PARAMETRIC LINKS FOR BINARY CHOICE MODELS: A FISHERIAN-BAYESIAN COLLOQUY

ROGER KOENKER AND JUNGMO YOON

Abstract. The familiar logit and probit models provide convenient settings for many binary response applications, but a larger class of link functions may be occasionally desirable. Two parametric families of link functions are investigated: the Gosset link based on the Student t latent variable model with the degrees of freedom parameter controlling the tail behavior, and the Pregibon link based on the (generalized) Tukey λ family with two shape parameters controlling skewness and tail behavior. Both Bayesian and maximum likelihood methods for estimation and inference are explored, compared and contrasted. In applications, like the propensity score matching problem discussed in Section 4, where it is critical to have accurate estimates of the conditional probabilities, we find that misspecification of the link function can create serious bias. Bayesian point estimation via MCMC performs quite competitively with MLE methods; however nominal coverage of Bayes credible regions is somewhat more problematic.

1. Introduction

In the classical binary response model the probability, π_i, of the occurrence of an event, $Y_i = 1$, rather than the event $Y_i = 0$, conditional on a vector of covariates x_i is expressed as,

$$g(\pi) = x_i^\top \beta.$$

The function g links the linear predictor to the probability and determines the shape of the quantal response. McCullagh and Nelder (1989) discuss four possible link functions for binary response models:

logit: $g(\pi) = \log(\pi/(1-\pi))$
probit: $g(\pi) = \Phi^{-1}(\pi)$
cloglog: $g(\pi) = \log(-\log(1-\pi))$
loglog: $g(\pi) = -\log(-\log(\pi))$.

They note the strong similarity of logit and probit citing Chambers and Cox (1967), and observe that the loglog link "is seldom used because its behavior is inappropriate for $\pi < 1/2$, the region that is usually of interest." This advice is generally in

Version: November 5, 2007. This research was partially supported NSF grant SES-05-44673. The authors, while retaining full culpability for any errors and misconceptions, would like to express their appreciation to Victor Chernozhukov, Trevor Hastie, Stephen Portnoy and an anonymous referee for valuable comments.

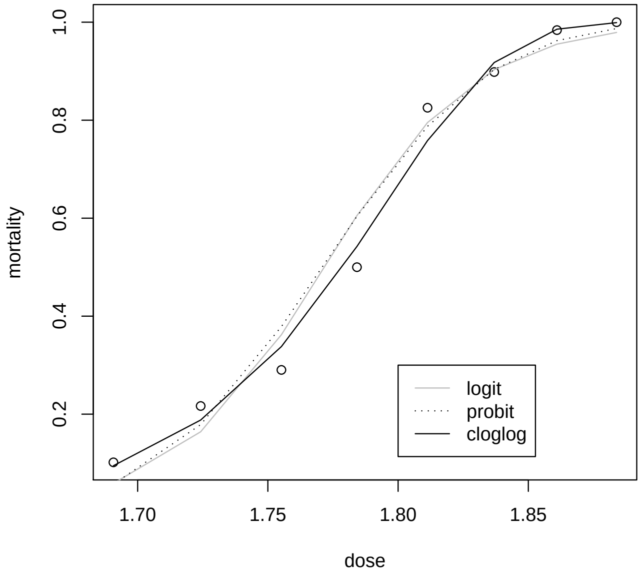

FIGURE 1. Three Dose Response Curves for Beetle Mortality.

accord with observed statistical practice where logit and probit seemed to be employed almost interchangeably and the log-log links are a relative rarity. The close proximity of the probit and logit link functions is frequently extrapolated to imply that *all* links are essentially indistinguishable. One objective of this paper is to correct this misapprehension. The other primary objective is to provide a case study comparing Fisherian and Bayesian approaches to estimation and inference in a relatively simple parametric setting.

Each link function corresponds to a latent variable model

$$y_i^* = x_i^\top \beta + u_i$$

with u_i iid, in which we observe $y_i = I(y_i^* > 0)$. Maximum likelihood estimation of the latent variable model implies a link function chosen as the quantile function of the u_i's. A well-known empirical example, e.g. Prentice (1976), illustrating the advantage of the cloglog link is illustrated in Figure 1. Observed mortality rates of adult flour beetles exposed to gaseous carbon disulphide are plotted at 8 distinct dosages, in units of CS_2 $mg/litre$ in logarithms (base 10), with about 60 beetles tested at each dose. As is evident in the figure the observed dose-response points are quite asymmetric. As expected the logit and probit fits are very similar, but the cloglog fit appears much better. This is confirmed by examining the log-likelihoods; logit and probit achieve -18.72 and -18.16 respectively, while cloglog attains -14.82.

A more exotic entry on the menu of links is the "cauchit" link,[1] given by the standard Cauchy quantile function,

$$g(u) = \tan(\pi(u - 1/2)).$$

The "cauchit" model is attractive when observed responses exhibit a few surprising values, observations for which the linear predictor is large in absolute value indicating that the outcome is almost certain, *and yet the linear predictor is wrong.* These binary "outliers" may be the result of a variety of easily imagined circumstances including data recording errors, but whatever their source both probit and logit are rather intolerant of them, while cauchit is more forgiving.

Having estimated the probit and cauchit models in a particular application, a natural question arises: Can we test for the suitability of one link versus the other? This question leads directly to the family of Gosset links considered in Section 2. Liu (2004) has previously suggested estimating such models using the EM algorithm, and there have been several Bayesian MCMC proposals following Albert and Chib (1993). We consider estimation and inference in the Gosset link model from both a classical Fisherian maximum likelihood viewpoint and from a Bayesian viewpoint and compare performance of the two approaches.

The Gosset link model enables us to account for symmetrically distributed heavy tails in the latent variable model for binary response. What about skewness? Pregibon (1980) proposed a "goodness of link" diagnostic for the logistic binary response model based on the generalized Tukey λ family link,

$$g(u) = \frac{u^{\alpha-\delta} - 1}{\alpha - \delta} - \frac{(1-u)^{\alpha+\delta} - 1}{\alpha + \delta}.$$

This link is logistic for $\alpha = \delta = 0$ and describes an attractive family of unimodal distributions for other values of α and δ, as illustrated in Figure 2. For $\delta = 0$ we have symmetric densities with α controlling the heaviness of the tails, while δ controls the skewness of the distribution. This family of link functions is considered in Section 3.[2] Since probit and logit link functions are nested in the Gosset and Pregibon link families respectively, we can use those broader models to test the validity of a model specification. Further details of the R implementation of the proposed links are given in Koenker (2006). And R code for all of the computations reported below is available from the url: `http://www.econ.uiuc.edu/ roger/research/links/links.html`.

It is sometimes argued, e.g. Hastie and Tibshirani (1987), that more flexible links for binary response models are unnecessary because more flexible specification of the

[1] The origins of the "cauchit" link are somewhat misty, but see Morgan and Smith (1992) for an empirical example in which it appears to perform better than the probit link.

[2] In preparing a revised version of this paper we discovered that Prentice (1976) had proposed an alternative two parameter family of link functions that has a convenient density function representation. Unfortunately, evaluation of the corresponding quantile and distribution functions appears to be less tractible than is the Pregibon family.

linear predictor can successfully imitate the misspecified link. This is formally true, but from a more pragmatic perspective there is something quite appealing about the simplicity of the univariate link, iid error latent variable model. With several covariates the necessary flexibility of the linear predictor is difficult to achieve. We should also note that there is an extensive literature on semi-parametric extimation of more flexible binary link models, see e.g. Manski (1975), Klein and Spady (1993) and Newton, Czado, and Chappell (1996). We regard the approach developed here as offering a relatively simple, parsimonious compromise between the conventional logit and probit specifications and these alternatives.

In the literature on average treatment effects, matching estimators based on the propensity score has received considerable recent attention. Since matching estimators depend crucially on the estimated probabilities of treatment given covariates, or propensity scores, the choice of the binary response link function is critical. In section 4, we compare the logit specification which is common in literature with our proposed alternative links.

