PARAMETRIC LINKS FOR BINARY CHOICE MODELS

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 suggested: 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 and contrasted.

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 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. The main objective of this paper is to correct this misapprehension.

Each link function corresponds to a latent variable model

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

Version: November 28, 2006. This research was partially supported NSF grant SES-05-44673.

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 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. 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 1. 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. 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.

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.

Further details of the R implementation of the proposed links are given in Koenker (2006).

[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.

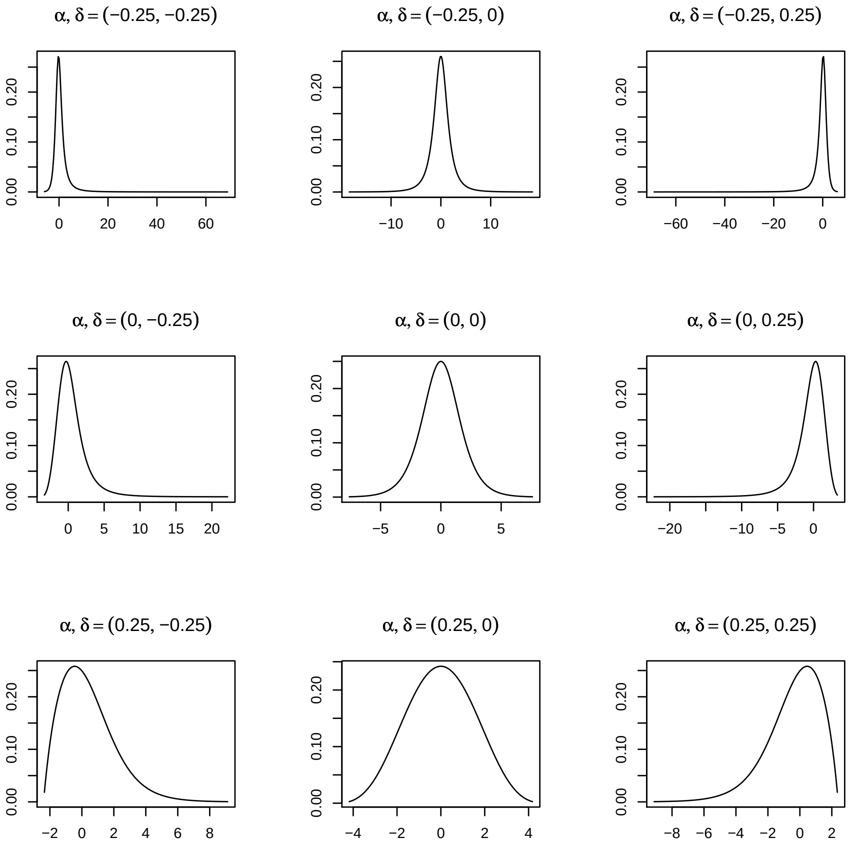

Figure 1. Some examples of the Pregibon (Tukey λ) densities.

2. The Gosset Link

The probit and cauchit link functions are naturally nested in 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 2. Optimization of the profiled likelihood is easily carried out with the aid of the Brent (1973) algorithm, or similar methods. 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 see e.g. Liu (2004), but to our knowledge there has been no attempt to explore the performance of conventional MLE methods for similar models.

Several caveats should be mentioned:

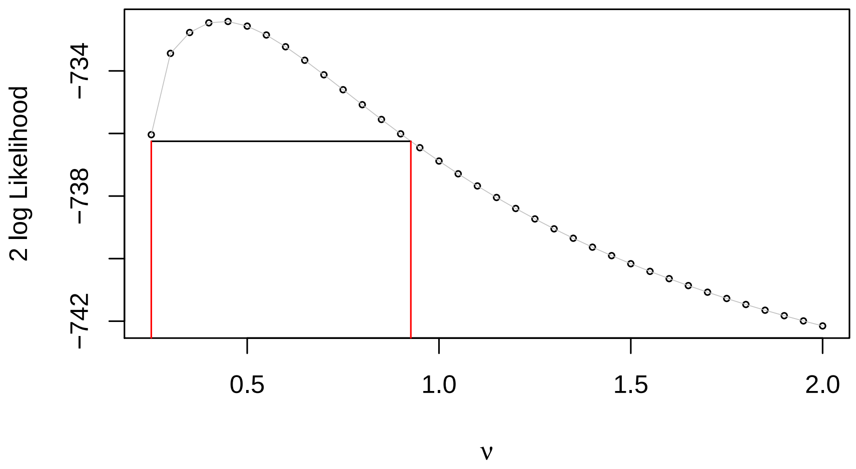

Figure 2. Profile likelihood for the Gosset link parameter ν for a model of quit behavior for a large U.S. manufacturing firm. The vertical lines indicate a 95% confidence interval for ν.

