Gregory Kordas
Last update: June 27, 2023

Lecture 9
Binary Choice Models

Abstract: We present parametric and semiparametric estimators of binary choice models and discuss biases resulting from misspecification.

1. Introduction

The classical binary choice model is given by

$$\begin{aligned} y_i^* &= x_i'\beta + u_i, \\ y_i &= 1\{y_i^* \geq 0\} \end{aligned}$$

where y_i^* is a latent (unobserved) continuous variable, y_i is a binary indicator, x_i is a k-vector of regressors, β is a conformable vector of coefficients to be estimated from data, and u_i is unobserved disturbance.

Example 1. Consider the decision of students to walk or take a bus to school. Let $U_i(1)$ be student's i utility of taking a bus, and $U_i(0)$ be the utility of walking to class, and assume that these are the only available choices. Let x_i be "distance to class", and assume that we can write

$$\begin{aligned} U_i(0) &= \alpha^0 + \gamma^0 x_i + \varepsilon_i^0, \\ U_i(1) &= \alpha^1 + \gamma^1 x_i + \varepsilon_i^1 \end{aligned}$$

where $\varepsilon_i^j, j = 0, 1$, are choice-specific error terms that capture taste variation across different students. Clearly, student i takes a bus if and only if $U_i(1) - U_i(0) \geq 0$, so letting y_i^* denote this *utility differential*, we can write

$$\begin{aligned} y_i^* &= U_i(1) - U_i(0) \\ &= (\alpha^1 - \alpha^0) + (\gamma^1 - \gamma^0)x_i + (\varepsilon_i^1 - \varepsilon_i^0) \\ &\equiv \beta_0 + \beta_1 x_i + u_i, \end{aligned}$$

with β_0, β_1 and u_i having the obvious definitions. Now, utilities, and therefore y_i^*, are not observed by the economist, but we do observe choices y_i given by

$$y_i = \begin{cases} 1, & \text{if } y_i^* \geq 0; \\ 0, & \text{otherwise.} \end{cases}$$

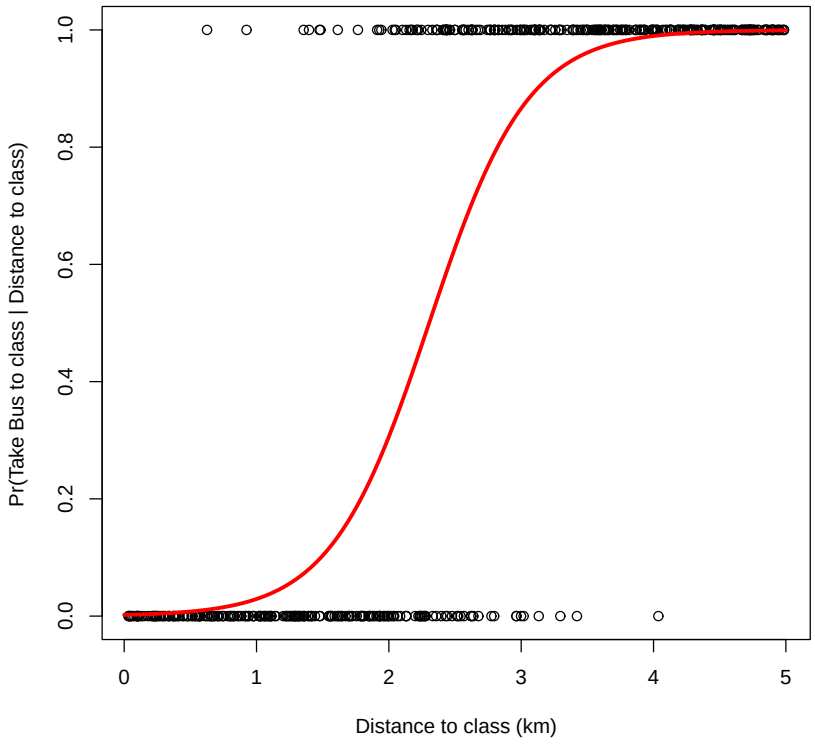

FIGURE 1. The decision to take a bus to class.

Since utility is ordinal, multiplying y_i^* with a positive constant $\sigma > 0$ would not affect the observed choice, so given data $\{y_i, x_i, i = 1, ..., n\}$, the best we can hope for is to identify β/σ, i.e., identify β "up to scale".

■

To fix the unidentified scale of the model we will assume that $u_i \sim$ i.i.d $F(0, 1)$, i.e., that the error variance is 1. Then the probability of $y_i = 1$ given x_i is

$$\begin{aligned} \Pr(y_i = 1|x_i) &= \Pr(y_i^* \geq 0|x_i) \\ &= \Pr(x_i'\beta + u_i \geq 0|x_i) \\ &= \Pr(u_i > -x_i'\beta|x_i) \\ &= 1 - F(-x_i'\beta). \end{aligned}$$

Some of the often used distributions, $F(\cdot)$ and corresponding link (quantile) functions, $F^{-1}(\cdot)$, are given below.

