf{W}\mathbf{X}\vec{\hat{\beta}}^{(m)} = \mathbf{X}'\mathbf{W}\mathbf{z}$, where $\mathbf{W} = diag\left(\frac{1}{\mathrm{var}(y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2\right)$, $diag\,(t)$ denotes the diagonal elements of the matrix $\mathbf{T}$ and $\mathbf{z}$ is a vector of order n with elements

$$z_i = \sum_{k=1}^{p} x_{ik}\beta_k^{(m-1)} + (y_i - \mu_i)\frac{\partial\eta_i}{\partial\mu_i} \quad i = 1, \ldots, \mathrm{n};$$

The maximum quasi-likelihood estimates for the regression parameters, $\vec{\hat{\gamma}}$, in the dispersion submodel are estimated from $\mathbf{V}'\mathbf{W}^*\mathbf{V}\vec{\hat{\gamma}}^{(m)} = \mathbf{V}'\mathbf{W}^*\mathbf{z}^*$ where $\mathbf{W}^* = diag\left(\frac{\alpha^2}{2\phi_i^2}\left(\frac{\partial\phi_i}{\partial\eta_i^*}\right)^2\right)$ and $\mathbf{z}^*$ is a vector with elements

$$\mathrm{z}_i^* = \sum_{l=1}^{p^*} v_{il}\hat{\gamma}_l^{(m-1)} + \frac{\left(d_i - \phi_i^{\alpha}\right)}{\alpha\phi_i^{(\alpha-1)}}\frac{\partial\eta_i^*}{\partial\phi_i}$$

by fixing ϕ_i and $\vec{\beta}$ and estimates of the nonlinear parameters α and τ at known value. The process is continued until convergence is achieved.

The variance of $\vec{\beta}$ under the generalized quasi-likelihood function is $\mathrm{cov}\left(\vec{\hat{\beta}}\right) = \left(\mathbf{l}_m'\mathbf{V}_m^{-1}\mathbf{l}_m\right)^{-1}$, where $\mathbf{l}_m = \left[\frac{\partial\mu_i}{\partial\beta_j}\right]_{i,j}$ is the vector of partial derivatives and $\mathbf{V}_m =$ diag$(\phi_i^{\alpha}\,\mathrm{V}_{\tau}((\mu_i))$. Similarly for the vector of estimates $\vec{\hat{\gamma}}$, $\mathrm{cov}\left(\vec{\hat{\gamma}}\right) = \left(\mathbf{l}_d'\mathbf{V}_d^{-1}\mathbf{l}_d\right)^{-1}$ where $\mathbf{l}_d = \left[\frac{\partial\phi_i}{\partial\gamma_1}\right]_{i,1}$ and $\mathbf{V}_d = \mathrm{diag}\left(\frac{\alpha^2}{2\phi_i^2}\right)$.

Results

The major interest is to determine which social factors contribute if any to Asian American women living in poverty. These social factors included whether she is married, her years of schooling, residence, whether she works, and how many children she has. These data are confined to those women living in the western region of the United States (i.e. California, Washington, Oregon, Arizona, etc.). There are a total of 639 Asian American women in our sample.

Studies on poverty have focused on whites and other racial minorities and few studies examine the likelihood of poverty for Asian Americans. In this study, the definition of an Asian American living in poverty follows the definition given by the U.S. Census Bureau. A woman is considered to live in poverty if she lives on her own with family income less than $7,500, if a woman lives with another family member with family income less than $10,000, if a woman lives with two other family members with family income less than $15,000, etc. This definition is based on 1998 figures and takes into account the family size. Of all the poor people eighteen and older, 62% are women and 38% are men (U.S. Census Bureau, 1999). The motivating factor that brought these data into focus is in part due to an emerging belief that there is a trend by which women represent an increasing proportion of the poor.

Previous research on other racial groups reveals that marital status, educational attainment, area of residence, employment status, and number of children are strong predictors of poverty. The increases in poverty among women are partly as a result of increases in unmarried women, and families headed by single mothers (Macionis, 2001).

Although people living in central cities are most likely to live in poverty, people living in suburban areas are least likely to live in poverty (Macionis, 2001). Asian American women living in metropolitan areas are less likely to live in poverty as compared to those living in non-metropolitan areas, although Asian Americans are least likely to live in non-metropolitan areas. Educational attainment and employment status are as expected significant predictors: the more educated women the less likely they live in poverty; no jobs translate into more poverty (Wilson, 1996). The number of children also has a positive impact on poverty: the more children a woman has it is more likely for her to live in poverty (Wilson, 1996).

In the binary models used to model poverty, variables are coded as follows. Marital status is coded 1 if a woman is unmarried (widowed, divorced, separated, or never married) and 0 if a woman is married. Educational attainment has four categories: “1” denotes less than high school; “2” denotes high school; “3” denotes some college; and “4” denotes college graduate and above. Area of residence is coded 1 if a woman lives in metropolitan areas and 0 if a woman lives in non-metropolitan areas. Employment status is coded 1 if a woman worked for pay and 0 otherwise.

There are three categories for number of children: “1” denotes no children; “2” denotes 1 to 3 children; and “3” denotes more than 3 children. Table 1 provides a percentage distribution of women living in poverty and the tabulation between poverty and each of the predictors.

Bivariate analyses of poverty and each predictor reveal that of 639 Asian American women in the sample, 23.2% live in poverty. A higher percentage of unmarried Asian American women lived in poverty compared to married Asian American women (26.5% vs. 19.9%). Women with high school education have the highest percentage living in poverty (41.3%). Women with college education and above have the lowest percentage living in poverty. Fewer Asian American women lived in non-metropolitan areas than in metropolitan areas (56 vs. 583). Those living in metropolitan areas have higher percentage living in poverty than those living in non-metropolitan areas (37.5% vs. 21.8%). Of employed women, only 18.8% lived in poverty while 30% of unemployed women lived in poverty. The number of children is not significant at the 0.05 level.

These bivariate results are consistent with those obtained from previous literature on poverty for other racial groups. However, simultaneous effects of these predictors on poverty are more informative if one is to adequately assess the different impacts. Thus a multivariate logistic regression model suitable for a 2 x 4 x 2 x 2 x 3 contingency table is required. The logistic regression model and the generalized quasi-likelihood function were compared in their use to analyze the data from U.S. Census Bureau’s 1998 Current Population Survey.

Applications of Binomial Logistic Regression Model

A logistic regression model with a binomial distribution and a logit link function was fitted to the 2 x 4 x 2 x 2 x 3 contingency table. This model was presented to determine the simultaneous impact of marital status, educational attainment, area of residence, employment status, and number of children on the probability that Asian American women live in poverty. Table 2 provides a summary of the results from the fit of such a maximum likelihood binomial logistic regression model. The odds ratios are obtained from the exponentiation of the parameter estimates. Unmarried Asian American women are 1.75 times as likely to be poor than married Asian American women. Educational attainment has a negative effect on poverty: It also seems that more educational years reduced the odds of living in poverty by 33.9%.

Asian American women living in nonmetropolitan areas are 1.63 times as likely to be poor than those living in metropolitan areas. Evidently, whether a woman has a job affects the likelihood of being poor: those without jobs are 1.56 times as likely to be poor than those with jobs. The impact of number of children on poverty is not significant. This could be due to the fact that poverty measure (whether a person lives in poverty) is adjusted by the family size.

Table 1. Percentage Distributions of Asian American Women Living in Poverty by Marital Status, Educational Attainment, Type of Residence, Employment, and Number of Children.

Variable	% in poverty	Number of Cases
Total	23.2%	639
*Marital Status***		
Married	19.9%	326
Unmarried	26.5%	313
Educational		
Less Than High School	22.5%	111
High School	41.3%	150
Some College	21.7%	184
College Graduate and	10.8%	194
*Area of Residence****		
Metropolitan	21.8%	583
Nonmetropolitan	37.5%	56
Employed?***		
Yes	18.8%	389
No	30.0%	250
Number of Children		
No children	21.5%	311
1-3 children	23.7%	296
4 and more children	34.4%	32

Note: **, significant at the .05 level and ***, significant at the .01 level (Pearson chi-square test).

