SAMPLE SELECTION BIAS IN CREDIT SCORING MODELS

John Banasik, Jonathan Crook
Credit Research Centre, University of Edinburgh
Lyn Thomas
University of Southampton

The Problem

We wish to estimate an accept-reject model applicable to ALL applicants but we only know the performance of those accepted in the past.

Conventional "Treatment" Reject Inference

Extrapolation

Multinomial regression (various)

Augmentation

Others

Weaknesses of “Reject Inference”

Extrapolation

(Hand & Henley 1993) Provided population model for all applicants is same as that for accepts only then

Direct posterior modelling (Logistic, Probit) does not give biased estimates of parameters.

Linear Discriminant Analysis will give biased estimates because LDA assumes p(X|G) and p(X|B) distributions are normal but they are not.

A More General Approach to Missing Data

Let $D=(D_0, D_m)$

D denotes whether a case has defaulted
subscripts: 0 denotes value is observed
m denotes missing

Let $S = \{0,1\}$

0 denotes value of D is missing
1 denotes value of D is observed

Little & Rubin's (1987) Categories

<u>Missing Completely At Random (MCAR)</u>: S is independent of D and X

<u>Missing at Random(MAR):</u> S depends on X but not on D

<u>Non-ignorably Missing (NIM):</u> S depends on D and on X

MCAR

D_0 values are a random sample from the D values. Estimated parameters from this sample does not give biased estimates of parameters applicable to population of all D values

MAR

Can be shown (Hand & Henley, Little& Rubin) that if sampling is conditional only on X then using only observed values of D will not give biased estimates of the parameters of a population model applicable to all D values

NIM

D depends on X values and on other variables (called them Z).
So S depends on X and Z.
Relating D_0 values only to X **WILL** give biased estimates of population of all applicants model because the chance of observation of D values depends on Z and X not just X. Z values appear in the likelihood function and so ignoring them by relating D_0 merely to X will lead to omitted variable bias. Known as “sample selection bias” in the econometric literature.

Sample Selection Techniques

We are interested in parameterising a model relating to a population where

$$Y_1 = \beta'_1 X_1 + \varepsilon_1 \qquad \ldots.(1)$$

We observe only those values of Y_1 when another variable, Y_2 is positive
Suppose Y_2 is a linear model

$$Y_2 = \beta'_2 X_2 + \varepsilon_2$$

Call observed values of Y_1, Y^*_1. Then:

$$\begin{array}{ll} Y_1 = Y^*_1 & \text{if } Y_2 > 0 \\ Y_1 \text{ is unobserved} & \text{if } Y_2 \leq 0 \end{array}$$

If we apply OLS to eqtn 1 using only the selected sample the regression line is

$$E(Y^*_1 \mid X_1, Y_2 > 0) = \beta'_1 X_1 + E(\varepsilon_1 \mid \varepsilon_2 > -\beta'_2 X_2)$$

which differs from the regression for the population of interest

$$E(Y_1) = \beta'_1 X_1$$

Further effect

Suppose there are variables in X_2 (and so affect Y_2) but they are not in X_1 (they do not affect Y_1)

Is possible that if estimate Eqtn 1 using only the selected sample these variables may appear statistically significant when in fact they do not affect Y_1.

Bivariate Probit Model with Sample Selection

Suppose Y_3 and Y_4 are continuous unobserved variables with

$$Y_3 = \beta'_3 X_3 + \varepsilon_3$$

$$Y_4 = \beta'_4 X_4 + \varepsilon_4$$

and $Y_3 = 1$ if $Y_3 > 0$
0 if $Y_3 \leq 0$

and $Y_4 = 1$ if $Y_4 > 0$
0 if $Y_4 \leq 0$

with $E(\varepsilon_3) = E(\varepsilon_4)$
$Cov(\varepsilon_3, \varepsilon_4) = \rho$
$(\varepsilon_3, \varepsilon_4) \sim$ bivariate normal

Also Y^*_3 is observed only if $Y^*_4 = 1$
Y^*_4 is observed in all cases

Appears to be the credit scoring problem

Information Observed

	Good	**Bad**	**Total**
Accepted	observed A	observed B	observed C
Rejected	not observed D	not observed E	observed F

Given that $(\varepsilon_3, \varepsilon_4) \sim$ bivariate probit the unconstrained observations with associated probabilities are

$Y^*_3 = 1\ Y^*_4 = 1$: $P(Y^*_3 = 1\ Y^*_4 = 1) = \Phi_B(\beta'_3 X_3, \beta'_4 X_4, \rho)$ Cell A

$Y^*_3 = 0,\ Y^*_4 = 1$: $P(Y^*_3 = 0,\ Y^*_4 = 1) = \Phi_B(-\beta'_3 X_3, \beta'_4 X_4, -\rho)$ Cell B

