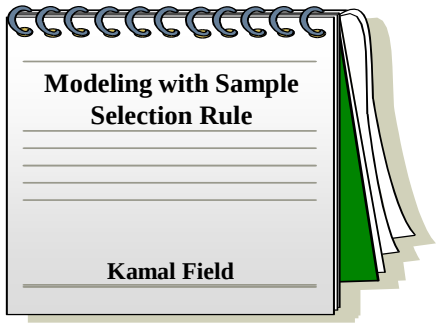
Statistical Modeling

with Sample Selection Rule

Presentation to the Advanced Analysis &
Data Mining Practitioners Forum on 18 January 2002

V. 1.0



- Our objective is to improve business opportunities by profitably satisfying customers' needs.
- In marketing, models are used to target potential customers to produce cost-effective mailing campaigns.
- The quality of models depends on the integrity of the data and on an efficient estimation procedure.
- Today, the aim is to create awareness to the sample selection problem and review a means of overcoming it.

Contents

1) Problem of sample selection rule
2) Implications of sampling selection rule
3) Analytical illustration of the problem
4) Application: Modeling risk / Reject inference

Problem of sample selection: Examples

Many modeling processes involve the use of an apparently "random" sample that is actually a censored sample. The following are examples:

A) Modeling expected market wage rate for women based on profile of women participating in labor force (i.e. women not in the labour force are not included).

B) Modeling expected spending for a particular event using sample that excludes non-respondents.

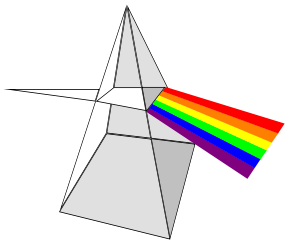
c) Modeling Credit Risk using a sample that excludes rejected individuals.

Definition: Censored Sample

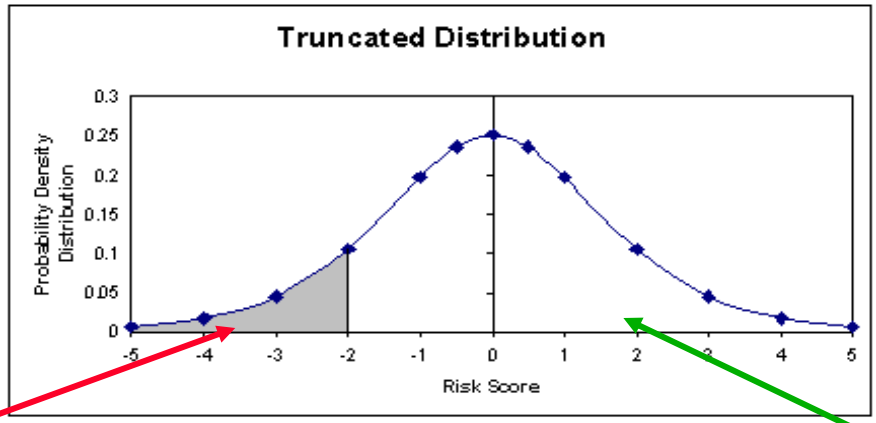
In all of the above examples there are one or more hidden selection rules that result in censored samples.

By a censored sample we mean a distribution for which we have available all of the explanatory variables **x** (independent variables) but only limited data for the dependent variable or outcome **y**.

The missing information may lead to a misspecification bias.