Model	**Error Distribution** $\mathbf{p = F(z)}$	**Link (Quantile) Function** $\mathbf{z = F^{-1}(p)}$
Probit	$\Phi(z) = \int_{-\infty}^{z} \phi(z)dz$	$\Phi^{-1}(p)$
Logit	$\Lambda(z) \equiv \dfrac{e^z}{1+e^z}$	$\Lambda^{-1}(p) = \log\left(\dfrac{p}{1-p}\right)$
Cauchy	$C(z) = \dfrac{1}{2\pi}\tan^{-1}(z)$	$C^{-1}(z) = \tan(\pi(p - \frac{1}{2}))$
Log-Log	$G_1(z) = 1 - e^{-e^z}$	$G_1^{-1}(p) = \log(-\log(1-p))$
C-Log-Log	$G_2(z) = e^{-e^z}$	$F^{-1}(p) = -\log(-\log(p))$

In what follows we will also assume that F is symmetric about zero, so that $\Pr(y_i = 1|x_i) = F(x_i')$ and $\Pr(y_i = 0|x_i) = 1 - F(x_i'\beta)$. If F is known (and symmetric about zero), the *log-likelihood of a random sample* $\{y_i, x_i, i = 1, ..., n\}$ is given by

$$\ell_n(b) = n^{-1}\sum_{i=1}^{n} y_i \log F(x_i'b) + (1 - y_i)\log[1 - F(x_i'b)]$$

and the MLE is defined by $\hat{\beta} = \operatorname{argmax}_{b\in\mathbb{B}} \ell_n(b)$. Let $F_i \equiv F(x_i'b)$, $f_i \equiv f(x_i'b)$, so that the *sample score* is

$$s_n(b) \equiv \frac{\partial \ell_n(b)}{\partial b} = n^{-1}\sum_{i=1}^{n} \frac{(y_i - F_i)f_i}{F_i(1-F_i)}x_i$$

and the *sample Hessian* is

$$H_n(b) \equiv \frac{\partial^2 \ell_n(b)}{\partial b \partial b'} = -n^{-1}\sum_{i=1}^{n} \frac{(y_i - F_i)^2 f_i^2}{F_i^2(1-F_i)^2}x_i x_i' + n^{-1}\sum_{i=1}^{n} \frac{(y_i - F_i)f_i'}{F_i(1-F_i)}x_i x_i'$$

The *population log-likelihood* is given by

$$\ell(b) \equiv E[\ell_n(b)] = \operatorname*{plim}_{n\to\infty} \ell_n(b) = E_x[F_0 \log F + (1 - F_0)(1 - \log F)]$$

where $F_0 \equiv F(x'\beta)$ and $F \equiv F(x'b)$. Differentiating $\ell(b)$ and rearranging we obtain the *population score*

$$s(b) \equiv \frac{\partial \ell(b)}{\partial \beta} = E_x\left[\frac{(F_0 - F)f}{F(1-F)}x\right],$$

and it is clear that $s(\beta) = 0$. The *population Hessian* is

$$H(b) \equiv \frac{\partial^2 \ell(b)}{\partial b \partial b'} = -E_x\left[\frac{(F_0 f' - f^2 - Ff')F(1-F) - (F_0 - F)f(f - 2Ff')}{F^2(1-F)^2}xx'\right]$$

and evaluating at $b = \beta$ we obtain

$$H(\beta) = -E_x \left[\frac{f_0^2}{F_0(1-F_0)} xx' \right].$$

Theorem 1. *Assume that*

(i) *F has a derivatives f and f', and $0 < F(u) < 1$ and $f(u) > 0$ for every u.*
(ii) *The parameter space $\mathbb{B}$ is an open bounded subset of $\mathbb{R}^k$.*
(iii) *$\{x_i\}$ is uniformly bounded in i and $n^{-1}\sum_{i=1}^n x_i x_i'$ converges in probability to a finite nonsingular matrix $E(xx')$.*

Then $\hat{\beta} \xrightarrow{p} \beta_0$ and

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N\left(0, E_x\Big[\frac{f(x'\beta)^2}{F(x'\beta)[1-F(x'\beta)]} xx'\Big]^{-1}\right).$$

The MLE $\hat{\beta}$ satisfies $s(\hat{\beta}) = 0$ and it may be computed by the Newton-Raphson iteration

$$\hat{\beta}_{[i+1]} = \hat{\beta}_{[i]} - [H_n(\hat{\beta}_{[i]})]^{-1} s_n(\hat{\beta}_{[i]}).$$

Alternatively, we could replace $H_n(\hat{\beta}_{[i]})$ by it's asymptotic analogue $H(\hat{\beta}_{[i]})$, in which case the resulting iteration is referred to as the *method of scoring*. Let $W = \text{diag}(f_i^2/(F_i(1-F_i)))$ be a $n \times n$ matrix, let $r = (y_i - F_i)/f_i$ be a $n \times 1$ vector of scaled residuals, and let X be the $n \times k$ design matrix. The method of scoring iteration is then given by

$$\hat{\beta}_{[i+1]} = \hat{\beta}_{[i]} + (X'WX)^{-1}X'Wr,$$

where all rhs quantities are evaluated at $\hat{\beta}_{[i]}$.

2. Interpretation of the Coefficients

In linear regression models the change in $E(y^*|x)$ affected by a small change in x_j is equal to β_j, that is, if $E(y^*|x) = x'\beta)$ then

$$\frac{\partial E(y^*|x)}{\partial x_j} = \beta_j, \quad j = 1, ..., k.$$

In binary response models, however, the conditional expectation $E(y|x)$ of the observed binary variable y given x is nonlinear in $x'\beta$, i.e.,

$$E(y|x) = 1 \cdot \Pr(y = 1|x) + 0 \cdot \Pr(y = 0|x) = F(x'\beta),$$

so in this model

$$\frac{\partial E(y|x)}{\partial x_j} = \beta_j f(x'\beta), \quad j = 1, ..., k.$$

These numbers are often called the *partial effects* of the model: given and error distribution F with density f, and a coefficient estimate $\hat{\beta}_j$, we compute $\hat{\beta}_j f(\bar{x}'\hat{\beta})$ for each $j = 1, ..., k$, where $\bar{x}$ is the vector of sample means of the covariates. The nice thing about these

quantities is that they are comparable across different models, whereas the $\hat{\beta}_j$'s themselves are not since they depend on the specific scale parameter used in each model.

Another way to standardize the coefficient estimates so as to compare them across different models is to scale them by $f(0)$, the error density evaluated at zero.

Model	$\mathbf{f(0)}$
Probit	$1/\sqrt{2\pi}$
Logit	$1/4$
Cauchy	$1/\pi$

Then, roughly speaking,

$$\frac{1}{\sqrt{2\pi}}\hat{\beta}_j^{\text{probit}} \approx \frac{1}{4}\hat{\beta}_j^{\text{logit}}$$

or

$$\hat{\beta}_j^{\text{logit}} \approx 1.6\hat{\beta}_j^{\text{probit}}.$$

Other models can be compared in a similar fashion.

3. Estimation Under Heteroskedasticity

Consider again the latent variable model

$$y_i^* = x_i'\beta + u_i,$$

where now u_i is heteroskedastic, that is $u_i = h(x_i)\upsilon_i$, with $h(x_i)$ is being a skedastic function and $\upsilon_i \sim$ i.i.d. F. If $y_i = 1\{y_i^* \geq 0\}$, and F is symmetric as it was assumed above, we have

$$\begin{aligned} E(y_i|x_i) &= \Pr(y_i = 1|x_i) \\ &= \Pr(u \geq -x_i'\beta) \\ &= \Pr\Big(\upsilon \geq -\frac{x_i'\beta}{h(x_i)}\Big) \\ &= F\Big(\frac{x_i'\beta}{h(x_i)}\Big). \end{aligned}$$

Since $h(\cdot)$ is a skedastic function its range should be strictly positive, so a natural parametrization is the exponential skedastic function given by

$$h(x_i) = \exp(x_i'\gamma).$$

Estimation of heteroskedastic probit or logit models is straightforward. All we have to do is to plug in this new expression into the likelihood. For the exponentially heteroskedastic model the log-likelihood is given by

$$\ell_n(b,c) = n^{-1}\sum_{i=1}^n y_i \log F\Big(\frac{x_i'b}{\exp(x_i'c)}\Big) + (1-y_i)\log\left[1 - F\Big(\frac{x_i'b}{\exp(x_i'c)}\Big)\right]$$

which we may maximize by numerical methods to obtain $\hat{\beta}$ and $\hat{\gamma}$.

4. Discriminant Analysis

4.1. **The normal DA model.** Let y denote a binary random variable which takes the values 0 and 1 and let X be a $k \times 1$ vector of related continuous random variables. Denote by $F(y, X)$ the joint distribution function of (y, X). The standard statistical problem of classification can be stated as: Given an observation x of attributes X which is generated by the two probability models indexed by y, namely $X|y = 0$ and $X|y = 1$, decide which population x belongs to.

The standard DA procedure assumes that the conditional distribution of $X|y$ is multivariate normal with mean μ_y and common variance Σ. More formally, let $F_D(X|y)$ denote the conditional distribution of $X|y$ and let $f_D(X|y)$ be the corresponding density function. The normal DA requires that

$$f_D(X|y) = (2\pi)^{-k/2}|\Sigma|^{-1/2} \exp\left[-\tfrac{1}{2}(X - \mu_y)'\Sigma^{-1}(X - \mu_y)\right]. \tag{1}$$

Under these conditions, the solution of the general classification problem takes the particularly simple form based on the linear discriminant function. To derive this function, let $f_X(X)$ denote the marginal density function of X and let

$$\pi_y = \int f_X(X)P(y|X)dX \tag{2}$$

be the marginal distribution of y or, in DA terminology, be the *a priori* probability of an observation x being a member of population y. Letting $F_L(y|X)$ denote the conditional distribution of $y|X$ and applying Bayes' formula yields

$$F_L(y|X) = \frac{f_D(X|y)\pi_y}{f_X(X)}. \tag{3}$$

Since

$$f_X(X) = \sum_y \pi_y f_D(X|y),$$

eq. (3) may be written as

$$F_L(y|X) = \frac{f_D(X|y)\pi_y}{\sum_y \pi_y f_D(X|y)} = \left(1 + \frac{\pi_0}{\pi_1}\frac{f_D(X|y=0)}{f_D(X|y=1)}\right)^{-1}. \tag{4}$$

Substituting the conditional densities of (1) into (4) and simplifying yields

$$F_L(y = 1|X) = (1 + \exp[-(\alpha + \beta' X)])^{-1}, \tag{5}$$

$$\alpha = -\tfrac{1}{2}(\mu_1 - \mu_0)'\Sigma^{-1}(\mu_1 + \mu_0) + \log(\pi_1/\pi_0), \tag{6}$$

$$\beta = \Sigma^{-1}(\mu_1 - \mu_0). \tag{7}$$

Eq. (5) demonstartes that the assumptions for normal DA yield a logistic conditional distribution for $y|X$. Because the converse is not true, logit analysis is a more robust

procedure. In fact, as Efron (1975) shows, logit analysis obtains under the general exponential family assumption for $F_D(X|y)$, Specifically, let

$$f_D(X|y) = g(\theta_y, \eta)h(X, \eta)\exp[\theta_y' X], \tag{8}$$

where η is an arbitrary nuisance parameter. Note that (1) is a special case of (8). The conditional density of $y|X$ under (8) is given by

$$\begin{aligned} F_L(y = 1|X) &= (1 + \exp[-(\alpha + \beta' X)])^{-1}, & (9)\\ \alpha &= -\log[g(\theta_0, \eta)/g(\theta_1, \eta)] + \log(\pi_1/\pi_0), & (10)\\ \beta &= \theta_1 - \theta_0. & (11) \end{aligned}$$

An observation x belongs to population $y = 1$ if

$$F_L(y = 1|X = x) \geq \tfrac{1}{2}, \tag{12}$$

else it belongs to population $y = 0$. This is equivalent to requiring that $\alpha + \beta' x \geq 0$ or that

$$-\tfrac{1}{2}\mu_1'\Sigma^{-1}\mu_1 + x'\Sigma^{-1}\mu_1 + \log \pi_1 \geq -\tfrac{1}{2}\mu_0'\Sigma^{-1}\mu_0 + x'\Sigma^{-1}\mu_0 + \log \pi_0. \tag{13}$$

In DA applications it is customary to compute for each candindate x the *linear discrimination function*

$$\delta_y(x) = -\tfrac{1}{2}\mu_y'\Sigma^{-1}\mu_y + x'\Sigma^{-1}\mu_y + \log \pi_y. \tag{14}$$

and classify it to the population $y = \arg\max_y \delta_y(x)$.

4.2. **Estimation of the normal DA model.**

5. Bias Under Misspecification

Now consider the situation in which our model is misspecified in that the link function we have chosen to represent the probability $\Pr(y = 1|x)$ is false, either because we have assumed the wrong error distribution and/or because the error is heteroskedastic and we have failed to correctly account for it in our estimation. In particular, assume that $\Pr(y_i = 1|x_i) = G(x_i'\beta)$ but we have falsely assumed that $\Pr(y_i = 1|x_i) = F(x_i'\beta)$. What are the properties of the MLE $\hat{\beta}$ under such misspecification?

The population log-likelihood in this case can be written as

$$\ell^*(b) \equiv E[\ell_n^*(b)] = \operatorname*{plim}_{n\to\infty} \ell_n^*(b) = E_x\Big\{G(x'\beta)\log F(x'b) + [1 - G(x'\beta)]\log[1 - F(x'b)]\Big\}.$$

It can be shown that since F and G do not coincide, the maximizer of this misspecified likelihood is not equal to β, but it is equal to some other vector β^*. The following theorem gives the result.

Theorem 2. *Under misspecification of the link function the MLE $\hat{\beta}$ converges in probability to β^*, the minimizer of $\ell^*(b)$, i.e., as $n \to \infty$,*

$$\hat{\beta} = \underset{b}{\operatorname{argmax}}\ \ell_n^*(b) \xrightarrow{p} \beta^* = \underset{b}{\operatorname{argmax}}\ \ell^*(b).$$

Furthermore, $\beta^* = \beta$, the true parameter vector if and only if x is jointly distributed according to an elliptical distribution (for example multivariate normal). Otherwise $\beta^* \neq \beta$, and the QMLE $\hat{\beta}$ is inconsistent.

Theorem 2 says that unless a very peculiar condition on the x's is satisfied, the quasi MLE (QMLE) $\hat{\beta}$ will be inconsistent for β. Consistency of $\hat{\beta}$ can only be guaranteed if the x's are multivariate normal, or multivariate t, or jointly distributed according to an *elliptical* distribution. This is interesting only in so far as it reminds us of the special place that normality holds in statistics, but it of no practical consequence in actual applications. In economic applications the x's are often strictly positive (income, consumption), skewed, counts (like years of schooling) and even dummy variables (like gender), making the elliptical distribution assumption completely unrealistic.

The inconsistency of the MLE in the binary model under misspecification of the error distribution stands in stark contrast to the continuous case in which violations of the normality assumption did not lead to inconsistency of the OLS estimator. The question that arises then is: how big could the bias due to misspecification be? The following simulations show that the bias could indeed be of the same order as the coefficient we try to estimate.

To compute the bias under misspecification we consider logit estimation under 9 data generating processes (DGP's). Let the model be given by

$$y = 1\{\beta_1 x_1 + \beta_2 x_2 + u \geq 0\}$$

where $x_1 \sim N(0,1)$, and $x_2 \sim \chi^2_4$ normalized to have mean zero and variance 1. The first 6 designs consider the effects of misspecification in homoskedastic models. The distributions of u considered are:

(1) Logistic independent of x.
(2) Standard normal independent of x.
(3) Uniform on $[-1,1]$ independent of x.
(4) Student-t with 2 degrees of freedom independent of x.
(5) Chi-squared with 4 degrees of freedom normalized to have median 0 independent of x.
(6) A 50-50 mixture of normal distributions $N(-3,1)$ and $N(3,1)$ independent of x.

The last 3 designs consider the effects of unaccounted heteroskedasticity in otherwise correctly specified models. In particular $u = h(x)v$, where v has the logistic distribution independent of x, and $h(x)$ is of the following form:

(7) $h(x) = 1 + \gamma|x_1 + x_2|$.
(8) $h(x) = 1 + \exp[\gamma(x_1 + x_2)]$.
(9) $h(x) = 1 + \gamma(x_1 + x_2)^2$.

All designs but design (1) are misspecified. Note that x is not elliptically distributed so we will avoid the case of trivial consistency under misspecification described in Theorem 2. Table 1 reports β_2^*/β_1^* for the different designs. It also presents two measures of discrepancy between F and G, the *mean absolute distance*

$$MAD = E_x|F(-x'\beta^*) - G(-x'\beta)|$$

and the *sup (maximum) distance*

$$SUP = \sup_x |F(-x'\beta^*) - G(-x'\beta)|.$$

Figure 1 graphs the various link functions corresponding to the designs considered. We see that there is no bias from assuming a logit link when the true link is normal or uniform. The bias however, becomes significant as we move to fat tailed, skewed and bimodal distributions. The bias resulting from unaccounted heteroskedasticity is also very severe especially for strongly heteroskedastic designs. Looking at the MAD and SUP measures of discrepancy between the assumed and true models, we see that the bias is highest when these measure of discrepancy are high, and it is lowest or non-existent when these discrepancies are small.

6. Diagnostic for the Logit Link Function

Using the Box-Cox transformation, Pregibon (1980) proposed the following generalizations of the logit link function

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

Note that

$$\lim_{\alpha,\delta\to 0} g(p;\alpha,\delta) = \log p - \log(1-p) \equiv \text{logit}(p),$$

so the logit link is a special case for $\alpha, \delta \to 0$. The parameters δ and α control the skewness and the fatness of the tails, respectively: $\delta = 0$ implies a symmetric link, while $\alpha = 0$ implies logistic tails. Expanding $g(p;\alpha,\delta)$ around $\alpha = \delta = 0$ we obtain

$$g(p;\alpha,\delta) = \text{logit}(p) + \alpha g_\alpha(p) + \delta g_\delta(p),$$

where

$$g_\alpha(p) = \frac{\partial g(p;\alpha,\delta)}{\partial \alpha} = \frac{1}{2}\Big[\log^2(p) - \log^2(1-p)\Big],$$

and

$$g_\delta(p) = \frac{\partial g(p;\alpha,\delta)}{\partial \delta} = -\frac{1}{2}\Big[\log^2(p) + \log^2(1-p)\Big].$$

TABLE 1. Asymptotic bias of a binary logit model under alternative forms of misspecification.

Model	Description	Distribution	β_2^*/β_1^*	MAD	SUP
(1)	Logistic	$L(0, \pi^2/3)$	1.00	0.00	0.00
(2)	Normal	$N(0,1)$	1.00	0.01	0.01
(3)	Uniform	$U[-1,1]$	1.00	0.03	0.07
(4)	Fat-tailed	$t(2)$	0.99	0.01	0.03
(5)	Skewed	$\chi^2(4)$	1.09	0.02	0.11
(6)	Bimodal	$.5N(-3,1), .5N(3,1)$	1.34	0.04	0.27
(7a)	Absolute Value	$\gamma = 0.2$	0.97	0.02	0.05
(8a)	Exponential	$\gamma = 0.2$	0.89	0.04	0.32
(9a)	Quadratic	$\gamma = 0.2$	0.85	0.05	0.42
(7b)	Absolute Value	$\gamma = 0.4$	0.94	0.03	0.12
(8b)	Exponential	$\gamma = 0.4$	0.78	0.09	0.49
(9b)	Quadratic	$\gamma = 0.4$	0.80	0.05	0.45
(7c)	Absolute Value	$\gamma = 0.8$	0.89	0.03	0.23
(8c)	Exponential	$\gamma = 0.8$	0.68	0.14	0.50
(9c)	Quadratic	$\gamma = 0.8$	0.76	0.05	0.43

This suggests the following LM-type test for checking the adequacy of the logit link function:

(i) Estimate a logit model and obtain fitted probabilities $\hat{p}_i = \Lambda(x_i'\hat{\beta})$.
(ii) Compute $g_\alpha(\hat{p}_i)$ and $g_\delta(\hat{p}_i)$.
(iii) Re-estimate the model adding $g_\alpha(\hat{p}_i)$ and $g_\delta(\hat{p}_i)$ as regressors.
(iv) Reject the logit link if the new regressors are significant.

Example 2.

Mortality of Adult Beetles after 5 hours Exposure to Gaseous Carbon Disulphide (Bliss, 1935). Assume that

$$\Pr(y = 1|x) = F(x;\theta) = \Psi\left(\beta(x - \mu)\right).$$

```
> summary(mod.glm)

Call:
glm(formula = sf ~ ldose, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5941  -0.3944   0.8329   1.2592   1.5940
```