2. The Gosset Link

The probit and cauchit link functions are naturally nested within the Student t family and the conventional glm iteratively weighted least squares, or method of scoring, described in McCullagh and Nelder (1989) seems to provide a simple, effective method of estimation of the parameters of the linear predictor for fixed values of the degrees of freedom parameter, ν. This approach yields a profile likelihood like the one illustrated in Figure 3. Optimization of the profiled likelihood is easily carried out with the aid of the Brent (1973) algorithm, or similar methods.[3] There is an extensive literature on the use of Student models for continuous response models where it has obvious robustness advantages. For binary response there have been several Bayesian MCMC investigations, but to our knowledge there has been no attempt to compare their performance to conventional MLE methods for similar models.

Several caveats regarding the Gosset model should be mentioned:

- When ν is moderate, say $\nu > 6$, it is difficult to distinguish Gosset models from probit and it is common to see profiled likelihoods that are monotone increasing over a wide range of ν. The logit model is well approximated by the Gosset model with ν equal to 7 or 8, see e.g. Liu (2004).
- When ν is near zero, say less than 0.2, evaluation of the likelihood becomes problematic due to the dramatically heavy tails of the distribution.

[3] Liu (2004) discusses several variants of the EM algorithm for estimating the Gosset model, but we are not aware of an available implementation of any of these methods. Using Finey's classic skin vasoconstriction data, Liu reports a point estimate of $\hat{\nu} = 0.11$, without any estimate of precision.On the same data we obtain $\hat{\nu} = 0.527$ with a confidence interval of $(0.43, 2.13)$ based on the asymptotic behavior of the likelihood ratio, as described below.

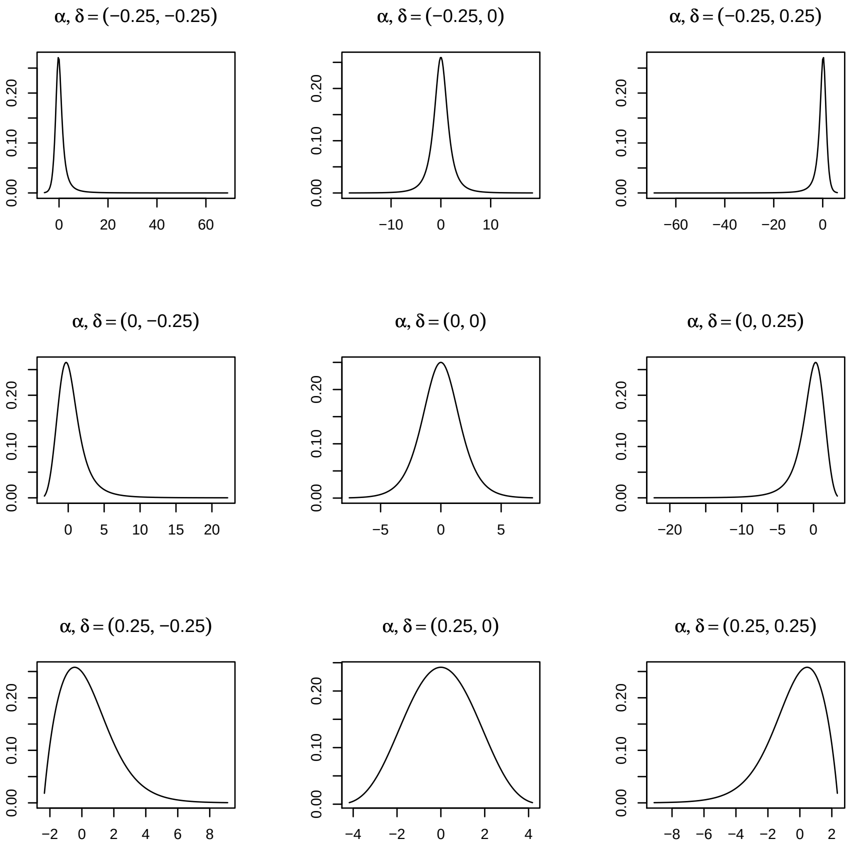

Figure 2. Some examples of the Pregibon (Tukey λ) densities.

- Any realistic hope of distinguishing Gosset models from binary data requires at least moderately large sample sizes. As a rough rule of thumb, we suggest that $n \geq 500$ seems reasonable.

Inference regarding ν can be based on the profiled likelihood and its classical χ^2 asymptotic approximation. A $(1-\alpha)$-confidence interval for ν is thus,

$$I = \{\nu \mid 2(\ell(\hat{\nu}) - \ell(\nu)) \leq \chi^2_{1,1-\alpha}\} \tag{2.1}$$

as illustrated in Figure 3. Inference about the parameters of the linear predictor *conditional on a particular* ν, is easily carried out, but unconditional inference is more of a challenge and perhaps better suited to Bayesian methods, which properly account for uncertainty about ν.

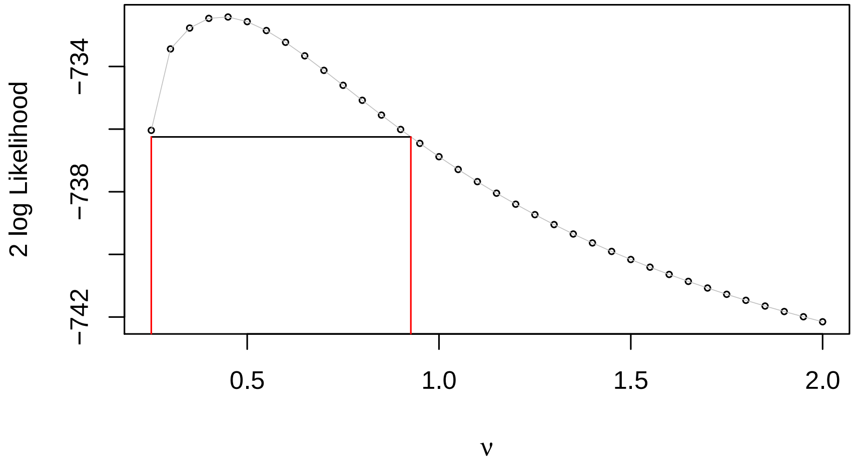

FIGURE 3. Profile likelihood for the Gosset link parameter ν for a model of quit behavior for a large U.S. manufacturing firm. The vertical lines indicate a 95% confidence interval for ν.

2.1. **An Application.** The profiled likelihood appearing in Figure 3 is based on a model of quit behavior for a sample of Western Electric workers,

$$g_{\nu}(\pi_i) = \beta_0 + \beta_1 SEX_i + \beta_2 DEX_i + \beta_3 LEX_i + \beta_4 LEX_i^2$$

where g_ν is the quantile function of the Student t distribution with ν degrees of freedom, π_i is the probability of quitting within 6 months of being hired; SEX_i is the gender of the employee, males coded 1; DEX_i is the score on a preemployment dexterity exam, and LEX_i is years of education. The explanatory variables in this example are taken from the study of Klein, Spady, and Weiss (1991), but the response variable was altered long ago to improve the didactic impact of the model as a class exercise. To this end, quit dates for each individual were generated according to a log Weibull proportional hazard model. In Table 1 we report estimates of the model using several link functions. The maximum likelihood estimate of ν is 0.432 for this example with a 95% confidence interval of $(0.27, 0.93)$ based on the asymptotic χ^2_1 theory of log-likelihood as illustrated in Figure 3.

As an indication of the difference between the fitted models, Figure 4 plots the fitted probabilities of the sample observations for the probit and optimal Gosset model against one another. This PP plot shows that the two models deliver dramatically different estimates of the quit probabilities, even though there is extremely high linear correlation between the estimates of the linear predictor in the two models.

Estimator	β_0	β_1	β_2	β_3	β_4	AIC
Probit	3.549	0.268	-0.053	-0.313	0.012	748.711
Logit	6.220	0.479	-0.094	-0.539	0.021	746.663
Cauchit	8.234	0.677	-0.125	-0.694	0.028	736.881
Gosset	20.353	1.675	-0.297	-1.728	0.069	732.409

Table 1. Estimation of the WECO model with several link functions

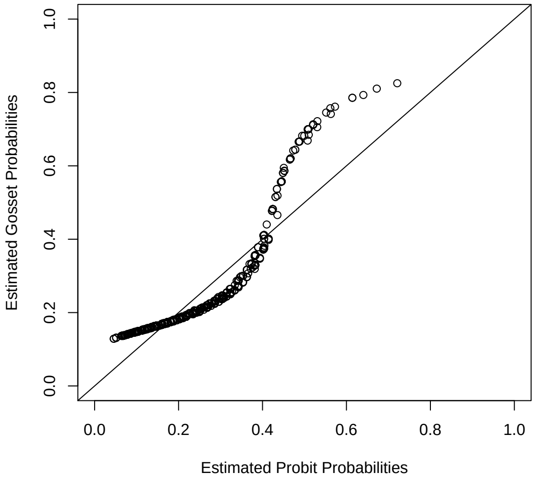

Figure 4. PP Plot of Fitted Probabilities of the Probit and Gosset Models for the WECO data: The solid line is the 45 degree line.

2.2. **Bayesian Methods for the Gosset Link.** The binary response model with Gosset link function attracted early attention in the Bayesian MCMC literature. The latent variable formulation of the model lends itself to Gibbs sampling methods. Thus, Albert and Chib (1993) consider the latent variables $\{Y_i^*\}$ which are assumed to be independent and normally distributed with means, $\{x_i^\top \beta\}$, and variances $\{\lambda_i^{-1}\}$. Observed responses are given by,

$$Y_i = I(Y_i^* > 0).$$

If the λ_i's are independent Gamma$(\nu/2, 2/\nu)$ random variables with density function

$$f(\lambda) = (\lambda^{\nu/2-1} e^{-\nu\lambda/2})/(\Gamma(\nu/2)(\nu/2)^{-\nu/2})$$

then the Y_i^* are independent Student t random variables with location $x_i^\top \beta$, scale 1, and degrees of freedom parameter ν. When we choose a uniform prior for the regression coefficients β, we have the following conditional distributions. The conditional posterior density of ν is

$$\nu|Y^*, \beta, \lambda, Y \sim \pi(\nu) \prod_{i=1}^{n} \frac{\lambda_i^{\nu/2-1} e^{-\nu\lambda_i/2}}{\Gamma(\nu/2)(\nu/2)^{-\nu/2}}.$$

where $\pi(\nu)$ is the prior density of ν. The conditional posterior density of independent random variables $\lambda_1, \ldots, \lambda_n$ is

$$\lambda_i|\beta, Y^*, \nu, Y \sim \text{Gamma}(\frac{\nu+1}{2}, \frac{2}{\nu + (Y_i^* - x_i^T\beta)^2}).$$

Given the foregoing, the conditional density of β can be expressed as,

$$\beta|Y^*, \lambda, \nu \sim N(\hat{\beta}, (X^\top \Lambda X)^{-1})$$

where $\hat{\beta} = (X^\top \Lambda X)^{-1} X^\top \Lambda Y^*$ and $\Lambda = \text{diag}(\lambda_i)$. Finally, the conditional distribution of latent variables is independent with

$$Y_i^*|\beta, \lambda, \nu, Y \sim N(x_i^T\beta, \lambda_i^{-1})$$

truncated on $(0, \infty)$ if $Y_i = 1$ or on $(-\infty, 0)$ if $Y_i = 0$.

The Gibbs sampling procedure can now be implemented by drawing samples sequentially from these conditional distributions. The only remaining difficulty is the somewhat non-standard conditional density of ν. Following Albert and Chib (1993) we discretize this density, that is, we evaluate the conditional density of ν at grid points $\{\nu_1, \cdots, \nu_L\}$ which are chosen according to the prior $\pi(\nu)$.

We first draw $\{\nu_1, \cdots, \nu_L\}$ from the prior distribution $\pi(\nu)$, and evaluate the conditional posterior probabilities of ν at each grid point, say, $\{p_1, \ldots, p_L\}$, and then draw one multinomial observation of ν from the discrete set $\{\nu_1, \cdots, \nu_L\}$ with probabilities $\{p_1, \ldots, p_L\}$. Parameterizing ν in terms of $\xi = 1/\nu$ we impose a uniform prior on the interval $[0, \bar{\xi}]$. We use an equally-spaced grid for ξ on the interval $[0, \bar{\xi}]$ as the most probable realization from the uniform distribution. In our simulations we will consider two values for the upper bound $\bar{\xi}$. Both choices favor low degrees of freedom ν since prior mass is concentrated near the lower bound $1/\bar{\xi}$.

2.3. A Simulation Exercise.

To explore the performance of both maximum likelihood and Bayes estimators for the Gosset link function we report the results of a small simulation experiment designed to evaluate both the accuracy of estimators and their associated confidence/crediblity intervals. We consider 3 model configurations,

$$g_\nu(\pi_i) = \beta_0 + \beta_1 x_i \qquad i = 1, \ldots, n$$

with $\nu = 1, 2$, and 6. In all cases x_i is iid Gaussian with mean zero and standard deviation 5. The linear predictor parameters are fixed at $\beta_0 = 0, \beta_1 = 1$. We consider two sample sizes $n = 500$ and $n = 1000$. Response observations are generated by the latent variable model

$$y_i^* = \beta_0 + \beta_1 x_i + u_i$$

with the u_i iid Student with ν degrees of freedom, and observed response $y_i = I(y_i^* > 0)$.

Criterion	$n = 500$			$n = 1000$		
	$\nu = 1$	$\nu = 2$	$\nu = 6$	$\nu = 1$	$\nu = 2$	$\nu = 6$
MLE						
Mean	1.177	2.837	14.499	1.019	2.231	12.500
Median	1.009	2.038	8.286	0.979	2.039	6.878
MAE	0.228	0.527	3.541	0.154	0.351	2.745
RMSE	1.426	3.386	14.403	0.250	1.008	12.401
Bayes-1						
Mean	0.845	1.811	4.496	0.898	1.902	5.657
Median	0.769	1.515	3.571	0.847	1.786	5.000
MAE	0.273	0.766	2.712	0.213	0.483	2.269
RMSE	0.323	1.050	3.383	0.257	0.710	3.024
Bayes-2						
Mean	0.981	1.862	4.832	0.957	1.898	5.770
Median	0.898	1.579	4.115	0.898	1.774	4.970
MAE	0.180	0.757	2.681	0.159	0.511	2.473
RMSE	0.305	1.071	3.739	0.197	0.663	4.100

Table 2. Performance of the maximum likelihood and Bayes estimators of ν: We report Mean, Median, Mean absolute error (MAE), and root mean squared error (RMSE) of the maximum likelihood and Bayes estimates. Results are based on 500 replications for both sample sizes $n = 500$ and $n = 1000$. Sample median of the posterior distribution of ν is used as a Bayes point estimate. For "Bayes-1", uniform prior for $\xi = 1/\nu$ is placed on an interval [0, 2]. A modified prior for $\xi = 1/\nu$ uniform on [0, 1.4] is used for "Bayes-2".

Table 2 reports the performance of estimators in terms of bias, mean absolute error (MAE), and root mean squared error (RMSE). Performance the maximum likelihood estimator is quite good for the $\nu = 1$ (Cauchy) and $\nu = 2$ cases, exhibiting small bias (within simulation error bounds) and respectable mean absolute and root mean squared error. For $\nu = 6$ case we see more bias upward and larger MAE and RMSE reflecting the difficulty of distinguishing Student distributions with larger degrees of freedom.

To compare performance of the Bayesian procedures with maximum likelihood we use the Gibbs sampling procedure with equally-spaced grid values with 100 points for $\xi = 1/\nu$ on the interval $[0, 2]$. We evaluate the conditional density of ν at grid points and draw a multinomial random variable according to the evaluated probabilities. Note that when ξ is uniform on the interval $[0, 2]$, the prior density for ν is proportional to $1/\nu^2$ on $[0.5, \infty]$, with half of prior mass is placed on $[0.5, 1]$ for ν. Therefore, the prior belief for the link function strongly prefers the lower degrees of freedom.

We generate chains of 20,000 draws for each realization of the simulation. The first 10,000 draws are discarded and every tenth draw was retained thereafter. The sample median of the retained draws is taken as a point estimate of ν, and 95% credible intervals were constructed from the 0.025 and 0.975 quantiles of the retained realizations for each chain.

In Table 2 under the heading "Bayes-1" we report performance measures of the Bayesian point estimates. We observe that the Bayes point estimates exhibit downward bias throughout the experiments. This is not surprising given the fact that our prior prefers lower degrees of freedom. While having a stong prior imposes a cost in terms of bias, there is also compensation in that it reduces the variability of the Bayes estimator. When we compare the MSE and MAE of two estimators, we observe that the Bayesian estimator outperforms the MLE in terms of MSE throughout the experiments, but they show the comparable performance in terms of MAE. This tendency is especially visible for the case of $\nu = 6$. It indicates that MLE tend to yield a few instances which go widely off the mark, but with a strong direction provided by the prior, the Bayes point estimator has less tendency to do so. For example, when $\nu = 6$, the Bayes estimates tend to be clustered around a low degrees of freedom, this seems to be largely attributable to the effect of the prior on ν which tends to heavily discount the likelihood of large values of ν.

To check the sensitivity of our results to the particular choice of the prior for ν we repeated the simulation with a uniform prior for $\xi = 1/\nu$ on the interval $[0, 1.4]$. The results are shown under the heading "Bayes-2". Here, the Bayes point estimator tends to do better in terms of bias, as we can expect from the prior which emphasize less on the region of lower degrees of freedom. Throughout the experiments, we observe that the bias of the Bayes point estimates are much improved.

Table 3 reports rejection frequencies of both the likelihood ratio test and the test based on 95% credible intervals for ν. Column entries represent fixed values of the true ν parameter, while row entries represent fixed values of the hypothesized parameter. Thus, diagonal table entries indicate size of the test, off-diagonal entries report power. Performance of the likelihood ratio test as reflected in rejection rates, frequency of non-coverage, for the test under our three configurations seems to be quite good. Nominal size of the tests is quite accurate, at sample size $n = 500$ there is power roughly one-half of distinguishing $\nu = 1$ from $\nu = 2$, and similar power for distinguishing $\nu = 2$ from $\nu = 6$; power increases to about 0.8 when the sample size increases to $n = 1000$. At these sample sizes we can distinguish $\nu = 1$ from $\nu = 6$ with very high probability.

Table 3 also reports non-coverage frequencies for the Bayesian credibility intervals for ν. These can be compared to the rejection frequencies of the likelihood ratio test. We count how often the 95% credibility interval fails to include the value specified

Frequency	$n = 500$			$n = 1000$		
	$\nu_0 = 1$	$\nu_0 = 2$	$\nu_0 = 6$	$\nu_0 = 1$	$\nu_0 = 2$	$\nu_0 = 6$
MLE						
$H_0:\ \nu_0 = 1$	0.062	0.530	0.988	0.056	0.842	1.000
$H_0:\ \nu_0 = 2$	0.458	0.056	0.516	0.776	0.070	0.808
$H_0:\ \nu_0 = 6$	0.930	0.522	0.010	1.000	0.814	0.042
Bayes-1						
$H_0:\ \nu_0 = 1$	0.308	0.446	0.954	0.278	0.804	1.000
$H_0:\ \nu_0 = 2$	0.766	0.224	0.302	0.896	0.140	0.692
$H_0:\ \nu_0 = 6$	0.982	0.754	0.174	0.998	0.902	0.100
Bayes-2						
$H_0:\ \nu_0 = 1$	0.072	0.470	0.952	0.130	0.802	1.000
$H_0:\ \nu_0 = 2$	0.756	0.256	0.372	0.886	0.174	0.726
$H_0:\ \nu_0 = 6$	0.972	0.754	0.168	1.000	0.886	0.098

Table 3. Rejection frequencies of the likelihood ratio test and the 95% Bayesian credibility intervals. : Column entries represent fixed values of the true ν parameter, while row entries represent fixed values of the hypothesized parameter. Thus, diagonal table entries indicate size of the test, off-diagonal entries report power. Results are based on 500 replications for each sample size. For the likelihood ratio test, we use 95% confidence interval based on χ^2_1 distribution. For the Bayes test, we count how often 95% credibility interval does not include the value specified in the null hypothesis.

in the null hypothesis. Just as before, "Bayes-1" corresponds to the choice of prior, $\xi = 1/\nu$ uniform on $[0, 2]$. Here too the strong prior influences the performance of the Bayesian procedure in terms of correct nominal coverage probability. For example, when the true value of parameter $\nu = 1$, the credible interval tends to exclude the true value 30 percent of the time. Coverage is somewhat better for $\nu = 2$ and $\nu = 6$, but still the prior tends to exaggerate our confidence in the credible interval. Of course, Bayesian credible intervals are not meant to be invariant procedures that attain nominal coverage for any fixed value of the population parameter, rather they are intended to represent what one should believe after observing the data *given the prior*. We repeated the simulation with a uniform prior for $\xi = 1/\nu$ on the interval $[0, 1.4]$. These results, "Bayes-2" show some improvement for the Cauchy, $\nu = 1$ case, indicating that the previous choice of the prior apparently placed too much mass on the interval of ν between 0.5 and 1, and thus too often obtained credibility intervals that excluded the value 1.

A more direct measure of performance of the alternative link functions is obtained by assessing the accuracy of the estimated success probabilities. To this end, we

Estimator	d_1			d_2			d_∞		
	$\nu = 1$	$\nu = 2$	$\nu = 6$	$\nu = 1$	$\nu = 2$	$\nu = 6$	$\nu = 1$	$\nu = 2$	$\nu = 6$
Probit	0.065	0.038	0.013	0.133	0.119	0.092	0.186	0.171	0.136
Cauchit	0.016	0.024	0.033	0.022	0.034	0.048	0.055	0.107	0.167
MLE	0.020	0.016	0.012	0.027	0.024	0.021	0.070	0.065	0.058
Bayes-2	0.019	0.018	0.014	0.026	0.028	0.026	0.069	0.081	0.078

Table 4. Performance of Several Binary Response Estimators : In each run of Monte carlo experiment, three performance measures, d_1, d_2, d_∞ are calculated for probit, cauchit, Gosset MLE, Bayes estimator with Gosset link function. Reported values are the sample mean of 500 replications for the sample size n= 500.

consider the family of performance measures,

$$d_p(\hat{F}, F) = (\int |\hat{F}(x^\top \hat{\beta}) - F(x^\top \beta)|^p dG(x))^{1/p}$$

We will consider the three conventional choices of $p \in \{1, 2, \infty\}$. In Table 4 we report estimates of these performance measures for four estimators: probit, cauchit, the Gosset MLE, and the Bayesian posterior coordinatewise median for the Gosset model with the more concentrated prior described above. Estimates of our performance measures are obtained by substituting the empirical measure of the x observations into the above expression.

Performance of the MLE and Bayes estimators are quite similar over these settings. There is a modest sacrifice of performance to the cauchit estimator when the model is Cauchy, and rather substantial gains over the probit estimator in all three cases.

3. The Pregibon Link

Symmetry of the link function may be inappropriate for some applications and a convenient two-parameter family of links is provided by the function

$$g(u) = \frac{u^{\alpha-\delta} - 1}{\alpha - \delta} - \frac{(1-u)^{\alpha+\delta} - 1}{\alpha + \delta}.$$

This is the parameterization used by Pregibon to derive his goodness-of-link score test. The Pregibon test provides a convenient one-step procedure to generate starting values, but to the best of our knowledge there has been no systematic effort to explore the behavior of the maximum likelihood and Bayes estimators for this family of link functions.

Inference based on the profiled likelihood for (α, δ) can be fairly easily carried out by plotting contours of the likelihood surface. This approach provides an alternative, albeit a slightly more computational demanding one, to Pregibon's score test. An example is shown in Figure 5 based on 500 observations from a simple bivariate

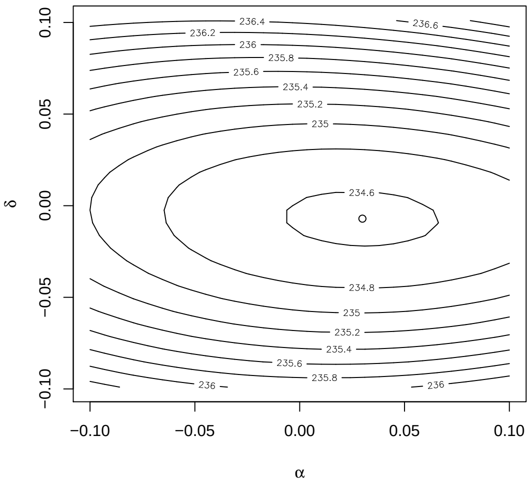

Figure 5. Contour plot of the profiled likelihood function for the Pregibon link model. The example is based on 500 observations from a simple bivariate logit model. Contours a label in AIC units so difference in contours can be compared directly to the quantiles of the χ^2_2 distribution.

logistic model. Obviously, in this example the data are not very informative about the parameters (α, δ); the conventional χ^2_2 theory gives a confidence region that includes most of the densities illustrated in Figure 2.

3.1. **Bayesian implementation for the Pregibon link.** As Figure 2 clearly shows, given (α, δ), both the density and the cumulative distribution function of the Pregibon link function $g(u)$ are well defined, although the actual computation is carried out by numerical procedures. Let $f(\cdot, \alpha, \delta)$ and $F(\cdot, \alpha, \delta)$ denote the density and distribution functions respectively.

The joint posterior of the model with the Pregibon link function is

$$\pi(Y^*, \beta, \alpha, \delta | Y) \sim \pi(\alpha, \delta)\pi(\beta) \cdot \prod_{i=1}^{n} M_i \; f(Y_i^* - x_i^\top \beta, \alpha, \delta)$$

where $M_i = \{I(Y_i^* > 0 | Y_i = 1)/(1 - F(-x_i^\top \beta, \alpha, \delta)) + I(Y_i^* < 0 | Y_i = 0)/F(-x_i^\top \beta, \alpha, \delta)\}$. We divide the joint posterior into three conditional posteriors which correspond to three blocks of parameters; the latent variables Y^*, the shape parameters of the Pregibon link α, δ, and finally the linear predictor β. Our Gibbs sampling procedure is determined by the following conditional distributions.

(i) Given Y, β, α, δ, the conditional distribution of Y_i^* is

$$Y_i^* | Y_i, \alpha, \delta, \beta \sim M_i \cdot f(Y_i^* - x_i^\top \beta, \alpha, \delta)$$

which means that when we observe $Y_i = 1$, $Y_i^* \sim F(\cdot - x_i^\top \beta, \alpha, \delta)$ truncated on $(0, \infty)$, and when $Y_i = 0$, $Y_i^* \sim F(\cdot - x_i^\top \beta, \alpha, \delta)$ truncated on $(-\infty, 0)$. If the outcome is $y_i = 1$, we draw $u \sim \text{Uniform}[F(-x_i^\top \beta, \alpha, \delta), 1]$ and invert it by $Y^* = F^{-1}(u, \alpha, \delta) + x_i^\top \beta$. If $y_i = 0$, we draw $u \sim \text{Uniform}[0, F(-x_i^\top \beta, \alpha, \delta)]$ and invert it in the same way.

(ii) Given Y^*, Y, β, the conditional distribution of α, δ is

$$\alpha, \delta | Y, Y^*, \beta \sim \pi(\alpha, \delta) \prod_{i=1}^{n} f(Y_i^* - x_i^\top \beta, \alpha, \delta)$$

This step is done by the Metropolis algorithm. We assume a flat prior, $\pi(\alpha, \delta) \propto 1$ inside a square $\alpha \in [-d, d], \delta \in [-d, d]$. In the t-th step of the Metropolis algorithm, given the current value of parameters $(\alpha_{t-1}, \delta_{t-1})$, we propose a candidate pair (α^*, δ^*), and calculate the ratio of the densities

$$r = \frac{\prod_{i=1}^{n} f(Y_i^* - x_i^\top \beta, \alpha^*, \delta^*)}{\prod_{i=1}^{n} f(Y_i^* - x_i^\top \beta, \alpha_{t-1}, \delta_{t-1})}$$

and set $(\alpha_t, \delta_t) = (\alpha^*, \delta^*)$ with probability $\min(r, 1)$, or keep the old values $(\alpha_t, \delta_t) = (\alpha_{t-1}, \delta_{t-1})$ otherwise.

(iii) Given Y^*, Y, α, δ, the conditional distribution of β is

$$\beta | Y, Y^*, \alpha, \delta \sim \pi(\beta) \prod_{i=1}^{n} f(Y_i^* - x_i^\top \beta, \alpha, \delta)$$

We approximate the above posterior for β by a multivariate normal distribution centered at the posterior mode. More specifically, assume a flat prior $\pi(\beta) \propto 1$ and let the logarithm of the posterior $l(\beta | Y, Y^*, \alpha, \delta) = \sum_{i=1}^{n} \log f(Y_i^* - x_i^\top \beta, \alpha, \delta)$. Let $\hat{\beta}$ be the mode of the log-posterior, and $l_{\beta\beta}(\hat{\beta})$ be the Hessian matrix of the log-posterior, evaluated at the mode, then the posterior distribution of β is approximated by $N(\beta | \hat{\beta}, [-l_{\beta\beta}(\hat{\beta})]^{-1})$. Once we find the mode and associated Hessian matrix, we generate β from the multivariate normal distribution.

The "Metropolis within Gibbs" approach was suggested by Tierney (1994), and further illustrated in, e.g. Gilks, Best, and Tan (1995) and Geweke and Tanizaki (2001). This is a standard technique in MCMC literature when the direct sampling from one or more of the conditional posteriors is difficult. The approximation of the posterior based on single or multiple modes of the posterior is suggested in Chapter 12 of Gelman, Carlin, Stern, and Rubin (2004). We could, as suggested by the referee, use another step of Metropolis algorithm with Gibbs here but our main concern of not using it was the possiblilty that the dimension of the β is potentially quite large. It is known that in large dimensional problems the Metropolis-Hastings algorithm may converge quite slowly, see e.g. Chapter 7 in Robert and Casella (2004) and Section 8 in Neal (2003).

In our simulations, we iterate the whole chain of Gibbs sampling 2000 times. In every Gibbs step, for the parameters (α, δ), the Metropolis algorithm with the length 100 is used. We discard the first 1000 realizations of Gibbs sampling and keep every fifth draw after the burn-in period. Bayesian inference is based on the posterior samples of parameters we obtained. When we compare the performance of the Bayes estimator with MLE, we use the coordinate-wise posterior median for the Bayes point estimate. For the credible set of the Pregibon link parameters (α, δ), we calculate the contours of a kernel density estimate of the bivariate density of (α, δ), of the posterior, and choose a contour curve which includes 95% of total mass.

3.2. **A Simulation Experiment.** In an effort to gain some further experience we undertook another small simulation exercise to evaluate maximum likelihood and Bayes estimators of the (α, δ) parameters of the generalized λ family. We compared behavior of the one-step Pregibon estimates, the maximum likelihood estimates, and Bayes point estimates of (α, δ). In Tables 5 and 6 we report performances of each estimator for four simulation configurations of the shape parameters: $(\alpha, \delta) = (0, 0), (\alpha, \delta) = (-.25, 0), (\alpha, \delta) = (-.15, .15)$ and $(\alpha, \delta) = (-.4, .3)$. We report median and mean bias and median absolute and mean squared error for two sample sizes $n = 500$ and $n = 1000$.

In the first configuration the one-step estimator is starting from a consistent estimate (the truth) so it and the MLE have the same asymptotic distribution. This is reflected in Tables 5 and 6 where the two estimators show almost the same performance in terms of bias and MSE. It is not surprising that the MLE improves substantially on the Pregibon one-step in the non-null cases, since the latter is seriously biased in those cases, due to the inconsistent starting value.

It is apparent that the skewness parameter δ is more precisely estimated than is α, the parameter that governs tail behavior. To some degree this is due to the fact that the scale of the coefficient vector of the linear predictor can compensate for variation in α. This can be made more precise by exploring the asymptotic behavior of the MLE. For fixed designs like those used in the simulation experiment, we compared asymptotic confidence regions for (α, δ) at the logit model for both known and unknown linear predictor vector β. Knowing β reduces the asymptotic standard error of $\hat{\alpha}$ by about half, but has almost no effect on the precision of $\hat{\delta}$. These findings are quite consistent with the simulation results reported in the tables. Indeed, there is a striking resemblance in both size and orientation between the confidence ellipses generated by the likelihood contours and their normal theory asymptotic counterparts.