- 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.
- 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.
- 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 2. 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.

2.1. **An Application.** The profiled likelihood appearing in Figure 2 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 2.

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

As an indication of the difference between the fitted models, Figure 3 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.

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$(\nu + 1)/2, 2/(\nu - (Y_i^* - x_i^\top \beta)^2))$ random variables then the Y_i^* are independent Student t random variables with location $x_i^\top \beta$, scale 1, and degrees of freedom parameter, ν. Finally if we take the prior density of ν to be $\pi(\nu)$, then the conditional posterior density of ν is

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

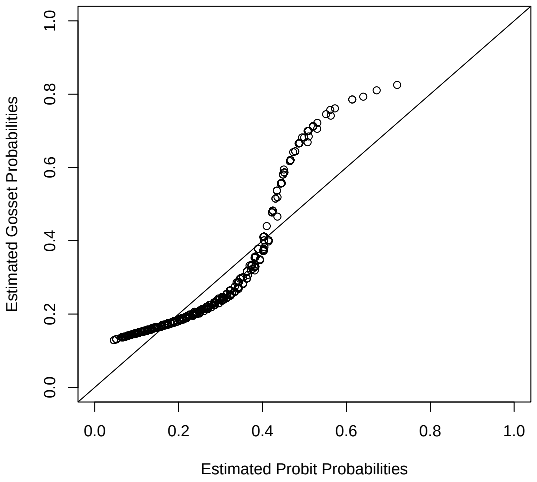

FIGURE 3. PP Plot of Fitted Probabilities of the Probit and Gosset Models for the WECO data: The solid line is the 45 degree line.

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 $\Lambda = diag(\lambda_i)$. 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 and draw ν from the set $\{\nu_1, \cdots, \nu_L\}$ according to the evaluated probabilities. Parameterizing ν in terms of $\xi = 1/\nu$ we impose a uniform prior on the interval $[0, \bar{\xi}]$. 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	1.013	2.200	5.782	0.998	2.162	6.246
Median	1.034	1.667	5.556	1.034	1.875	5.556
MAE	0.284	0.947	2.110	0.187	0.680	2.128
RMSE	0.439	1.549	2.698	0.241	1.648	3.390
Bayes-2						
Mean	1.084	2.107	5.789	1.027	2.080	6.515
Median	1.034	1.579	5.263	1.034	1.765	5.556
MAE	0.227	0.909	2.051	0.154	0.601	2.318
RMSE	0.422	1.221	2.683	0.195	0.842	4.095

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 grid values for $\xi = 1/\nu$ on the interval $[0, 2]$, we generate chains of 50,000 draws for each realization of the simulation. The first 25,000 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. It will be seen that the results are somewhat mixed with the Bayesian estimator outperforming the MLE both in terms of bias and MSE for the case of $\nu = 6$. 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. Note that with $1/\nu$ uniform on the interval $[0, 2]$, half of prior mass is placed on $[0.5, 1]$ for 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, in terms of bias and mean squared error, we do not observe much difference, but we see in Table 3 that the choice of prior is crucial to obtaining correct coverage from the Bayesian credibility regions.

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						
$H_0 :\ \nu_0 = 1$	0.132	0.548	0.972	0.114	0.836	1.000
$H_0 :\ \nu_0 = 2$	0.652	0.140	0.506	0.854	0.102	0.776
$H_0 :\ \nu_0 = 6$	0.948	0.656	0.056	1.000	0.822	0.030
Bayes-2						
$H_0 :\ \nu_0 = 1$	0.034	0.506	0.980	0.038	0.828	1.000
$H_0 :\ \nu_0 = 2$	0.636	0.148	0.508	0.842	0.112	0.790
$H_0 :\ \nu_0 = 6$	0.950	0.634	0.046	1.000	0.846	0.036

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.