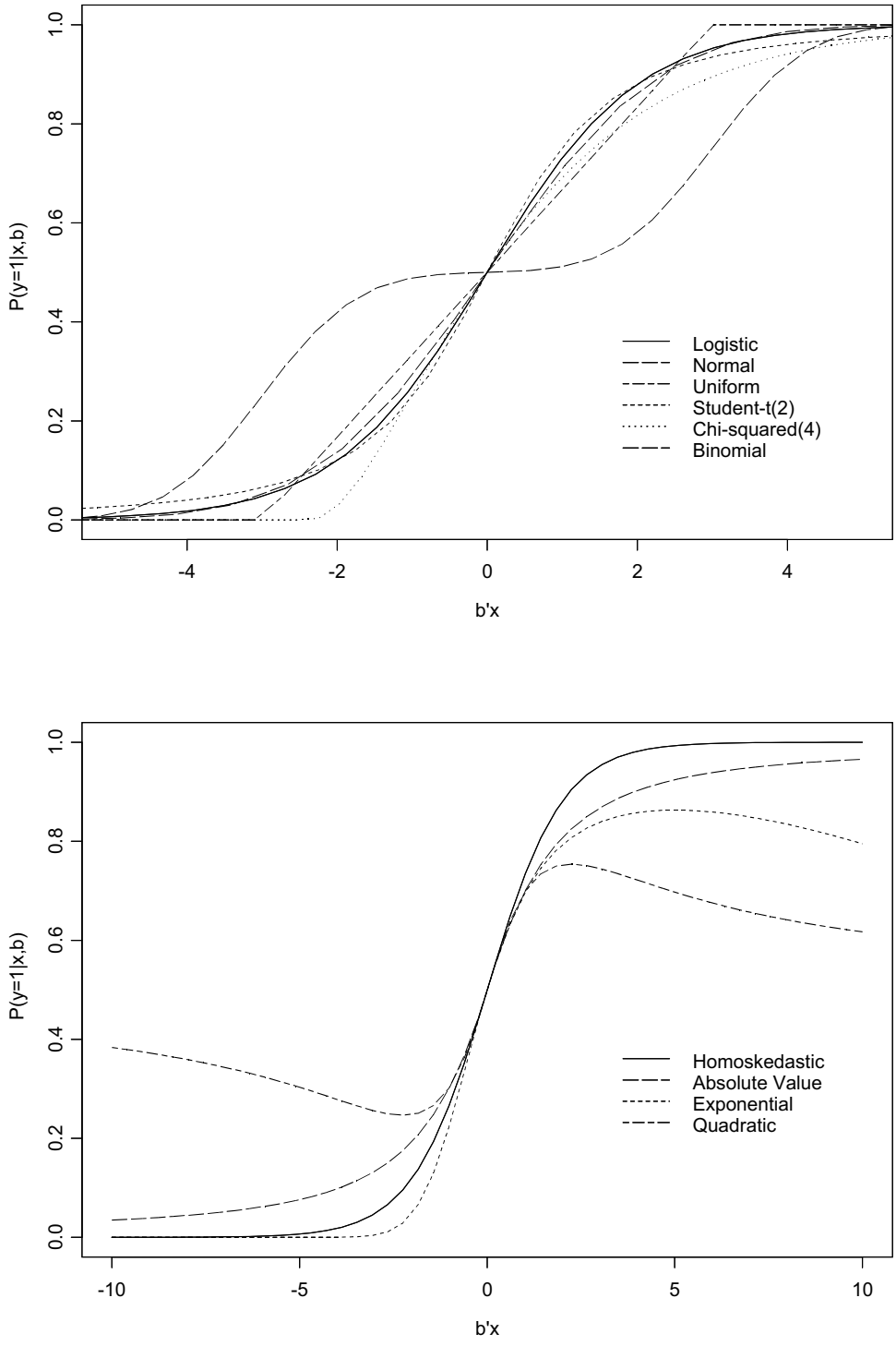

Figure 2. Link Functions Used in the Simulations.

```
Coefficients:
```

TABLE 2. Mortality of Adult Beetles after 5 hours Exposure to Gaseous Carbon Disulphide (Bliss, 1935)

Log dosage	*Number exposed*	*Number killed*	*Logit fit*
1.6907	59	6	3.45
1.7242	60	13	9.84
1.7552	62	18	22.45
1.7842	56	28	33.89
1.8113	63	52	50.10
1.8369	59	53	53.29
1.8610	62	61	59.22
1.8839	60	60	58.74

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -60.717      5.181  -11.72   <2e-16 ***
ldose         34.270      2.912   11.77   <2e-16 ***
---


    Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:  11.232  on 6  degrees of freedom
AIC: 41.43

Number of Fisher Scoring iterations: 4
```

In this model, the 50% response dose (termed the ED5O) is equal to μ. Here

$$\hat{\mu} = -(-60.171/34.270) = 1.772 \text{ CS}_2\text{mg/litre}.$$

■

7. Semiparametric Estimation

The large biases due to misspecification have motivated a lot of research in estimators that remain consistent under weaker than parametric assumptions. Perhaps the most famous of these semiparametric estimators of the binary choice model is the maximum score estimator.

Let $y = 1$ if $y^* \geq 0$ and $y = -1$ if $y^* < 0$, and define the *sample score function* by

$$S_n(b) = n^{-1} \sum_{i=1}^{n} y_i \operatorname{sgn}(x_i' b)$$

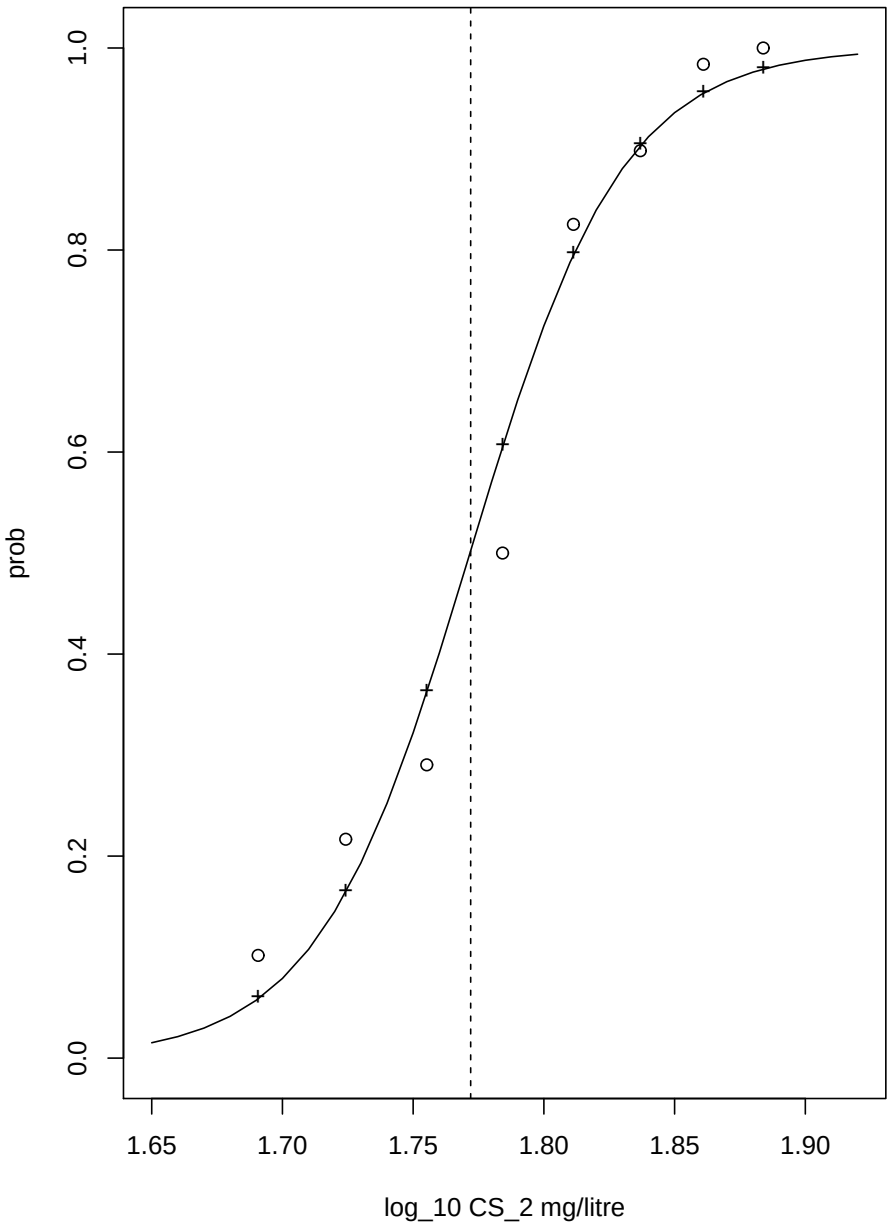

FIGURE 3. Probability of mortality as a function of log-dosage, observed and logit-fit, vertical line at ED50 = 1.772 CS_2 mg/litre.

Let $B = \{b \in \mathbb{R}^k : ||b|| = 1\}$ be the unit ball in $\mathbb{R}^k$. The *maximum score estimator* is the maximizer of the sample score function oven the unit ball B i.e.,

$$\hat{\beta}_{MS} = \underset{\beta \in B}{\operatorname{argmax}}\ S_n(b).$$

This estimator has a very intuitive interpretation: find b that maximizes the matches between the observed sign of y_i^* and the resulting sign of the index $x_i'b$.

Theorem 3. *If $Med(u|x) = 0$ then the maximum score estimator $\hat{\beta}_{MS}$ is consistent for β. However, it converges at the very slow $n^{1/3}$ (cube root) rate to a very complicated non-normal distribution.*

From Theorem 3 we see that the maximum score estimator is a *median regression estimator for the binary choice model*. Judging from the simplicity of the idea of MS estimation it is perhaps surprising to find that the asymptotics of this estimator are so complicated. In fact the proof that this estimator converges at a cube-root rate and the derivation of its asymptotic distribution requires the most advanced methods available today – *empirical process methods*. These methods have been called the "nuclear weapons of statistics", and I will discuss them in Econ 721.

8. Empirical Application: Titanic

On April 15, 1912, the *RMS Titanic* sank on its maiden voyage from Southampton, England, to New York City. It was carrying 2,201 passengers and crew, of which 711 survived and 1,490 drowned. The table below classifies the passengers according to a) the CLASS they were travelling, b) their GENDER, and c) their AGE.

Table 3. Dataset

	Child				Adult			
	Survived		Drowned		Survived		Drowned	
Class	Male	Female	Male	Female	Male	Female	Male	Female
1st	5	1	0	0	57	140	118	4
2nd	11	13	0	0	14	80	154	13
3rd	13	14	35	17	75	76	387	89
Crew	0	0	0	0	192	20	670	3

We are interested in testing the hypothesis that all passengers had the same probability of survival, against the alternative that at least one of the variables CLASS, GENDER and AGE were important in determining the survival probability of a passenger.