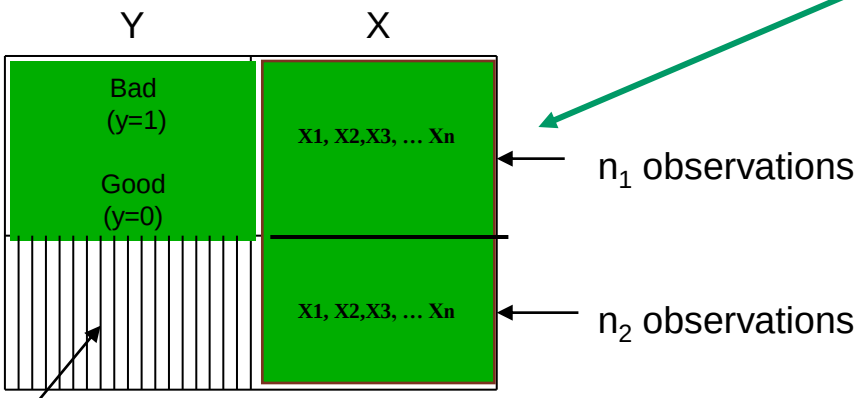


Reject Inference



Rejected

Accepted



n_1 observations

n_2 observations

Y is not observed for the Rejects, which is indicated by the shaded area

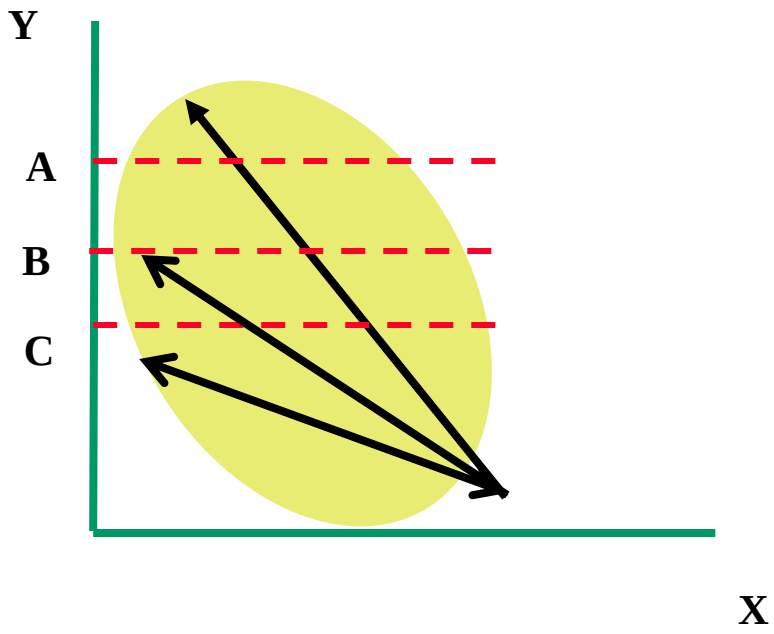


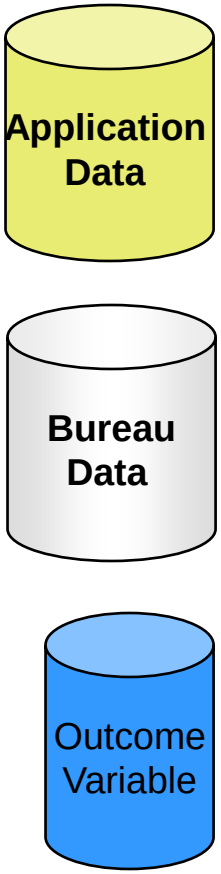
Illustration of Impact of Truncated Sample on Estimated slope

Example:Application on Reject Inference

Consider a case of building a risk scorecard for an Auto Loan & the following information is provided:

1. Application data
2. Bureau data
3. Outcome variable (default & non-default)

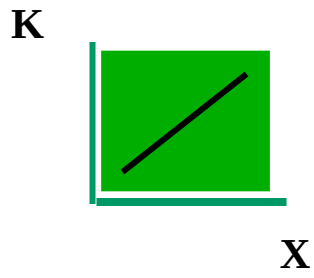
As the outcome variable is available only to those applied and were accepted and not available for those that applied and were rejected. Therefore, the available sample is a censored sample and could lead to invalid inference about risk scoring.