$Y^*_4 = 0$: $P(Y^*_4 = 0) = \Phi_U(\beta'_4 X_4)$ Cell F

where Φ_B is CDF of bivarariate normal distribution with density $\phi_B(\beta'_3 X_3, \beta'_4 X_4, \rho)$
Φ_U is CDF of univariate normal distribution with density $\phi_U(\beta'_4 X_4)$

Notice

One cause of $\rho=0$ is if at least one variable is omitted from *both* the default and the AR equation or if the omitted variables are correlated

Sample selection bias may occur unless $\rho=0$

If accept-reject decision is deterministic (no overrides) and if the accept-reject model can be estimated perfectly then $\rho=0$ and no sample selection bias occurs

But

overrides do occur

the accept-reject model may not be perfectly estimatable (eg may have been estimated by neural network algorithm)

Previous Studies

Boyes, et al (1989)

Used bivariate probit to estimate parameters in a default equation but do not compare the predictive performance of this model with one based on aacepts only using a holdout of all applicants

Greene (1992, 1998)

Compares the predicted conditional probabilities ($P(Y^*_3 \mid X_3, Y_4=1)$ with unconditional probabilities - but does not compare the predictive performance using a sample of all applicants

Other Relevant Studies

Methodology: Initial Analysis – Data

1. Complete sample comprises **12208** applicants virtually all of whom are accepted, distinguishing those who would normally be accepted.

2. Course classification of **46** variables (typically 6 categories each):

   Classification designed to be equally applicable to

   - All applicants
   - Accepted applicants in isolation

   Each variable to be considered alternatively as

   - Weights of evidence
   - Sets of binary variables

3. Training sample obtained by proportional stratified sampling to enhance comparability with holdout sample.

   - 2/3 of all applicants in the training sample
   - 2/3 of all **accepted** applicants in the training sample
   - 2/3 of all good applicants in the training sample
   - 2/3 of all good **accepted** applicants in the training sample

Methodology: Initial Analysis – Model Estimation and Evaluation

1. All explanatory variables deployed to estimate logistic regression model using both types of variable (weights of evidence and binary sets) on the basis of training cases alone for both
   - All applicants
   - Accepted applicants only

   Weights of evidence separately estimated for both applicant categories on the basis of training cases.

2. Each model used to predict corresponding holdout cases to assess the possible extent of over-fitting.

3. The model for accepted applicants only used to predict all holdout cases to assess its general applicability.

Preliminary indications are that

- Models based on accepted applicants only seem generally applicable
- Reject inference affords very modest prospects for improved predictive performance.

Table 2: Prediction Performance for Original Models based on Weights of Evidence

Training Sample - Accepted

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	3540	577	4117	86.0%
Bad	577	719	1296	55.5%
Total	4117	1296	5413	78.7%

Holdout Sample - Accepted

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	1756	317	2073	84.7%
Bad	339	343	682	50.3%
Total	2095	660	2755	76.2%

Holdout Sample - All Cases

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	1954	707	2661	73.4%
Bad	480	928	1408	65.9%
Total	2434	1635	4069	70.8%

Training Sample - Accepted

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	4259	1062	5321	80.0%
Bad	1062	1756	2818	62.3%
Total	5321	2818	8139	73.9%

Holdout Sample - All Cases

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	2098	563	2661	78.8%
Bad	570	838	1408	59.5%
Total	2668	1401	4069	72.2%

Holdout Sample - All Cases

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	2098	563	2661	78.8%
Bad	570	838	1408	59.5%
Total	2668	1401	4069	72.2%

Table 3: Prediction Performance for Original Models based on Binary Variables

Training Sample - Accepted

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	3579	538	4117	86.9%
Bad	538	758	1296	58.5%
Total	4117	1296	5413	80.1%

Holdout Sample - Accepted

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	1748	325	2073	84.3%
Bad	339	343	682	50.3%
Total	2087	668	2755	75.9%

Holdout Sample - All Cases

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	1970	691	2661	74.0%
Bad	486	922	1408	65.5%
Total	2456	1613	4069	71.1%

Training Sample - Accepted

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	4311	1010	5321	81.0%
Bad	1010	1808	2818	64.2%
Total	5321	2818	8139	75.2%

Holdout Sample - All Cases

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	2078	583	2661	78.1%
Bad	578	830	1408	58.9%
Total	2656	1413	4069	71.5%

Holdout Sample - All Cases

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	2078	583	2661	78.1%
Bad	578	830	1408	58.9%
Total	2656	1413	4069	71.5%

Table 4: Comparative Prediction Performance of the Bivariate Probit Model

Ref		Bivariate Probit With Selection				Probit (Accepts only)		Probit (All cases)	
	Variables Included	Number of Cases	Rho	Signif	AUROC	Number of Cases	AUROC	Number of Cases	AUROC
Binary Variables									
LL35	All	5413	-0.219	0.09	0.7747	5413	0.7722	8139	0.7811
LL30	Those in SW	5413	0.103	0.25	0.7715	5413	0.7729	8139	0.7778
LL31	SW less 3	5413	0.166	0.07	0.7695	5413	0.7720	8139	0.7764
Weights of Evidence									
LL36	All	5413	-0.066	0.63	0.7816	5413	0.7815	8139	0.7836
LL33	Those in SW	5413	0.071	0.44	0.7782	5413	0.7780	8139	0.7802
LL34	SW less 3	5413	0.056	0.55	0.7772	5413	0.7770	8139	0.7791

Methodology: Subsequent Analysis – General Approach

Subsequent analysis empirically examines two aspects of the foregoing conclusions.

- Perhaps the limited scope for reject inference occurs only because the **acceptance threshold** is too low.
- The **variable sets** for models previously used to **accept or reject** applicants may diverge from those of subsequent models used to predict how **good or bad** will be the accepted applicants.

To examine these aspects

1. The data set was recast to distinguish **5 acceptance thresholds**, one for each couple of deciles.
   - The same proportional stratified sampling was adopted.
   - The original course classification was retained.
   - Weights of evidence were re-estimated for each training sample.
2. Distinct Accept-Reject and Good-Bad models were constructed on the basis of distinct explanatory variable sets.

Methodology: Subsequent Analysis – Distinct Variable Sets

The **Accept-Reject** model was based on 2540 Scottish applicants. The remaining 9668 English applicants comprised 6446 training cases to estimate a **Good-Bad** model and 3222 holdout cases.

1. An eligible pool of variables for the **Accept-Reject** model was discerned by stepwise regressions where **Accept-Reject** was the dependent variable, one each for Scotland, England, and the UK.
2. A variable surviving in any of these three was entered into a stepwise regression on **Scottish** data where **Good-Bad** was the dependent variable.
3. The **Scottish model** based only on significant variables was used to provide a **ranking of acceptability** used to group English applicants into **5 bands**.
4. An English model was discerned on the basis of stepwise regression where **Good-Bad** was the dependent variable. This determined the variable set to be used for each of the five bands.

Table 5: Variables Included in the Accept-Reject and Good-Bad Equations

Ref	Variable	Good-Bad Equation	Accept-Reject Equation
43	B1		✓
20	Time and Present Add		✓
33	B2		✓
40	Weeks since last CCJ		✓
22	No of children under 16		✓
15	B3	✓	✓
11	Television area code	✓	✓
32	B4	✓	✓
31	P1	✓	✓
35	P2	✓	✓
6	B5	✓	✓
19	Accommodation Type	✓	✓
17	Age of applicant (years)	✓	✓
26	Has Telephone	✓	✓
23	P3	✓	✓
34	B6	✓	✓
21	Type of Bank/Building Soc Accts	✓	
16	B7	✓	
28	Current Electoral Role Category	✓	
38	P4	✓	
25	Occupation Code	✓	
36	B8	✓	
30	Years on Electoral Role at current address	✓	
48	No of searches in last 6 months	✓	
47	B9	✓	
7	B10	✓	
9	B11	✓	
46	B12	✓	
44	B13	✓	
27	P5	✓	

Bn = bureau variable n; P = proprietary variable n; ✓ denotes variable is included

Table 6: Prediction Performance for Models based on Weights of Evidence

Training Sample - Band 1

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	1079	71	1150	93.8%
Bad	71	68	139	48.9%
Total	1150	139	1289	89.0%

Holdout Sample - Band 1

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	538	37	575	93.6%
Bad	32	38	70	54.3%
Total	570	75	645	89.3%

Holdout Sample - Band 5

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	1475	671	2146	68.7%
Bad	289	787	1076	73.1%
Total	1764	1458	3222	70.2%

Training Sample - Band 2

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	1991	198	2189	91.0%
Bad	198	191	389	49.1%
Total	2189	389	2578	84.6%

Holdout Sample - Band 2

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	979	115	1094	89.5%
Bad	99	96	195	49.2%
Total	1078	211	1289	83.4%

Holdout Sample - Band 5

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	1447	699	2146	67.4%
Bad	249	827	1076	76.9%
Total	1696	1526	3222	70.6%

Training Sample - Band 3

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	2626	407	3033	86.6%
Bad	407	427	834	51.2%
Total	3033	834	3867	79.0%

Holdout Sample - Band 3

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	1305	212	1517	86.0%
Bad	190	227	417	54.4%
Total	1495	439	1934	79.2%

Holdout Sample - Band 5

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	1534	612	2146	71.5%
Bad	291	785	1076	73.0%
Total	1825	1397	3222	72.0%

Training Sample - Band 4

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	3076	638	3714	82.8%
Bad	638	804	1442	55.8%
Total	3714	1442	5156	75.3%

Holdout Sample - Band 4

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	1528	329	1857	82.3%
Bad	306	415	721	57.6%
Total	1834	744	2578	75.4%

Holdout Sample - Band 5

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	1607	539	2146	74.9%
Bad	348	728	1076	67.7%
Total	1955	1267	3222	72.5%

Training Sample - Band 5

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	3432	861	4293	79.9%
Bad	861	1292	2153	60.0%
Total	4293	2153	6446	73.3%

Holdout Sample - Band 5

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	1705	441	2146	79.5%
Bad	408	668	1076	62.1%
Total	2113	1109	3222	73.6%

Holdout Sample - Band 5

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	1705	441	2146	79.5%
Bad	408	668	1076	62.1%
Total	2113	1109	3222	73.6%

Table 7: Prediction Performance for Models based on Binary Variables

Training Sample - Band 1

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	1087	63	1150	94.5%
Bad	63	76	139	54.7%
Total	1150	139	1289	90.2%

Holdout Sample - Band 1

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	530	45	575	92.2%
Bad	33	37	70	52.9%
Total	563	82	645	87.9%

Holdout Sample - Band 5

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	1568	578	2146	73.1%
Bad	418	658	1076	61.2%
Total	1986	1236	3222	69.1%

Training Sample - Band 2

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	1998	191	2189	91.3%
Bad	191	198	389	50.9%
Total	2189	389	2578	85.2%

Holdout Sample - Band 2

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	975	119	1094	89.1%
Bad	92	103	195	52.8%
Total	1067	222	1289	83.6%

Holdout Sample - Band 5

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	1425	721	2146	66.4%
Bad	249	827	1076	76.9%
Total	1674	1548	3222	69.9%

Training Sample - Band 3

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	2646	387	3033	87.2%
Bad	387	447	834	53.6%
Total	3033	834	3867	80.0%

Holdout Sample - Band 3

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	1316	201	1517	86.8%
Bad	180	237	417	56.8%
Total	1496	438	1934	80.3%

Holdout Sample - Band 5

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	1506	640	2146	70.2%
Bad	254	822	1076	76.4%
Total	1760	1462	3222	72.3%

Training Sample - Band 4

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	3111	603	3714	83.8%
Bad	603	839	1442	58.2%
Total	3714	1442	5156	76.6%

Holdout Sample - Band 4

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	1538	319	1857	82.8%
Bad	292	429	721	59.5%
Total	1830	748	2578	76.3%

Holdout Sample - Band 5

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	1612	534	2146	75.1%
Bad	331	745	1076	69.2%
Total	1943	1279	3222	73.2%

Training Sample - Band 5

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	3458	835	4293	80.5%
Bad	835	1318	2153	61.2%
Total	4293	2153	6446	74.1%

Holdout Sample - Band 5

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	1718	428	2146	80.1%
Bad	399	677	1076	62.9%
Total	2117	1105	3222	74.3%

Holdout Sample - Band 5

Actual	Predicted Good	Predicted Bad	Total	Percent Correct
Good	1718	428	2146	80.1%
Bad	399	677	1076	62.9%
Total	2117	1105	3222	74.3%

Table 8: Comparative Prediction Performance of the Bivariate Probit Model for each Band

Ref			Bivariate Probit With Selection				Probit (Accepts only)		Probit (All cases)	
	Weights of Evidence	Variables Included	Number of Cases	Rho	Signif	AUROC	Number of Cases	AUROC	Number of Cases	AUROC
LL39	Band 1	All	1289	-0.150	0.39	0.7661	1289	0.7649		
	Band 2		2578	-0.199	0.20	0.7866	2578	0.7858		
	Band 3		3867	-0.055	0.60	0.7973	3867	0.7969		
	Band 4		5156	-0.069	0.52	0.8025	5156	0.8024		
	Band 5								6446	0.8027
LL42	Band 1	As Table 5	1289	-0.029	0.85	0.7842	1289	0.7840		
	Band 2		2578	-0.003	0.98	0.7940	2578	0.7940		
	Band 3		3867	-0.046	0.62	0.8016	3867	0.8015		
	Band 4		5156	0.048	0.61	0.8038	5156	0.8039		
	Band 5								6446	0.8050
LL43	Band 1	As Table 5 less 2	1289	0.182	0.23	0.7866	1289	0.7865		
	Band 2		2578	0.088	0.49	0.7931	2578	0.7929		
	Band 3		3867	0.061	0.42	0.8001	3867	0.8003		
	Band 4		5156	0.028	0.79	0.8013	5156	0.8014		
	Band 5								6446	0.8015

Conclusions

The scope for improved predictive performance by any form of reject inference is modest

The scope depends on the cut-off score adopted.

Use of the bivariate probit model with selection only marginally improves predictive performance although this depends on the variables included in the model.