Table 2 Parameter estimates, Standard errors, and Odds Ratios For Binomial Logistic Regression Model.

Covariate	Parameter	Standard	Odds Ratios
Intercept	-.705	.557	.494
Marital Status			
Unmarried	.559	.239	1.749
Married			
Educational	-.273	.089	.761
Area of Residence			
Metropolitan	-.638	.305	.528
Nonmetropolitan			
Employment Status			
Employed	-.446	.201	.640
Not Employed			
Number of Children	.487	.194	1.63

It is imperative to know, prior to accepting the odds ratios as obtained, whether or not there is a good fit with the model: the extent to which the fitted values of the response variable under the model compare with the observed values. If the agreement between the observations and the corresponding fitted values is good, the model may be acceptable (Collett, 1991). To examine the fit, the likelihood ratio with the covariates in the model, $\hat{L}_c$, is compared with the likelihood ratio with the saturated model, $\hat{L}_f$. The deviance,

$$D = -2\log(\hat{L}_c / \hat{L}_f) = -2[\log \hat{L}_c - \log \hat{L}_f],$$

where $\hat{L}_c$ is obtained based on the predicted probability of the event under the model with covariates while $\hat{L}_f$ is obtained based on the observed proportions of the event provides such a measure.

The deviance from the model with covariates is 138.81 with 74 degrees freedom. The ratio of the deviance to the degrees of freedom (1.87) is substantially greater than one. Thus, there is a strong likelihood that over-dispersion is present and the assumption of the binomial variability may not be valid (Collett, 1991). Such results suggest that the data exhibit overdispersion. Thus there is a significant amount of variation unaccounted for. This indicates that Φ is greater than 1 in the variance function where $\text{var}(y) = \Phi \dfrac{p_i}{1-p_i}$. Thus, the assumption that Φ is equal to 1 in the logistic regression model is not valid. Thus it is evident that the data are over-dispersed.

Overdispersion arises because of clustering in the population (McCullagh and Nelder, 1989). Overdispersion could be present due to the fact that unobserved heterogeneity operates at the level of groups rather than individuals (Allison, 1999). It may also be an account of the cost of living differences between metropolitan and nonmetropolitan cities.

Given the presence of such overdispersion, a quasi-likelihood model was chosen to analyze the data. The quasi-likelihood model allows us to estimate the parameters in the model and determine its significance without specifying the distribution function while accounting for the overdispersion. The model is fully determined since the link and variance functions are sufficient for fitting the model. Once these are specified, the same iterative procedure that is used for fitting the other families can be used to estimate the linear parameters. This is readily available in SPLUS.

Applications of Generalized Quasi-Likelihood Function

Using the logistic regression model to fit the data left indication that overdispersion was present. The overdispersion may be due to the fact that some variables tend to produce clustering in poverty and thus some unobserved heterogeneity affects the fit of the model. To account for any such extra variation, a joint modeling of the mean and the variance using the generalized quasi-likelihood function was used. Quasi-likelihood estimation makes it possible to estimate relationships without fully knowing the random component of model.

The difference between a quasi-likelihood function and a maximum likelihood function is analogous to the comparison between normal-theory regression models and least squares regression estimates. As least-squares estimation and normal theory models give identical regression parameter estimates so does quasi-likelihood and maximum likelihood procedures. However, least-squares estimation relies on second moment assumptions for its variance whereas normal-theory models rely on full distributional assumptions.

Under quite general conditions, quasi-likelihood estimates are consistent and asymptotically normal (Agresti, 1990). Quasi-likelihood estimators still retain relatively high efficiency as long as the degree of overdispersion is moderate (Cox, 1983; Firth, 1987). Thus, quasi-likelihood function allows us to estimate the dispersion parameter in moderately over-dispersed regression models. We applied these principles to the present data under investigation.