Estimation

Stage 1. Estimate the Reject Inference equation:

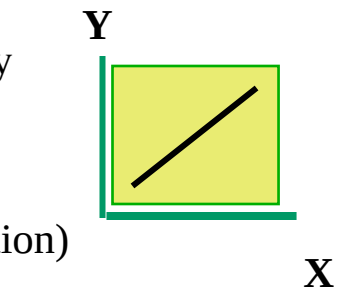
To quantify the conditional error term or the missing term (using probit analysis).



Stage 2. Estimate Risk equation:

To build the ‘Risk Scorecard’ where the missing term is included as an explanatory variable (using logistic regression).

The results will establish whether the missing term is a significant. If the term is significant then we can be sure that the misspecification bias (due to sample selection) has been removed.



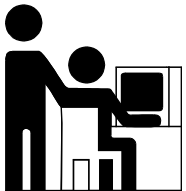
Stage 1: Probit Analysis K = g(X)+μ(accept=1 & reject=0)

Variable	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	-0.2586991	0.177474	2.124821	0.1449	Intercept
MARRIED	-0.2343255	0.014011	279.7145	0.0001	If Married or Couple
CHILD1	-0.129546	0.013168	96.78199	0.0001	If there is one or more child
CAR_AGE2	0.94569885	0.015409	3766.786	0.0001	If car age between 2 and 5 years
CAR_AGE3	0.013915	4153.985	4153.985	0.0001	If car age greater than 5 years
AGE1	0.49833612	0.014104	1248.364	0.0001	If age <= 25 and <= 35
AGE2	0.5712426	0.01521	1410.536	0.0001	If age > 35
T_JOB1	0.22536374	0.0162	193.5249	0.0001	If time in job <= 6 months
T_JOB2	0.1362321	0.073587	3.427306	0.0641	If time in job > 6 & <= 18
T_JOB3	0.49800772	0.012342	1628.067	0.0001	If time in job > 18
COSIG	0.22910111	0.015861	208.6347	0.0001	If cosigner wife/husband or dealer
Reject	-0.3046869	0.027498	122.7746	0.0001	If Rejected internally or externally
Request	-0.1124623	0.013746	66.93459	0.0001	If there is an open request internal or external
W0	-0.3738968	0.02304	263.3523	0.0001	If Weak negative internal or external
S0	-0.8858912	0.024399	1318.272	0.0001	If strong negative internal or external
ODAYS_1	-0.1071887	0.177011	0.366689	0.5448	If number of day since oldest ongoing transaction is within one year
ODAYS_2	-0.0434911	0.177038	0.060349	0.8059	If number of day since oldest ongoing transaction is greater than one year

Second Stage: Logistic Regression Y = f (X)+ ξ (Good=1 & Bad=0)

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	
INTERCPT	-3.4745	0.4737	53.8062	0.0001	.
MARRIED	-0.4993	0.0823	36.7978	0.0001	If Married or Couple
CHILD1	-0.2356	0.0668	12.4437	0.0004	If there is one or more child
CAR_AGE2	0.8549	0.2327	13.4956	0.0002	If care age between 2 and 5 years
CAR_AGE3	1.3745	0.2268	36.7210	0.0001	If care age >5 years
AGE1	0.1118	0.1288	0.7529	0.3855	If age <= 25 and <= 35
AGE2	0.0784	0.1457	0.2897	0.5904	If age > 35
T_JOB1	0.5788	0.0807	51.3783	0.0001	If time in job <= 6 months
T_JOB23	0.0340	0.3089	0.0121	0.9125	If time in job > 6 & <= 18
T_JOB_3	-0.3351	0.1212	7.6440	0.0057	If time in job > 18
COSIG	-0.1361	0.0908	2.2484	0.1338	If cosigner wife/husband or dealer
Reject	0.4403	0.1274	11.9352	0.0006	If Rejected internally or Externally
Request	0.3945	0.0630	39.2362	0.0001	If there is an open request internal or External
W0	0.7440	0.1106	45.2551	0.0001	If Weak negative internal or External
S0	0.3577	0.2415	2.1933	0.1386	If strong negative internal or External
ODAYS_12	-0.0368	0.0901	0.1672	0.6826	If number of day since oldest ongoing transaction is within one year
ODAYS_34	-0.1724	0.0893	3.7287	0.0535	If number of day since oldest ongoing transaction is greater than one year
LAMBDA	1.1324	0.4597	6.0684	0.0138	The missing variable