We consider a logit model, so that

$$\log\left(\frac{p}{1-p}\right) = x'\beta.$$

For a dummy (0/1) variable x_j, we have

$$\log\left(\frac{p}{1-p}\right)\Big|_{x_j=1} - \log\left(\frac{p}{1-p}\right)\Big|_{x_j=0} = \beta_j$$

i.e., β_j is the change in the log-odds induced by changing the value of the dummy variable x_j from 0 to 1. It follows that, for dummy variables, the odds ratio is simply e^{β_j}. Simmilarly, the relative risk among two different types of passengers j and j' is given by $e^{\beta_j - \beta_{j'}}$.

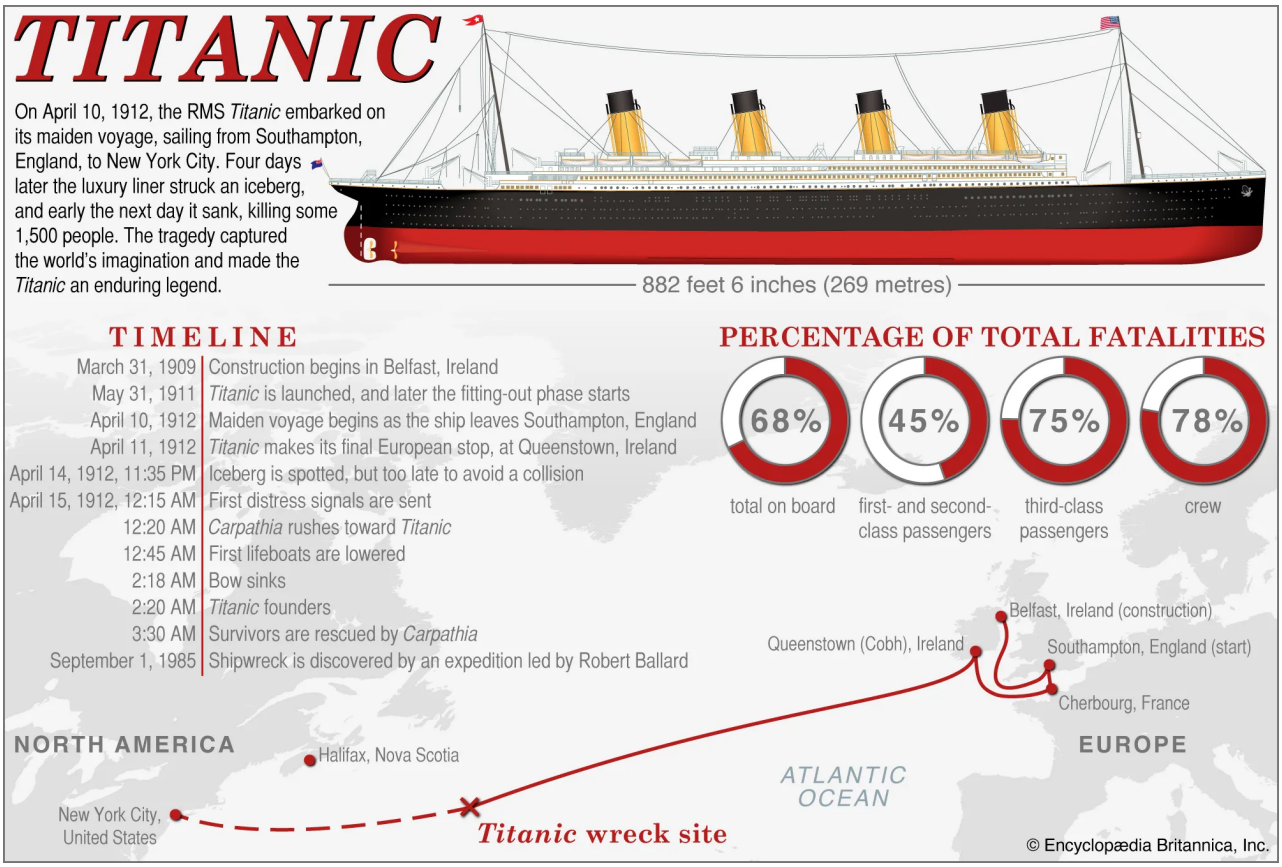

Figure 4. .

We estimate saturated models (i.e., full set of dummies) without an intercept. Equivalent results can be obtained by excluding a dummy category and including an intercept, but their interpretation would be relative to the category we chose to exclude and thus more complicated.

Table 4. Logit Model 1

Variable	Coefficient	Std. Error	Odds Ratio	Std. Error
Child	1.062	.277	2.8908	.705
Female	2.420	.136	11.247	1.579
1st Class	-0.376	.126	0.6864	.093
2nd Class	-1.394	.129	0.2480	.039
3rd Class	-2.154	.144	0.1160	.015
Crew	-1.234	.080	0.2912	.023

Table 3 reports the estimation results of a simple model without interactions, both in terms of slopes β and in terms of log-odds e^{β}. We see that women's odds of survival were about 11 times those for men, and that the children's odds of survival were 3 times those of adults. Also, 1st class passengers were $0.6864/0.2480 = 2.77$ times more likely to survive than 2nd class passengers, $0.6864/0.1160 = 5.92$ times more likely to survive

than 3nd class passengers, and $0.6864/0.2912 = 2.36$ times more likely to survive than the crew members. Interestingly enough, crew members were about as likely to survive as 2nd class passengers, and $0.2912/0.1160 = 2.51$ times more likely to survive than 3nd class passengers! Apparently, the crew members did their best to survive (the captain Edward Smith did, however, "go down with the ship", as the expression goes).

TABLE 5. Fitted Probabilities of Survival – Logit Model 1

	Child		Adult	
Class	Male	Female	Male	Female
1st	0.6649	0.9571	0.4070	0.8853
2nd	0.4176	0.8897	0.1987	0.7361
3rd	0.2512	0.7904	0.1040	0.5661
Crew	–	–	0.2255	0.7661

TABLE 6. Fitted Probabilities of Survival – Logit Model 2

	Child		Adult	
Class	Male	Female	Male	Female
1st	0.7005	0.9768	0.3343	0.9724
2nd	0.3950	0.8935	0.1229	0.8751
3rd	0.4407	0.4970	0.1447	0.4521
Crew	–	–	0.2227	0.8696

References

Dawson, R. J. M. (1995). "The 'Unusual Episode' Data Revisited", *Journal of Statistics Education*, 3.

> Is it right I ask, is it even prudence,
> To bore thyself and bore the students?
>
> — *Question put by Mephistopheles to Faust*