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 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 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 performance of the Bayesian procedure is very good for the case of $\nu = 6$ and somewhat worse in terms of correct nominal coverage probability for the $\nu = 1$ and $\nu = 2$ cases. For example, when the true value of parameter $\nu = 1$, the credible interval tends to exclude the true value too often. 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.

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	0.020	0.018	0.013	0.028	0.027	0.024	0.071	0.077	0.069

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 $n = 500$.

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 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

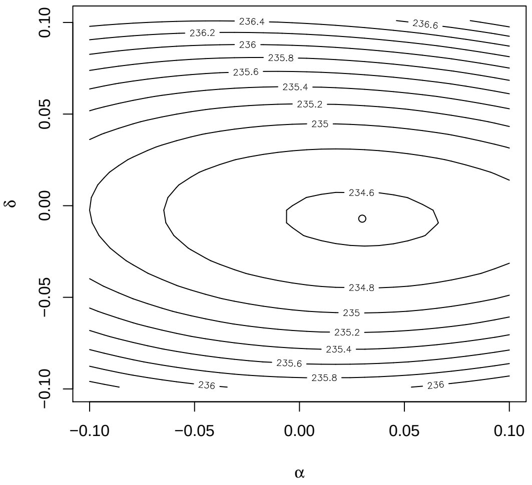

Figure 4. Contour plot of the profiled likelihood function for the Pregibon link model. The example is based on 500 observations from a simple bivariate logit model. Contours a label in AIC units so difference in contours can be compared directly to the quantiles of the χ^2_2 distribution.

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 4 based on 500 observations from a simple bivariate 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 1.

3.1. **Bayesian implementation for the Pregibon link.**

As Figure 1 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

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

where $M_i = \{I(Y_i^* > 0)I(Y_i = 1) + I(Y_i^* < 0)I(Y_i = 0)\}$. 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.

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.

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

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	0.040	-0.215	-0.118	-0.378	-0.001	0.006	0.158	0.308
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	0.040	-0.277	-0.148	-0.380	0.002	0.009	0.157	0.305
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	0.186	0.139	0.162	0.193	0.054	0.065	0.066	0.068
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	0.237	0.209	0.222	0.274	0.080	0.097	0.095	0.107

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.

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 estimator. Estimation of the shape parameter α is more variable than that of skewness parameter δ. With a smaller sample size (Table 5), however, Bayes point estimates outperform maximum likelihood estimates in terms of bias. But those differences are attenuated as the sample size increases as shown in Table 6.

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	0.024	-0.255	-0.146	-0.425	-0.001	-0.001	0.160	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	0.024	-0.274	-0.160	-0.425	-0.004	-0.005	0.158	0.312
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	0.109	0.101	0.111	0.138	0.042	0.036	0.044	0.056
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	0.167	0.143	0.161	0.209	0.061	0.060	0.066	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.

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 exclude the origin far too often, more than half the time for $n = 500$ and more than one third of time for $n = 1000$, when the nominal coverage probability is supposed to be 0.05. This is not due to the bias; Tables 5 and 6 show that the Bayes point estimates are quite good in terms of bias. Rather, it is due to the fact that the credible sets obtained from the posterior are far too strongly concentrated.

Several experiments were conducted to explore ways to remedy this problem. Expanding the number of Metropolis iterations failed to improve the coverage; expanding the length of the Gibbs sampling chain did provide a modest improvement. For example, expanding the length of the chain from 2000 to 20000 reduced exclusion

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	0.512	0.740	0.836	0.984	0.388	0.776	0.894	0.998

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.

frequency from 0.514 to 0.307 when we maintained the selection of every 5th draw, and to .271 when we instead kept all the post-burn-in draws. Adjusting the concentration of the prior around the true parameter can be an effective remedy, however this has the obvious drawback that it is not invariant to the choice of centering of the prior distribution. This tendency for the Bayesian credibility intervals to be too concentrated is disturbing and merits further investigation.