For all four configurations, Bayes point estimates correspond with MLEs nicely, and both maximum likelihood and Bayes estimators performed better than one-step

Criterion	α				δ			
	$(0,0)$	$(-.25,0)$	$(-.15,.15)$	$(-.4,.3)$	$(0,0)$	$(-.25,0)$	$(-.15,.15)$	$(-.4,.3)$
Mean								
One-step	0.160	-0.027	0.131	0.296	0.009	0.007	0.179	0.401
MLE	0.108	-0.184	-0.077	-0.320	0.007	0.009	0.172	0.330
Bayes	-0.043	-0.365	-0.256	-0.493	0.001	0.003	0.163	0.318
Median								
One-step	0.105	-0.072	0.099	0.240	0.002	0.004	0.157	0.347
MLE	0.093	-0.213	-0.107	-0.333	0.002	0.010	0.169	0.319
Bayes	-0.033	-0.362	-0.237	-0.523	0.001	-0.001	0.162	0.317
MAE								
One-step	0.112	0.178	0.249	0.640	0.056	0.035	0.085	0.140
MLE	0.158	0.157	0.166	0.187	0.059	0.068	0.065	0.076
Bayes	0.154	0.196	0.180	0.202	0.055	0.066	0.069	0.069
RMSE								
One-step	0.276	0.262	0.340	0.761	0.126	0.098	0.140	0.306
MLE	0.241	0.230	0.252	0.316	0.112	0.104	0.106	0.143
Bayes	0.238	0.269	0.273	0.249	0.081	0.094	0.100	0.109

Table 5. Performance of one-step, maximum likelihood, and Bayes estimators of Pregibon link parameters (α, δ) for sample size $n = 500$. Results are based on 500 replications.

estimator. Estimation of the shape parameter α is more variable than that of skewness parameter δ. For the bias in shape parameter α, we observe that maximum likelihood estimates tend to choose distributions whose tails are thinner than the true one, on the other hand, the Bayes point estimates are inclined to thicker tails. In terms of MSE and MAE, both Bayes and maximum likelihood estimates performs comparably.

In Table 7 we report rejection frequencies of the (logit) null hypothesis that $(\alpha, \delta) = (0, 0)$. The line labeled GOL gives the rejection frequencies for the Pregibon goodness-of-link test, while the LR test reports the frequency that

$$2(\ell(\hat{\beta}, \hat{\alpha}, \hat{\delta}) - \ell(\tilde{\beta}, 0, 0)) > 5.99$$

where $\tilde{\beta}$ is usual logit estimator of β. The likelihood ratio test performs somewhat better than the Pregibon score test in terms of its ability to detect departures from the logit model, at the price of some inflation in size.

The last row reports the non-coverage frequencies of the 95% credible sets. It will be noted that for the logistic (null) case, $(\alpha, \delta) = (0, 0)$, the Bayes credible sets tend to exclude the origin more often than the nominal coverage probability which is supposed to be 0.05. When the sample size $n = 500$, the rejection frequency is 0.11, whereas with larger sample size $n = 1000$, the rejection frequency is 0.09. We have conducted several experiments to further explore the behavior of Bayesian credible

Criterion	α				δ			
	$(0,0)$	$(-.25,0)$	$(-.15,.15)$	$(-.4,.3)$	$(0,0)$	$(-.25,0)$	$(-.15,.15)$	$(-.4,.3)$
Mean								
One-step	0.082	-0.099	0.075	0.228	0.005	0.002	0.159	0.356
MLE	0.054	-0.241	-0.105	-0.387	0.003	0.002	0.159	0.314
Bayes	-0.023	-0.336	-0.190	-0.474	-0.001	-0.003	0.164	0.317
Median								
One-step	0.049	-0.111	0.054	0.199	0.003	0.003	0.143	0.332
MLE	0.037	-0.240	-0.116	-0.378	0.001	0.001	0.157	0.310
Bayes	-0.027	-0.321	-0.189	-0.480	0.000	-0.003	0.161	0.314
MAE								
One-step	0.073	0.139	0.204	0.599	0.038	0.020	0.063	0.108
MLE	0.092	0.097	0.092	0.118	0.041	0.043	0.047	0.056
Bayes	0.098	0.130	0.107	0.148	0.041	0.045	0.042	0.054
RMSE								
One-step	0.158	0.162	0.260	0.668	0.069	0.042	0.100	0.195
MLE	0.150	0.138	0.166	0.188	0.062	0.062	0.073	0.081
Bayes	0.150	0.199	0.169	0.206	0.063	0.066	0.073	0.082

Table 6. Performance of one-step, maximum likelihood, and Bayes estimators of Pregibon link parameters (α, δ) for sample size $n = 1000$. Results are based on 500 replications.

Test	$n = 500$				$n = 1000$			
	$(0,0)$	$(-.25,0)$	$(-.15,.15)$	$(-.4,.3)$	$(0,0)$	$(-.25,0)$	$(-.15,.15)$	$(-.4,.3)$
GOL	0.050	0.168	0.416	0.890	0.056	0.354	0.684	0.996
LR	0.084	0.260	0.480	0.882	0.074	0.488	0.716	0.944
Bayes	0.108	0.368	0.446	0.872	0.092	0.484	0.726	0.984

Table 7. Performance of Pregibon goodness-of-link (GOL), likelihood ratio (LR) tests, and Bayesian credible sets (Bayes) of the logistic hypothesis $H_0 : (\alpha, \delta) = (0, 0)$: The table entries report rejection frequencies for the two tests under the null and two configurations of the alternative hypothesis. All entries are based on 500 replications of the test.

sets. For example, expanding the length of the Gibbs sampling chain from 2000 to 30000 reduced exclusion frequency to 0.075 for $n = 1000$. In terms of the power of the test based on confidence and credible regions, both maximum likelihood and Bayesian methods show comparable results.

Table 8 is comparable to Table 4. It compares the performance of estimators in terms of the estimated probabilities. As the Pregibon link parameters deviate from

Estimator	**n=500**				**n=1000**			
	(0,0)	(-.25,0)	(-.15,.15)	(-.4,.3)	(0,0)	(-.25,0)	(-.15,.15)	(-.4,.3)
d_1								
Logit	0.013	0.020	0.023	0.042	0.009	0.017	0.021	0.041
MLE	0.017	0.019	0.019	0.021	0.012	0.013	0.013	0.015
Bayes	0.018	0.021	0.021	0.021	0.013	0.015	0.014	0.015
d_2								
Logit	0.019	0.027	0.031	0.054	0.014	0.023	0.028	0.052
MLE	0.026	0.026	0.028	0.029	0.018	0.018	0.019	0.020
Bayes	0.027	0.029	0.030	0.030	0.019	0.021	0.020	0.022
d_∞								
Logit	0.044	0.063	0.071	0.120	0.031	0.055	0.064	0.116
MLE	0.063	0.062	0.067	0.072	0.042	0.043	0.044	0.050
Bayes	0.066	0.070	0.074	0.075	0.045	0.050	0.047	0.054

Table 8. Performance of Several Binary Response Estimators : Estimated probabilities with Pregibon link families.

the null logistic case further, the fitted probabilities from both maximum likelihood and Bayes estimators are much closer to the true probabilities than those obtained from logit model.

4. Propensity Score Matching Methods

In this section we reanalyze data from the National Supported Work (NSW) Demonstration experiment. The effectiveness of the job training program is measured by the post-intervention income levels of program participants. Several authors, including Dehejia and Wahba (1999), Dehejia and Wahba (2002), and Smith and Todd (2005), have explored the use of propensity score matching methods for estimating average treatment effect using this data.

Most empirical studies use logit or probit model to estimate the propensity score, and it is usually claimed that these models produce similar results. But as we have already argued this should not be taken as evidence that a broader class of link functions should also be summarily dismissed. Since the matching estimator of the average treatment effect crucially depends on the first step estimation of the propensity score, as Shaikh, Simonsen, Vytlacil, and Yildiz (2005) argued, the misspecified propensity score may lead to inconsistent estimates of the average treatment effect.

Since the logistic model was used in all the foregoing studies involving the NSW data, we decided to investigate whether logistic specification can be justified within the larger class of Pregibon link models. This class is not nearly as general as some semi-parametric estimators that have been proposed in the literature, but more narrowly targeted alternatives are sometimes preferably to more omnibus methods when sample sizes are moderate.

The literature on the propensity score matching methods focuses on several related robustness issues. LaLonde (1986) has argued that applied to the NSW data, various econometric estimators produce widely different results, and fail to replicate benchmarks obtained from an experimental sample, highlighting the risk involved in using observational studies. In contrast Dehejia and Wahba (1999) have argued that one can still find reasonably good estimators with the non-experimental data. They showed that a matching method based on the propensity score produced results very close to the benchmark from the experiments.

There has been considerable controversy regarding the sensitivity of propensity score matching methods to sample selection and covariate specification. Smith and Todd (2005) have noted that Dehejia and Wahba's results are highly sensitive to their choice of a particular subsample from LaLonde's original data. The propensity score matching method performed remarkably well in Dehejia and Wahba's subsample, but did poorly in both LaLonde's sample and Smith and Todd's subsample. In addition, the outcome of the matching method varied substantially with the choice of covariates used in estimating the propensity score. Dehejia (2005) has responded that one has to choose different set of variables in a different set of data, and showed that if one carefully chooses the right set of variables, the matching method based on the propensity score performs well in all samples.

We would prefer not to take a position on these controversial aspects of the specification and focus attention instead on the simpler issue of the choice of link function. We adopt the set of covariates chosen by Dehejia (2005) in each treatment-control group and estimate the propensity score employing the Pregibon link function and test whether logistic model is justified; the tests indicate that the logistic specification is implausible.

We considered three versions of the treatment sample: the LaLonde (1986) original treatment sample, the Dehejia and Wahba (1999) sample which is a proper subsample of LaLonde (1986), and the Smith and Todd (2005) sample, which is a proper subsample of the Dehejia and Wahba sample. For the control group, we use full PSID sample in LaLonde (1986). The last two rows of Table 9 show the sample size in each treatment and control group. Two covariate selections were considered: the set of covariates chosen in Dehejia (2005), which is called Dehejia specification[4] and a simpler specification which include only linear terms in the covariates without indicators, quadratics, and interaction terms.[5] The unbalanced sample sizes of the control and

[4] The following variables are used in each sample. LaLonde sample : $re75$ (real income in 1975), $married$, $black$, $hispanic$, age, $school$, $black \cdot school$, $hispanic \cdot re75$, $nodegree \cdot school$. Dehejia-Wahba sample : $re74$, $re75$, $married$, $black$, $hispanic$, age, $school$, $married \cdot u75$, $nodegree \cdot u74$ ($u74$ = indicator taking 1 when $re74 = 0$). Smith-Todd sample : $re74$, $re75$, $married$, $black$, $hispanic$, age, $school$, $hispanic \cdot school$, $re74^2$. For further details, see Table 2 in Dehejia (2005).

[5] The linear specification includes the following variables. $married$, $black$, $hispanic$, age, $school$, $nodegree$, $re74$, $re75$.

Estimator	**LaLonde**		**Dehejia-Whaba**		**Smith-Todd**	
	Dehejia	Linear	Dehejia	Linear	Dehejia	Linear
MLE						
α	-0.196	-0.334	-0.494	-0.634	-1.260	-1.286
δ	0.056	0.145	0.244	0.311	0.887	0.906
p-value	0.387	0.174	0.006	0.001	0.007	0.005
Bayes						
α	-0.046	-0.236	-0.366	-0.570	-1.712	-0.620
δ	-0.040	0.107	0.220	0.307	1.857	0.354
p-value	0.030	0.001	0.001	0.001	0.001	0.001
Number of obs.						
control group	2490	2490	2490	2490	2490	2490
treatment group	297	297	185	185	108	108

Table 9. Comparison of Logistic and Pregibon Models for the NSW Application: The table gives point estimates and "p-values" for tests of the logit model against the more general Pregibon specification for three choices of sample and two specifications of the covariates, as discussed in the text. The "p-values" for the likelihood ratio tests are based on the classical asymptotic χ^2 theory; while for the Bayesian MCMC methods they are computed from the contours of the bivariate (α,δ) posterior density plot.

treatment groups contributes to the difficulty of estimating average treatment effects, as well as the estimation of the propensity score.

In Table 9, we report point estimates of the parameters of the Pregibon link function, and p-values of "tests" of the logistic hypothesis for the various sample and covariate specifications. The Bayes credibility regions assign negligible probability to the logistic hypothesis; as do the the likelihood ratio tests except in the case of the LaLonde sample.

Figure 6 compares estimated propensity scores from logit model with the propensity scores from the maximum likelihood estimates of the Pregibon model. The 45 degree line corresponds to a perfect match in terms of fitted probabilities. Departures from the 45 degree line indicate that the propensity scores differ. The top two panels illustrate results for the Dehejia-Wahba treatment group, the bottom panels depict results for the Smith-Todd treatment group. We only plot the propensity scores of the treatment group; propensity scores of the control groups display similar tendencies. Systematic discrepancies between the logit and Pregibon specifications of the link may result in misleading propensity scores and consequently misleading estimates of treatment effects.

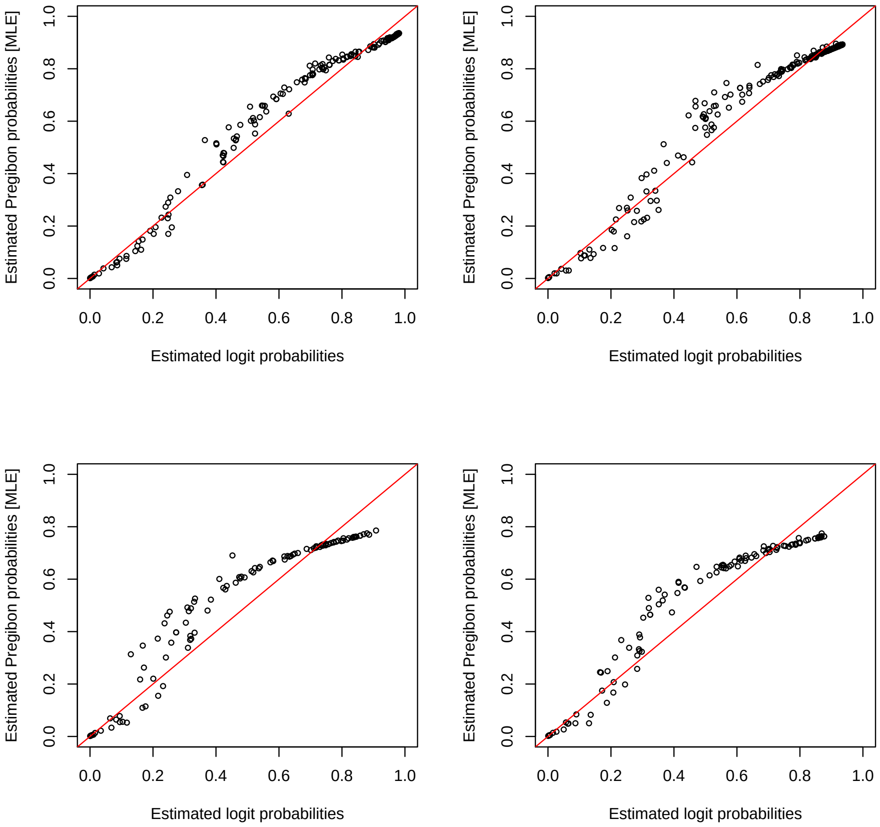

Figure 6. PP Plots of the estimated probabilities of the logit and Pregibon models. The solid line is 45 degree line. The Top-left is the results from the Dehejia-Wahba treatment group/Dehejia specification. The Top-right hand side is for Dehejia-Wahba treatment group/linear specification. The bottom-left is Smith-Todd treatment group/Dehejia specification, and the bottom-right is Smith-Todd treatment group/linear specification. We plot propensity scores of the treatment group. The propensity score of the control groups display the same tendency.

5. Conclusion

Some simple methods for introducing parametric link functions for binary response models have been described and evaluated in some very limited simulation experiments. Much more general semi-parametric methods for analyzing binary response are, of course, already available in the literature, however sometimes intermediate, parametric methods can also provide additional insight. The Gosset and Pregibon models span a reasonably large class of plausible link functions that significantly expand on the traditional logit/probit options. Both maximum likelihood and Bayes point estimators perform well in our investigations. Bayesian credibility regions produced by our MCMC methods were somewhat optimistic in both simulation settings in the sense that observed coverage was somewhat less than the nominal level of the intervals. However, in the Gosset setting this could be attributed to the lack of equivariance induced by the prior, while in the Pregibon setting with a less informative prior results were quite comparable to the likelihood ratio intervals.

Appendix A. Implementation of MCMC for Parametric Links

We describe briefly in this appendix the implementation of MCMC method with the Pregibon link function. Full details of the implementation are available from the url: `http://www.econ.uiuc.edu/ roger/research/links/ links.html`. The order of the presentation follows the three blocks of the Gibbs sampling discussed in page 11-12. First, to generate the latent variable Y^*, given the values of other parameters $\beta, (\alpha, \delta)$, we define a function

```
latent<-function(y,x,beta,ab,tol = 1e-10){
   yhat <- cbind(1,x)%*%beta
   w    <- pPregibon(-yhat,ab[1],ab[2])
   ww   <- pmax(tol,pmin(w,(1-tol)))
   umin <- ifelse(y,ww,0)
   umax <- ifelse(y,1,ww)
   u    <- runif(length(yhat),umin,umax)
   z    <- qPregibon(u,ab[1],ab[2])+yhat
  return(z)
  }
```

Second, to draw the Pregibon shape parameters (α, δ), we use the Metropolis algorithm with length `M` within Gibbs sampling. The shape parameters in Pregibon distribution (α, δ) are defined on square domain $[-d, d]^2$. A new candidate (α', δ') is generated from a bivariate uniform distribution on a box, centered at the current values $((\alpha, \delta)$, extended to both direction by `L`, that is, the proposal distribution is $p(\alpha', \delta' | \alpha, \delta) \sim U[\alpha - L, \alpha + L]^2$.

```
proposal<-function(ab, d = 1, L= .02){
     a<-runif(1,min=max(-d,(ab[1]-L)),max=min(d,(ab[1]+L)))
     b<-runif(1,min=max(-d,(ab[2]-L)),max=min(d,(ab[2]+L)))
   return(c(a,b))
  }
```

The function `Metropolis` performs the Metropolis iterations `M` times inside each Gibbs step. For the parameters in Metropolis algorithm, such as, `M`, `d`, and `L`, appropriate values should be chosen by users.

```
Metropolis <- function(x, z, M = 100, beta, ab, tol=1e-10){
    ehat <- z - cbind(1,x)%*%beta
    u    <- runif(M)
    for(i in 2:M){
        proposed <- proposal(ab)
        R0 <- sum(log(dPregibon(ehat,proposed[1],proposed[2]) + tol))
        R1 <- sum(log(dPregibon(ehat,ab[1],ab[2]) + tol))
        adj<- max(R0,R1)
        R  <- exp(R0-adj)/exp(R1-adj)
        if(u[i] < R) ab <- proposed
       }
   return(ab)
  }
```

The results we reported in our paper are based on the following choices. Our default length of the Metropolis iteration is 100. The default value for the domain of Pregibon parameters is d= 1. For the proposal distribution, we have tried several values for L from 0.01 to 0.05, but the results in this paper are based on L= 0.02. We have tried both shorter or longer chains, but results seem to be stable.

Finally, for the regression coefficients β given Y^*, α, δ, we define the logarithm of the conditional posterior with uniform prior, and use R's nlm function to find the mode of log-posterior and its Hessian matrix. Since the default of nlm routine is minization, we define a negative log-posterior and minimize it. The function mvrnorm from the R package MASS, generates a random vector from the specified multivariate normal distribution.

```
f <- function(theta, x, z, a1, b1, tol=1e-10){
   sum(-log(dPregibon((z-theta[1]-theta[2]*x),a1,b1) + tol))
   }
newbeta <- function(x, z, ab, beta){
    out<-nlm(f,beta,x=x,z=z,a1=ab[1],b1=ab[2],hessian=TRUE)
    return(mvrnorm(n=1,mu=out$estimate,Sigma=solve(out$hessian)))
  }
```

To initiate the chain, we need starting values for the parameters. In our Monte carlo studies, for the initial value of β, we use the outcome of the logistic regression, and for the initial values of (α, δ), we simply start from the origin. We tried other ways of choosing initial values of α, δ, such as generating them from a uniform distribution, but it did not change the results. So we have the following initialization before Gibbs iteration. Given all three blocks of the conditional posterior sampling, we can now run Gibbs sampling as follows:

```
# Initialization
beta0 <- glm(y~x,family=binomial(link="logit"))$coef
ab0   <- c(0,0)
betav <- matrix(0,2,G)
```

```
abv   <- matrix(0,2,G)

# Main MCMC Loop
for(i in 1:G){
   z     <- latent(y,x,beta0,ab0)
   beta1 <- newbeta(x,z,ab0,beta0)
   ab1   <- Metropolis(x,z,M,beta1,ab0)
   betav[,i] <- beta1
   abv[,i]   <- ab1
   beta0     <- beta1
   ab0       <- ab1
  }
```

References

Albert, J. H., and S. Chib (1993): “Bayesian analysis of binary and polychotomous response data,” *Journal of the American Statistical Association*, 88, 669–679.

Brent, R. (1973): *Algorithms for Minimization without Derivatives*. Prentice-Hall.

Chambers, E. A., and D. R. Cox (1967): “Discrimination between alternative binary response models,” *Biometrika*, 54, 573–578.

Dehejia, R. (2005): “Practical propensity score matching: a reply to Smith and Todd,” *Journal of Econometrics*, 125, 355–364.

Dehejia, R. H., and S. Wahba (1999): “Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs,” *Journal of the American Statistical Association*, 94, 1053–1062.

——— (2002): “Propensity Score Matching Methods for Nonexperimental Causal Studies,” *The Review of Economics and Statistics*, 84, 151–161.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004): *Bayesian Data Analysis*. Chapman-Hall, Boca Raton, Florida, second edn.

Geweke, J., and H. Tanizaki (2001): “Bayesian estimation of state-space models using the Metropolis-Hastings algorithm within Gibbs sampling,” *Computational Statistics and Data Analysis*, 37(2), 151–170.

Gilks, W. R., N. G. Best, and K. K. C. Tan (1995): “Adaptive Rejection Metropolis Sampling within Gibbs Sampling,” *Applied Statistics*, 44(4), 455–472.

Hastie, T., and R. Tibshirani (1987): “Non-parametric Logistic and Proportional Odds Regression,” *Applied Statistics*, 36, 260–276.

Klein, R., and R. Spady (1993): “An efficient semiparametric estimator for binary response models,” *Econometrica*, 61, 387–421.

Klein, R., R. Spady, and A. Weiss (1991): “Factors Affecting the Output and Quit Propensities of Production Workers,” *The Review of Economic Studies*, 58, 929–954.

Koenker, R. (2006): “Parametric Links for Binary Response,” *R News*, 6, 32–34.

LaLonde, R. J. (1986): “Evaluating the Econometric Evaluations of Training Programs with Experimental Data,” *The American Economic Review*, 76, 604–620.

Liu, C. (2004): “Robit regression: a simple robust alternative to logistic and probit regression,” in *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, pp. 227–238. Wiley, Chichester.

Manski, C. F. (1975): “Maximum Score Estimation of the Stochastic Utility Model of Choice,” *Journal of Econometrics*, 3, 205–228.

McCullagh, P., and J. A. Nelder (1989): *Generalized linear models (Second edition)*. Chapman & Hall Ltd.

Morgan, B. J. T., and D. M. Smith (1992): “A note on Wadley’s problem with overdispersion,” *Applied Statistics*, 41, 349–354.

Neal, R. M. (2003): “Slice sampling,” *The Annals of Statistics*, 31(3), 705–767, With discussions and a rejoinder by the author.

Newton, M. A., C. Czado, and R. Chappell (1996): “Bayesian Inference for Semiparametric Binary Regression,” *Journal of the American Statistical Association*, 91, 142–153.

Pregibon, D. (1980): “Goodness of link tests for generalized linear models,” *Applied Statistics*, 29, 15–24.

Prentice, R. L. (1976): “Generalization of the Probit and Logit Methods for Dose Response Curves,” *Biometrics*, 32, 761–768.

Robert, C. P., and G. Casella (2004): *Monte Carlo statistical methods*, Springer Texts in Statistics. Springer-Verlag, New York, second edn.

Shaikh, A., M. Simonsen, E. Vytlacil, and N. Yildiz (2005): “On the Identification of Misspecified Propensity Score,” `http://www.stanford.edu/ ashaikh/webfiles/matching.pdf`.

Smith, J. A., and P. E. Todd (2005): “Does matching overcome LaLonde’s critique of nonexperimental estimators?,” *Journal of Econometrics*, 125, 305–353.

Tierney, L. (1994): “Markov chains for exploring posterior distributions,” *The Annals of Statistics*, 22(4), 1701–1762, With discussion and a rejoinder by the author.

University of Illinois at Urbana-Champaign

University of Illinois at Urbana-Champaign