The mean submodel has first and second moments as

$$E(y|x) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}$$

and $\operatorname{var}(y_i) = \Phi V(\mu_i)$ respectively, where Φ is the overdispersion parameter. Systematic components consist of marital status, educational attainment, type of residence, employment status and number of children. The model was fitted to the data using several different link functions including logit, log, and complementary log-log. For the variance functions, choices were made from μ, $\mu(1-\mu)$, and the constant.

Based on the goodness of fit statistics, the mean submodel with a log link and μ as the variance function gave the best fit. The log link corresponds to multiplicative effects of the covariates. The "μ" variance function is equivalent to Φ as the coefficient of variation of the response (McCullagh & Nelder, 1989). The regression coefficient estimates for the mean and the dispersion submodel are given in the first two columns of Table 3.

The dispersion submodel was also fitted with different link and variance functions. The choices for link functions included identity and square root and the choices for variance functions included the constant, μ, and μ^2 (the squared coefficient of variation). Based on the goodness-of-fit statistics (mostly, how much deviance relative to the degrees of freedom), the dispersion model with square root link function and μ the variance function was chosen.

Some parameter estimates from the generalized quasi-likelihood model from Table 3 are similar in value to the corresponding values of Table 2 when the binomial logistic regression model was applied. In the generalized model, there are two variables significant at the .05 level. Education has a negative effect on poverty, thus the more educated they are the less likely they are in poverty, while the more children in the household increased the odds of Asian women living in poverty. The deviance from the generalized quasi-likelihood model suggests that the overdispersion is accounted for and the model is a good fit.

The response variable of the dispersion submodel is the square of the residual. Residuals are one principal tool for assessing how well a model fits the data. They can be used to assess the importance and relationship of a term in the model as well as to search for anomalous values. For generalized linear models, residuals can also help assess and verify the form of the variance as a function of the mean response.

There are different kinds of residuals that can be employed. First the deviance residuals,

$$r_i^D = \text{sign}(\mathrm{y}_\mathrm{i} - \mu_i)\sqrt{d_i}$$

where d_i is the contribution of the *i*th observation to the deviance. The deviance is

$$D = \sum\nolimits_i (r_i^D)^2$$

These residuals are useful detecting observations with unduly large influence on the fitting process, since they reflect the same criterion as used in the fit. Secondly, there is the Pearson residuals,

$$r_r^P = \frac{y_i - \mu_i}{\sqrt{V(\mu_i)}} \text{ and } \chi^2 = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{V(\mu_i)}$$ is the chi-square statistic.

The dispersion submodel has as its response variable the squares of the residuals (the difference between observed values and fitted values). If the mean submodel fits the model well, then there may not be a need to model the deviance and none of the parameter estimates in the dispersion model may be significant. An examination of the parameter estimates and standard errors from the dispersion submodel in Table 3 suggests that the form of the variance as a function of the mean response is appropriate in our model and there are almost no anomalous values in our model. The mean deviance for the dispersion model is 2.05.

Table 3. Parameter estimates and (standard errors) for Generalized Quasi-likelihood model.

Covariate	Mean Submodel		Dispersion Submodel	
	Parameter Estimate	Standard Errors	Parameter Estimate	Standard Errors
Intercept	-1.128*	.504	1.698**	.414
Marital Status				
Unmarried	.388	.220	-.034	.201
Married				
Educational	-.206*	.084	-.104	.069
Area of Residence				
Metropolitan	-.412	.264	.234	.209
Non-metropolitan				
Employment Status				
Employed	-.315	.191	.216	.199
Not Employed				
Number of Children	.338*	.172	-.345*	.138

Note: * at the .05 level, and ** at the .01 level.

Conclusion

Generalized linear models such as binomial logistic regression and Poisson regression are very widely used in social, economic, and medical research. While the binomial logistic regression is easy to use and interpret, we need to look for an alternative if there is overdispersion in our data.