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	0.022	0.041	0.009	0.017	0.021	0.041
MLE	0.017	0.019	0.018	0.020	0.012	0.013	0.013	0.014
Bayes	0.018	0.020	0.019	0.021	0.013	0.014	0.014	0.015
d_2								
Logit	0.020	0.027	0.030	0.053	0.014	0.023	0.028	0.052
MLE	0.026	0.027	0.026	0.028	0.018	0.018	0.018	0.020
Bayes	0.027	0.027	0.027	0.029	0.019	0.019	0.020	0.021
d_∞								
Logit	0.044	0.063	0.069	0.117	0.031	0.056	0.064	0.117
MLE	0.061	0.062	0.063	0.069	0.042	0.043	0.043	0.047
Bayes	0.064	0.064	0.064	0.071	0.045	0.045	0.046	0.051

Table 8. Performance of Logit, ML, and Bayes Estimators : Sample discrepancy between true probabilities and the estimated probabilities with Pregibon link families.

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 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.

Estimator	**LaLonde**		**Dehejia-Wahba**		**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	-0.436	-0.164	-0.559	-0.571	-1.301	-1.031
δ	0.163	0.060	0.291	0.281	0.334	0.644
p-value	0.001	0.008	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.

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[2] and a simpler specification which include only linear terms in the covariates without indicators, quadratics, and interaction terms.[3] The unbalanced sample sizes of the control and 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 5 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.

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 estimation methods perform well in our investigations. The Bayesian credibility regions produced by our MCMC methods were however consistently too concentrated; limited experimentation to see whether this could be remedied by extending the length of the Markov chains exhibited only limited success. Modification of the prior to assign more mass near the origin is another obvious treatment, but one that seems not entirely satisfactory.

[2] 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).

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

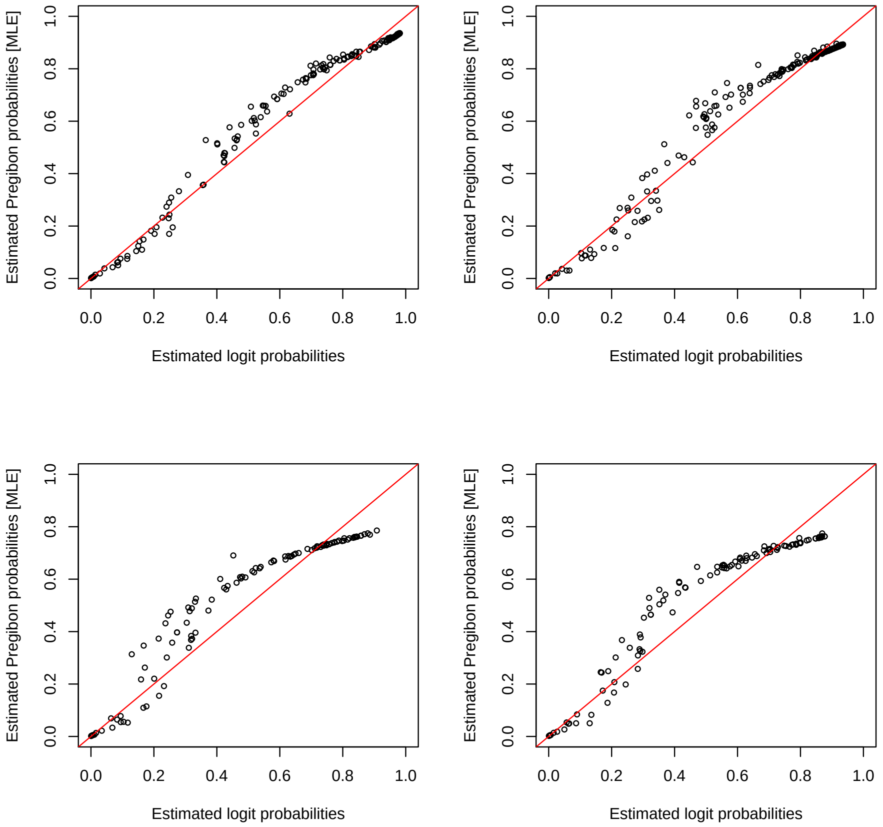

Figure 5. PP Plots of the estimated probabilities of the logit and Pregibon models. The solid line is 45 degree line. The Top-left is the results from the Dehejia-Wahba treatment group/Dehejia specification. The Top-right hand side is for Dehejia-Wahba treatment group/linear specification. The bottom-left 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.

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.

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*, forthcoming.

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.

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.

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

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.

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN