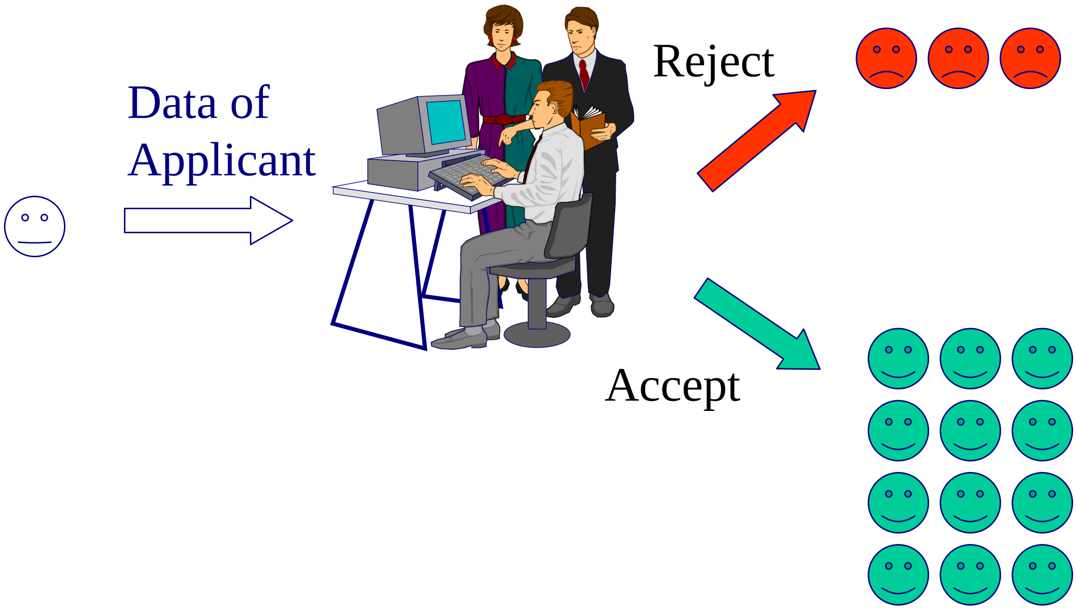
An Overview of Model Based Reject Inference for Credit Scoring

BANFF Credit Risk Conference 2003

Ad Feelders
Institute for Information and Computing Sciences
Utrecht University
The Netherlands

Credit Scoring



Assumptions and Notation

For each applicant, we observe a vector of variables/attributes $\mathbf{x} = (x_1, \ldots, x_k)$.

The outcome $y \in \{0, 1\}$ of the loan is observed for the accepted applicants, but missing for the rejected applicants.

We define an auxiliary variable a, with $a = 1$ if the applicant is accepted and $a = 0$ if the applicant is rejected. Note that y is observed if $a = 1$ and missing if $a = 0$.

Observed data

x_1 x_2 $\ldots\ldots$ x_k	a	y
	1	
	1	
	1	
	1	
	1	
	0	
	0	
	0	
	0	
	0	

Missing Completely at Random

Acceptance does not depend on characteristics of the applicant, nor on the outcome of the loan:

$$P(a = 1 \mid \mathbf{x}, y) = P(a = 1).$$

For example:

- Toss a coin.
- Accept (or reject) all applicants.

Missing at Random (MAR)

Given the characteristics of the applicant, acceptance does not depend on the outcome of the loan:

$$P(a = 1 \mid \mathbf{x}, y) = P(a = 1 \mid \mathbf{x}).$$

For example: reject if $g(\mathbf{x}) < c$, otherwise accept.

Missing Not at Random (MNAR)

Given the characteristics of the applicant, acceptance still depends on the outcome of the loan:

$$P(a = 1 \mid \mathbf{x}, y) \neq P(a = 1 \mid \mathbf{x}).$$

For example: reject if $g(\mathbf{x}) < c$, otherwise accept, but sometimes override this decision on the basis of attributes that are not in $\mathbf{x}$.

Why is this distinction important?

In credit scoring we are primarily interested in modelling the outcome mechanism, i.e. the relation between probability of default and characteristics of the applicant.

When MAR applies, we don't have to include the missing data mechanism (accept/reject decision) into the model to obtain valid results with respect to the outcome mechanism.

The missing data mechanism is *ignorable*.

The Ignorable Case

For example: reject if $g(\mathbf{x}) < c$, otherwise accept.

The real problem here is that the acceptance rule is *deterministic*. We have no observation at all of y in the reject region $(g(\mathbf{x}) < c)$. Therefore we have to *extrapolate* into the reject region.

Better would be: reject with probability 0.95 if $g(\mathbf{x}) < c$, otherwise accept. Bit expensive perhaps?

The Ignorable Case: function estimation

Complete case analysis is not generally valid under MAR, but will work if we use a function estimation approach (e.g. logistic regression). Why?

If MAR applies, then $y \perp\!\!\!\perp a \mid \mathbf{x}$ so

$$\begin{aligned} P(y = 1 \mid \mathbf{x}, a = 1) &= P(y = 1 \mid \mathbf{x}, a = 0) \\ &= P(y = 1 \mid \mathbf{x}), \end{aligned}$$

Hence we can use $P(y = 1 \mid \mathbf{x}, a = 1)$ to estimate $P(y = 1 \mid \mathbf{x})$.

The Ignorable Case: density estimation

Complete case analysis is not generally valid under MAR, and will *not* work if we use a density estimation approach (e.g. linear discriminant analysis). Why not?

We use Bayes' rule:

$$P(y = 1 \mid \mathbf{x}) = \frac{P(y = 1)p(\mathbf{x} \mid y = 1)}{P(y = 0)p(\mathbf{x} \mid y = 0) + P(y = 1)p(\mathbf{x} \mid y = 1)}$$

These quantities will all be distorted when estimated from the accepted loans only, because acceptance depends on $\mathbf{x}$ and (marginally) on y as well.

Density Estimation: Example

Suppose $p(x \mid y = 0) = N(2, 1)$ and $p(x \mid y = 1) = N(6, 1)$, $P(y = 0) = P(y = 1) = 1/2$

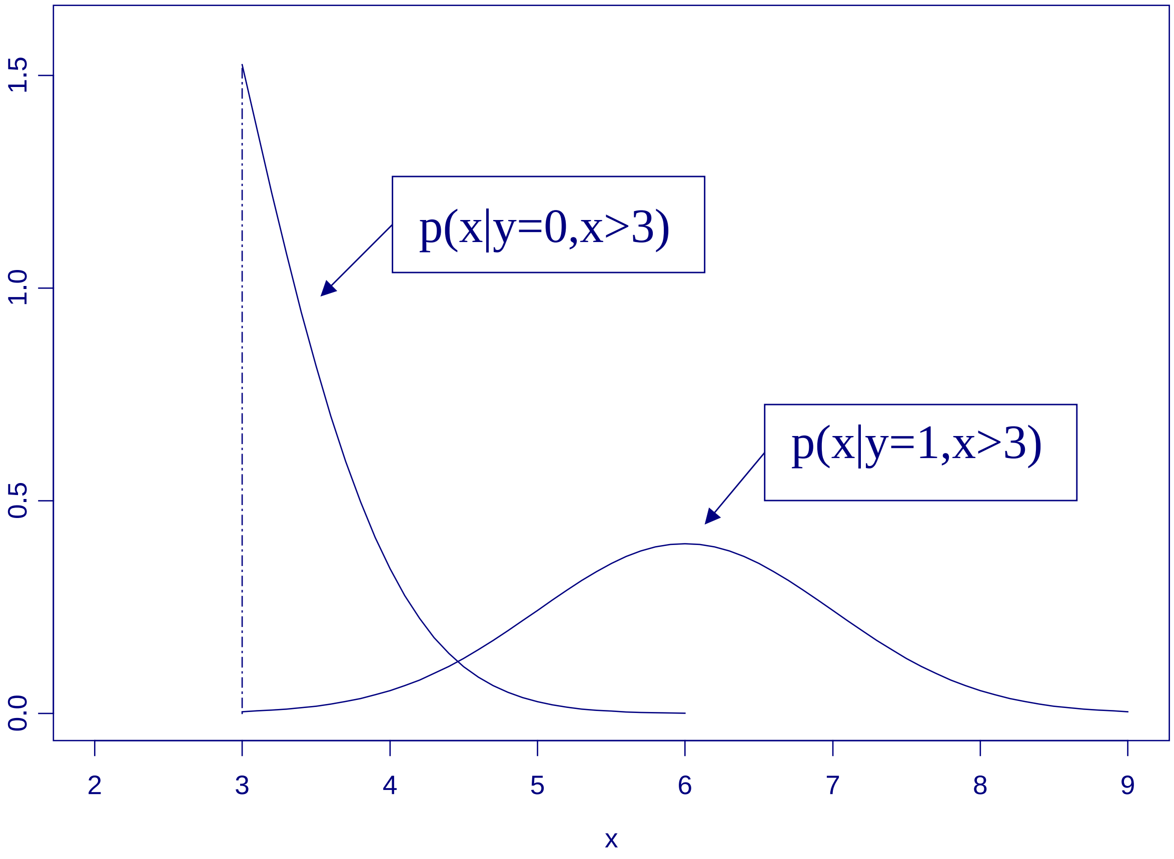
Applicant is accepted if $x > 3$.

$E[x|y = 0, x > 3] \approx 3.53$, $E[x|y = 1, x > 3] \approx 6.00$.

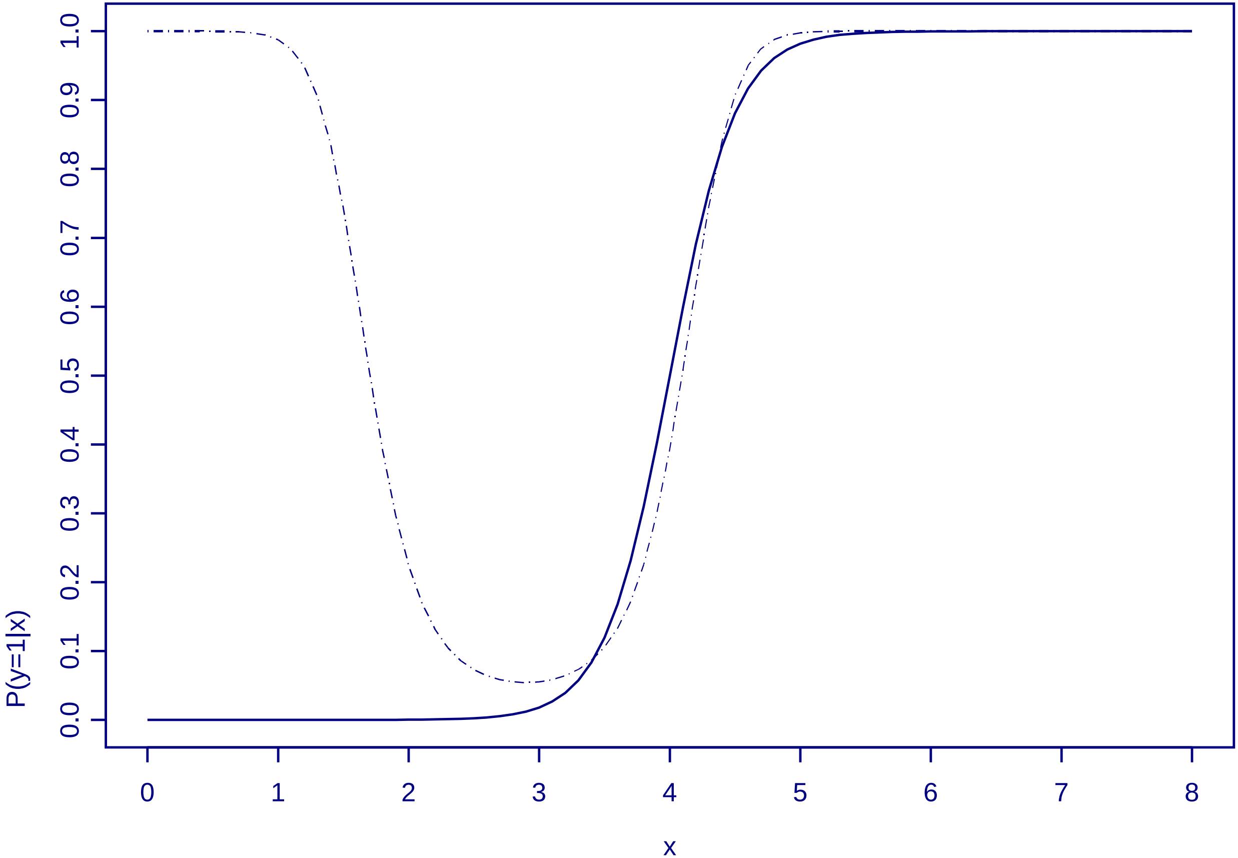
$\text{Var}[x|y = 0, x > 3] \approx 0.2$, $\text{Var}[x|y = 1, x > 3] \approx 0.99$.

$P(y = 0|x > 3) \approx 0.14$ and $P(y = 1|x > 3) \approx 0.86$.

Density Estimation: Example



Density Estimation: Example



The Ignorable Case: density estimation

- We can avoid this bias, by including the rejected applicants into the estimation process.
- Use a mixture distribution formulation of the problem.
- Estimate with EM-algorithm.

Function estimation vs density estimation

	Total	Accept	Reject
QDA^{ri}	.249	.250	.233
QLR^{ri}	.256	.253	.287
QDA	.245	.249	.209
QLR	.247	.251	.221
QDA^{ri}/QLR^{ri}	.97	.99	.81
QDA/QLR	.99	.99	.94

Relative performance of quadratic discriminant analysis (QDA) and quadratic logistic regression (QLR), with $n = 150$ and 10% rejects

Function estimation vs density estimation

	Total	Accept	Reject
QDA^{ri}	.239	.243	.201
QLR^{ri}	.241	.244	.221
QDA	.238	.243	.196
QLR	.239	.244	.197
QDA^{ri}/QLR^{ri}	0.99	1.00	0.91
QDA/QLR	1.00	1.00	0.99

Relative performance of quadratic discriminant analysis (QDA) and quadratic logistic regression (QLR), with $n = 500$ and 10% rejects

Ignorable case: function estimation or density estimation?

Function estimation:

- Can use complete case analysis (no bias).
- Not fully efficient.

Density estimation:

- Rejects can (must!) be included in estimation process.
- How to specify component densities?

The nonignorable case

Distribution of outcome can be written as:

$$P(y \mid \mathbf{x}) = P(y \mid \mathbf{x}, a = 1)P(a = 1 \mid \mathbf{x}) + P(y \mid \mathbf{x}, a = 0)P(a = 0 \mid \mathbf{x}).$$

Sampling process identifies:

- Acceptance/Rejection probability: $P(a \mid \mathbf{x})$
- Outcome conditional on acceptance: $P(y \mid \mathbf{x}, a = 1)$

but is uninformative on

- Outcome conditional on rejection: $P(y \mid \mathbf{x}, a = 0)$

Ignorability assumption

Assume that

$$P(y \mid \mathbf{x}, a = 1) = P(y \mid \mathbf{x}, a = 0)$$

Now $P(y \mid \mathbf{x})$ coincides with the observable distribution $P(y \mid \mathbf{x}, a = 1)$.

We already looked at this case.

Bounds on $P(y = 1 \mid \mathbf{x})$

Without making any assumptions, we can compute bounds on $P(y = 1 \mid \mathbf{x})$:

$$P(y = 1 \mid \mathbf{x}, a = 1)P(a = 1 \mid \mathbf{x}) \leq$$
$$P(y = 1 \mid \mathbf{x}) \leq$$
$$P(y = 1 \mid \mathbf{x}, a = 1)P(a = 1 \mid \mathbf{x}) + P(a = 0 \mid \mathbf{x}).$$

The width of the interval is equal to the rejection probability at $\mathbf{x}$.

Bounds on $P(y = 1 \mid \mathbf{x})$: Example

Suppose

$$P(a = 1 \mid \mathbf{x}) = 0.8,\ P(a = 0 \mid \mathbf{x}) = 0.2$$
$$P(y = 1 \mid \mathbf{x}, a = 1) = 0.75$$

Then $0.8 \times 0.75 \leq P(y \mid \mathbf{x}) \leq 0.8 \times 0.75 + 0.2$

So $P(y \mid \mathbf{x}) \in [0.6, 0.8]$

Can we tighten the bounds?

Assume

$$P(y = 1 \mid \mathbf{x}, a = 0) \leq P(y = 1 \mid \mathbf{x}, a = 1)$$

(begging the question?)

Then

$$\begin{aligned} & P(y = 1 \mid \mathbf{x}, a = 1) P(a = 1 \mid \mathbf{x}) \leq \\ & P(y = 1 \mid \mathbf{x}) \leq \\ & P(y = 1 \mid \mathbf{x}, a = 1) \end{aligned}$$

So $P(y \mid \mathbf{x}) \in [0.6, 0.75]$

Bivariate Probit Model with Sample Selection

Consists of

- selection equation (models the accept/reject decision)
- Outcome equation (models the outome of the loan)

Model:

$$\begin{aligned} a_i^* &= \mathbf{x}_i \alpha + \varepsilon_i \\ y_i^* &= \mathbf{x}_i \beta + \upsilon_i \quad \text{for } i = 1, 2, \ldots, n \end{aligned}$$

a_i^* and y_i^* are unobserved numeric variables.

Bivariate Probit Model with Sample Selection

We do observe:

$$a_i = \begin{cases} 0 \text{ if loan rejected } (a_i^* < 0) \\ 1 \text{ if loan accepted } (a_i^* \geq 0) \end{cases}$$

$$y_i = \begin{cases} 0 \text{ if bad loan } (y_i^* < 0) \\ 1 \text{ if good loan } (y_i^* \geq 0) \end{cases}$$

y_i is only observed if $a_i = 1$.

Bivariate Probit Model with Sample Selection

The disturbances $(\varepsilon_i, \upsilon_i)$ are assumed to follow a bivariate normal distribution:

$$\begin{pmatrix} \varepsilon_i \\ \upsilon_i \end{pmatrix} \sim N(\mu, \Sigma) \qquad \mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

If $\rho = 0$ then

$$P(y = 1 \mid \mathbf{x}, a = 1) = P(y = 1 \mid \mathbf{x}, a = 0),$$

and MAR applies after all.

Bivariate Probit Model with Sample Selection

On the other hand, if $\rho > 0$ then

$$P(y = 1 \mid \mathbf{x}, a = 1) > P(y = 1 \mid \mathbf{x}, a = 0),$$

i.e. at any point $\mathbf{x}$, the probability of a good loan is *higher* among the accepts than among the rejects.
If $\rho < 0$, then

$$P(y = 1 \mid \mathbf{x}, a = 1) < P(y = 1 \mid \mathbf{x}, a = 0),$$

i.e. at any point $\mathbf{x}$, the probability of a good loan is *lower* among the accepts than among the rejects.

Bivariate Probit: empirical results and theoretical properties

- Boyes et al.; Greene; Jacobson and Roszbach found *negative* values for ρ !
- Ash and Meester; Banasik, Crook and Thomas: only marginal improvement of predictive performance.
- Relies on (and is highly sensitive to) untestable assumption of normality.
- Parameters poorly identified.

Opinions/Discussion

- You control the missing-data mechanism: make sure it's ignorable.
- Avoid bivariate probit model with sample selection (at least perform a sensitivity analysis).
- Sample from the reject region to avoid complete extrapolation.
- If data is sufficient, use function estimation approach rather than density estimation approach.