Association of Predicted Probabilities and Observed Responses
Concordant = 74.1%

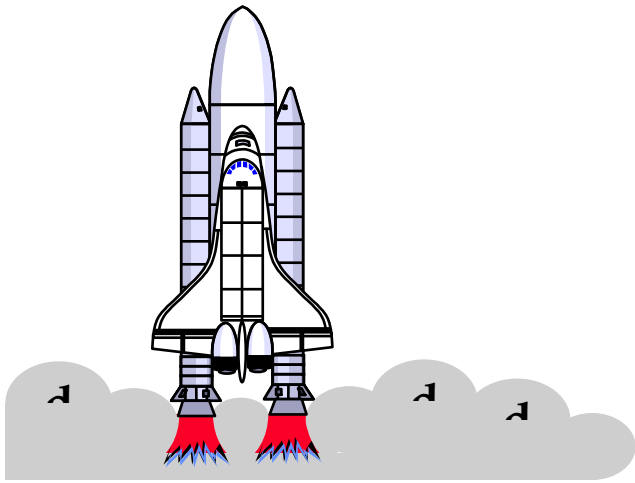


Conclusion

- The use of a test of sample censoring is a potential mechanism to improve modeling and an extra tool for our analytical tool boxes

- Modeling without allowing for the sample section rule may lead to misspecification and biased regression analysis.

- At the very least, the concept of censored samples is something we often need to bear in mind



Appendix

Analytical Illustration:

Analytically, censored samples could lead to biased estimates. Heckman (1979) developed a methodology to illustrate impact of the problem.

let

$$K = X\beta + \mu \qquad (1)$$ (accept/reject)

$$Y = X\theta + \xi \qquad (2)$$ (default/non-default)

where

$K_i = 1$ if $z > 0$, otherwise $K_i = 0$
z is the expected profit from issuing a loan
$Y_i = 1$ for default, $Y_i = 0$ for non- default,
X vector of explanatory variables, θ is vector of parameters, $[\mu, \xi]$ follow a bivariate normal distribution
Y_i is observed when $K_i = 1$.

For a random sample the population regression equation:

$$\mathbf{Y = X\theta + \xi}$$

where

$$\mathbf{E(Y) = E(X\theta)} \qquad \textbf{(since } \mathbf{E(\xi) = 0)}$$

The expected value of the error term is zero

But for a sample with selection rule:

$$\mathbf{E(Y \mid} \text{ sample selection rule}\mathbf{)} = \mathbf{E(X\theta) + E(\xi \mid} \text{ sample selection rule}\mathbf{)}$$

or

$$\mathbf{E(Y \mid K_i > 0)} = \mathbf{E(X\theta) + E(\xi \mid K_i > 0)}$$
$$= \mathbf{E(X\theta) + E(\xi \mid X_i\beta + \mu_i > 0)}$$
$$= \mathbf{E(X\theta) + E(\xi \mid \mu_i > -X_i\beta)}$$

If the term $\mathbf{E(\xi \mid \mu_i > -X_i\beta)}$ not equal to 0 & then this may cause violation to the standard assumption for unbiased regression analysis.

If μ is correlated with ξ then the term $\mathbf{E}\,(\,\xi \mid \mu_i > -\mathbf{X}_i\boldsymbol{\beta}\,)$ will be non zero and the regression results would be biased.

One way of dealing with this issue is to cast it as a problem of missing a relevant variable.

By treating the term **E (ξ | μ_i > - $X_i\beta$)** as the new variable **'missing variable'** and include it to the rest of variables in the regression equation, we may be able to remove the biased.

By examining the statistical property of the missing term we will establish whether the new variable is a significant variable or not.