When the data are over-dispersed, due to heterogeneity or the clustering effect at the group level, it is necessary to model the overdispersion. Quasi-likelihood models allow you to model such overdispersion as the estimation process assumes only a form for the functional relationship between the mean and the variance. Further they allow us to simultaneously model the mean and the variance without accounting for any distributional assumptions.

Quasi-likelihood models were used to model the data from U.S. Census Bureau's 1998 Current Population Surveys. Data pertaining to Asian American women who lived in the western region of the United States showed that covariates such as marital status, educational attainment, area of residence, employment status, and number of children are not all predictors when modeling poverty, as with other ethnic and racial groups. Use of the binomial logistic regression model showed the presence of overdispersion. Quasi-likelihood functions were used to model that overdispersion. Several link functions and variance functions were examined to identify a model with the best fit. For these data, a mean submodel with the log as the link function and μ as the variance function and a dispersion submodel with square root as the link function and μ as the variance function fit well. Thus, the binomial logistic regression models overstated the effects of the covariates, in part due to the unaccounted extravariation.

References

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons.

Allison, P. D. (1999). *Logistic Regression Using the SAS System.* Cary: SAS Institute.

Aitkin, M. (1987). Modeling variance heterogeneity in normal regression using GLIM. *Applied Statistics, 36,* 332-339.

Bain, L. J. & Engelhardt, M. (1987). *Introduction to Probability and Mathematical Statistics*, Duxbury.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, MA: MIT.

Carroll, R. J. & Ruppert, D. (1982). Robust estimation in heteroscedastic linear models. *Annals of Statistics*, *10,* 429-441.

Collett, D. (1991). *Modeling Binary Data.* London: Chapman & Hall.

Cox, D. R. (1983). Some Remarks on Overdispersion. *Biometrika*, *70,* 269-274.

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, *81*, 701-721.

Firth, D. (1987). On the Efficiency of Quasi-Likelihood Estimation. *Biometrika*, *74*, 233-245.

Macionis, J .J. (2001). *Sociology* (8th ed). Upper Saddle River: Prentice Hall.

McCullagh, P. (1983). Quasi-likelihood functions. *Annals of Statistics*, *11*, 59-67.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman & Hall.

Moore, D. F. & Tsiatis, A. (1991). Robust Estimation of the Variance in Moment Methods for Extra-Binomial and Extra-Poisson Variation. *Biometrics*, *47*, 383-401.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society*, *A135* 370-384.

Nelder, J. A. & Pregibon, D. (1987). An Extended quasi-likelihood function. *Biometrika*, *74*, 221-232.

Pregibon, D. (1984). Review of *Generalized Linear Models* by McCullagh and Nelder. *Annals of Statistics, 12*, 1589-1596.

Smyth, G. K. (1989). Generalized Linear Models with Varying Dispersion. *Journal of Royal Statistical Society* Ser. B *51*, 47-60.

S-Plus 4 Guide to Statistics. (1997). Seattle: Data Analysis Products Division, MathSoft.

Stroud, T. W. F. (1971). On obtaining large sample tests from asymptotically normal estimators. *Annals of Statistics*, *42*, 1412-1424.

U.S. Census Bureau. (1998). *Statistical Abstract of the United States 1998.* Washington, D.C.: U.S. Government Printing Office.

U.S. Census Bureau. (1999). *Poverty in the United States 1998*. P60-207. Washington, D.C.: U.S. Government Printing Office.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss Newton method. *Biometrika*, *61*, 439-447.

Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics*, 31, 144-148.

Wilson, J. R. (1989). Chi-square tests for overdispersion with multiparameter estimates. *Applied Statistics*, 38, 441-453.

Wilson, W. J. (1996). *When Work Disappears: The World of the New Urban Poor*. New York: Alfred A. Knopf.

Yanez, N. D. & Wilson, J. R. (1995). Comparison of quasi-likelihood models for overdispersion. *Australian Journal of Statistics*, *37*, 217 -231.